Journal of Applied Crystallography

ISSN 0021-8898

Serial Crystallography with Multi-stage Merging of 1000's of Images

Herbert J Bernstein,[a]* Lawrence C Andrews,[b] James Foadi,[c] Martin R. Fuchs,[d] Jean Jakoncic,[d] Sean McSweeney,[d] Dieter K. Schneider,[d] Wuxian Shi,[e] John Skinner,[d] Alexei Soares[d] and Yusuke Yamada [f]*

[a]School of Chemistry and Materials Science, Rochester Institute of Technology, Rochester, NY USA, [b]retired, Kirkland, WA USA , [c]Diamond Light Source, Chilton UK, [d]Brookhaven National Laboratory, Upton, NY USA, [e]Case Western University, Cleveland, OH USA, and [f]Photon Factory, Tsukuba,Ibaraki JP. Correspondence e-mail: [a]*hjbsch@rit.edu, [f]*yusuke.yamada@kek.jp

KAMO and Blend provide particularly effective tools to automatically manage the merging of large numbers of data sets from serial crystallography. The requirement for manual intervention in the process can be reduced by extending Blend to support additional clustering options to increase the sensitivity to differences in unit cell parameters and to allow for clustering of nearly complete datasets on the basis of intensity or amplitude differences. If datasets are already sufficiently complete to permit it, apply KAMO once, just for reflections. If starting from incomplete datasets, one applies KAMO twice, first using cell parameters. In this step either the simple cell vector distance of the original Blend is used, or the more sensitive NCDist, to find clusters to merge to achieve sufficient completeness to allow intensities or amplitudes to be compared. One then uses KAMO again using the correlation between the reflections at the common HKLs to merge clusters in a way sensitive to structural differences that may not perturb the cell parameters sufficiently to make meaningful clusters.

Many groups have developed effective clustering algorithms that use a measurable physical parameter from each diffraction still or wedge to cluster the data into categories which can then be merged to, hopefully, yield the electron density from a single protein iso-form. What is striking about many of these physical parameters is that they are largely independent from one another. Consequently, it should be possible to greatly improve the efficacy of data clustering software by using a multi-stage partitioning strategy. Here, we have demonstrated one possible approach to multi-stage data clustering. Our strategy was to use unit-cell clustering until merged data was of sufficient completeness to then use intensity based clustering. We have demonstrated that, using this strategy, we were able to accurately cluster data sets from crystals that had subtle differences.

1. Introduction

KAMO(Yamashita *et al.*, 2017*b*) (Yamashita *et al.*, 2017*a*) (Hasegawa *et al.*, 2017) and Blend(Foadi *et al.*, 2013) provide particularly effective tools to automatically manage the merging of large numbers of data sets from serial crystallography. The requirement for manual intervention in the process can be reduced by extending Blend to support additional clustering options to increase the sensitivity to differences in unit cell parameters and to allow for clustering of nearly complete datasets on the basis of intensity or amplitude differences. If datasets are already sufficiently complete to permit it, apply KAMO once, just for reflections. If starting from incomplete datasets, one applies KAMO twice, first using cell parameters. In this step either the simple cell vector distance of the original Blend is used, or the more sensitive NCDist, to find clusters to merge to achieve sufficient completeness to allow intensities or amplitudes to be compared. One then uses KAMO again using the correlation between the reflections at the common HKLs (Assmann *et al.*, 2016) to merge clusters in a way sensitive to structural differences that may not perturb the cell parameters sufficiently to make meaningful clusters.

X-ray free-electron lasers (XFELS) have pioneered effective crystallography data collection from large numbers of crystals. (Colella & Luccio, 1984) (Neutze *et al.*, 2000) Serial crystallography, an essential technique at x-ray free electron laser light sources, has become an important technique at synchrotrons (Giordano *et al.*, 2012) (Liu & Hendrickson, 2013) (Rossmann, 2014) (Standfuss & Spence, 2017), especially at newer high intensity synchrotron beamlines. The data may be organized either as XFEL-like still images or as thousands of wedges of data produced from very large numbers of crystals. The stills and wedges need to be carefully organized into reasonably homogeneous clusters of data that can be merged for processing. This is going to be one of the common tools to assemble complete data from many partial wedges in MR, SAD, ligand studies and to sort classes of crystals for studies of

dynamics, binding, interactions, *etc.*. KAMO includes cluster analysis based both on cell parameters and on reflection correlation coefficients.

In this paper we discuss the issues involved in improving the sensitivity of both approaches to clustering, using, as an example, 999 5° wedges from lysozyme in four iso-forms:

- NAG: native with N-acetylglucosamine (NAG) soaked in,
- Benz.: native with benzamadine soaked in, and
- Benz.+NAG: native with both NAG and benzamadine soaked in.
- Native: no ligands

As we will see, although the cell parameters are changed sufficiently to allow recognition of the NAG soak, it is difficult to filter the benzamadine soak simply on the basis of cell parameter changes, suggesting the desirability of switching from cell-based clustering to reflection-based clustering as early in the process as possible.

2. Limits of conventional clustering

Since our goal is to expand the capabilities of existing clustering techniques, we began by applying a conventional clustering strategy to diffraction data from lysozyme mirco-crystals containing various combinations of known small molecule binders. Lysozyme micro-crystals suitable for acoustic harvesting (Soares *et al.*, 2011) were grown using batch crystallization by dissolving 120 mg/ml lysozme in 0.2M sodium acetate pH 4.6 (Hampton Research HR7-110) and combining with equal parts precipitant (10% ethylene glycol + 12% sodium chloride) (Roessler *et al.*, 2016). The resulting slurry of 5-10 um crystals was divided into four aliquots. Three of the four aliquots were then equilibrated overnight with an equal volume of 0.5 M solutions of, respectively, benzamadine, NAG, and benzamadine plus NAG. These two small molecules are known to bind tetragonal lysozyme crystals (Yin *et al.*, 2014). The fourth aliquot was diluted with an equal volume of mother liquor but contained no ligands.

The diffusion rate for benzamadine and NAG within lysozyme crystals is approximately $1\mu m/s$ (Cole *et al.*, 2014). To prevent cross-contamination of crystals with neighboring iso-forms, crystals could not be mixed with different iso-forms for more than 1s before diffusion was halted by plunge cryo-cooling in LN_2. To accomplish this, we deposited $5\mu L$ of crystal slurry from each aliquot onto a separate agarose support (Cuttitta *et al.*, 2015). We used acoustic sound pulses to harvest $2.5nL$ of crystal slurry from each of the four lysozyme aliquots, and separately position them on a micro-mesh (MiTeGen M3-L18SP-10) such that none of the droplets was in contact with any other (Figure 1A). Crystal containing droplets were threaded through small apertures to prevent cross-contamination (Foley *et al.*, 2016). We then swept the non-crystal containing side of the micro-mesh against a sponge moistened with cryo-protectant (mother liquor + 20% glycerol) and, in one smooth motion, immediately cryo-cooled the micro-mesh in LN_2. In addition to cryo-protection, this also mixed the crystals together into one contiguous field. The same procedure was repeated for a micro-mesh containing only two lysozyme iso-forms, Benz. + NAG and native. Serial diffraction data were then obtained in 5 degree wedges from 100 crystals on each micro-mesh.

The software package KAMO was then used in default configuration to partition the diffraction data from micro-meshes containing four lysozyme iso-forms into four clusters, and the diffraction data from micro-meshes containing two lysozyme iso-forms into two clusters. Each cluster of data was then merged, and then phased using the known structure of lysozyme. The atomic model was then refined using refmac (Winn *et al.*, 2003), and an omit difference map was examined using coot in the region where the ligands are expected to bind to the protein surface (Emsley & Cowtan, 2004). The omit difference map was contoured at 1.5 sigma and displayed using pymol (DeLano, 2002). The omit maps calculated from the four-way clustering data was not observed to closely match any of the four lysozyme iso-forms known to have been acoustically harvested onto the micro-meshes (data not shown). We concluded from this result that the clustering algorithm was not sufficiently sensitive to differentiate these four classes of very similar crystals using only variations in the observed unit cell parameters. However, the omit maps calculated from the two-way clustering data were a good fit to the expected lysozyme iso-forms (Fig.1). We concluded from this result that the two-ligand iso-form was sufficiently different from the native iso-form that unit cell based clustering could be successful. To do the four-way split, intensity-based clustering was added to the process.

3. Clustering on Cell Parameters

Stills and wedges of very low completeness are more appropriate for cell parameter clustering, rather than reflection clustering, because pairs of images with very few commensurate reflections may still provide reasonable estimates of unit cells but not provide enough data to compute a meaningful distance between sets of reflections.

The default Blend approach to clustering on cell parameters is to use $[a, b, c, \alpha, \beta, \gamma]$ as a six vector, drop the columns without significant variance, and use the Euclidean distance calculated from the remaining columns. This approach does not deal as effectively with the discontinuities produced by experimental error and ambiguities in reduction (*e.g.* between Type I and Type II cells and near the cubics) as the Andrews-Bernstein NCDist algorithm (Andrews & Bernstein, 2014), which allows slightly larger clusters of truly similar datasets to be formed, working in the space G^6 formed using Niggli reduction in the six-dimensional space formed by the metric tensor with the last three components doubled, $[a^2, b^2, c^2, 2bc\,cos(\alpha), 2ac\,cos(\beta), 2ab\,cos(\gamma)]$.

In our test case of 999 datasets of lysozyme with NAG and benzamadine soaks 998 clusters are found with completeness ranging from 40% to 100%. The top levels of the two dendrograms are shown in Figs. 2 3.

The dendrograms are qualitatively similar but, for this test data, the discrimination of the clustering changes. For the original Blend algorithm, the largest clusters that are 100% native,

100% NAG, 100% benzamadine and 100% NAG+benzamadine contain 4, 15, 5, and 10 datasets, respectively. For the NCDist clustering, the largest clusters that are 100% native, 100% NAG, 100% benzamadine and 100% NAG+benzamadine contain 9, 15, 8, and 7 datasets, respectively. This provides a better base for switching over from cell clustering to reflection clustering.

4. Clustering on Reflections

In a regime of high completeness (say, 90%) different datasets can have enough reflections at common hkl's to generate a satisfactory similarity or distance for clustering. If the data has been scaled, a R-value can be used as a distance, but, for unscaled data, the preferred approach is to use a Pearson Correlation Coefficient (*CC*) as a measure of similarity, i.e. having a larger value for sets of reflections that are similar and a smaller value for sets of reflections that are dissimilar. The Pearson Correlation Coefficient is essentially the cosine of the angle between vectors of data. The lack of common scaling is dealt with by subtracting the mean (μ) of each vector from each component and dividing by the norm of each to get two unit length vectors:

$$data\_set_1 = [F_{1,hkl_1}, F_{1,hkl_2}, ...]$$

$$data\_set_2 = [F_{2,hkl_1}, F_{2,hkl_2}, ...]$$

$$vec_1 = [F_{1,hkl_1} - \mu_1, F_{1,hkl_2} - \mu_1, ...]$$

$$vec_2 = [F_{2,hkl_1} - \mu_2, F_{2,hkl_2} - \mu_2, ...];$$

$$CC(data\_set_1, data\_set_2) = \frac{vec_1.vec_2}{||vec_1||||vec_2||}$$

In order to extend the range of applicability of *CC*, we convert it to a distance,

$$SFdist(data\_set_1, data\_set_2) = ||\frac{vec_1}{||vec_1||} - \frac{vec_2}{||vec_2||}||$$

which is related to *CC* by

$$SFdist(data\_set_1, data\_set_2)^2 = 2 - 2CC$$

Having this as a distance allows a simple adaptation to cases of completeness lower than 90% by adding a penalty to the distance for each unmatched reflection.

5. Impact of choices in clustering

unambiguous benzamadine-only, NAG-only, and benzamadine+NAG clusters are shown in the omit difference maps of the NAG site in clusters 28, 43 and 62 in Figs. 4, 5 and 6, respectively, and then omit difference maps of the Benzamidine site in clusters 28, 43 and 62 are in Figs. 7, 8 and 9, respectively. These are the results of two-stage KAMO clustering of the test data using NCDist cell-parameter based clustering to get to at least 10% completeness and then SFDist reflection-based clustering on the resulting 107 non-overlapping clusters.

The impact of using clustering on reflections for larger clusters can be seen by looking at how well-represented reasonably pure clusters are. In Figs. 10 and 11, we have represented the purity of native, NAG, benzamadine and benzamadine+NAG species NCDist and SFDist.

The extreme variations in the SFDist results suggest two important lessons.

- It is best to use a reflection-based clustering starting from datasets that are small enough to still be likely to be pure species, *i.e.* use cell-based clustering only just far enough to get to completeness that the reflection-based clustering can handle.
- It is not necessarily desirable to continue clustering to the largest of the "best" possible clusters. Smaller clusters of sufficient quality for processing are more likely to be pure species.

6. Discussion

Because micro-crystals are expected to react quickly and uniformly to changes in their environment, serial crystallography is a desirable tool for examining the plasticity with which protein crystals respond to external perturbations. In some cases the external perturbation can be physical, such as conformational changes induced by light (Young et al., 2016). In other cases proteins are perturbed by chemical means (Fromme, 2015). It is often not possible to to draw a sharp boundary between diffraction images from different protein iso-forms without the assistance of some type of clustering software. In response to this, many groups have developed effective clustering algorithms that use a measurable parameter from each diffraction still or wedge to cluster the data into categories which can then be merged to, hopefully, yield the electron density from a single protein iso-form. Examples of measurable parameters that have been used for this purpose include unit cell dimensions (Foadi *et al.*, 2013) (Zeldin *et al.*, 2015), and diffraction intensities (Assmann *et al.*, 2016) (Diederichs, 2017). What is striking about many of these physical parameters is that they are largely independent from one another. Consequently, it should be possible to greatly improve the efficacy of data clustering software by using a multi-stage partitioning strategy. Here, we have demonstrated one possible approach to multi-stage data clustering. Our strategy was to use unit-cell clustering until merged data was of sufficient completeness to then use intensity based clustering. We have demonstrated that, using this strategy, we were able to accurately cluster data sets from crystals that had subtle differences.

Acknowledgements

This is a preliminary report on work in progress. Work supported in part by

- US Dept. of Energy, Office of Science, DE-AC02-98CH10886 and E-SC0012704
- US NIH National Institute of General Medical Sciences, P41RR012408, P41GM103473, and P41GM111244
- HJB supported in part by Dectris, Ltd.

References

Andrews, L. C. & Bernstein, H. J. (2014). *J. Appl. Crystallogr.* **47**(1), 346 – 359.

Assmann, G., Brehm, W. & Diederichs, K. (2016). *J. Appl. Crystallogr.* **49**(3).

Cole, K., Roessler, C. G., Mule, E. A., Benson-Xu, E. J., Mullen, J. D., Le, B. A., Tieman, A. M., Birone, C., Brown, M., Hernandez, J. *et al.* (2014). *PloS one*, **9**(7), e101036.

Colella, R. & Luccio, A. (1984). *Optics communications*, **50**(1), 41 – 44.

Cuttitta, C. M., Ericson, D. L., Scalia, A., Roessler, C. G., Teplitsky, E., Joshi, K., Campos, O., Agarwal, R., Allaire, M., Orville, A. M. *et al.* (2015). *Acta Crystallogr.* D**71**(1), 94 – 103.

DeLano, W. L. (2002). *CCP4 Newsletter On Protein Crystallography*, **40**, 82 – 92.

Diederichs, K. (2017). *Acta Crystallogr.* D**73**(4), 286 – 293.

Emsley, P. & Cowtan, K. (2004). *Acta Crystallogr.* D**60**(12), 2126 – 2132.

Foadi, J., Aller, P., Alguel, Y., Cameron, A., Axford, D., Owen, R. L., Wes Armour, d. D. G. W., Iwataa, S. & Evans, G. (2013). *Acta Crystallogr.* D**69**, 1617 – 1632.

Foley, B. J., Drozd, A. M., Bollard, M. T., Laspina, D., Podobedov, N., Zeniou, N., Rao, A. S., Andi, B., Jackimowicz, R., Sweet, R. M. *et al.* (2016). *Journal of laboratory automation*, **21**(1), 115 – 124.

Fromme, P. (2015). *Nature chemical biology*, **11**(12), 895 – 899.

Giordano, R., Leal, R. M., Bourenkov, G. P., McSweeney, S. & Popov, A. N. (2012). *Acta Crystallographica Section D: Biological Crystallography*, **68**(6), 649 – 658.

Hasegawa, K., Yamashita, K., Murai, T., Nuemket, N., Hirata, K., Ueno, G., Ago, H., Nakatsu, T., Kumasaka, T. & Yamamoto, M. (2017). *Journal of Synchrotron Radiation*, **24**(1).

Liu, Q. & Hendrickson, W. (2013). *Acta Crystallogr.* D**69**(7), 1314 – 1332.

Neutze, R., Wouts, R., van der Spoel, D., Weckert, E. & Hajdu, J. (2000). *Nature*, **406**(6797), 752 – 757.

Roessler, C. G., Agarwal, R., Allaire, M., Alonso-Mori, R., Andi, B., Bachega, J. F., Bommer, M., Brewster, A. S., Browne, M. C., Chatterjee, R. *et al.* (2016). *Structure*, **24**(4), 631 – 640.

Rossmann, M. G. (2014). *IUCrJ*, **1**(2), 84 – 86.

Soares, A. S., Engel, M. A., Stearns, R., Datwani, S., Olechno, J., Ellson, R., Skinner, J. M., Allaire, M. & Orville, A. M. (2011). *Biochemistry*, **50**(21), 4399 – 4401.

Standfuss, J. & Spence, J. (2017). *IUCrJ*, **4**(Pt 2), 100.

Winn, M. D., Murshudov, G. N. & Papiz, M. Z. (2003). *Methods in enzymology*, **374**, 300 – 321.

Yamashita, K. *et al.*, (2017*a*).
URL: *https://github.com/keitaroyam/yamtbx/blob/master/doc/kamoen.md*

Yamashita, K. *et al.* (2017*b*). *in preparation*.

Yin, X., Scalia, A., Leroy, L., Cuttitta, C. M., Polizzo, G. M., Ericson, D. L., Roessler, C. G., Campos, O., Ma, M. Y., Agarwal, R. *et al.* (2014). *Acta Crystallogr.* D**70**(5), 1177 – 1189.

Zeldin, O. B., Brewster, A. S., Hattne, J., Uervirojnangkoorn, M., Lyubimov, A. Y., Zhou, Q., Zhao, M., Weis, W. I., Sauter, N. K. & Brunger, A. T. (2015). *Acta Crystallogr.* D**71**(2), 352 – 356.

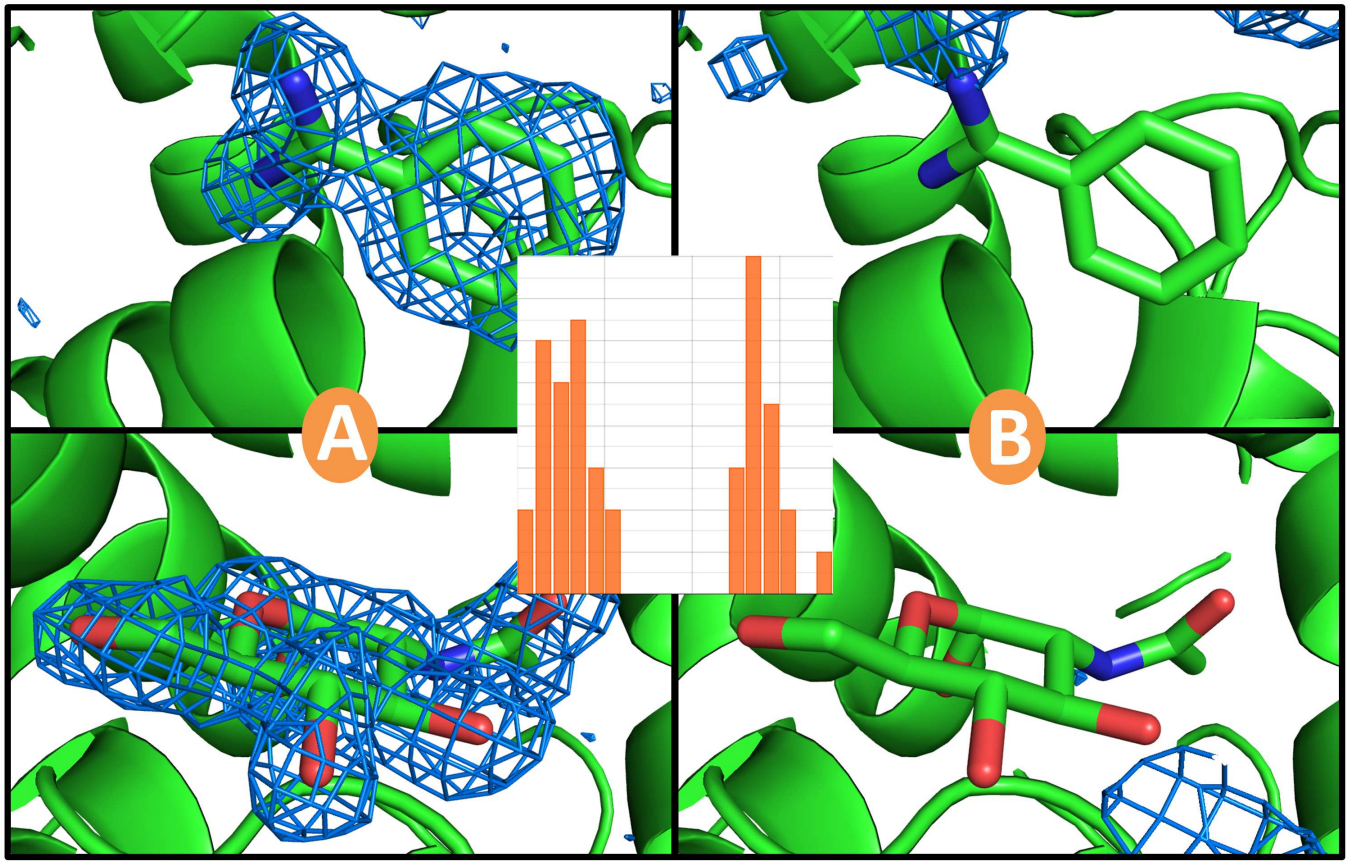

Figure 1

Electron density maps calculated after two-way clustering of diffraction data obtained from micro-meshes that contained a mixture of native crystals (no ligands; figure 1B) and double bound crystals (benzamadine + NAG; figure 1A). The omit difference maps are contoured at 1.5 sigma in the region expected to contain benzamadine (top) and NAG (bottom). The historgram cluster on the left represents the unit cell dimensions of the cluster of crystal data sets that yielded the omit difference map shown in A. Similarly, the histogram cluster on the right represents the unit cell dimensions of the cluster of crystal data shown in B. Clearly the clustering algorithm was able to accurately partition the data for this simple two-way split.

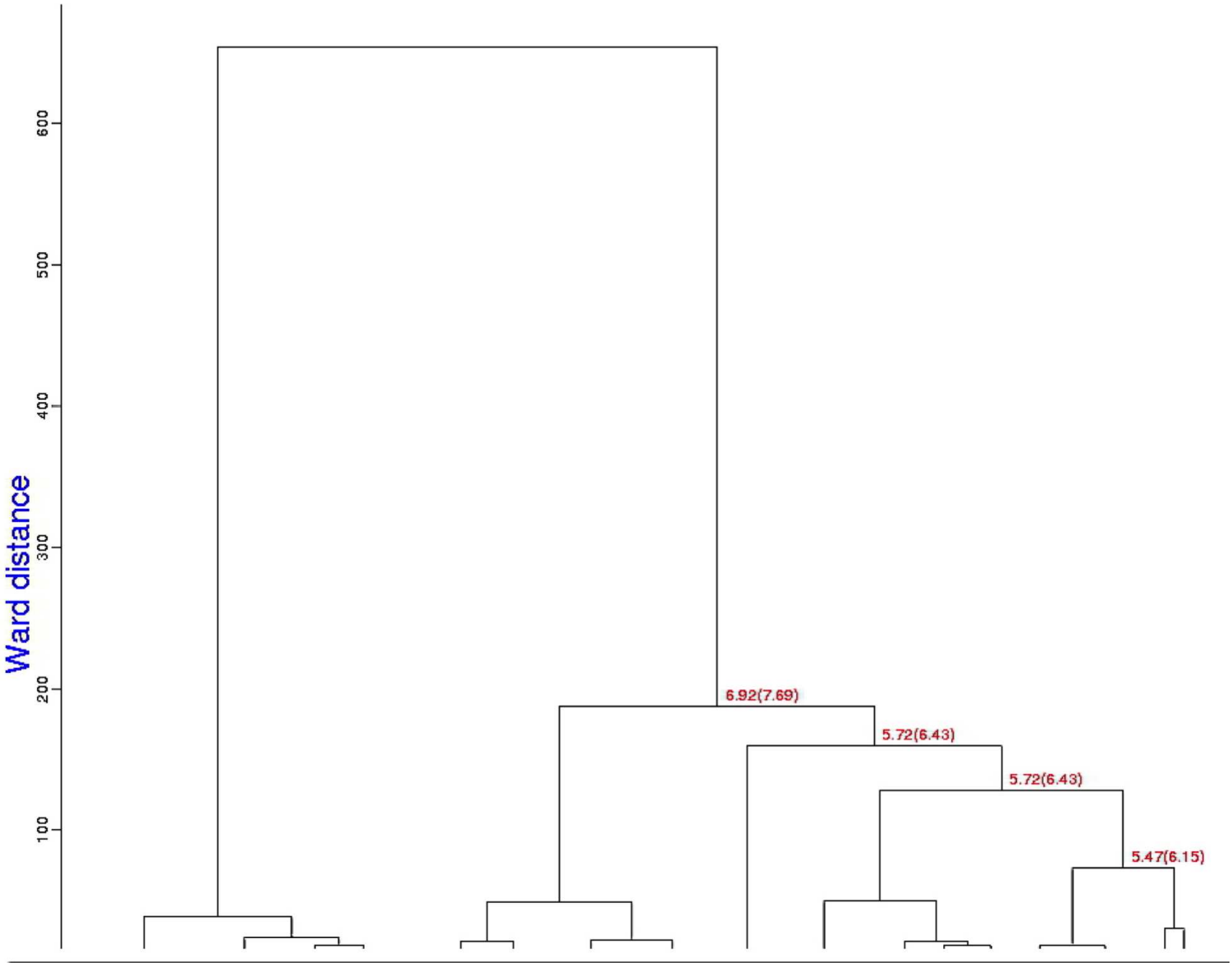

Figure 2
This dendrogram on presents the top levels of Blend clustering using the original Blend cell-parameters Euclidean distance function.

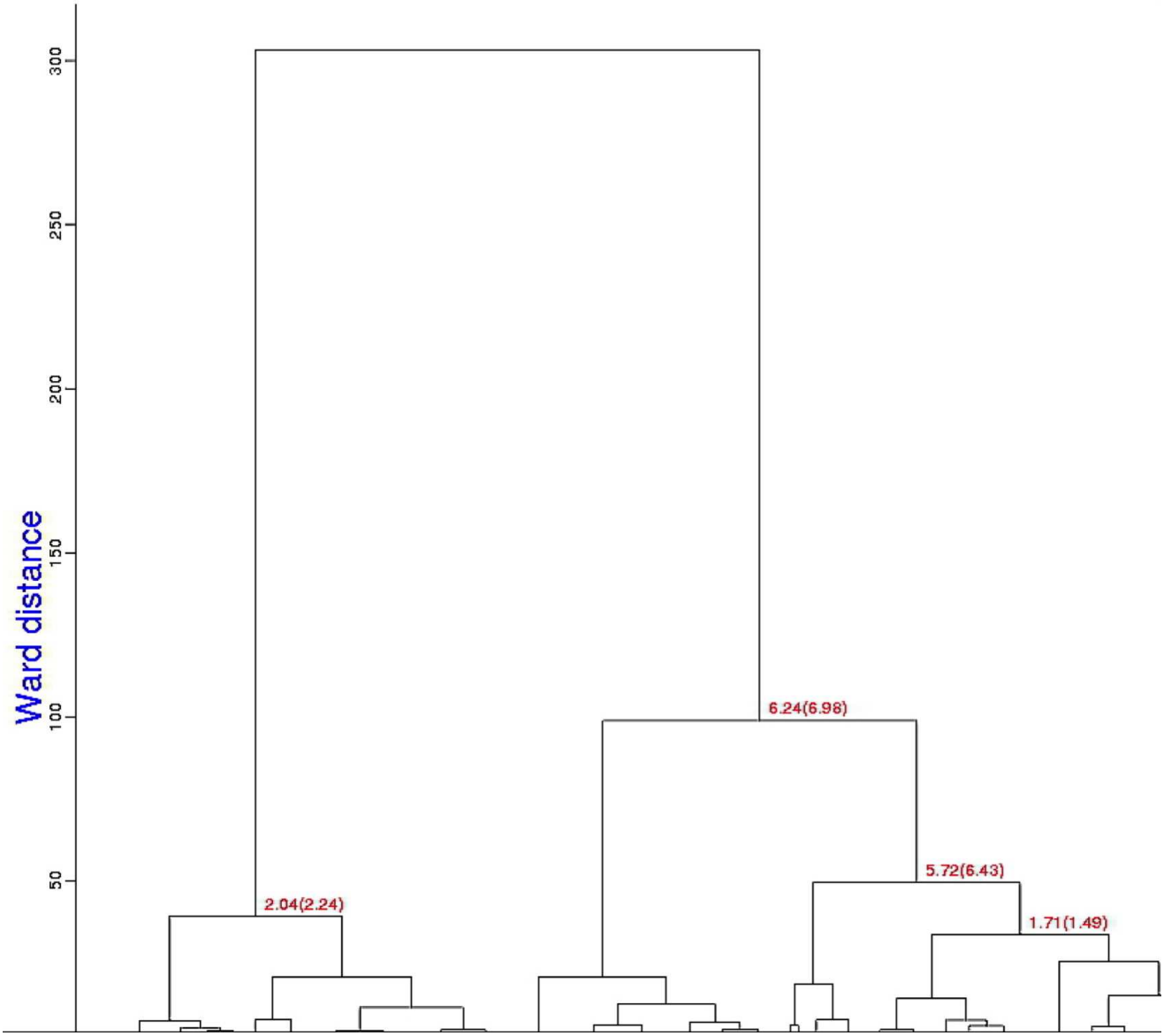

Figure 3
This dendrogram presents the top levels of Blend clustering using the more sensitive Andrews-Bernstein Niggli-Cone-Distance (NCDist) algorithm.

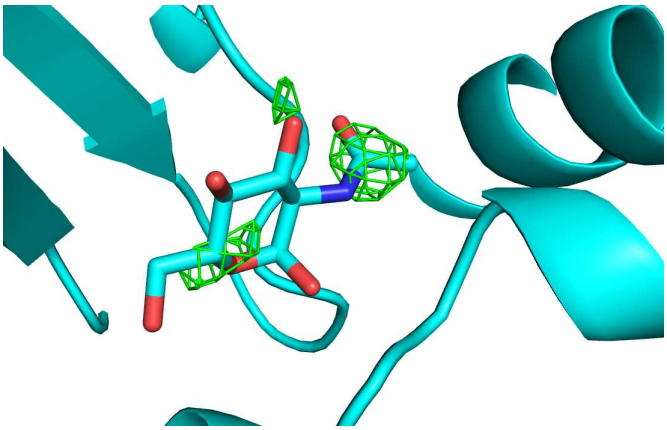

Figure 4
Omit difference maps of the NAG site in cluster 28 of a two-stage clustering with KAMO using cell parameters and NCDist to get to at least 10% completeness and then CC clustering with SFDist.

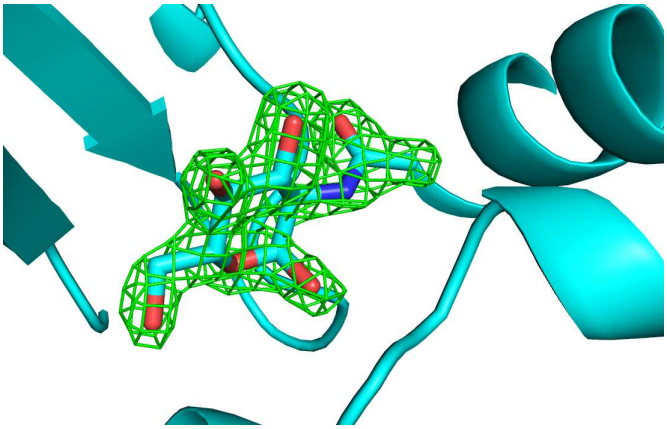

Figure 5
Omit difference maps of the NAG site in cluster 43 of a two-stage clustering with KAMO using cell parameters and NCDist to get to at least 10% completeness and then CC clustering with SFDist.

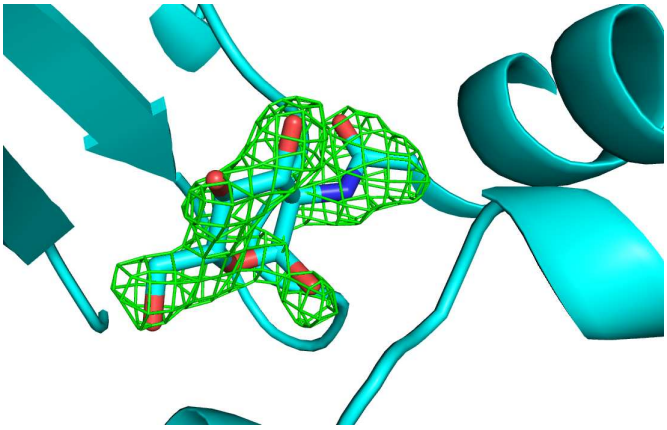

Figure 6
Omit difference maps of the NAG site in cluster 62 of a two-stage clustering with KAMO using cell parameters and NCDist to get to at least 10% completeness and then CC clustering with SFDist.

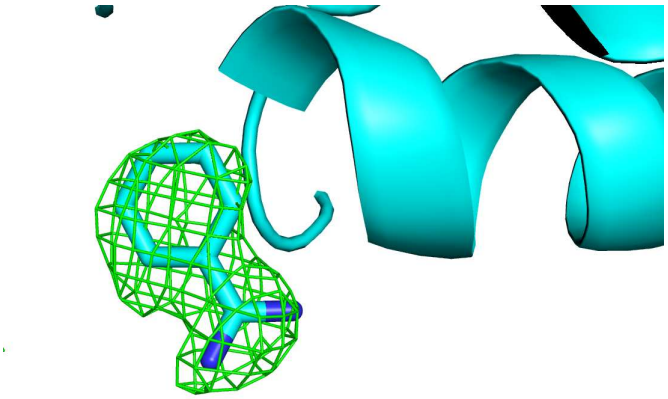

Figure 7
Omit difference map of the Benzamidine site in cluster 28 of a two-stage clustering with KAMO using cell parameters and NCDist to get to at least 10% completeness and then CC clustering with SFDist.

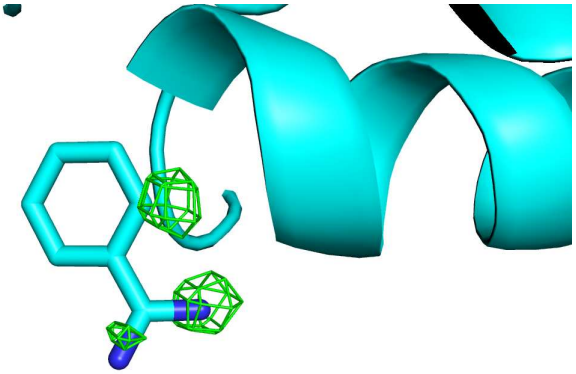

Figure 8
Omit difference map of the Benzamidine site in cluster 43 of a two-stage clustering with KAMO using cell parameters and NCDist to get to at least 10% completeness and then CC clustering with SFDist.

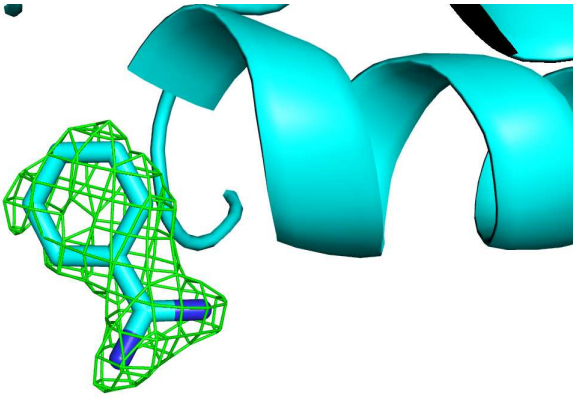

Figure 9
Omit difference map of the Benzamidine site in cluster 62 of a two-stage clustering with KAMO using cell parameters and NCDist to get to at least 10% completeness and then CC clustering with SFDist.

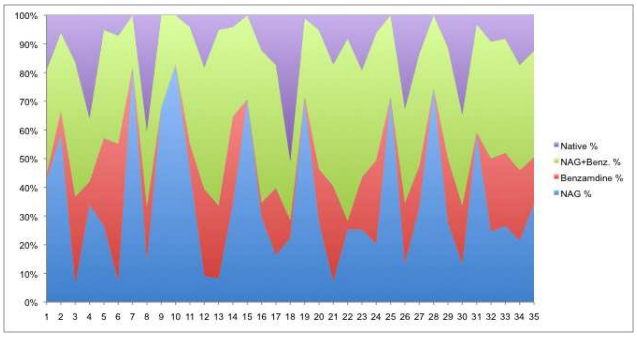

Figure 10
Color charts of the 35 largest dataset clusters for the NCDist clustering. From top to bottom the color blocks are the native soak, the NAG+benzamadine soak, the benzamadine soak and the NAG soak. If one color reaches nearly from the bottom to the top at a given position, that cluster is a nearly pure species.

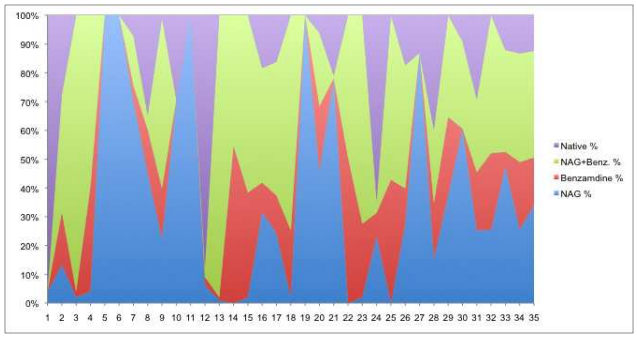

Figure 11
Color charts of the 35 largest dataset clusters for the SFDist clustering. From top to bottom the color blocks are the native soak, the NAG+benzamadine soak, the benzamadine soak and the NAG soak. If one color reaches nearly from the bottom to the top at a given position, that cluster is a nearly pure species. That is the case for each soak on the left end of this SFDist chart.