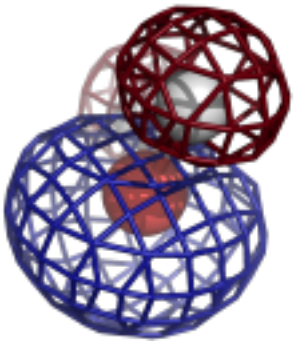
VOLUME FIVE



COMPUTATIONAL CRYSTALLOGRAPHY NEWSLETTER

JULY MMXIV

FAB ELBOW, MAP CONNECTIVITY, CDL

Table of Contents

Editor

Nigel W. Moriarty, NWMoriarty@LBL.Gov

PHENIX News

New programs

Guided Ligand Replacement

Adding ligands to an *apo* structure can now be performed using a similar ligand fit into a similar protein. The guiding protein-ligand complex can be sourced from the RCSB via web services, from a local file directory or supplied by the user.

New features

Conformation Dependent Library

The CDL has been added to *Phenix* and tested extensively. High resolution structures were used to determine restraints of the protein backbone based on the φ,ψ torsion angles. The parameter `cdl=True` activates its use. More details can found in the article on page 42 of this issue.

Automatic linking

All programs based on model geometry in *Phenix* (e.g. `phenix.real_space_refine`, `phenix.geometry_minimization` and `phenix.refine`) have moved to automatic determination of metal coordination and covalent bonding. Currently not the default, the parameter `link_all=True` will activate the linking based on residue type and distance cutoff. Finer control can be achieved by using the linking type switches (`link_metals`, `link_rna_dna`, `link_residues` and `link_carbohydrates`) and the corresponding distance cutoff parameters. The procedure covers covalently linked ligands and does simple validation of the carbohydrate polymers.

Crystallographic meetings and workshops

5th Murnau conference on Structural Biology – Focus Topic: Signal Transduction
Sept 10-13, 2014

Location: Murnau am Staffelsee, Germany
www.murnauconference.de/2014/index.html

Expert advice

Fitting Tip #8 – Acetyl Groups are like Peptides: Planar and *trans*

Jane Richardson, Jeff Headd, and Nigel W. Moriarty,
Duke University, J&J & LBNL

Acetylation is quite a common protein modification. Disturbingly, however, a recent study (Genshaft 2013) found that even for crystal structures involving the histone acetyl-lysine (3-letter code "ALY") so important in control of gene expression over 25% were modeled with extremely implausible high-energy conformations. (If it makes you feel better, over 50% were implausible in NMR models.) Clearly our automated tools have some catching up. But in the meantime it is very easy for you to apply user expertise: as we will show, acetyl groups have the same chemistry and conformational limitations as the familiar peptide bond, so just make sure they are planar and *trans* in your deposited structure. Figure 1 shows front and side views of an acetyl group at 1.2Å resolution, where the density clearly distinguishes between the carbon and the oxygen. The *trans* dihedral between the Cα-analog atoms is marked by a green line. Note that the N–H and C–O bonds in the peptide plane are also *trans* to each other and the planarity is quite exact.

Acetyl-lysine

For 109 acetyl examples in the Cambridge Structural Database (CSD; Allen 2002) of high-resolution small molecule crystal structures, every one was planar and *trans* (ω = 180° +/- 4°; Genshaft 2013). This agrees with energy calculations for *cis* versus *trans* acetyl and the high barrier between them, and the same regularity is seen for well-ordered high-resolution examples in protein crystals, such as the figures here. Genshaft et al. saw no significant correlation of acetyl-lysine error rate with refinement software, or even with resolution (since high-B examples are not adequately constrained by density even at high resolution). However in our informal survey, most multi-acetyl protein structures were either nearly random or all correct. Presumably this means that so far none of the automated software gets acetyls right reliably, but many crystallographers do understand acetyl geometry.

Acetyl N-termini

Acetyl N-termini ("ACE") are not usually as functionally important as acetyl-lysines, but they are an order of magnitude more common. Figure 2 shows map and model for a 0.8-Å resolution acetyl N-terminus, where the analogy with a standard peptide linkage is even more obvious than in figure 1. For crystal structures with a modeled ACE group, Genshaft et al. found 17% with highly strained conformations.

Confirmation by all-atom contacts

When there are other atoms nearby, then looking at the all-atom contacts for H-bonds and clashes is another very helpful diagnostic, or confirmation, of the correct orientation. Within the validation section of the Phenix GUI (Echols 2012), or on the MolProbity web site (Chen 2010), you can do this

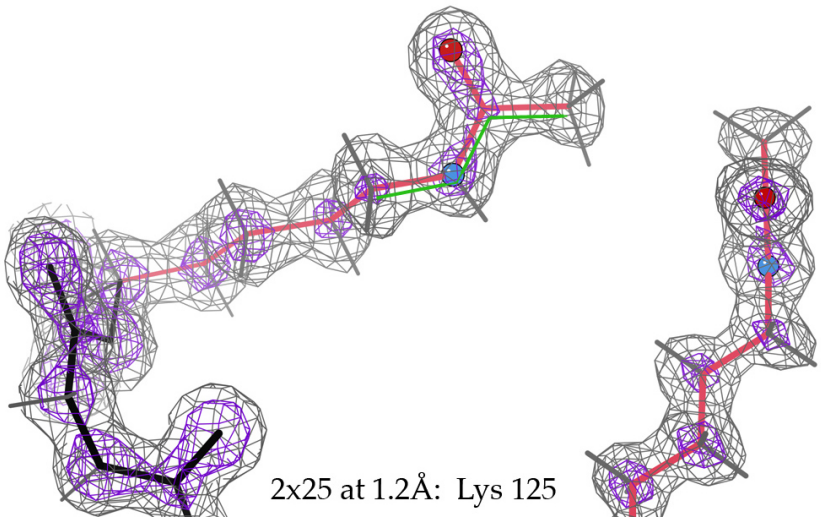


Figure 1: An N-acetyl-lysine at high resolution: (A) front view, showing that the C–N–C-C dihedral (green line) is trans, confirmed by higher density at the O than at the methyl; (B) side view, to show the tight planarity (2x25; Lammers 2010)

easily by launching the multi-criterion kinemage in KiNG (Chen 2009). For the somewhat-related case of non-covalent acetate ligands ("ACT"), all-atom contacts are the best way of correctly fitting the triangular electron density at mid resolutions (see figure 3).

Issues in refinement

In response to this problem, we have re-examined the handling of acetyl restraints in Phenix, which historically has produced almost no highly twisted cases (defined in Genshaft 2013 as >30° from planarity, and occurring from other programs), but a fair number of *cis* (0° +/- 30°). This is because Phenix has always had a restraint to planarity. However, since early 2013 it's had a strong term preferring *trans* for both ALY and the link to ACE, but no way to fix an initial fit of *cis*. We are working on a system that would check for highly strained torsions, including acetyls, and try flipping them.

Conclusion

Acetyl groups on lysine side-chains or on N-termini are fairly often modeled wrong, either by people or by software. But all you need to remember to get them right is that they have the same chemistry and conformational properties as a peptide -- so, as you know, they are very close to planar and essentially always *trans*.

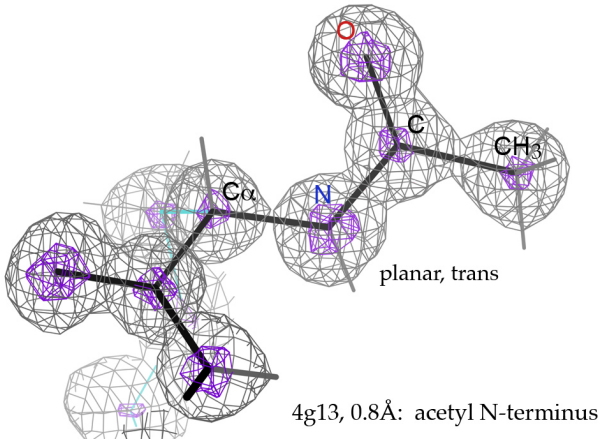


Figure 2: An acetyl N-terminus at high resolution, model and electron density (4g13; Gessmann 2012).

References

Allen FH (2002) "The Cambridge Structural Database: a quarter of a million crystal structures and rising", Acta Crystallogr. B58: 380-388

Chen VB, Davis IW, Richardson DC (2009) "KiNG (Kinemage, Next Generation: A versatile interactive molecular and scientific visualization program", Protein Sci 18:2403-240

Chen VB, Arendall WB III, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010) "MolProbity: all-atom structure validation for macromolecular crystallography", Acta Crystallogr D66:12-21

Echols N, Grosse-Kunstleve RW, Afonine PV, Bunkoczi G, Chen VB, Headd JJ, McCoy AJ, Moriarty NW, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Adams PD (2012) "Graphical tools for macromolecular crystallography in Phenix", *J Applied Crystallogr* **45**: 581-586

Genshaft A, Moser JA, D'Antonio EL, Bowman CM, Christiansen DW (2013) "Energetically unfavorable amide conformations for N6-acetyllysine side chains in refined protein structures", *Proteins: Struct Funct Bioinf* **81**: 1051-1057

Gessmann R, Axford D, Evans G, Bruckner H, Petratos K (2012) "The crystal structure of samarosporin at atomic resolution", J Pept Sci 18: 678-684

Lammers M, Neumann H, Chin JW, James LC (2010) "Acetylation regulates cyclophilin A catalysis, immunosuppression, and HIV isomerization", *Nat Chem Biol* **6**: 331-337

Symersky J, Li S, Carson M, Luo M (2003) "Structural genomics of Caenorhabditis elegans: triosephosphate isomerase", *Proteins: Struct Funct Bioinf* **51**: 484-486

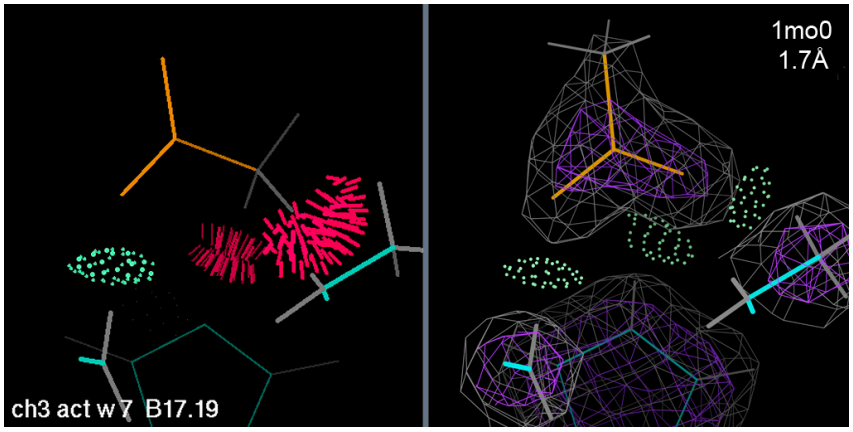


Figure 3: A solvent acetate, all-atom contacts before and after refitting (1mo0; Symersky 2003).

Connectivity analysis tools in *CCTBX*

Oleg V. Sobolev[a], Pavel V. Afonine[a] and Paul D. Adams[a,b]

[a]*Lawrence Berkeley National Laboratory, Berkeley, CA 94720*

[b]*Department of Bioengineering, University of California at Berkeley, Berkeley, CA 94720*

Correspondence email: OSobolev@lbl.gov

Introduction

Connected regions in crystallographic maps defined at a given contouring level are an important feature of the map and map analysis. Studies of this feature may have various uses in crystallography: from *ab initio* phasing procedures (Lunin *et. al.,* 2000), various methods of map improvement and interpretation to analysis of connectivity of reflection data in reciprocal space (Urzhumtseva & Urzhumtsev, 2011). Robust implementation of connectivity analysis algorithm (Lunina *et al.,* 2003) in *CCTBX* (Grosse-Kunstleve *et al.* 2002) is presented.

A crystallographic map is a three-dimensional array of numbers representing the contents of a unit cell or a smaller region (asymmetric unit or a box). These, for instance, may be a distribution of electron or nuclear density, Fourier maps, solvent or other masks. One can contour the map at a particular level and study the regions that are above this level. This is the usual way for Fourier map interpretation during manual model building procedures. One can think of connected regions as peaks (also sometimes referred to as 'blobs') in a map. In the present work we report the implementation of an established connectivity analysis algorithm and extend it to determine the volume of all connected regions, coordinates, and the value of the maximum point for each region.

Implementation

The connectivity algorithm is implemented as a C++ module, `maptbx/connectivity.h`, in *CCTBX* as a separate class. Corresponding functionality is available from Python via the boost.python bindings (Abrahams & Grosse-Kunstleve, 2003) in `maptbx/boost_python/maptbx_ext.cpp`.

The runtime of the algorithm linearly scales with the number of grid points of the map. Typical runtimes are 1.85 seconds for a large map of size 400*400*400 points and 0.03 seconds for medium map of size 100*100*100 points. Calculation of the region volume, values and coordinates of maximum point in each region during the analysis can be performed at no significant additional computational cost.

The definition of connected region is based on the definition of neighbors for a particular point. In the present implementation we use six neighbors for a point with `(x,y,z)` coordinates: `(x-1,y,z)`, `(x+1,y,z)`, `(x,y-1,z)`, `(x,y+1,z)`, `(x,y,z-1)`, `(x,y,z+1)`. Alternative definitions of neighbor points may use 18 neighbor points (varying 2 of 3 coordinates) or 26 neighbor points (varying all 3 coordinates). We believe that implementation of 18- or 26-point neighbor scheme is unnecessary at this point.

Examples

1. Basics

The following code snippet illustrates the basic manipulations with the connectivity object (`co` in what follows): analysis of connected regions on the map and obtaining supplementary data (volumes, coordinates of maxima and values of maxima). To instantiate a connectivity object one needs to pass the map data and the threshold level to its constructor (see schema 1). In this example we obtained four arrays from connectivity object:

- `result_of_connectivity_analysis` – a 3D integer array of the same dimension as `map_data` filled with numbers 0, 1, 2, … N, that enumerate the N connected regions. Each grid point contains 0 if the input map values

```
>>> co = maptbx.connectivity(map_data=map_data, threshold=100)
>>> result_of_connectivity_analysis = co.result()
>>> region_volumes = list(co.regions())
>>> print region_volumes
[975696, 12152, 12152]
>>> coordinates_of_maximum_points = list(co.maximum_coors())
>>> print coordinates_of_maximum_points
[(74, 62, 62), (20, 20, 20), (60, 60, 60)]
>>> values_of_maximum_points = list(co.maximum_values())
>>> print values_of_maximum_points
[99.8855747054916, 1569.9055625594979, 1569.9055625594979]
```

Schema 1: Initialization and basic commands of connectivity analysis object.

```
>>> co = maptbx.connectivity(map_data=cmap, threshold=5)
>>> resulting_mask = co.volume_cutoff_mask(volume_cutoff=10)
```

Schema 2: Commands to obtain mask for filtering out peaks with volume smaller than 10 on a threshold equals 5.

are less than the threshold value and a non-zero value for connected regions. Each connected region is assigned its own unique integer number.

- `region_volumes` – 1D integer array of length N+1. Each element, *i*, contains the volume of *i*-th region. The zeroth element contains the volume of the under-threshold region.
- `coordinates_of_maximum_points` – 1D array of length N+1 of tuples. Similarly to volumes, contains the coordinates of maximum point.
- `values_of_maximum_points` – 1D float array of length N+1, contains the values of corresponding maximum points.

As noted previously the input data to the connectivity analysis procedure is a three-dimensional array with numbers. Therefore coordinates of the maximum points are provided as array indexes of these grid points. The volume of the region is the number of grid points with values greater than the specified threshold. The `result_of_connectivity_analysis` may be used further to obtain various kinds of masks.

2. Volume cutoff

An example of useful application of connectivity analysis procedure is to eliminate small map peaks at a particular threshold level (Afonine *et al.*, 2014, in preparation) is illustrated in schema 2. Here we obtain the `co` object the same way as in the first example, and then call the `volume_cutoff_mask` method with the `volume_cutoff` parameter to provide the smallest size of peaks that should be kept in the `resulting_mask`. A binary 3D mask with 0 where the value in `cmap` are less than the threshold and where the volume of connected regions are less or equal to `volume_cutoff` and 1 everywhere else. It should be stressed that in this case we obtain a binary mask of 0 and 1 although the connectivity analysis results are still available via `co.result()`. Finally to apply the filtration to the data one has to multiply map data `cmap` by `resulting_mask`.

3. Noise elimination based on analysis of two cutoff levels.

A more elaborate noise elimination procedure is used in the "Feature enhanced map" tool implemented in *Phenix* package (Adams *et al.*, 2010) as the *phenix.fem* program (Afonine et al, 2014, in preparation) that analyzes connectivity regions at two cutoff levels simultaneously.

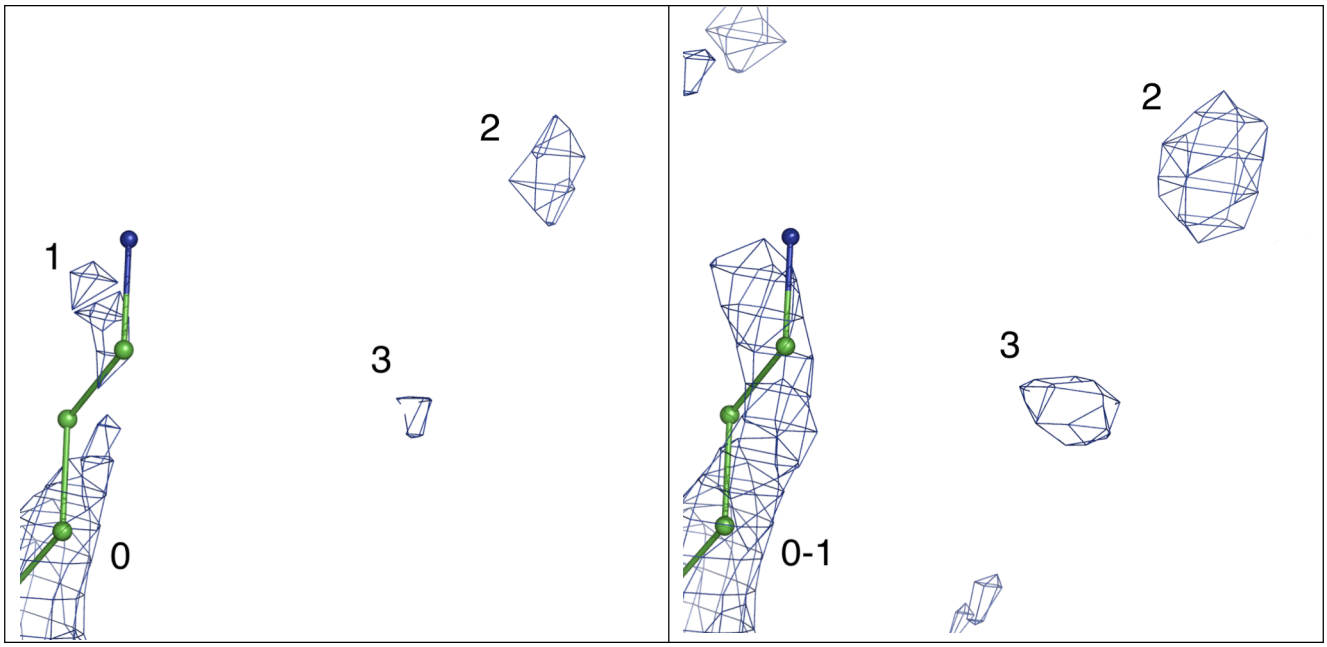


Figure 1: Schematic illustration of noise peaks elimination. Small-volume peaks at threshold *t* (left; peaks 1,2,3) and *t*-δ (right; peaks 2 and 3). Peaks 2 and 3 will be eliminated.

```
>>> co1 = maptbx.connectivity(map_data=cmap, threshold=25)
>>> co2 = maptbx.connectivity(map_data=cmap, threshold=22)
>>> resulting_mask = co2.noise_elimination_two_cutoffs(
      connectivity_object_at_t1=co1,
      elimination_volume_threshold_at_t1=12,
      zero_all_interblob_region=True)
```

Schema 3: Commands to obtain mask for noise filtering using two cutoffs and volume estimation of good peaks on the first cutoff.

Map peaks that have smaller volume than the typical volume of a reliably placed atom (e.g. water) at some threshold level *t* can be considered as noise peaks and ideally should be eliminated. Moreover, removing them at a lower level, *t*-δ, could be even better because more noise peaks would be removed. Nevertheless at the *t*-δ cutoff level some peaks previously scheduled for deletion could merge with good peaks. Such peaks should be retained. This procedure is schematically illustrated in figure 1. Figure 1(a) shows peaks 0,1,2,3 at the *t* cutoff level. Naïvely, peaks 1,2,3 could be deleted based on their small volume. However, figure 1(b) shows the same map on *t*-δ cutoff level where peak 1 has merged with molecule-related (not scheduled for deletion) peak 0 and should therefore be retained. Since peaks 2 and 3 did not merge with any region not scheduled for deletion, they will be zeroed using the shape as they appear at *t*-δ level. Particular threshold levels can be determined based on a trial-and-error approach to achieve best performance of the procedure. Information between peaks below the *t*-δ threshold level may be zeroed or preserved at the values in the original map. The code to do this filtration is shown in schema 3.

Here we created two connectivity objects with the same map data but with different threshold levels for determining connected regions. Then we call the `noise_elimination_two_cutoffs` function of the connectivity object with the lower threshold, provide the first connectivity object (with the higher threshold), volume of blobs that should be considered as good on

threshold t and a boolean parameter to specify whether to keep (`zero_all_interblob_region =False`) or mask out (`zero_all_interblob_region=True`) data between peaks at the t-δ threshold. The result is a 3D mask with the same dimensions as the original map. It is filled with 0 for map points that should be deleted and 1 for map points that should be retained. To apply the filtration one need to multiply the initial map data `cmap` by `resulting_mask`.

Conclusion

We have implemented a fast and reliable search of connected regions along with determination of their volume, values and position of maximum point at an arbitrary cutoff level. This module could provide a strong foundation for the development of various map-processing algorithms. The tools described here are available as a part of *CCTBX*.

References

Abrahams D. and Grosse-Kunstleve R.W., C/C++ Users Journal, July, 2003.

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. and Zwart, P. H. (2010). *Acta Cryst.* D**66**, 213-221.

Grosse-Kunstleve, R.W., Sauter, N.K., Moriarty, N.W. and Adams, P.D. (2002). *J. Appl. Cryst.* **35**, 126-136.

Lunin, V. Yu., Lunina, N. L. & Urzhumtsev, A. G. (2000). Acta Cryst. A**56**, 375-382.

Lunina, N.L., Lunin, V.Y. & Urzhumtsev, A. (2003). Acta Cryst. D**59**. 1702-1755.

Urzhumtseva, L. & Urzhumtsev, A. (2011). *J. Appl. Cryst.* **44**. 865-872.

phenix.fab_elbow_angle: Fragment Antigen-Binding elbow angle calculation tool

Youval Dar[a], Pavel V. Afonine[a], Nigel W. Moriarty[a] and Paul D. Adams[a,b]

[a]*Lawrence Berkeley National Laboratory, Berkeley, CA 94720*

[b]*Department of Bioengineering, University of California at Berkeley, Berkeley, CA 94720*

Correspondence email: ydar@lbl.gov

Introduction

Antibody (Ab) used by white blood cell and the immune system to identify and fight antigens (Antibody generators) such as viruses and bacteria. They are large Y-shape proteins (figure 1). Fragment Antigen-Binding (Fab) are the tips of the Y-shape Ab allowing only a specific antigen to bind. The Fab is cleaved from the complete Ab to allow studying of the active portion without interference from other portions of the molecule. The Fab are composed of heavy and light chains, each with a constant and a variable regions. In a schematic illustration (figure 1.A) the constant and variable portions of the Fab are aligned at angle 180°. Figure 1.B shows a common situation, where the angle is not 180°. The angle between the variable and constant regions is called the Fab elbow angle. To find that angle, the light chain of each region is aligned onto the heavy chain by a rotation. The corresponding rotation axes are pseudo-dyad axes (imperfect dyad symmetry axes) and the Fab elbow angle is defined as the angle between these two axes. The elbow angle is an important characteristics of this class of protein structures and is typically reported in corresponding structural reports (Stanfield *et al.*, 2006). We have added a function to *CCTBX* (Grosse-Kunstleve *et al.*, 2002) that calculates Fab elbow angle.

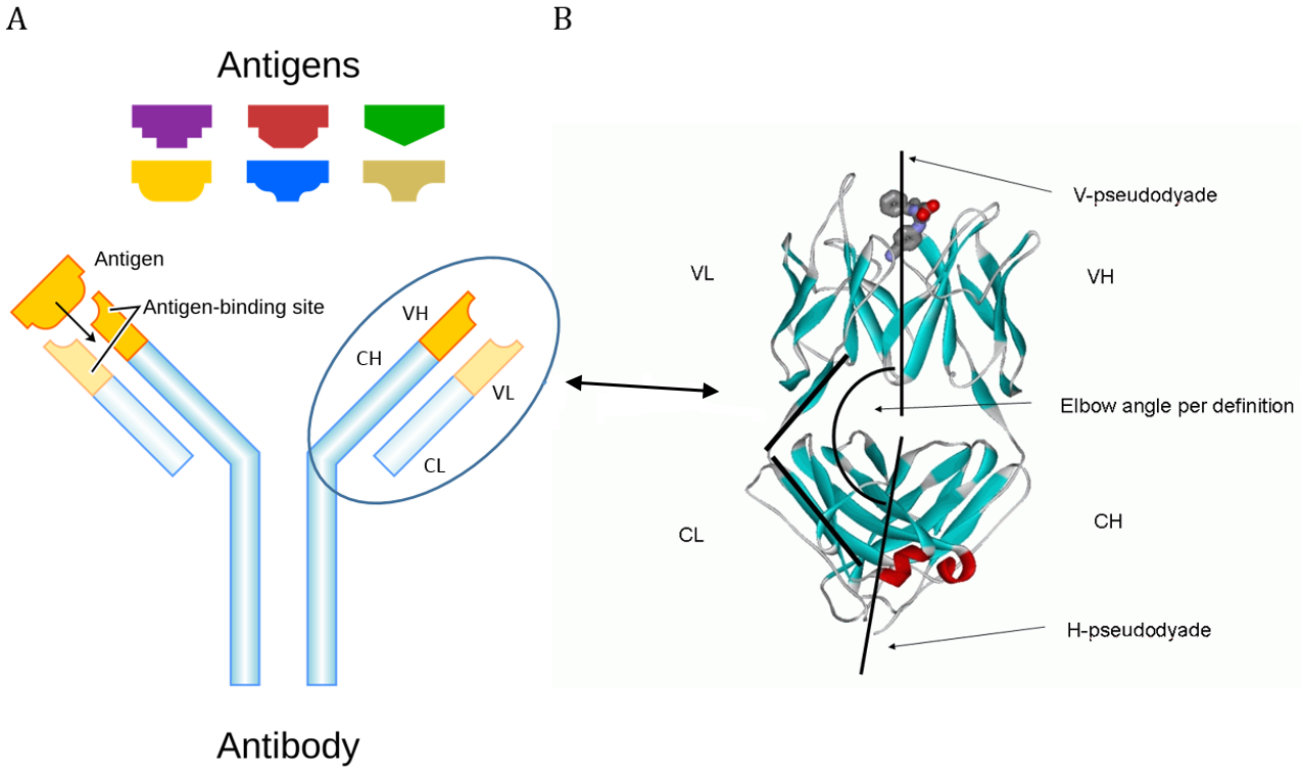


Figure 1: (A) Antigen, Wikipedia, Antigen, http://upload.wikimedia.org/wikipedia/commons/2/2d/Antibody.svg. By Fvasconcellos 19:03, 6 May 2007 (UTC) [Public domain], via Wikimedia Commons. (B) Fab angle from (Zemla A. 2005) (Web server addition). VL – Variable Light, VH – Variable Heavy, CL – Constant Light, CH – Constant Heavy.

Since the Fab is not the complete Ab its orientation with respect to the rest of the Ab is unknown. As a result, the angle between the two pseudo-dyad axes does not define a unique orientation between the variable portion of the Fab and rest of the Ab. This causes an ambiguity in the definition and interpretation of the Fab elbow angle. Factors affecting the Fab elbow angle determination are:

- Choice of limit residues (residues separating the variable and constant domains of the heavy and light chains).
- Selection of residues to align.

We used two methods to resolve the angle ambiguity in a way that matches published work (Stanfield *et al.*, 2006) and a similar function available in PyMol (Schrodinger, 2010; DeLano, 2002). The first method uses a known Fab structure as a reference while the second uses only the pseudo-dyad axes to determine the angle. Preferring not to rely on a particular structure as a reference, we have incorporated only the second method to *CCTBX*.

Method

Following (Stanfield *et al.*, 2006) we consider only the two dimensional ambiguity in elbow angle calculation, namely checking if the resulting angle is α or $(360-\alpha)$.

The program steps for finding the pseudo-dyad axes and the angle between them are as follows:

- Using *CCTBX* tools we get the rotation matrices that superpose light segments onto heavy ones.
- The atoms selected to be superposed are $C\alpha, C$ and N that do not have alternative locations. The selection strings for each are listed:
  - For the variable segment:
    chain *chain_ID* and resseq *1:Limit* and pepnames and (name ca or name n or name c) and altloc " "
  - For the constant segment:
    chain *chain_ID* and resseq *Limit+1:end* and pepnames and (name ca or name n or name c) and altloc " "
- The normalized eigenvectors, corresponding to eigenvalues of unity of those rotation matrices, represent the pseudo-dyad axes of the variable and the constant domains, $\hat{V}$ and $\hat{C}$ (Figure .B).
- $\cos(\alpha) = \hat{V} \cdot \hat{C}$ (α can be either between 90° and 180°, or between 180° and 270°).

Known protein as reference

Following the implementation described in (Stanfield *et al.*, 2006), we used PDB (Berman *et al.*, 2000; Bernstein *et al.*, 1977) protein *1bbd* as a reference using this procedure:

- Align Fab constant-heavy segments of a *1bbd* with the tested protein.
- Use the method described above and find the angle β between the two variable heavy parts.
- If $\alpha + \beta > 180°$ then choose the angle α or $(360-\alpha)$, whichever is larger than 180°, otherwise choose the smaller.

Geometrical method

Define $\hat{Z} = \hat{V} \times \hat{C}$ and $\hat{X} = \hat{C} \times \hat{Z}$ as described in figure 2, where $\hat{V}, \hat{C}$ are the pseudo-dyad axes of the Fab variable and constant portion: $\hat{X}, \hat{C}, \hat{Z}$ form a right-hand coordinate system. We can decide whether the Fab angle is α or $(360-\alpha)$ by comparing $\hat{V}$ with $\hat{X}$ or $\hat{Z}$. Because we are turning a three-dimensional problem to a two-dimensional problem, the choice is somewhat arbitrary without further refinement of the Fab angle definition. Figure 2.A suggests that comparing $\hat{V}$ with $\hat{X}$ may be a reasonable choice for resolving the angle ambiguity, while figure 2.B implies that comparing $\hat{V}$ with $\hat{Z}$ could be a better choice. The choice of reference axis $(\hat{X}, \hat{Z})$ can affect whether the angle we choose is larger or smaller than 180°. In *CCTBX* we used $\hat{Z}$ as a reference, to keep in line with PyMol and (Stanfield *et al.*, 2006) test cases results.

Results

Table 1 compares the elbow angle calculation between *CCTBX*, PyMol elbow angle calculation functionality and values reported in (Stanfield *et al.*, 2006). Table 2 shows how a

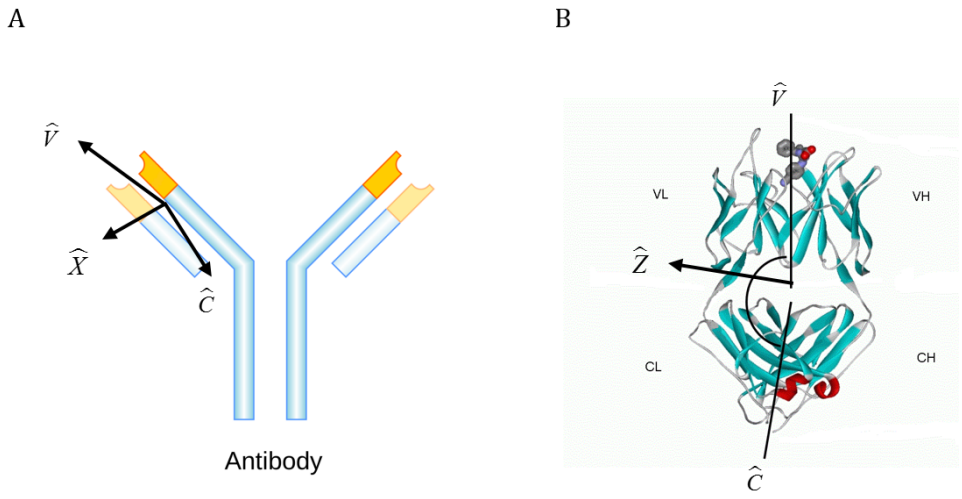


Figure 2: Geometric method. Define coordinate system to describe Antibody. $\hat{V}$ and $\hat{C}$ are the pseudo-dyad axes of the variable and constant segments respectively. In (A) $\hat{V}$ and $\hat{C}$ are drawn in a slight angle to demonstrate the geometrical approach.

Table 1: Compare the Fab angle (degrees) for five structures as reported by (Stanfield et al 2006) in blue, PyMol elbow angle in white and is CCTBX Fab angle in yellow. L and L.H are the limit light and limit heavy residues, the residue number separating the constant and variable portions of the chains.

1bbd			7fab			1dba			1plg			1nl0		
L.L	L.H	angle	L.L	L.H	angle	L.L	L.H	angle	L.L	L.H	angle	L.L	L.H	angle
114	118	127	104	117	132	107	113	183	112	117	190	107	113	220
114	118	125	104	117	126	107	113	176	112	117	189	107	113	221
114	118	133	104	117	123	107	113	172	112	117	187	107	113	214

small difference in the choice of the limit residues. For example, ±2 residues can affect the angle by up to 4° for *1nl0* when Limit-Heavy (L.H.) changes from 113 to 114. In other cases the angle can be insensitive to the change of four residues in L.H., see *1dba*. When the angle is close to 180°, the choice of limit residue can affect the decision between α or $(360 - \alpha)$ see *1dba* when L.L. is 109, changing L.H from 112 to 113 causes the angle to change from 193° to 172°

Additional source of differences between the elbow angle calculation results might stem from different superposition tools used to overlay the Variable-Light (VL) onto the Variable-Heavy (VH) and the Constant-Light (CL) onto the Constant-Heavy (CH) rotation matrices.

Implementation in *CCTBX*

The Fab elbow angle calculation routine assumes labels for the heavy and light chains to be "*H*" and "*L*", and the limits (residue numbers separating the constant and variable segments) to be *113* and *107* based on Kabat and Chothia numbering (Wu & Kabat, 1970; Chothia & Lesk, 1987).

Command line implementation

phenix.fab_elbow_angle pdb_file_name [light=L] [heavy=H] [limit_l=107] [limit_h=113]

CCTBX

from mmtbx.utils.fab_elbow_angle import fab_elbow_angle
fab = fab_elbow_angle(pdb_hierarchy= pdb_hierarchy,limit_light=107,limit_heavy=113)
fab_angle = fab.fab_elbow_angle

Conclusion

Functionality to calculate the Fab elbow angle was added to *CCTBX*. For a set of *1bbd,7fab, 1dba, 1plg* and *1nl0* structures it yields results consistently similar to those produced by PyMol and (Stanfield *et al.*, 2006).

Table 2: Exploring the effect of different limits-heavy and limit-light choices relative to the limits in Table 1

		1bbd	7fab	1dba	1plg	1nl0
ΔL.L	ΔL.H	angle	angle	angle	angle	angle
-2	-1	126	126	172	187	214
-2	0	126	126	172	187	214
-2	1	126	126	172	187	214
-2	2	126	126	172	187	214
-1	-2	133	123	172	187	214
-1	0	126	126	172	187	214
-1	1	126	126	190	187	214
-1	2	126	126	172	187	214
0	-2	133	123	172	195	210
0	-1	133	123	172	187	214
0	1	126	126	172	187	214
0	2	126	126	172	187	214
1	-2	133	123	197	195	210
1	-1	133	123	172	195	210
1	0	133	123	180	187	214
1	2	126	126	172	187	214
2	-2	123	119	185	195	210
2	-1	133	123	193	195	210
2	0	133	123	172	195	210
2	1	131	123	172	187	214

References

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.

Chothia, C. & Lesk, A. M. (1987). *Journal of Molecular Biology* **196**, 901-917.

DeLano, W. L. (2002). *The PyMOL Molecular Graphics System,* http://www.pymol.org.

Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. (2002). *J Appl Crystallogr* **35**, 126-136.

Schrodinger, LLC (2010).

Stanfield, R. L., Zemla, A., Wilson, I. A. & Rupp, B. (2006). *J Mol Biol* **357**, 1566-1574.

Wu, T. T. & Kabat, E. A. (1970). *J Exp Med* **132**, 211-&.

Details of the Conformation-Dependent Library

Nigel W. Moriarty[a], Paul D. Adams[a,b] and P. Andrew Karplus[c]

[a]*Physical Biosciences Department, Lawrence Berkeley National Laboratory, Berkeley, CA 94720 USA*
[b]*Department of Bioengineering, University of California Berkeley, Berkeley, CA, 94720 USA*
[c]*Department of Biochemistry and Biophysics, Oregon State University, Corvallis, OR 97331 USA*

Correspondence email: NWMoriarty@lbl.gov

Introduction

Crystallographic refinement of protein models is generally performed using a geometry restraints library based on the Engh and Huber work (Engh and Huber 1991; Engh and Huber 2001). This is a single value library (SVL) in the sense that the internal coordinates for each type (based on atom names) are restrained to a single value regardless of the molecular environment. Clearly, not all bonds between atom name pairs are given the same restraint values for every amino acid residue as shown in table 1. Indeed, alanine, for example, has different values for some of its restraints because its side chain terminates at Cβ. The only bond that is unchanged throughout the SVL is a C=O double bond which is set to 1.231 Å, while the only angle unchanged throughout is N–C=O which is set to 123.0 degrees. Certain features of the peptide linkage have special values for residues that precede a proline, as for instance, the angle C–N–Cα is adjusted by 0.9 degrees and the C–N peptide bond is elongated by 0.012 Å. An analysis of the estimated standard deviations (ESD) reveals similar characteristics.

A number of studies have shown that the SVL does not adequately represent the reality of the geometry of the polypeptide chains, but that the geometric parameters should instead be described by 'ideal geometry functions" explicitly allowing the ideal bond lengths and angles to vary with the conformation of the peptide. This was predicted over 30 years ago based on quantum mechanics calculations of small peptides (Schäfer et al. 1984; Klimkowski et al. 1985) and then verified in empirical observations of crystal structures of small peptides (Schäfer and Cao 1995)) and high-resolution crystal structures of proteins (Karplus 1996; Jiang et al. 1997). In 2008, it was proposed (Karplus et al. 2008) that the inadequacy of the SVL paradigm was responsible for the perplexing observation

Table 1: Backbone bond and angle restraints based on Engh & Huber (1991) specifying the residues associated with each restraint value and, in the case of the some of the peptide linking restraints, the class (preceding PRO or not preceding PRO) restraint values. Values are in Å and degrees.

Restraint	Value	Residues
Cα–N	1.466	PRO
	1.451	GLY
	1.458	ALA ARG ASN ASP CYS GLN GLU HIS ILE LEU LYS MET PHE SER THR TRP TYR VAL
C=O	1.231	all
C–Cα	1.516	GLY
	1.525	ALA ARG ASN ASP CYS GLN GLU HIS ILE LEU LYS MET PHE PRO SER THR TRP TYR VAL
C–N	1.341	preceding PRO
	1.329	not preceding PRO
Cα–Cβ	1.521	ALA
	1.540	ILE THR VAL
	1.530	ARG ASN ASP CYS GLN GLU HIS LEU LYS MET PHE PRO SER TRP TYR
C–N–Cα	122.6	preceding PRO
	121.7	not preceding PRO
Cα–C=O	119.0	PRO
	120.8	ALA ARG ASN ASP CYS GLN GLU GLY HIS ILE LEU LYS MET PHE SER THR TRP TYR VAL
N–Cα–C	112.2	ASN
	112.5	GLY
	111.8	PRO
	111.2	ALA ARG ASP CYS GLN GLU HIS ILE LEU LYS MET PHE SER THR TRP TYR VAL
N–C=O	123.0	all
Cα–C–N	116.9	preceding PRO
	116.2	not preceding PRO
C–Cα–Cβ	110.5	ALA
	109.1	ILE THR VAL
	110.1	ARG ASN ASP CYS GLN GLU HIS LEU LYS MET PHE PRO SER TRP TYR
N–Cα–Cβ	110.4	ALA
	103.0	PRO
	111.5	ILE THR VAL
	110.5	ARG ASN ASP CYS GLN GLU HIS LEU LYS MET PHE SER TRP TYR

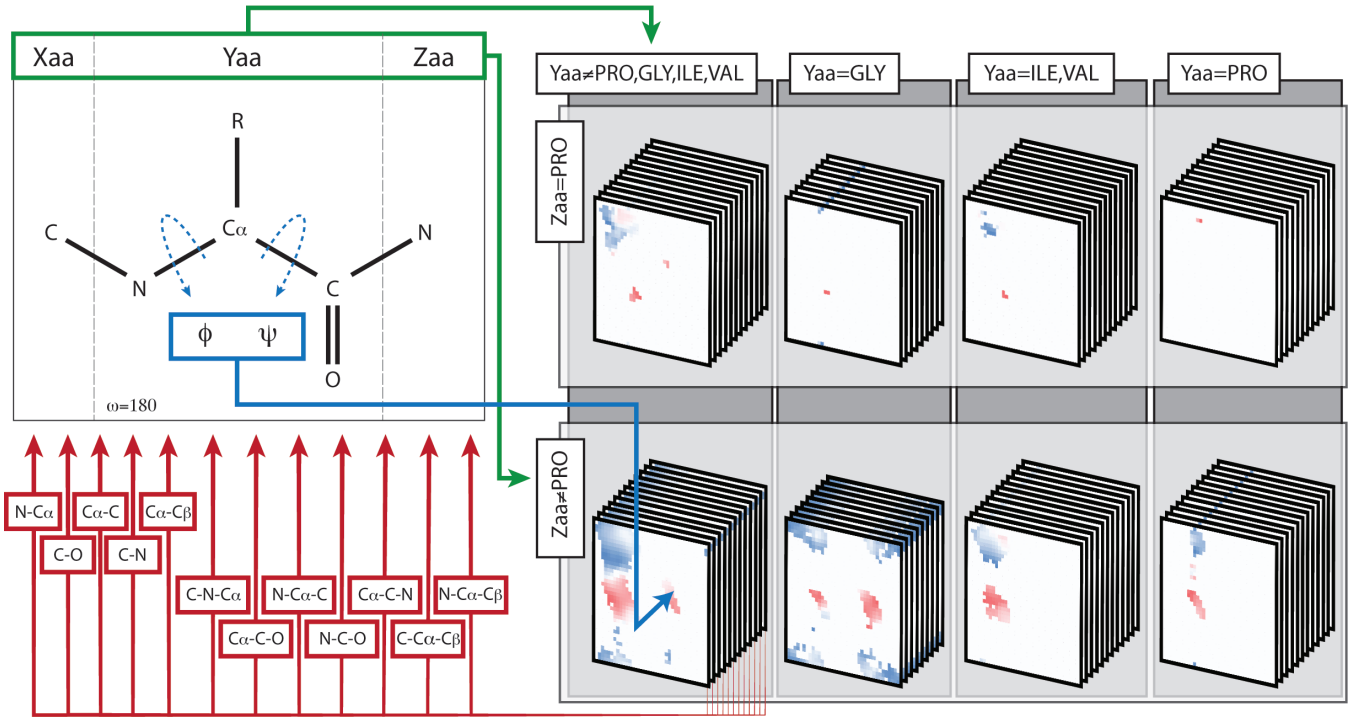


Figure 1: Diagrammatic representation of information content of the backbone CDL showing how a central residue (Yaa) and its C-terminal neighbour (Zaa) define one of eight residues classes (green lines), and the f,y-angles of the residue specify which restraint values to obtain from that class of residue (blue line) for each of the up to seven backbone bond angles and five backbone bond lengths (red lines). The colouring scheme for the CDL plots of each residue type has a common background color for simplicity.

that ultra-high resolution crystal structures consistently showed had much larger discrepancies with the restraint libraries than was expected (Jaskolski et al. 2007; Stec 2007).

The Conformation-Dependent Library (CDL) for the protein backbone (Berkholz et al. 2009; Tronrud, Berkholz, and Karplus 2010; Tronrud and Karplus 2011) is an approach that begins to address these complexities by allowing changes to the target backbone bond and angle values based on the φ,ψ torsion angles of a residue, and it has recently been implemented by Moriarty et al (Moriarty et al. 2014) in *Phenix* (Adams et al. 2010; Adams et al. 2011). The conclusions include that the CDL generates more accurate protein crystal structures and that the dramatic improvement in the geometric residuals that derives from use of the CDL comes with no draw-backs; this reinforces the conclusion that conformation-dependent ideal geometry functions truly do constitute a more accurate representation of reality than the conventional single-value ideal geometry targets.

A graphical representation of the procedure for loading the appropriate restraint values is shown in figure 1. The procedure requires three protein residues in order to determine the φ,ψ angles of the central residue. The current version of the CDL (v1.2) is limited to the largest class of residue triplets: those with both the ω torsions in the trans configuration. Each residue triplet can be further divided into those with/without the last residue (Zaa in figure 1) being a proline. This is conceptually equivalent to preceding or not preceding PRO in table 1. Figure 1 displays the lookups for Zaa=PRO in the top row and Zaa≠PRO in the bottom row. Each of these groupings is divided into four categories based on the identity of the central residue, Yaa. Explicitly, these groups are GLY; PRO; ILE

```
>> mmtbx.cdl_lookup residue_names="ALA,ALA,ALA" phi_psi_angles="90,90"

CDL values returned for specific tripeptide of sequence aa1-aa2-aa3 with the
central residue having (phi,psi) angles of (xxx°,xxx°)

Tripeptide class: NonPGIV_nonxpro

CDL values
  statistical type, number : B 218
  C(-1) - N(0)  - Ca(0)    :    122.38      1.81
  N(0)  - Ca(0) - Cb(0)    :    110.44      1.53
  N(0)  - Ca(0) - C(0)     :    113.16      1.24
  Cb(0) - Ca(0) - C(0)     :    110.01      1.80
  Ca(0) - C(0)  - O(0)     :    119.28      1.21
  Ca(0) - C(0)  - N(+1)    :    118.32      1.35
  O(0)  - C(0)  - N(+1)    :    122.37      1.38
  C(-1) - N(0)             :      1.3306    0.0148
  N(0)  - Ca(0)            :      1.4560    0.0132
  Ca(0) - Cb(0)            :      1.5319    0.0178
  Ca(0) - C(0)             :      1.5226    0.0141
  C(0)  - O(0)             :      1.2354    0.0133
```

Scheme 1: CDL lookup command-line example and output.

or VAL; and all of the rest. Once the residue triplet class has been determined, the lookup has been conceptually narrowed to the 12 restraint/Standard Deviation (SD) value pairs (nine for GLY). Each of the restraint and SD pairs is selected from a function consisting of a 36 by 36 array of values corresponding to the φ,ψ torsion angle ranges of -180 to 180 degrees binned into 10x10 degree boxes. The stacks of pages in figure 1 represent the restraint value functions for the set of 12 (or 9 for GLY) backbone bonds and angles. The values of φ,ψ (determined by the geometry of the triplet) is used to lookup the grid point for all restraints and SD. Each grid is colour-coded to represent the deviation of the value from the global "average" value which is represented as white in these images and provides the value returned for φ,ψ-bins that do not have sufficient observations to give a reliable CDL value .

Tools for accessing CDL

The CDL database of restraint values contains sets of 12 restraints and SD value pairs (nine value pairs for GLY). Accessing them in *Phenix* is done using a dedicated and thoroughly tested python module. This allows quick and easy modifications (including future CDL versions). A simple command-line tool has been added to *Phenix* that returns the restraint values. See schema 1 for an example of the command and the output. A user can easily view some restraint value pairs, and those interested in programming can view the code in the open-source *CCTBX* (Grosse-Kunstleve et al. 2002) and use it as an example to create their own scripts.

Difficulties with the hyper-dimensional optimisation space

A macromolecular refinement involves a multitude of parameters that are optimized simultaneously. This leads to a very complex potential surface that hinders the goal of finding the global minima. The starting model can have an impact on the final results. The differences can be numerical (e.g. rounding errors) or "real" (local minima with a high barrier potential). An additional complication is the weight between the experimental data and the geometry terms. The range of usable values can be quite large resulting in various differences in the final result.

Many algorithmic changes to protein refinement have been implemented in the current packages or are being tested at any

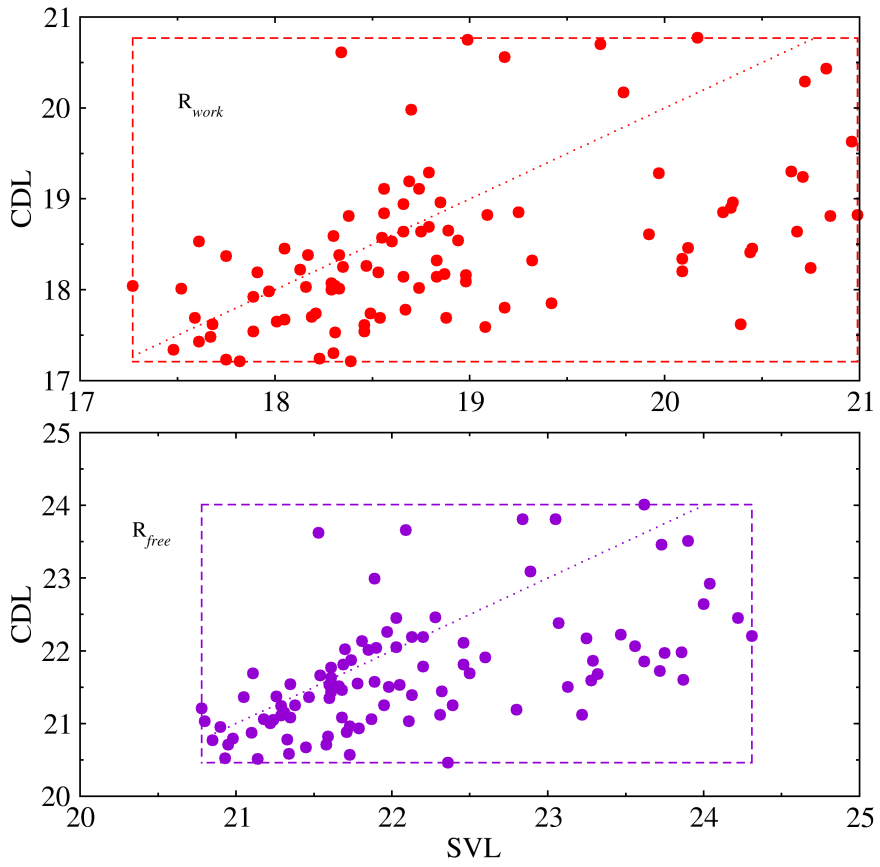


Figure 2: Comparisons of 100 randomly shaken model refinement R-factors.

point in time by the software developers. Typically, each change is tested on a certain protein model and data to determine if it is any improvement. This approach can also be used by the crystallographer to determine the best set of procedures for their particular system. Developers extend the tests to include a large number of systems to understand the parameters that control the improvements and to get statistics that validate the method as a real improvement and not a coincidence. This approach was taken by us (Moriarty et al. 2014) when investigating the effect of the CDL in *Phenix* (Adams et al. 2010; Adams et al. 2011).

An alternative approach to testing involves more extensive explorations of a single system. For instance, the effect of the starting geometry on the results (both with and without the algorithmic change) can be examined by running a number of refinements using a different random seed for each and "shaking" the geometry randomly by a reasonable amount. The values of the R-factors and the geometry rmsd will not be identical even after many refinement steps to a converged result. The spread of the values can be viewed as providing an estimate of the uncertainty in these values. This simulates the noise that the model may contain from many sources and also tests the convergence radius of the algorithmic change. The variance in the results requires a more precise approach to verifying that a change in the refinement algorithm really does improve the statistics. This technique was modified slightly for the tests.

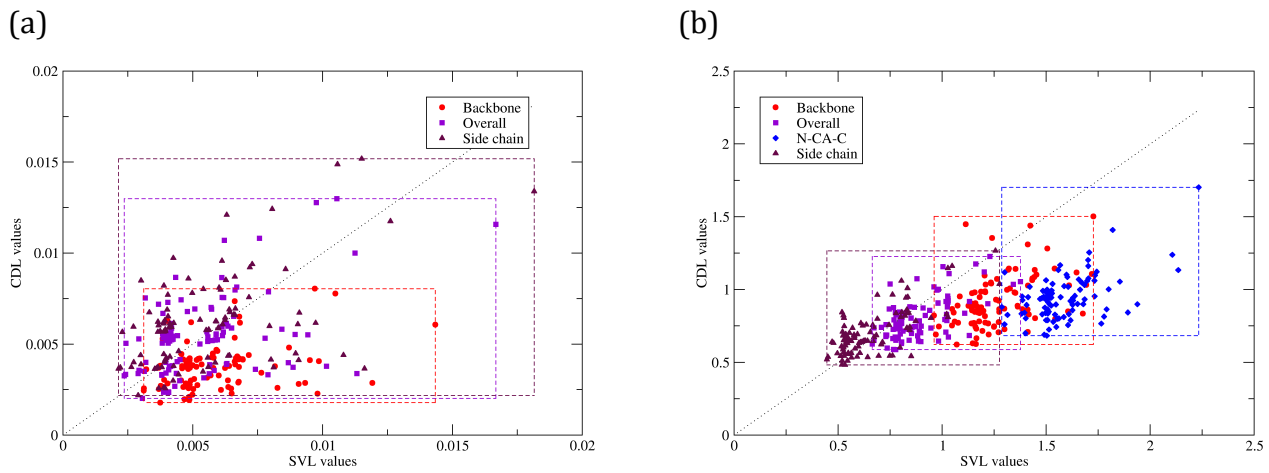


Figure 3: Comparisons of 100 randomly shaken model refinement rmsd values for bond (left) and angles (right). The plots are further divided into backbone, side chain and overall restraints.

Tests

Several PDB codes taken from the earlier proof-of-principle CDL implementation papers (Tronrud, Berkholz, and Karplus 2010; Tronrud and Karplus 2011) were chosen as representative. For each model, a set of 100 random structures were produced using the `phenix.refine` (Afonine et al. 2012) parameter, `modify_start_model.sites.shake`, set to 0.3 Å. These initial 100 models were saved so that parallel refinement could be performed that differed only by the parameter controlling the use of the CDL being set to False or True. Matching the starting geometries allows the direct comparison of what changes by a certain parameter choice. Note that the random number seed is the same for each refinement to restrict the input differences to only the geometry and the algorithmic choices.

The final values that are presented include the R-factors and the overall rmsd values for bonds and angles. Because the CDL operates on the main chain atoms via changing the bond and angle ideal values and estimated standard error of each geometry restraint, the rmsd for the main chain atoms is also presented. In addition, the Molprobity clashscore and overall score are plotted. In all plots, the values for the CDL refinements are plotted on the vertical axis versus the SVL on the horizontal axis. Each point is the result of using the same random starting geometry with and without the benefit of the CDL. Consequently, if the point is below the y=x line (included in all figures) then the CDL value is less than the SVL value. In the case of rmsd and validation scores, this is an improvement. In general, there are usually points on both sides of the y=x line. To assist in visualising the trends, a dashed box surrounds each group of points.

Figure 2 displays the R-factors for the refinements of the perturbed models of `3eln`. All following figures relate to the `3eln` example. The top portion is the R_{work} values plotted with the CDL values plotted on the vertical axis and the SVL on the horizontal. The limits of the CDL are approximately 17.2–20.8% and SVL is 17.5–21.0%. This means that the spreads are 3.6 and 2.5%, respectively. The R_{free} values has ranges of 20.5–24.0% and 20.8–24.3% meaning the CDL has generally slightly lower R factors.

Similar variance plots can be produced for the bond and angle rmsd. The bond data is presented in figure 3a. The most striking feature is that box of spreads for the backbone rmsd values is mostly below the y=x trend

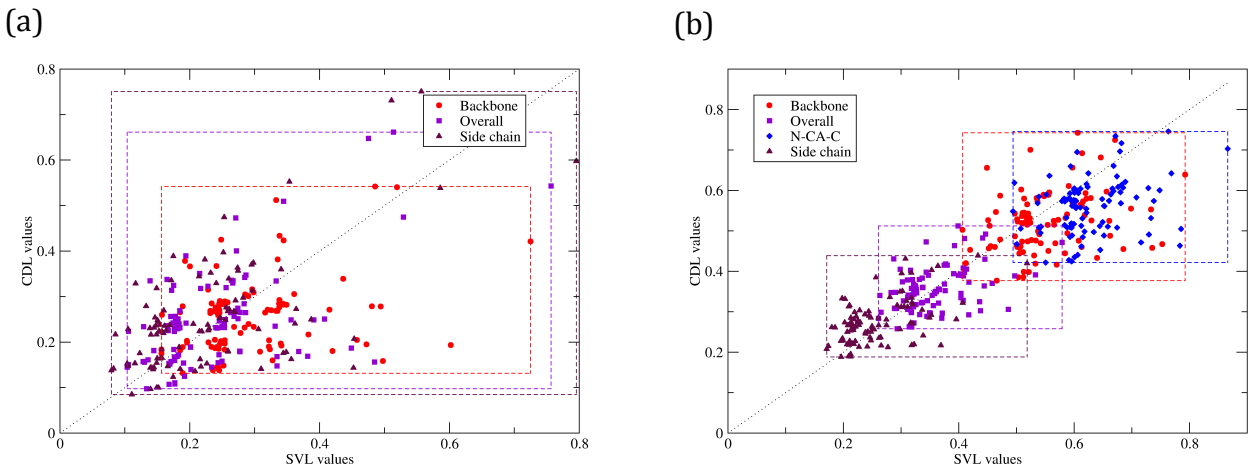


Figure 4: Comparisons of 100 randomly shaken model refinement rmsZ values for bond (left) and angles (right). The plots are further divided into backbone, side chain and overall restraints.

indicating that, in general, the bond rmsd values are reduced by the use of CDL. The rmsd value box is slightly below the trend. Interestingly, the backbone and overall rmsd have a similar distribution pattern within the box.

The angle rmsd data is plotted in figure 3b. In addition to the overall and backbone rmsd values, the rmsd of the N-C_α-C angle is also plotted. The spread boxes have a higher fraction of the values below the y=x trend than the corresponding bond boxes. In fact, the majority of the backbone and all of the N-C_α-C angle rmsd data are below the line. The angle rmsd values for the side chain are also plotted and show a relatively even spread.

In the CDL library formalism, not only are the ideal target values changed based on φ,ψ but the force constant of the restraint is also adjusted. The CDL SD value is always smaller than the corresponding SVL value. Taken alone this could be the sole reason for the decrease in rmsd results. However, as shown in figure 2, the R-factors are not unduly affected.

A common statistical measure of deviation from a mean by a population is the *Z* score. The *Z* value of an observation is defined as the

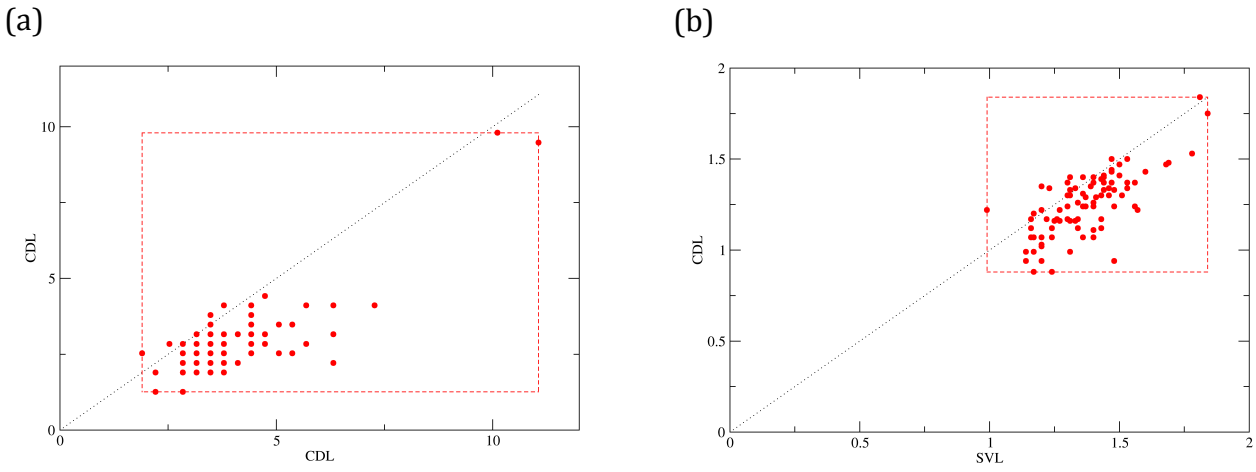


Figure 5: Comparisons of 100 randomly shaken model refinement validation scores – clashscore (left) and Molprobity (right).

number of standard deviations of a population of observations from the mean of the population. The formula is

$$Z = \frac{x - \bar{x}}{\sigma} \quad (1)$$

where x is the observation and σ is standard deviation. The *Z* score gives a dimensionless measure of the deviations based on the spread of the population. In a similar fashion to rmsd values, a *Z* value close to zero means a small deviation.

Applying the *Z* score formalism to geometry deviation can be performed as follows. Each geometry restraint has an ideal value and an SD. This provides the information needed to calculate the *Z* scores. The root mean square of the *Z* score (rms*Z*) can be calculated using

$$\mathrm{rms}Z = \sqrt{\frac{\sum_{j=1}^{N} Z_j^2}{N}} \quad (2)$$

In the context of bond and angle restraints, the smaller an rms*Z* value the closer the observations are to the ideal values. However, if the rms*Z* is close to 1.0, the distribution of the observed values has a similar (Gaussian) distribution to the distribution that provided the mean and standard deviation used in calculating the *Z* scores.

Figure 4 shows the bond and angle rms*Z* values, respectively. The *Z* scores for the SVL are calculated using the ideal values and ESD values from the standard restraints provided by the GeoStd (Moriarty and Adams) based on the Monomer Library (Vagin et al. 2004). For the CDL, the appropriated ideal and SD values are taken from the database for the backbone restraints and from the SVL for side chain.

The distributions for the bond rms*Z* values are generally less then 0.5 with a slight preference for the CDL. Interestingly, the angle rms*Z* values have the side chain scores close to 0.2 while the backbone scores are in the 0.4–0.7 range indicating that the backbone has a distribution that more closely matches the database distribution in both the CDL and SVL cases than the side chain distributions.

Validation of CDL structures is an ongoing area of research. Figure 5 shows the Molprobity clashscore (a) and overall score plots (b). In general, the spread of values is not unreasonable giving the variation in the input geometry with the scores generally better for the CDL refined models.

Conclusion

As was concluded in the *Phenix* CDL implementation paper based on parallel refinements of 27,587 models (Moriarty et al. 2014) and with the earlier papers based on refinements of only a few representative models (Tronrud, Berkholz, and Karplus 2010; Tronrud and Karplus 2011) the CDL improves the geometry of the models without adversely effecting the R-factors.

Acknowledgements

This work was supported in part by National Institutes of Health (NIH) grant R01-GM083136 (to PAK), by the NIH Project 1P01 GM063210 (to PDA), and the Phenix Industrial Consortium. This work was further supported in part by the US Department of Energy under Contract No. DE-AC02-05CH11231.

References

Adams, Paul D., Pavel V. Afonine, Gabor Bunkoczi, Vincent B. Chen, Ian W. Davis, Nathaniel Echols, Jeffrey J. Headd, et al. 2010. "PHENIX: A Comprehensive Python-Based System for Macromolecular Structure Solution." *Acta Crystallographica Section D-Biological Crystallography* 66 (February): 213–21. doi:10.1107/S0907444909052925.

Adams, Paul D., Pavel V. Afonine, Gabor Bunkoczi, Vincent B. Chen, Nathaniel Echols, Jeffrey J. Headd, Li-Wei Hung, et al. 2011. "The Phenix Software for Automated Determination of Macromolecular Structures." *Methods* 55 (1): 94–106. doi:10.1016/j.ymeth.2011.07.005.

Afonine, Pavel V., Ralf W. Grosse-Kunstleve, Nathaniel Echols, Jeffrey J. Headd, Nigel W. Moriarty, Marat Mustyakimov, Thomas C. Terwilliger, Alexandre Urzhumtsev, Peter H. Zwart, and Paul D. Adams. 2012. "Towards Automated Crystallographic

Structure Refinement with Phenix.refine." *Acta Crystallographica Section D-Biological Crystallography* 68 (April): 352–67. doi:10.1107/S0907444912001308.

Berkholz, Donald S., Maxim V. Shapovalov, Roland L. Dunbrack, Jr., and P. Andrew Karplus. 2009. "Conformation Dependence of Backbone Geometry in Proteins." *Structure* 17 (10): 1316–25. doi:10.1016/j.str.2009.08.012.

Engh, RA, and R Huber. 2001. "Structure Quality and Target Parameters." In *International Tables for Crystallo-graphy*, edited by MG Rossmann and E Arnold, F:382–92. Dordrecht: Kluwer Academic Publishers.

Engh, RA, and R. Huber. 1991. "Accurate Bond and Angle Parameters for X-Ray Protein-Structure Refinement." *Acta Crystallographica Section A* 47 (July): 392–400. doi:10.1107/S0108767391001071.

Grosse-Kunstleve, R. W., N. K. Sauter, N. W. Moriarty, and P. D. Adams. 2002. "The Computational Crystallography Toolbox: Crystallographic Algorithms in a Reusable Software Framework." *Journal of Applied Crystallography* 35 (February): 126–36. doi:10.1107/S0021889801017824.

Jaskolski, Mariusz, Miroslaw Gilski, Zbigniew Dauter, and Alexander Wlodawer. 2007. "Stereochemical Restraints Revisited: How Accurate Are Refinement Targets and How Much Should Protein Structures Be Allowed to Deviate from Them?" *Acta Crystallographica Section D-Biological Crystallography* 63 (May): 611–20. doi:10.1107/S090744490700978X.

Jiang, Xiaoqin, Ching-Hsing Yu, Ming Cao, Susan Q. Newton, Erich F. Paulus, and Lothar Schäfer. 1997. "Ø/ψ-Torsional Dependence of Peptide Backbone Bond-Lengths and Bond-Angles: Comparison of Crystallographic and Calculated Parameters." *Journal of Molecular Structure* 403 (1–2): 83 – 93. doi:http://dx.doi.org/10.1016/S0022-2860(96)09390-8.

Karplus, P. A. 1996. "Experimentally Observed Conformation-Dependent Geometry and Hidden Strain in Proteins." *Protein Science* 5 (7): 1406–20.

Karplus, P. A., M. V. Shapovalov, R. L. Dunbrack, Jr., and D. S. Berkholz. 2008. "A Forward-Looking Suggestion for Resolving the Stereochemical Restraints Debate: Ideal Geometry Functions." *Acta Crystallographica Section D-Biological Crystallography* 64 (3): 335–36. doi:10.1107/S0907444908002333.

Klimkowski, V.J., L Schäfer, F.A. Momany, and C. Van Alsenoy. 1985. "Local Geometry Maps and Conformational Transitions between Low Energy Con-Formers of N-Acetyl-N'-Methyl Glycine Amide: An Ab Initio Study at the 4–21G Level with Gradient Relaxed Geometries." *Journal of Molecular Structure* 124: 143–53.

Moriarty, Nigel W., and Paul D. Adams. *GeoStd*. http://sourceforge.net/projects/geostd.

Moriarty, Nigel W., Dale E. Tronrud, Paul D. Adams, and P. Andrew Karplus. 2014. "Conformation-Dependent Backbone Geometry Restraints Set a New Standard for Protein Crystallographic Refinement." *FEBS Journal*, n/a–n/a. doi:10.1111/febs.12860.

Schäfer, L., and M. Cao. 1995. "Predictions of Protein Backbone Bond Distances and Angles from First Principles." *Theochem-Journal of Molecular Structure* 333 (3): 201–8.

Schäfer, L., V. J. Klimkowski, Frank A. Momany, H. Chuman, and C. Van Alsenoy. 1984. "Conformational Transitions and Geometry Differences between Low-Energy Conformers of N-Acetyl-N′-Methyl Alanineamide: An Ab Initio Study at the 4-21G Level with Gradient Relaxed Geometries." *Biopolymers* 23 (11): 2335–47. doi:10.1002/bip.360231115.

Stec, Boguslaw. 2007. "Comment on - Stereochemical Restraints Revisited: How Accurate Are Refinement Targets and How Much Should Protein Structures Be Allowed to Deviate from Them? By Jaskolski, Gilski, Dauter & Wlodawer (2007)." *Acta Crystallographica Section D-Biological Crystallography* 63 (October): 1113–14. doi:10.1107/S0907444907041406.

Tronrud, Dale E., Donald S. Berkholz, and P. Andrew Karplus. 2010. "Using a Conformation-Dependent Stereochemical Library Improves Crystallographic Refinement of Proteins." *Acta Crystallographica Section D-Biological Crystallography* 66 (7): 834–42. doi:10.1107/S0907444910019207.

Tronrud, Dale E., and P. Andrew Karplus. 2011. "A Conformation-Dependent Stereochemical Library Improves Crystallographic Refinement Even at Atomic Resolution." *Acta Crystallographica Section D-Biological Crystallography* 67 (August): 699–706. doi:10.1107/S090744491102292X.

Vagin, A. A., R. A. Steiner, A. A. Lebedev, L. Potterton, S. McNicholas, F. Long, and G. N. Murshudov. 2004. "REFMAC5 Dictionary: Organization of Prior Chemical Knowledge and Guidelines for Its Use." *Acta Crystallographica Section D-Biological Crystallography* 60 (December): 2184–95. doi:10.1107/S0907444904023510.