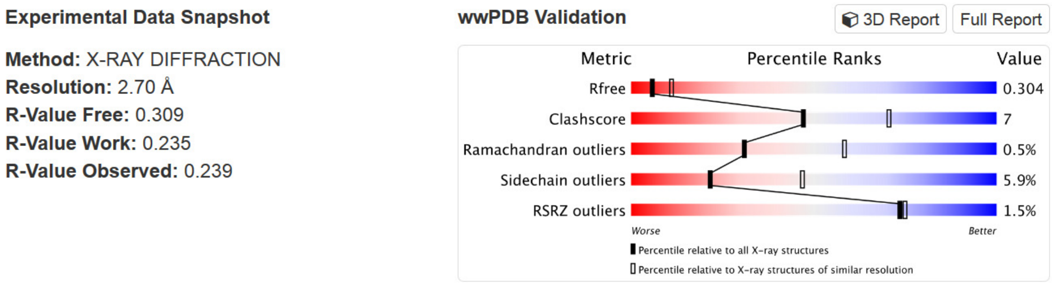
6w9c – Papain-like Protease Domain of nsp3

What is the best that can be done with lousy data?

Author: Dale E. Tronrud

Initial Impressions



Experimental Data Snapshot

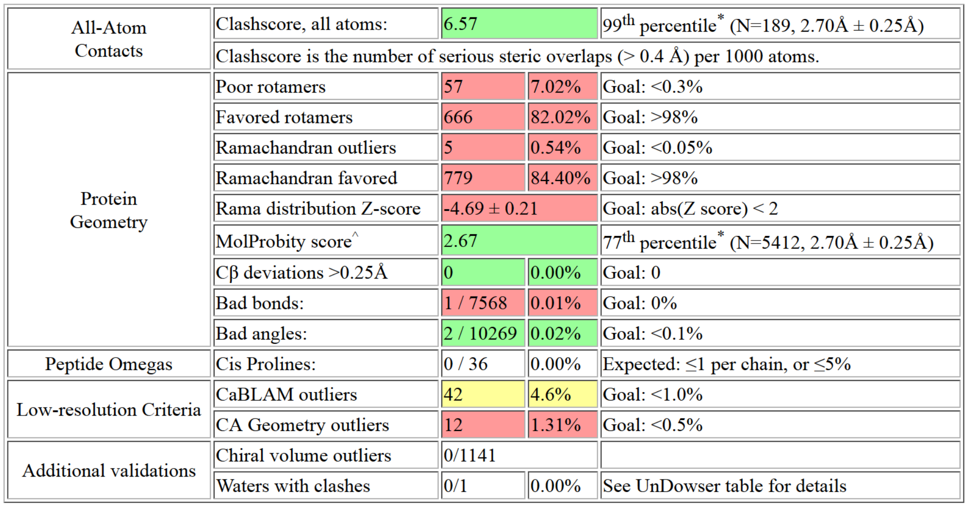
Method: X-RAY DIFFRACTION
Resolution: 2.70 Å
R-Value Free: 0.309
R-Value Work: 0.235
R-Value Observed: 0.239

wwPDB Validation

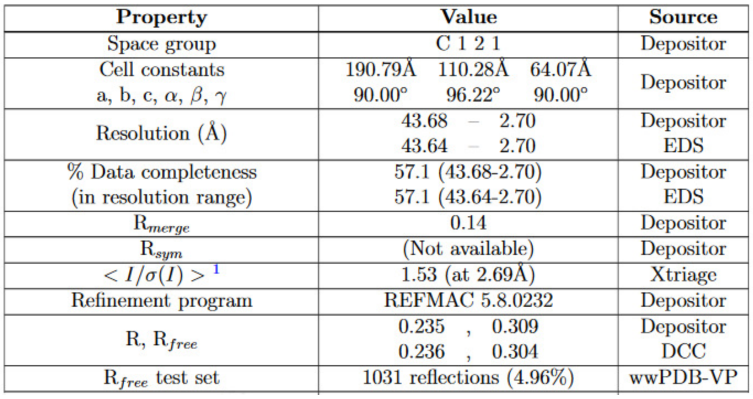
3D Report Full Report

Metric	Percentile Ranks	Value
Rfree		0.304
Clashscore		7
Ramachandran outliers		0.5%
Sidechain outliers		5.9%
RSRZ outliers		1.5%

▮ Percentile relative to all X-ray structures
▯ Percentile relative to X-ray structures of similar resolution



All-Atom Contacts	Clashscore, all atoms:	6.57		99th percentile* (N=189, 2.70Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	57	7.02%	Goal: <0.3%
	Favored rotamers	666	82.02%	Goal: >98%
	Ramachandran outliers	5	0.54%	Goal: <0.05%
	Ramachandran favored	779	84.40%	Goal: >98%
	Rama distribution Z-score	-4.69 ± 0.21		Goal: abs(Z score) < 2
	MolProbity score^	2.67		77th percentile* (N=5412, 2.70Å ± 0.25Å)
	Cβ deviations >0.25Å	0	0.00%	Goal: 0
	Bad bonds:	1 / 7568	0.01%	Goal: 0%
	Bad angles:	2 / 10269	0.02%	Goal: <0.1%
Peptide Omegas	Cis Prolines:	0 / 36	0.00%	Expected: ≤1 per chain, or ≤5%
Low-resolution Criteria	CaBLAM outliers	42	4.6%	Goal: <1.0%
	CA Geometry outliers	12	1.31%	Goal: <0.5%
Additional validations	Chiral volume outliers	0/1141		
	Waters with clashes	0/1	0.00%	See UnDowser table for details



Property	Value	Source
Space group	C 1 2 1	Depositor
Cell constants a, b, c, α, β, γ	190.79Å 110.28Å 64.07Å 90.00° 96.22° 90.00°	Depositor
Resolution (Å)	43.68 – 2.70 43.64 – 2.70	Depositor EDS
% Data completeness (in resolution range)	57.1 (43.68-2.70) 57.1 (43.64-2.70)	Depositor EDS
R_{merge}	0.14	Depositor
R_{sym}	(Not available)	Depositor
$< I/\sigma(I) >$[1]	1.53 (at 2.69Å)	Xtriage
Refinement program	REFMAC 5.8.0232	Depositor
R, R_{free}	0.235 , 0.309 0.236 , 0.304	Depositor DCC
R_{free} test set	1031 reflections (4.96%)	wwPDB-VP

This model has three levels of problems.

- Its diffraction pattern was only measured to 2.7 Å resolution.
- The diffraction data were not measured and processed as well as possible.
- Its agreement with those data, and other chemical requirements, is poorer than most models at this resolution.

Generally the first of these problems is of the least concern. The resolution of the data set is usually limited by the quality of the crystal, but in some cases high resolution data which can be measured is not. While this does appear to be the case for this data set, this is not the worst problem with the model.

The second problem becomes more apparent with further digging. In the PDB Validation report, shown in third place above, we see that the data completeness is only 57%. That is extraordinarily low. There must have been some serious problem during data collection for that level of completeness to be accepted.

The third problem indicates that significant improvement can be made in interpreting these data. One should always be able to create a model that is consistent with both the diffraction data and our knowledge of chemistry. All molecules must have good chemistry!

If we are going to spend the time improving the refinement of this model, we should first look into the data collection problems and ensure that we are working with the best interpretation of the diffraction images.

Reprocessing the Images

Before attempting any serious refinement I wanted to see if the data could be improved by reprocessing the images. Anything that could improve the resolution, or completeness (especially!), would help. Kay_Diederichs was recruited to the project and his help was very useful. I also had discussions with the Global Phasing people who have their own effort to improve the data for Coronavirus structures.

The images can be found at https://proteindiffraction.org/project/IDP51000_6w9c.The citation is https://doi.org/10.18430/m36w9c.

The opinions of both Kay and Global Phasing are that this data set was collected using a protocol that was improper in many ways. The most egregious was that too short of a sweep of phi angles was performed and this resulted in the low completeness. The other errors probably led to this choice.

Basically the conclusion is that the phi range of each image was way too large for this type of detector, and the exposures were too long resulting in radiation damage almost from the first image. Two sweeps were collected -- oddly enough starting from the same point. Some radiation damage can be omitted by ignoring the latter images, but the damage is present in all images.

Kay integrated the images using xdsgui and sent the results to me. Since the scattering is very anisotropic I decided to run the data through Global Phasing's STARaniso server to impose an elliptical resolution cutoff. I used their reports to identify the point in each sweep where the merging statistics begin to rapidly rise, and removed all images after those points. I found a compromise which lowered the merging stats but only resulted in the loss of 375 reflections (out of about 18,500). Kay showed that the resolution could be extended to 2.6 Å (from 2.7 Å) and maybe all the way to 2.5 Å. The completeness of the new data set when the elliptical cutoff was applied dropped to 44.5%. This is even lower than the deposited data but my hope is that the remaining reflections are of better quality.

I called my data set "Two Sweeps - Truncated Data". The stats reported by STARaniso are:

```
                                             Overall  InnerShell  OuterShell
   -------------------------------------------------------------------------
      Low resolution limit                    41.748      41.748       2.621
      High resolution limit                    2.502       7.466       2.502

      Rmerge  (all I+ & I-)                    0.096       0.038       0.610
      Rmerge  (within I+/I-)                   0.081       0.035       0.566
      Rmeas   (all I+ & I-)                    0.121       0.048       0.804
      Rmeas   (within I+/I-)                   0.111       0.048       0.764
      Rpim    (all I+ & I-)                    0.073       0.028       0.517
      Rpim    (within I+/I-)                   0.075       0.033       0.509
      Total number of observations             42480        2307        1973
      Total number unique                      18495         925         925
      Mean(I)/sd(I)                              7.1        19.4         1.4
      Completeness (spherical)                  40.4        52.1        15.7
      Completeness (ellipsoidal)                44.5        52.1        16.7
      Multiplicity                               2.3         2.5         2.1
      CC(1/2)                                  0.992       0.996       0.569

      Anomalous completeness (spherical)        16.8        37.2         1.8
      Anomalous completeness (ellipsoidal)      18.4        37.2         0.0
      Anomalous multiplicity                     1.6         1.5         1.6
      CC(ano)                                  0.054       0.061       0.110
      |DANO|/sd(DANO)                          0.798       0.824       0.922
```

The full report is in "Two Sweeps - Truncated_ STARANISO anisotropy & Bayesian estimation server.pdf". The run of STARaniso was identified as “cLDhZJlan4c6wZ94”. I named the final structure factor file “Two Sweeps - Truncated.mmcif”, or “Two Sweeps - Truncated.mtz”. Since a great many programs have difficulties with file names that contain spaces, I also named these data “6w9c_KD_DET”. Later Andrea and Yunyun transferred the original free R flags from the depositor’s file to this data set and the name became “6w9c_KD_DET_rfreeflags.mtz”.

Inspection of Model

The deposited model was created via Molecular Replacement (Phases) using the SARS-CoV analog model 5Y3Q and refined with REFMAC to R Values of 23.5%/30.9%.

The crystal contains three molecules in the asu with a 3-fold direct ncs axis pointing along the z coordinate axis. Near the axis the proteins have well ordered structure and the ncs agreement is very good. Far from the axis the symmetry breaks down and so does the structure. At the distal tips there is a Zinc binding site and each clearly contains multiple conformations, which makes them very unclear. The A chain, for example doesn't even have a zinc atom.

Each molecule actually has two zinc binding sites. The first is well ordered and sits directly on the ncs three-fold axis. The zinc is held by three, equivalent, cystine residues and the fourth ligand is, presumably, a water molecule. The zinc is in alternative conformations both above and below the plane of the SG atoms. One conformation was not modeled and neither is the water molecule. This site is probably a crystallization artifact.

The second site is at the distal extreme of the protein. It is a four-cystine zinc binding site. Only the B and C chains have a zinc atom in the model, and neither of them has all four cystines connected.

The bond lengths and angles of the SG-ZN ligation were set incorrectly. In the central zinc two of the SG-ZN distances are around 2.4 Å while the third is an amazing 2.7 Å! It is very strange to have this amount of asymmetry when the structure is sitting right on a nearly perfect 3-fold symmetry axis. The CB-SG-ZN angle only a little more than 70°! (It should be around 107°.) In the C chain there are only three links (although there are four cystine residues available) and their distances range from 2.85 to 2.22 Å. Two of the angles are 90° or less.

The well-ordered parts of the protein, itself, do not contain any gross errors. The poor geometry statistics indicate that there are probably many small corrections that should be made – principally side chain rotomer changes. The regions of the protein out near the distal Zinc binding site are very bad. The electron density there is very weak and the protein chain does not conform to the expected secondary structure rules.

Interesting Things about nsp3/Papain-like Protease

The principal catalytic residue for the protease functionality is Cys 111. This residue is assisted by His 272 and Trp 108.

Rather near by is a totally buried chloride ion which was called residue 502 in 6w9c but I have renamed 401. In the deposited model only the B chain contained this chloride, but it is pretty clear to me that it is present in all three. This is consistent with my later conclusion that the noncrystallographic symmetry of most of the protein is nearly perfect.

There are two zinc binding sites in this protein. One is located directly on the 3-fold ncs symmetry axis and shares a single zinc among all three protein chains. There are three protein ligands in this site with each chain providing residue Cys 270. In 6w9c this zinc was placed in the C chain and named 401, which was also the residue number of the chloride ion in the B chain. (!) I kept this zinc in the C chain but renamed it to 502. I believe that the zinc bound at this site is an artifact of crystallization.

The second zinc site is located at the distal end of the protein. The literature says that this region forms a "zinc finger". The zinc is held in place by four cystine residues – 189, 192, 224, and 226. You will note that the first pair of cystines form a tight circle when bound to the zinc, and an even more constrained cycle for the last pair. It is difficult to imagine much flexibility in the entire binding site, and yet the density for each of the three chains in this region is terrible. Either this entire end of the molecule is quite mobile, but moving as a rigid group, or the dose of radiation has freed the zinc ion in varying amounts in each chain. No zinc was even modeled n the A chain of 6w9c, while B was built with a zinc only liganded by 189 and 192. The zinc in the C chain was built with all four ligands.

The zinc in this site was named 501 in the B chain but 402 in the C chain. I added a zinc ion to the A chain, linked to 189 and 226, and ensured that all three were named 501.

The electron density for the distal tip of the protein, which I presume is the zinc finger but I'm not sure how big a zinc finger is, is very weak. This region of the protein consists of residues 188 through 195 and residue 218 through 231. Those two loops cover little more than the zinc binding site, but do not look like a formal "domain". I would say that the entire domain consists of residue 178 through 240 and 307 to the Carboxyl terminus at 315.

One could define the domain to be everything beyond 178, but the loop I'm excluding seems to pack tightly to the central domain and is better ordered than the region I'm specifying.

There is an amide terminal domain that ends at residue 74.

Initial Refinement Goals

Andrea wants a "best practices" refinement of 6w9c using the newly integrated data from Kay and myself. She wants to show how much improvement can be achieved just by applying good practices to a bad data set and poor refinement. Therefore, she doesn't want any references to higher resolution or better models. I am to pretend that no other models of this protease exist.

I edited the coordinates from 6w9c a bit before starting refinement. I wanted a unified description of nsp3, or at least what I'm calling "nsp3b", for the second protein unit of this undivided polyprotein. I renamed the Zinc atoms and the Chlorine in the hope that my new names will make better sense in future refinements. No model building was performed prior to refinement.

How Important is the Noncrystallographic Symmetry?

To get a feel for the importance of restraints on noncrystallographic symmetry (ncs) I ran a series of refinements using the TNT refinement package. I chose this program because it was the easiest for me to set up this test. This program is very limited in that it does not use Maximum Likelyhood methods, but since I'm not planning to use the resulting coordinates for anything this is not significant.

Run	wght	Rwrk	Rfree	Bond	Angle	ncs
6w9c	-	28.95%	38.68%	0.023	1.848	0.935
tnt-1	0	22.36%	38.92%	0.034	0.668	0.973
ncs-10	10	26.09%	36.49%	0.004	0.645	0.285
ncs-20	20	26.79%	36.14%	0.004	0.628	0.149
ncs-40	40	27.35%	35.72%	0.003	0.582	0.071
ncs-80	80	27.20%	35.56%	0.003	0.505	0.029

This first line is just the statistics for the deposited model (after my editing) using the new data set and TNT's geometry library. My first refinement was a run of TNT w/o ncs restraints. These two models provide the baseline against which the other refinements are compared. You are free to conclude from the difference in free R after the basic TNT run, that TNT sucks when working at this resolution and a data set this incomplete.

The "ncs-*" refinements are a series of refinement runs with increasing weight on the noncrystallographic symmetry. It is quite clear that the tighter I force the model to obey the symmetry the lower the free R. This trend continues even when I tighten the force to get very, very, good agreement.

This test shows that noncrystallographic symmetry restraints are critical for the refinement of this data set. There is no indication that the deposited model was created with ncs restraints. This omission may explain the unusually high free R of the deposited model. The ncs restraints will add information to the refinement that specifically compensates for the low completeness of the diffraction data set. There is hope that the areas of the three protein chains which obey the symmetry (all except the distal Zinc binding site, but including the active site) should be able to be well modeled.

Refinement with Buster/TNT

I ran a series of refinement tests using Buster/TNT to test the importance of both noncrystallographic symmetry restraints and the inclusion of explicit hydrogen atoms (in riding positions). Buster does not allow one to easily vary the weight of the ncs restraints, resulting in a binary choice.

```
Run     Rwrk   Rfree  Bond  Angle  ncs
6w9c    23.5%  30.9%                  Deposited values
6w9c    24.05% 31.09% 0.013 1.325  ---    Buster values/KD Data
plain1  15.97% 27.32% 0.010 1.142  no
Hydro1  16.58% 27.38% 0.010 1.081  no
ncs1    16.36% 25.31% 0.010 1.168  yes
hydNCS  18.07% 25.81% 0.010 1.070  yes
```

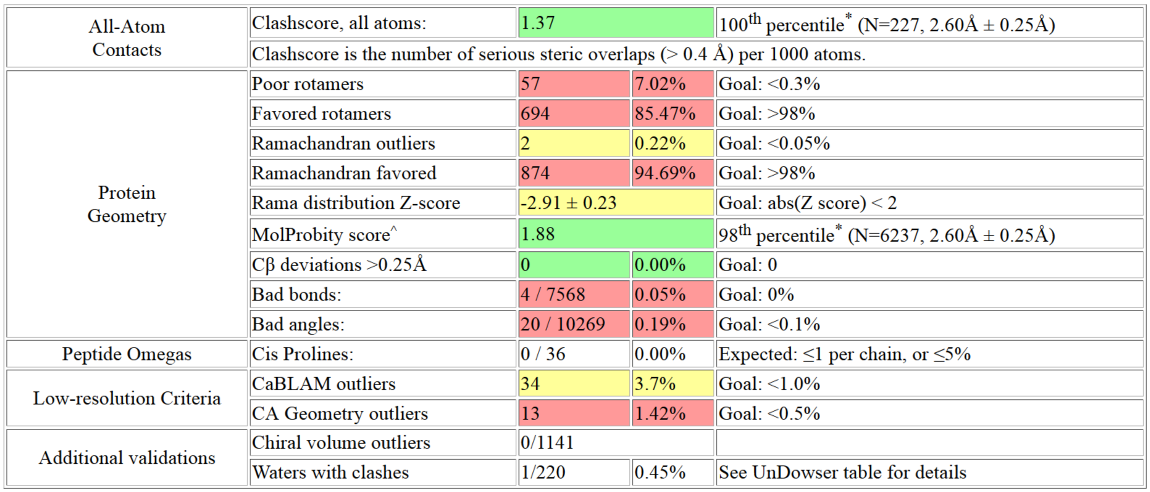
The first two lines of this table show that Buster is able to reproduce the reported R values with sufficient precision. There are expected to be differences because of alternative means of calculating and scaling structure factors, but more importantly, we are using a different data set derived from the same images. Apparently Kay's new processing didn't make huge changes to the data set.

"plain1" is the control run of Buster without ncs restraints and hydrogen atoms. This model has a lower free R than the deposited model but the reason for this lower value is not clear. It could be due to Buster being a better program than Refmac, or it could be due to Kay's new version of the data. Future tests would have to be performed to resolve that question, but this issue is not important to the current project.

"Hydro1" is identical to "plain1" except that riding hydrogen atoms have been added. The statistics in the table do not show any significant change, but examination of the Molprobity report (not shown) shows a decrease in the number of "clashes" when hydrogen atoms are included in the refinement.

"ncs1" is the refinement w/o hydrogen atoms but with ncs restraints. "hydNCS" is the refinement with both. Adding the restraints causes a drop of free R by about two percentage points, although part of this gain is lost with the inclusion of hydrogen atoms. Here is the Molprobity report for "hydNCS". The "Clashscore, all atoms" has improved from 6.57 to 1.37 and the "MolProbity score" from the 77th percentile to the 98th.

There is certainly too much red and yellow in the Molprobity report, but we are just starting.



All-Atom Contacts	Clashscore, all atoms:	1.37		100th percentile* (N=227, 2.60Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	57	7.02%	Goal: <0.3%
	Favored rotamers	694	85.47%	Goal: >98%
	Ramachandran outliers	2	0.22%	Goal: <0.05%
	Ramachandran favored	874	94.69%	Goal: >98%
	Rama distribution Z-score	-2.91 ± 0.23		Goal: abs(Z score) < 2
	MolProbity score^	1.88		98th percentile* (N=6237, 2.60Å ± 0.25Å)
	Cβ deviations >0.25Å	0	0.00%	Goal: 0
	Bad bonds:	4 / 7568	0.05%	Goal: 0%
	Bad angles:	20 / 10269	0.19%	Goal: <0.1%
Peptide Omegas	Cis Prolines:	0 / 36	0.00%	Expected: ≤1 per chain, or ≤5%
Low-resolution Criteria	CaBLAM outliers	34	3.7%	Goal: <1.0%
	CA Geometry outliers	13	1.42%	Goal: <0.5%
Additional validations	Chiral volume outliers	0/1141		
	Waters with clashes	1/220	0.45%	See UnDowser table for details

First Round Model Building

My principal goal in this round of model building is to build in as much ncs compliance as possible. I have decided to examine the ncs averaged map that results from the recent Buster refinement and modify one of the chains to fit the averaged map, and then replace the other two chains with copies.

When looking through the model I superimposed all three chains. In nearly all residues the main chains line up very well. The breakage of ncs occurs for a fraction of the side chains and for many of those there was no real justification for a disparity. I think unifying those should cause little problem. There are some places where there are differences between the three chains, but I think they are very few. I united all side chains and will worry about the differences at a later point.

The most significant differences occur right at the distal zinc binding site. The average density is very bad in this area, and so are the individual maps. I decided to stick with the C chain version since it was the most complete, but even its density was so poor around 225 and 228 that I deleted those residues. I did leave in the SG atom since its position is defined by the locations of the other SG atoms and the Zinc. I hope Buster will be able to handle the geometry restraints despite the incomplete model.

The biggest problem region, among those with some kind of density, is the strands of β-sheet near the zinc binding site. These strands are composed of the residue ranges 184-188, 221-217, and 229-233. Oddly enough, these strands seem to be frame-shifted but not shifted by an entire residue. They appear to be off by about, maybe, an Ångstrom or so. I had real difficulty getting Coot to fix this problem. Coot's real-space refinement will not correct such an error because the atoms are all locked into incorrect positions. I worked the hardest on 229-233 and think I have that one placed correctly.

I had to delete atom 234:N to break the chain. (I had already deleted residues on the other end as noted above.) I could then use Coot's manual rigid body tool to shift the range to where I wanted. I added back the deleted atom manually after I finished model building. The cross-strand hydrogen bonding in this area is not good and much more work is required in this area.

The final changes I made were to ensure that the chloride ion near the active site that was only in the B chain has been replicated in the other two chains. They had water molecules instead, which was quite unreasonable, especially since the ncs is so good. I'm not sure Buster imposes ncs on solvent. I should ask.

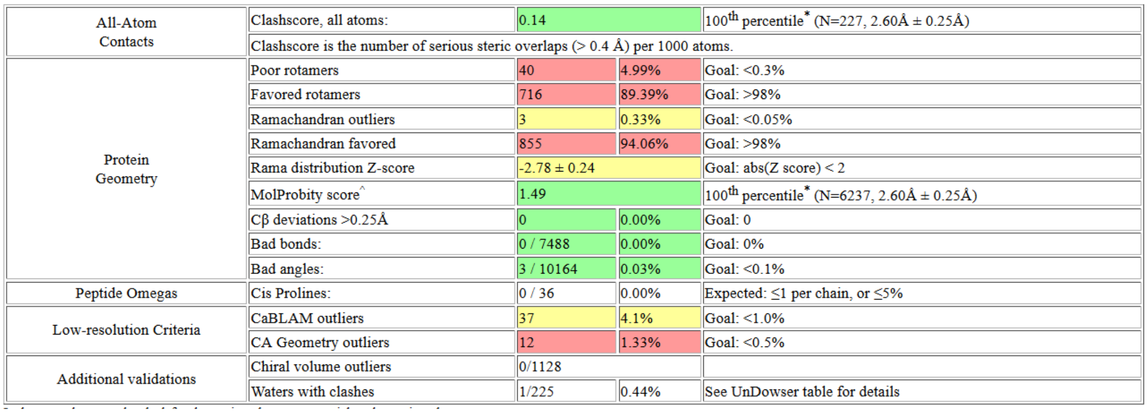
In general, the density in the problem regions is strongest for the B chain.

Buster: Round 2

Not much to say here. The only changes to Buster's configuration was to adjust the restraints on the distal zinc site to match the new model.

```
Run      Rwrk   Rfree  Bond  Angle  ncs
6w9c     23.5%  30.9%                   Deposited values
6w9c     24.05% 31.09% 0.013 1.325  ---   Buster values/KD Data
hydNCS   18.07% 25.81% 0.010 1.070  yes
model-1  24.20% 27.05% 0.031 2.385  yes    Now perfect ncs
```

```
refine-2 22.84% 26.06% 0.007 0.776  yes
```



All-Atom Contacts	Clashscore, all atoms:	0.14		100th percentile* (N=227, 2.60Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	40	4.99%	Goal: <0.3%
	Favored rotamers	716	89.39%	Goal: >98%
	Ramachandran outliers	3	0.33%	Goal: <0.05%
	Ramachandran favored	855	94.06%	Goal: >98%
	Rama distribution Z-score	-2.78 ± 0.24		Goal: abs(Z score) < 2
	MolProbity score^	1.49		100th percentile* (N=6237, 2.60Å ± 0.25Å)
	Cβ deviations >0.25Å	0	0.00%	Goal: 0
	Bad bonds:	0 / 7488	0.00%	Goal: 0%
	Bad angles:	3 / 10164	0.03%	Goal: <0.1%
Peptide Omegas	Cis Prolines:	0 / 36	0.00%	Expected: ≤1 per chain, or ≤5%
Low-resolution Criteria	CaBLAM outliers	37	4.1%	Goal: <1.0%
	CA Geometry outliers	12	1.33%	Goal: <0.5%
Additional validations	Chiral volume outliers	0/1128		
	Waters with clashes	1/225	0.44%	See UnDowser table for details

The free R went up a tiny amount, but the geometry is tighter and the MolProbity report looks significantly better. Of course, the ncs is also much tighter since I started refinement with a model and perfect ncs.

There was one difficulty. Coot seems to distort the geometry of hydrogen atoms when it performs real-space refinement. The bond lengths drop from Buster's 1.1 Å to about 0.9 Å. Coot's bonds are so short that it draws bonds between 1-3 related hydrogen atoms. Oddly enough Buster does not correct these bond lengths. To get a consistent set of hydrogen atoms, I deleted them all and used the Buster command (actually a borrowed Duke program) hydrogenate.

Model Building: Round 2

The goal of this round of model building was to examine every residue of all three chains and look for changes which need to be made to break the noncrystallographic symmetry. I examined Buster's model in Coot using its Calculate.NCS Maps… command. This doesn't work as well as I'd like since Coot doesn't calculate the averaged map over enough space to cover all the individual chains. This shortcoming means that I was unable to compare the model to the averaged map in some regions. In those places I simply assumed that the model was optimized to the average, since that is what Buster was supposed to do, and a disagreement between the placement and the local map indicates a symmetry violation.

I didn't record the particular residues I changed. There were several dozen.

I made a major change near C|267. I flipped the peptide bonds on either side. This is an experiment to see what Buster does, and I need to compare the new interpretation with that of the other two chains.

The beta strands near 234 continues to appear to be shifted by a fraction of a residue. Apparently my work in the last round of model building didn't do the job. I didn't have any new ideas this time.

I made a small change to the distal Zinc binding site. Previously I had only built the SG atoms for 224 and 226, since their locations are defined by the parts of the Zinc locus already built. I can see more density near 226 in each chain so I have added the CB atom.

A final touch up was the addition of the N and CA atoms of 315 to each chain. There is weak density for them but, more importantly, the locations of these atoms are implied by the preceding residue and the peptide bond geometry. I think everyone should be building these when the density of a chain dribbles out.

Buster: Round 3

Run	Rwrk	Rfree	Bond	Angle	ncs	
6w9c	23.5%	30.9%				Deposited values
6w9c	24.05%	31.09%	0.013	1.325	---	Buster values/KD Data
hydNCS	18.07%	25.81%	0.010	1.070	yes	
model-1	24.20%	27.05%	0.031	2.385	yes	Now perfect ncs
refine-2	22.84%	26.06%	0.007	0.776	yes	
model-2	23.76%	26.17%	0.011	0.968	no	Now broken a little
refine-3	18.86%	24.79%	0.008	0.886	yes	

This round of refinement was run with the same options as before. The model building and refinement resulted in a pretty good improvement in the free R.

My big change at C|267 didn't work out well. The geometry at 267 and 265 was bad enough, even after refinement, that some of the biggest outliers in the model were here. I decided that this couldn't stand. I went back to try other arrangements of the peptide bonds at this residue, and examined the other chain's versions. With each attempt I reran the Buster refinement before making a final judgment.

The final model made this residue similar to a beta sheet, with the carbonyl oxygen pointing in opposite directions on either side. This turned the follows residues into a beta turn, but only the B chain has a short enough hydrogen bond between the strands. The other peptide still doesn't make any hydrogen bonds, but that was true in all attempts.

There are a number of regions where the main chain doesn't make good hydrogen bonds. I have tried to come up with solutions, but my imagination has failed me.

Model Building: Round 3

My goal is to build a water model that complies with the noncrystallographic symmetry. So far I have started with the deposited water molecules, and just accepted the changes that Buster made. Neither of these included any consideration of the symmetry of the three chains in their decisions.

I'm told that Buster will enforce ncs restraints if the water molecules are separated into the three chains, and the equivalent water molecules given identical names. Global Phasing supplies no tools to do this. I don't know of anyone who does. I think I can accomplish this using TNT and a lot of manual intervention.

I took the results of the latest refinement, which contains 255 water molecules in the W chain. In the TNT control file I defined a single cluster of constrained ncs that covers all of each chain, along with the Zinc and Chloride ions. The TNT's **gather** command then produces a single molecule that is the average of the three chains, and a file with the ncs transformations.

Next I edited the coordinate file with the water molecules to change the chain name from W to "Protein", the name of the cluster. Using the **scatter** command and the ncs transformations creates copies of the water molecule transformed to overlay A, B, and C. Then I deleted the original water molecules in W from the .pdb file and appended these. By the way, I also renamed the water molecule so that they are numbered starting from 1001.

Now I have a set of waters which are completely compliant with the ncs. A|1001 is equivalent to B|1001 and C|1001. Unfortunately, very often Buster has picked water molecules that are equivalent to another w/o matching numbers. These have to be filtered out.

I opened this coordinate file in Coot with the new map and walked over each water molecule in B. If there is another water molecule near the current one, mark it to be killed. Also kill the water with the same number in the other two chains. This clears out the original duplicates. I also killed water molecules w/o any density and/or are far from any hydrogen bond partner. To save time I killed all waters with numbers greater than 1080.

Buster isn't very good a filtering water, so I ended up killing a lot of them. I'm now down to a little over one hundred water molecules.

This protocol, basically, implements my idea of expanding water by the ncs using a quorum of one. Any water molecule in the original list has now been replicated to the other two chains, even if the density for that chain doesn't support a water. In previous tests a quorum of one was superior to any other value.

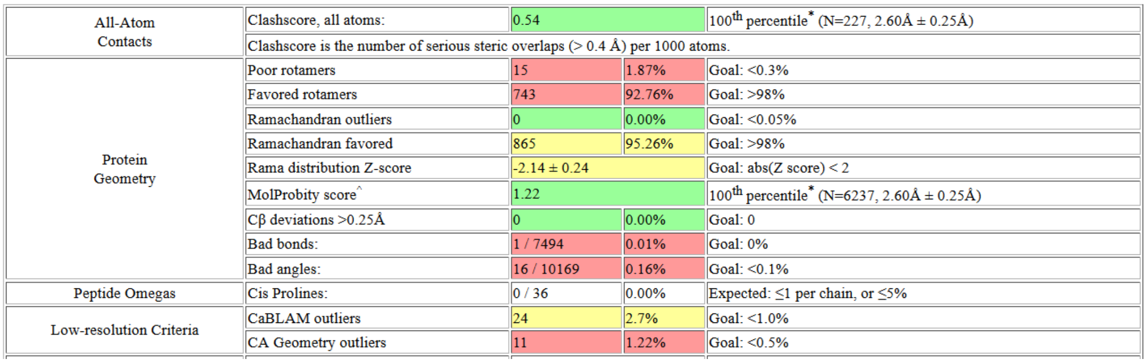
Buster: Round 4

Run	Rwrk	Rfree	Bond	Angle	ncs	
6w9c	23.5%	30.9%				Deposited values
6w9c	24.05%	31.09%	0.013	1.325	---	Buster values/KD Data
hydNCS	18.07%	25.81%	0.010	1.070	yes	
model-1	24.20%	27.05%	0.031	2.385	yes	Now perfect ncs
refine-2	22.84%	26.06%	0.007	0.776	yes	
model-2	23.76%	26.17%	0.011	0.968	no	Now broken a little
refine-3	18.86%	24.79%	0.008	0.886	yes	
model-3	22.92%	26.54%	0.008	0.886	no	Symmetrized water, truncated
refine-4	20.75%	25.53%	0.007	0.830	yes	

This is just another round of refinement to clean up the model with more rigorous water model. To keep Buster from messing with my solvent model I turned off its solvent rebuilding step. I ran a control where Buster was allowed to mess with the water. That test resulted in a drop of nearly one percent in free R with the addition of about a hundred more water molecules. I didn't really like the density or the hydrogen bonding geometry of these additions so I rejected this run. (I'm none too happy with the water I've built but I'm sticking with that.)

The only other change from the previous refinement is that I corrected an error in my refinement command. I had, inadvertently, been refining with the hydrogen atom occupancies set to zero. In this round these atoms scatter X-rays! This change resulted in an increase in the free R by a couple tenth of a percent, which I consider insignificant.

I don't believe that another round of model building will be useful, so this is the final model.



All-Atom Contacts	Clashscore, all atoms:	0.54		100th percentile* (N=227, 2.60Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	15	1.87%	Goal: <0.3%
	Favored rotamers	743	92.76%	Goal: >98%
	Ramachandran outliers	0	0.00%	Goal: <0.05%
	Ramachandran favored	865	95.26%	Goal: >98%
	Rama distribution Z-score	-2.14 ± 0.24		Goal: abs(Z score) < 2
	MolProbity score^	1.22		100th percentile* (N=6237, 2.60Å ± 0.25Å)
	Cβ deviations >0.25Å	0	0.00%	Goal: 0
	Bad bonds:	1 / 7494	0.01%	Goal: 0%
	Bad angles:	16 / 10169	0.16%	Goal: <0.1%
Peptide Omegas	Cis Prolines:	0 / 36	0.00%	Expected: ≤1 per chain, or ≤5%
Low-resolution Criteria	CaBLAM outliers	24	2.7%	Goal: <1.0%
	CA Geometry outliers	11	1.22%	Goal: <0.5%

Summary

With this additional refinement of the 6w9c model of the Papain-like Protease Domain of nsp3 we have reprocessed the data starting from the deposited images, reduced the amount of radiation damage in the data, increased the high resolution data cutoff from 2.7 to 2.6 Å, imposed strict noncrystallographic symmetry restraints to compensate for the extremely low completeness of the data, and manually improved the model's fit to the electron density maps and expected hydrogen bonding patterns. Despite forcing the model to be consistent with all these additional requirement (but probably *because* we did this) the free R of the new model has dropped by about 5 percentage points while decreasing the difference between the working and free R's by about 2 percentage points.

In addition, the new model has a better match to the expected chemical properties of a protein, as reflected in the MolProbity report. The "Clashscore" has dropped from 6.57 to 0.54 and the MolProbity score has changed its ranking from the 77th percentile to the 100th. In addition, the number of "poor rotamers" has dropped from 7.02% to 1.87% and the number of "favored rotamers" increased from 82% to 93%.

All is not well and this is still a very poor model. The number of "bad angles" has increased from 2 to 16 and there are still a number of quality indicators in the MolProbity summary that are marked in yellow and red. While I would argue that two of the sixteen bad angles are false positive, nearly all of the rest are Ca-Cb-Cg angles that are too large in Aspartic Acid residues. These are caused by the difficulty of placing side chains of these types in maps this poor.

And the electron density map for this data set is very poor. The density is weak, especially in the N and C terminal domains, has broken connectivity and false connections as well. While NCS averaging the map does improve it and overcomes some of the problems due to the low completeness of the diffraction data, it also is rather poor.

The distal (and non-artifactual) Zinc binding site is nearly unidentifiable in two of the three asymmetric units. All that can be seen in those chains is that there is an unusual density of electrons in the region where the Zinc and four Sulfurs are expected to be.

I have included water molecules in the model, which is difficult even in complete 2.6 Å maps. While these atoms have been placed in peaks in the map by the Buster/TNT refinement program, these peaks are often indistinct and have poor geometry for their hydrogen bonds. The model has been constructed so that the NCS symmetry can be imposed on these freely floating molecules as well as the protein, but I do not believe they are very trustworthy.

Fortunately the active site is located in the central domain where the map is at its best. In this region I can generally identify the locations and conformation of the side chains and the main chain carbonyl bumps. Even here the hydrogen bonding geometry does not look good to my trained eye.

This model reflects my best effort to interpret these data, but it should not be trusted in its details.

Acknowledgements

Refinement was performed using computer access donated by the Guillemin Lab at the University of Oregon. The principal software packages used were Coot, Buster/TNT, and CCP4.
.