

Atomic Fluidity in Macromolecular Refinement

Dale E. Tronrud

The Light & Electron Works
Springfield, Oregon, USA

January 21, 2024

Abstract

A method is proposed which might be able to allow a traditional refinement program correct otherwise intractable errors like peptide flips and rotomer changes. It basically increases the radius of convergence in situations where there is atomic model congestion.

0 *Prolog*

A question from James Holton has dredged up memories of a refinement issue solution that I thought of back in the 1990's, when TNT development was progressing apace. Implementing this idea would have required fundamental changes to TNT's structure so it never was implemented and tested. (As was the case with quite a number of other solutions I developed at that time and all together they eventually led to the ossification of TNT and the death of its development ten years later. Any significant improvement I wanted to make required a rewrite of the entire package.)

While I'm writing this now in 2024 the text is constructed as it would have been back thirty years ago. I'm not making any claim that I wrote down any of this back then, nor do I have any proof that I thought of this back then. It has just been kicking around in my gray matter since then.

I'm also not swearing anyone to secrecy about this. I can't write it up as a paper since I don't have an implementation and I can't demonstrate that it works or is actually useful in any way. I had hoped to do something with the idea eventually, but I think that ship has sailed. While I'm sitting down now to write in response to James I plan to pass this pdf on to other people I think might be interested. They are free to pass it on to anyone else. If someone makes use of the idea maybe they'll be kind enough to give me a 'personal communication' citation.

As you'll soon find out, if you didn't already know, I see no virtue in brevity. Most of you can just skim the first section. Section 2 is when things get interesting.

1 The Problem at Hand

The focus of my professional life is the development of the TNT Refinement Package (Tronrud et al. 1987). TNT is a gradient descent optimization program which means it improves an atomic model by making small changes to the model's parameters to improve its fit to a set of standards that include diffraction data and geometric restraints. At TNT's very core is the Calculus and this requires that each change of a parameter within one iteration must be very small and each small step must be an improvement. It does not optimize the model by moving parameters in the most "steeply downhill" direction since it uses some second derivative information, but, still, each step must be to a lower place than the previous.

A number of errors that commonly occur in preliminary protein models cannot be corrected by TNT because of this design. Often these models have problems that can only be fixed by rotating a group of atoms about a torsion angle. TNT, and other programs like it, cannot transform a model with such an error into the correct one because the intermediate states fit the density far worse than the incorrect starting configuration. TNT cannot go uphill, even a little bit.

Three methods of correcting this type of problem have been implemented and are commonly used. The first is to expect the crystallographer to identify and correct the error while using a graphical model-building program. Humans are very powerful pattern recognition instruments and, if the error is actually a simple rotation about one torsion angle, are highly reliable at finding and fixing these problems. If they are both properly trained and

paying attention, that is. Long experience has shown, however, that users cannot be trusted to fix these errors.

The second method is to write software that attempts to perform the same actions as the human, *i.e.* identify the existence of this type of error and edit the model to correct it. This solution works well, but creating such software is outside my skill set and doesn't fit into the way TNT is developed and used.

The third method is to implement a refinement algorithm which isn't purely gradient descent and can go up hill. I hate simulated annealing. I hate simulated annealing. I hate simulated annealing. I am never going to write code into TNT to do it.

This report describes a fourth method which I am unaware of being used in macromolecular crystallography. In it atoms are not required to move long distances but, instead, the atom names move.

Two classes of errors that are both common and problematic are peptide flips and the rotameric conformation of the end of Leucine side chains. In this document I will consider these as prototypes for the entire class of errors that require large rotations of a pair of atoms about about a bond vector.

2 Atom Identities as Variables

Let's look in detail at the problem of Leucine side chains. In one of the common rotomers of Leucine the CD1 and CD2 atoms are further from the main chain than CG (a conformation I call "ears up"). One could build a model with the other side chain atoms in the same place but these atoms closer to the main chain ("ears down"). Ears down is never seen in very high resolution structures and yet Leucine side chains with this conformation are build in early models.

When TNT is presented with a model containing this error (and the resolution is around 2 Å or better) it will allow the bond angles containing these atoms to widen a bit to allow the fit to the density to improve. How this all plays out depends on the relative weighting of density fit vrs ideal bond angles but if the density wins the side chain will pass through flat and adopt an ears up conformation with better density fit *and* good bond angles.

The consequence of this behavior is that the pseudochirality of the CG atom changes and this pisses off the user. Even though the model fits all the

restraints in the TNT library, the labels on the CD1 and CD2 atoms still have to be switched. (For you nonTNT aficionados, TNT does not restrain chirality but, instead, only warns the user with the expectation that they will edit the coordinate file to exchange the labels or otherwise do what is required.)

Other refinement programs have functional chirality restraints, either chiral volumes or improper dihedrals, and the sliding of atoms into a chiral flip is forbidden. The users are happy but the difference map is not. Either way the user has to do something to fix the model.

TNT’s solution, and its shortcoming, does bring to mind that the remaining problem has to do with the labeling of atoms and not really the issue of moving atoms around in space. Basically, we can smoothly move atoms around and into better positions, without transiently going up hill, if we allow atom labels to be flexible.

Allowing this flexibility isn’t as simple as it is with an x coordinate since a label isn’t, itself, variable like a floating point number. TNT has to calculate derivatives of anything that is changeable in the model. My proposal is to say that an “ear” does not *have* a label, but has a probability, P_1 , that it has the label CD1 and probability, $1 - P_1$, that it has CD2. The other “ear” has the swapped probabilities. For most Leucine side chains P_1 will be nearly equal to one and there is no ambiguity, but when the difference density is bad and/or the bond angles are distorted one could design the restrains to lead to a more ambiguous situation.

For TNT the behavior would be fairly simple. P_1 would become closer to 0.5 as the arrangement of atoms around CG becomes flat. If the pseudochirality flips, P_1 would simply drop below 0.5, continuously. Now the difference density and bond angle restrains would drive the side chain to a more pyramidal form and P_1 would drop to close to zero. The side chain would flip without the requirement for a rotation about the CB-CG torsion angle and the “chirality” would never be incorrect.

Since TNT has no chirality restraint it would be reasonable to have a very sharp transition of P_1 from 1 to zero as CG moves through planarity. For programs with chirality restraints, there would be two terms, one for each labeling and weighted by P_1 , and these two restraints would cancel when CG is flat and P_1 is 1/2. The side chain would be free to go whichever direction the difference density indicated and, once asymmetry is re-established, P_1 would start to drive the labels, and chiral restraints, to the values consistent with that configuration. In this case, setting the fuzziness of P_1 ’s variation

and the relative weights of the chiral restraint, the bond angles, and difference density would be more delicate and probably dependent on the resolution of the diffraction data.

2.1 Peptide Flips

A peptide bond that needs to be flipped will usually have severe distortions of bond angles and bond lengths. The atoms in the peptide bond cannot simply rotate about the bond's axis because the N and O atoms would have to move into regions that usually have no density and TNT can't move atoms up hill, if you recall. Correcting the error by shifting the atomic positions of the two atoms directly to their correct locations would require moving them across the axis of the peptide bond and distortions of geometric restraints would be enormous. The atoms would, in addition, have to pass through each other and that would also be horrible violations of a number of geometric restraints.

Imagine instead that the labels of N and O are fluid and controlled by a parameter in the model called P_N . In a happy peptide bond, P_N would be either one or zero, it doesn't matter, and one atom would be definitely identified as N and the other O. If the peptide bond is flipped, however, the fit to density and geometry restraints would all be unhappy. Putting in the concept of P_N , with a properly contrived calculation of its value based on all this unhappiness, would result in a value less than one. Now geometry restraints for a model with swapped labels would start to contribute. The principal difference would be that the bond to CA would start to appear between it and what used to be, certainly, identified as O. Having a partial bond to what used to be purely O and a weaker bond between CA and what used to be certainly N would allow the "O" atom to move more toward the peptide bond axis and the "N" atom to move further out. Since the flipped confirmation is assumed here to be the correct one, these shifts should improve the fit to density and drive P_N further and further from one. My expectation is that the arrangement of atoms in this peptide bond can be fixed by continuous, tiny, shifts that are all downhill!

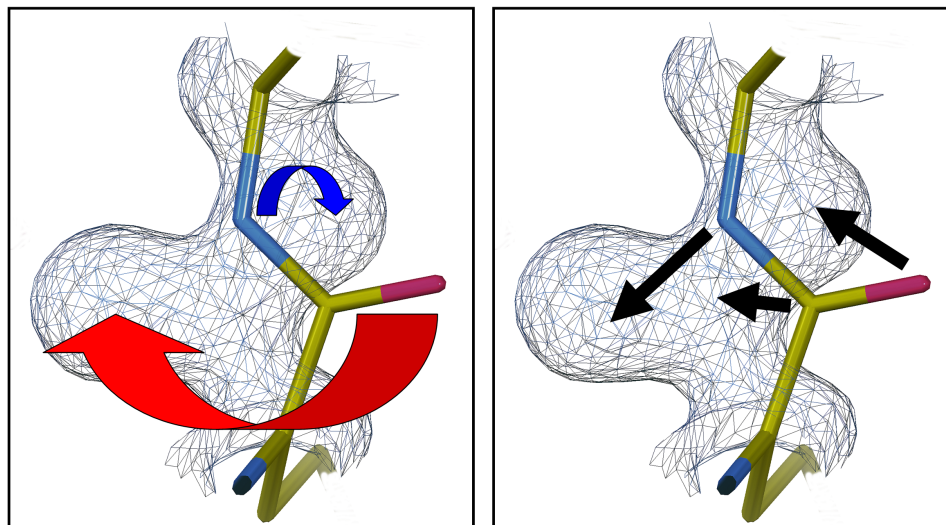


Figure 1: **Comparison of Shifts Required by Rotation and Label Swap.** Each panel shows electron density for a peptide bond with the superimposed model flipped. On the left is indicated the large rotation required to correct the model. The right panel shows the smaller shifts which bring the carbonyl oxygen into the density for the amide nitrogen and the amide nitrogen atom shifted into the density for the carbonyl oxygen. These shifts and an exchange of labels is all that it required to correct the model.

loop_ _chem_comp_bond.comp_id _chem_comp_bond.atom_id_1 _chem_comp_bond.atom_id_2 _chem_comp_bond.type _chem_comp_bond.value_dist _chem_comp_bond.value_dist_esd						loop_ _chem_comp_angle.comp_id _chem_comp_angle.atom_id_1 _chem_comp_angle.atom_id_2 _chem_comp_angle.atom_id_3 _chem_comp_angle.value_angle _chem_comp_angle.value_angle_esd					
BCL	CAB	C3B	coval	1.500	0.020	BCL	CBB	CAB	C3B	125.000	3.000
BCL	CAB	OBb	coval	1.220	0.020	BCL	OBb	CAB	C3B	122.000	3.000
BCL	CAB	CBB	coval	1.540	0.020	BCL	OBb	CAB	CBB	113.000	3.000

Table 1: **Bond Length and Angle Definitions for the Acetaldehyde Group in BCL.** The standard geometry values for the acetaldehyde group demonstrate considerable asymmetry. These differences can be used as clues to which atom is which when the diffraction data is of sufficient quality to make the distinction.

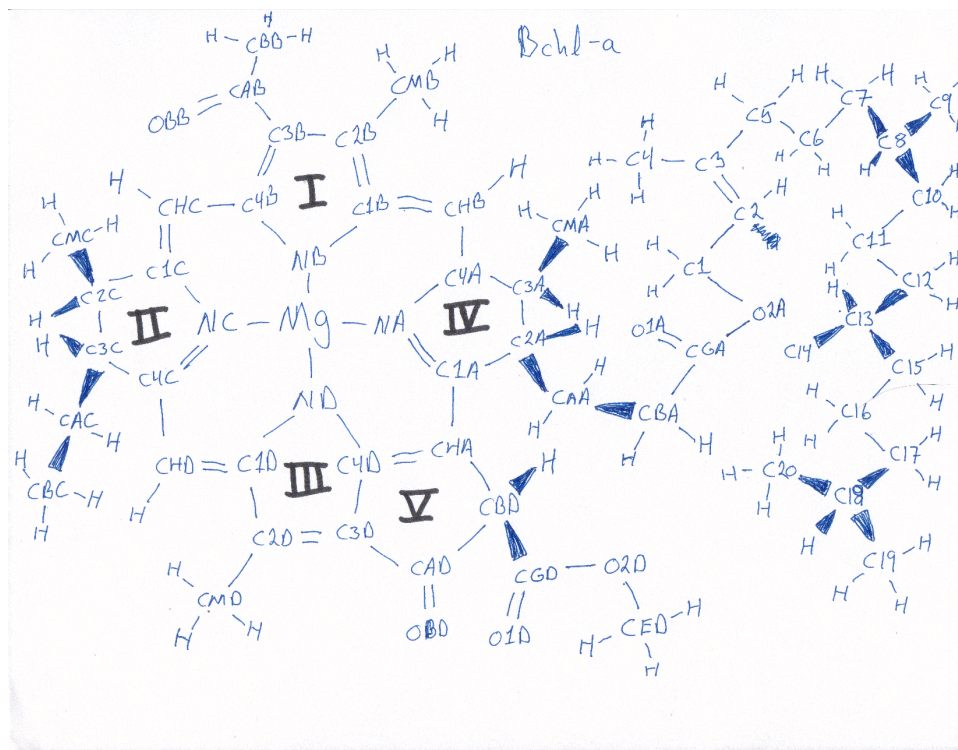


Figure 2: **Chart Defining the Atom Labels for Bacteriochlorophyll-a.** The part of BCL I'm concerned with today is at the top on the left. This is an acetaldehyde group bonded to the edge of the conjugated ring system, and extends the conjugation. (Sorry for the quality but this is how I keep track such things.)

3 A Really Simple Example

On the edge of the conjugated ring system in Bacteriochlorophyll-a is an acetaldehyde group. (See Fig 3.) One branch of this group is an oxygen atom labeled OBB and the other branch has a carbon atom labeled CBB. This group is flat and participates in the conjugation so all the atoms lie, pretty much, in the plane of the porphyrin ring. The group, however, is linked via a single bond and can rotate by 180° to swap the positions of the two branches. In the bacteriochlorophyll containing protein (now named “FMO”) (Tronrud et al. 1986, 2009) there are seven well-ordered copies of this molecule and both orientations occur within the set. The problem is that it is rather hard to decide from the density which branch is the OBB atom.

Information is available which can be used to determine which atom is which. In the ideal geometry both the bond lengths and angles (Table 3) are asymmetric and provide clues. At high resolution I have actually seen positive and negative difference map peaks over mislabeled atoms because the true carbon atom has significantly fewer electrons than the true oxygen. I was easily able to label these atoms by using this information in my 2009 1.3 Å structure determination, but I was much less reliable in 1986 when working with 1.9 Å resolution data.

The problem I hope to overcome is that it is impossible for TNT to correct a mislabeling of these atoms even with sufficient diffraction data that make the choice clear. This situation is a good candidate for continuously variable atom labeling.

I’ll name the two atoms in question #1 and #2 and the probability that atom #1 is the carbon is P_{CBB} . Of course, the probability that atom #2 is the carbon is $1 - P_{CBB}$. The least-squares residual equation for the geometry restraints is

$$\begin{aligned}
 P_{CBB} & \left((b_o[CAB, CBB] - b_c[CAB, \#1])^2 + \right. & (1) \\
 & (b_o[CAB, OBB] - b_c[CAB, \#2])^2 + \\
 & (\tau_o[CBB, CAB, C3B] - \tau_c[\#1, CAB, C3B])^2 + \\
 & (\tau_o[OBB, CAB, C3B] - \tau_c[\#2, CAB, C3B])^2 + \\
 & \left. (\tau_o[OBB, CAB, CBB] - \tau_c[\#2, CAB, \#1])^2 \right) + \\
 (1 - P_{CBB}) & \left((b_o[CAB, CBB] - b_c[CAB, \#2])^2 + \right.
 \end{aligned}$$

$$\begin{aligned}
& (b_o[CAB, OBB] - b_c[CAB, \#1])^2 + \\
& (\tau_o[CBB, CAB, C3B] - \tau_c[\#2, CAB, C3B])^2 + \\
& (\tau_o[OBB, CAB, C3B] - \tau_c[\#1, CAB, C3B])^2 + \\
& (\tau_o[OBB, CAB, CBB] - \tau_c[\#1, CAB, \#2])^2)
\end{aligned}$$

where $b[?, ?]$ is a bond length, $\tau[?, ?, ?]$ a bond angle, and with “ o ” a library value and “ c ” calculated from the model. When $P_{CBB} = 1$ the residual is simply that with atom #1 labeled with CBB and when $P_{CBB} = 0$ atom #2 has that label. There is nothing in this function to prefer one labeling over the other.

Asymmetrical geometric restraints can be successfully imposed on a model with fixed labels. (*i.e.* regularization) If the labels are swapped and geometry idealized a perfect fit can still be achieved. A model which has two labels for each of the problematic atoms, with equal probability, must be symmetric and therefore can never perfectly obey the restraints. This means that the geometry residual function is maximal when P_{CBB} is 0.5 creating a barrier to exchanging labels. Only diffraction data can introduce a ramp, when the difference map residuals are added to the geometric, and drive a label swap, and even then only when the map is of sufficiently high resolution.

At the start, a mislabeled model will be subject to asymmetrical geometry restraints which are not compatible with the, also asymmetric, electron density. TNT will try its best to fit all the restraints but, with the labels held fixed, it can only achieve a frustrated state. Allowing the labels to become fluid opens a pathway through which the labels change and the atoms can slide into the other asymmetric state which is now consistent with the map.

Even if the geometry restraints were symmetric the difference in scattering factor between a carbon and an oxygen atom would be able to drive P_{CBB} to either zero or one. A success like that isn’t guaranteed. It would probably far more difficult to distinguish between OD1 and ND2 in asparagine residues since there is only one electron difference in density and the motion of the atoms is usually higher. I think this method could be, on the other hand, used to flip histidine side chain rings. I’ve certainly seen difference density maps for incorrect histidine rings that show positive peaks where the nitrogen atoms should be placed.

Adding the parameter, P_{CBB} , to a model subject to low resolution refinement will only result in ill-determined parameters, not only P_{CBB} but the B factors too since they are connected to the scattering factors which, themselves, are tied to the labels. The management of ill-determined parameters

is a problem for which I have no solution. My guess is that the principle clue that the diffraction data is insufficient to define the labeling is when P_{CBB} does not refine to a value close to either extreme value.

4 Complicating Matters

- Code will have to be written to define all the labeling ambiguities and create the code for calculating the label-probability weighted residuals.
- When there is geometric asymmetry the labeling probability and the positions of the atoms will be connected parameters and they will not effectively refine to optimal values unless they are refined in the same calculation. TNT is pretty good at simultaneously refining parameters of different types but other refinement programs don't seem to have this skill.
- Need to devise a detector for when the diffraction data aren't good enough to define the labels. We need to know when the labeling is not defined by the data.
- Prior to deposition the label probabilities have to be projected to certainty. If the probability of a label is 60% we need to just adopt that label.

References

- Tronrud, D. E., Schmid, M. F. and Matthews, B. W.: 1986, Structure and X-ray amino acid sequence of a bacteriochlorophyll *a* protein from *prothelocochloris aestuarii* refined at 1.9 Å resolution, *J. Mol. Biol.* **188**, 443–454.
- Tronrud, D. E., Ten Eyck, L. F. and Matthews, B. W.: 1987, An efficient general-purpose least-squares refinement program for macromolecular structures, *Acta Cryst.* **A43**, 489–501.
- Tronrud, D. E., Wen, J., Gay, L. and Blankenship, R. E.: 2009, The structural basis for the difference in absorbance spectra for the FMO antenna protein from various green sulfur bacteria, *Photosyn Res* **100**(2), 79–87.