Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

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)

Boguslaw Stec

Infectious and Inflammatory Diseases Center, Burnham Institute for Medical Research, 10901 North Torrey Pines Road, La Jolla, California 92037, USA

Correspondence e-mail: bstec@burnham.org

Received 29 June 2007
Accepted 21 August 2007

Crystallographic refinement is an important subject. At the end of the 1980s, Petsko wrote that

> it is well to remember that just over 10 years ago it was commonly felt that protein structures could not even be refined

(Ringe & Petsko, 1986). With the emergence of crystallographic programs, this seemingly impossible task has become routine and even automated.

The problem of aiding the refinement process with stereochemical constraints in the absence of sufficient experimental data is an important issue. Precisely for this reason, the paper by Jaskolski *et al.* (2007) is welcome and perhaps even overdue. It summarizes developments over the past several years, including the emergence of high-resolution protein crystallography and the explosive growth of the PDB. However, despite warnings that it is for 'the cooks rather than the chefs', the paper does not discuss the wider ramifications of the problem, the elucidation of which might clearly benefit the 'cooks'.

Among many positive aspects, the paper by Jaskolski *et al.* (2007) contains interesting observations that should be much more widely implemented, such as the scaling problem in the determination of cell dimensions that leads to protein models that are either 'squeezed' or 'expanded'. It would be important to follow this up with more studies and to resolve the issue of to what degree standard indexing methods introduce anisotropy in the scaling problem. It is also possible that the fastest growing crystal directions introduce more disorder, which results in less than perfect lattice-period determination. Well behaved stereochemical parameters, such as the C═O bond, may serve as a probe to indicate directional distortion.

The general recommendations of the paper, while not dramatic, appear to be well supported by other studies. However, several important factors not covered by Jaskolski and coworkers that are highly relevant to the problem of constraints and derivation of reliable protein models should be mentioned. We focus here on four factors (neglecting many others that are perhaps equally important), such as the dual (solid–liquid) nature of protein structure, context-dependent stereochemistry, the change of paradigm for protein-structure determination (classical–quantum) when higher resolution data are available and finally the approximate nature of symmetry.

In the last several years it has become rather obvious that proteins have a special nature that appears to be a non-ergodic glassy state that combines the features of two different states of matter: the solid and the liquid states (Fenimore *et al.*, 2004; Teeter *et al.*, 2001). This dual nature of proteins is reflected in protein crystal structures by two contrasting features. Firstly, the temperature factors behave in accord with a rigid-body motion (protein as a solid; Kuriyan & Weis, 1991). Secondly, this model always breaks down with excessive 'motion' (high temperature factors) and the existence of disorder (protein as liquid). Jaskolski *et al.* (2007) are aware of the fact that most high-resolution structures have substantial disorder and try to separate ordered from disordered elements in order to draw their conclusions. However, it is not obvious at all that such idealities are unique or common to both dynamic and static structures.

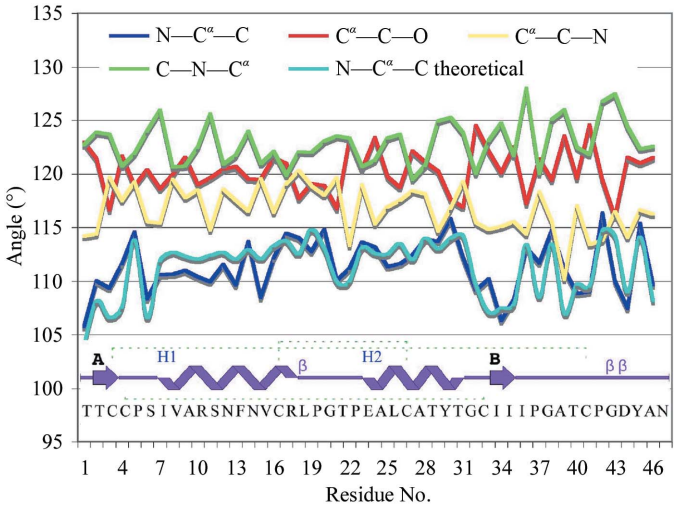

Figure 1
The planar angles as obtained from a model of crambin refined at 0.83 Å resolution by the full-matrix least-squares method without constraints (the disordered elements were minimally constrained; Stec *et al.*, 1995). The N—C^{α}—C theoretical curve represents the result of *ab inito* quantum-mechanical optimization as obtained by Van Alsenoy *et al.* (1998). There is a remarkable agreement between the high-resolution crystal structure and the result of the quantum-mechanical optimization. Below the curves, the excerpt from the PDB sequence viewer is shown to visualize the secondary structure (purple wavy line) and the disulfide bonds (green broken line). A clear correlation between the general behavior of the planar angles and the secondary-structure elements is visible.

There is no convincing experimental evidence that the idealities that the authors discuss have universal applicability, even for well ordered parts of the protein (solid-state components). In fact, there is emerging evidence that this is not the case (Esposito *et al.*, 2000). Recent results suggest that protein stereochemistry is context-dependent. This view is illustrated by the correlation of the results obtained by Stec *et al.* (1995) with quantum-mechanical optimization of crambin. Such a comparison showed the systematic dependence of several parameters (Van Alsenoy *et al.*, 1998) on the secondary structure (Fig. 1). The well established methodology of the χ^2 test also provided evidence that the distributions of some of the stereochemical parameters are not unimodal (Stec *et al.*, 1995; Vlassi *et al.*, 1998). Had Jaskolski and coworkers used this test, it would have confirmed non-Gaussian distributions of several other variables beyond those visible in Fig. 4 of their paper (the N—C^{α}—C angle). The elongation of the C=O bond as well as corrections to the C—N bond length and N—C^{α}—C angle have also been recommended by two previous studies (Stec *et al.*, 1995; Vlassi *et al.*, 1998).

The higher the resolution, the more disorder is observed, and the protein structure is less uniquely defined as a solid. Therefore, despite reaching higher apparent accuracy, it becomes more difficult to find the universality that the authors seek. At an average resolution (~2 Å), the accuracy of the idealities is not essential to obtain a reliable protein model (the expected error in a model would be ~0.2 Å). At high resolution the problem becomes complex and the highest resolution structures usually do not provide the needed resource [*e.g.,* as mentioned in the paper, the questionable quality of crambin at 0.54 Å resolution (PDB code 1ejg) as well as others]. At high resolution, methods such as least squares become less-than-perfect tools, because numerical implementation of matrix algebra methods does not handle singularities well. Atoms that are located closer than the resolution of the data are randomly shifted (sometimes at the cost of worsening *R*), leading to large stereochemical distortions. This particular problem calls for special methods that couple restraints with nonlinear optimization. For example, methods that combine *ab initio* quantum-mechanical optimization with crystallographic refinement are being developed, and such methods need to be mentioned (Yu *et al.*, 2005; Zarychta *et al.*, 2007; Volkov *et al.*, 2007).

Finally, the symmetry of a crystal lattice is only as good as our conventions. The symmetry determination only holds up to the accuracy of our analysis as expressed in $R_{\rm merge(sym)}$, which is usually above 5% (significantly higher than the 'small-molecule' standards). This leads us to believe that the symmetry of most deposited models in the PDB is not strictly obeyed, as the solvent is not expected to obey the ideal symmetry. In order to perform comparisons such as those attempted in this paper, it would be important to compare only the structures of proteins with similar size, the same space group and very well modeled solvent. The details of such a selection go beyond the analysis of Jaskolski and coworkers and beyond these remarks.

In conclusion, the paper by Jaskolski and coworkers is a valuable contribution to the complex subject of crystallographic refinement and a starting point for discussion on improving protein models in general (Furnham *et al.*, 2006). Such a discussion has already been initiated by a valuable session (01.07 Computational Methods) of the recent American Crystallographic Association meeting in Salt Lake City, USA.

References

Esposito, L., Vitagliano, L., Sica, F., Sorrentino, G., Zagari, A. & Mazzarella, L. (2000). *J. Mol. Biol.* **297**, 713–732.
Fenimore, P. W., Frauenfelder, H., McMahon, B. H. & Young, R. D. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 14408–14413.
Furnham, N., Blundell, T. L., DePristo, M. A. & Terwilliger, T. C. (2006). *Nature Struct. Mol. Biol.* **13**, 184–185.
Jaskolski, M., Gilski, M., Dauter, Z. & Wlodawer, A. (2007). *Acta Cryst.* D**63**, 611–620.
Kuriyan, J. & Weis, W. I. (1991). *Proc. Natl Acad. Sci. USA*, **88**, 2773–2777.
Ringe, D. & Petsko, G. A. (1986). *Methods Enzymol.* **131**, 389–433.
Stec, B., Zhou, R. & Teeter, M. M. (1995). *Acta Cryst.* D**51**, 663–681.
Teeter, M. M., Yamano, A., Stec, B. & Mohanty, U. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 11242–11247.
Van Alsenoy, C., Yu, C. H., Peeters, A., Martin, J. M. L. & Schafer, L. (1998). *J. Phys. Chem. A*, **102**, 2246–2251.
Vlassi, M., Dauter, Z., Wilson, K. S. & Kokkinidis, M. (1998). *Acta Cryst.* D**54**, 1245–1260.
Volkov, A., Messerschmidt, M. & Coppens, P. (2007). *Acta Cryst.* D**63**, 160–170.
Yu, N., Yennawar, H. P. & Merz, K. M. (2005). *Acta Cryst.* D**61**, 322–332.
Zarychta, B., Pichon-Pesme, V., Guillot, B., Lecomte, C. & Jelsch, C. (2007). *Acta Cryst.* A**63**, 108–125.

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Numerology *versus* reality: a voice in a recent dispute

Mariusz Jaskolski,[a] Miroslaw Gilski,[a] Zbigniew Dauter[b] and Alexander Wlodawer[c]*

[a]Department of Crystallography, Faculty of Chemistry, A. Mickiewicz University and Center for Biocrystallographic Research, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland, [b]Synchrotron Radiation Research Section, Macromolecular Crystallography Laboratory, NCI, Argonne National Laboratory, Biosciences Division, Building 202, Argonne, IL 60439, USA, and [c]Protein Structure Section, Macromolecular Crystallography Laboratory, NCI at Frederick, Frederick, MD 21702, USA

Correspondence e-mail: wlodawer@ncifcrf.gov

Received 5 October 2007
Accepted 8 October 2007

We have recently published a paper in this journal aimed at suggesting what values of root-mean-square deviations (r.m.s.d.s) of bond lengths and angles should be expected in well refined protein structures (Jaskolski *et al.*, 2007). It seems that some of our recommendations, which were in our opinion straightforward and non-controversial, have nevertheless generated considerable discussion (Stec, 2007; Tickle, 2007). Whereas both of these papers criticize some of the recommendations presented by us, the conclusions reached in them are quite contradictory, as will be pointed out below. We humbly admit that our recommendations appear to be in conflict with *some* previous experimental and theoretical work in this area, especially that of Tickle and coworkers (Tickle *et al.*, 1998), and that they may indeed lack very strict 'either experimental or theoretical basis' (Tickle, 2007). Our suggestions were based on quite straightforward analysis of the restraint libraries of Engh & Huber (1991, 2001) as well as of the structures deposited in the Protein Data Bank (PDB; Berman *et al.*, 2000) and Cambridge Structural Database (Allen, 2002). We were guided by our practical experience in refining and validating a large number of various crystal structures during about 30 years of our activity in the field. Indeed, we often tend to rely on experience rather than on elaborate numerical calculations. The latter sometimes are very sophisticated and absolutely correct mathematically, but may not be highly relevant if some of the underlying assumptions are not exactly fulfilled. It is our feeling that this may be the case presented in the analysis by Tickle (2007).

The results derived by Tickle are based on optimization of r.m.s.d.s of stereochemical parameters relative to their standard target values through maximization of the free log-likelihood (LL_{free}; Lunin & Skovoroda, 1995) in the refinement of a few protein models. These results show that the r.m.s.d.(bonds) should be as small as 0.01 Å or less, whereas we suggested a target value of about 0.02 Å (Jaskolski *et al.*, 2007). However, demanding that model stereochemistry should so precisely reproduce the library standards would require that those standards be absolutely correct and that the variability of geometrical parameters in various parts of protein structures be minimal. It seems that this point was not taken into account by Tickle. The almost universally utilized Engh & Huber (1991, 2001) library, also used by Tickle, was based on data from the crystal structures of amino acids and small peptides. The uncertainty in most types of bond lengths summarized by Engh and Huber is higher than 0.02 Å. There is no reason to expect that their variability should be smaller in larger proteins. It seems to be illogical to demand that the stereochemistry of protein structures should reproduce the library values with higher precision than the accuracy of these values themselves. Moreover, as pointed out by Stec (2007), there is 'emerging evidence that ... protein stereochemistry is context-dependent', so that some geometrical parameters may have more than one preferred value depending, for example, on the secondary structure, in analogy to the rotamers of side chains. In such a situation, a single target, as used in the refinement programs, will not agree with any of the truly preferred values. This again suggests that the geometrical parameters of protein models should not be too tightly restrained to some predefined values.

While we are on the subject of numerology, we would like to raise some additional points. Another well known example of the tendency to blindly rely on numerical calculations, regardless of reality, is the estimation of unit-cell parameters by the program *HKL*-2000 (Otwinowski & Minor, 1997). The values for unit-cell dimensions that are found in the files produced by this program are in the form 123.456 Å, suggesting that the precision of the measurements is 0.001 Å. Any experimenter realises that such precision is absolutely unrealistic and that the estimated unit-cell parameters of macromolecular crystals are much less accurate. Such numerical results come from the refinement of various parameters during data merging and only reproduce the intrinsic precision of this numerical process. Unfortunately, such results 'officially' printed out by the program are usually accepted as 'true' values and proliferate throughout the whole structure-solution, refinement and deposition process. In reality, the estimation of unit-cell dimensions also depends on the crystal-to-detector distance and X-ray wavelength, which normally cannot be determined with a meaningful accuracy of six digits.

Another related example of meaningless precision is provided by the addition of trailing zeros to a variety of parameters of the protein structures deposited in the PDB. Thus, resolution limits of 1.800–45.000 Å, a redundancy of 11.000 and an $R_{\rm merge}$ of 0.09700 (this particular example was taken from the *remediated* file 1rb1, but similar numbers are found in most if not all other deposits) seem to clash with common sense. It must be stressed that these meaningless zeros are added by the deposition software and not by the providers of the coordinates.

The above examples seem to fall into the category of very elaborate numerology (Dauter & Baker, 2007). The tendency to believe more in very sophisticated numerical calculations rather than common sense based on experience is not restricted to humans. Such individuals may be compared to Rabbit, a friend of Winnie-the-Pooh, as evidenced by the following conversation (Milne, 1928):

> 'Rabbit's clever,' said Pooh thoughtfully.
> 'Yes,' said Piglet, 'Rabbit has Brain.'
> 'I suppose,' said Pooh, 'that that's why he never understands anything.'

References

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.
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.
Dauter, Z. & Baker, E. N. (2007). *Acta Cryst.* D**63**, 275.
Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.
Engh, R. A. & Huber, R. (2001). *International Tables for Crystallography*, Vol. *F*, edited by M. G. Rossmann & E. Arnold, ch. 18.3. Dordrecht: Kluwer Academic Publishers.
Jaskolski, M., Gilski, M., Dauter, Z. & Wlodawer, A. (2007). *Acta Cryst.* D**63**, 611–620.
Lunin, V. Yu. & Skovoroda, T. P. (1995). *Acta Cryst.* A**51**, 880–887.
Milne, A. A. (1928). *The House at Pooh Corner*, ch. 8. London: Methuen.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Stec, B. (2007). *Acta Cryst.* D**63**, 1113–1114.
Tickle, I. J. (2007). *Acta Cryst.* D**63**, 1274–1281.
Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998). *Acta Cryst.* D**54**, 243–252.