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.