Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

On the propagation of errors

Mariusz Jaskolski

Center for Biocrystallographic Research, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland, and Department of Crystallography, Faculty of Chemistry, A. Mickiewicz University, Poznan, Poland

Correspondence e-mail: mariuszj@amu.edu.pl

Received 9 May 2013
Accepted 2 June 2013

The policy of the Protein Data Bank (PDB) that the first deposition of a small-molecule ligand, even with erroneous atom numbering, sets a precedent over accepted nomenclature rules is disputed. Recommendations regarding ligand molecules in the PDB are suggested.

Small-molecule ligands in macromolecular structures are often of key importance, even if sometimes somewhat neglected, for the understanding of the functioning of their complexes. It is therefore reassuring that a recent effort has aimed at the validation of ligand structures (Weichenberger *et al.*, 2013; Pozharski *et al.*, 2013), although on the other hand it is disappointing to see that misinterpretation or overinterpretation of ligand structure and binding is not an infrequent phenomenon. Here, I would like to sensitize the community of 'structure consumers' and depositors, and foremostly the Protein Data Bank (PDB; Berman *et al.*, 2000) itself, to a much more trivial (and easy to fix), but in fact potentially confusing, problem encountered with small-molecule ligands, which is the prevalence of incorrect atom numbering or of an incorrect chemical form in the PDB. We first encountered this problem when depositing a *trans*-zeatin complex with cytokinin-specific binding protein (CSBP; Pasternak *et al.*, 2006; PDB entry 1flh). To any chemist familiar with atom-numbering rules (IUPAC, 2004), *trans*-zeatin, which is an adenine derivative, should be numbered as in Fig. 1(*a*). However, the PDB required us to renumber the atoms as in Fig. 1(*b*), arguing that there had been a previous deposition using that (incorrect) scheme. Thus, we were required to adopt an incorrect chemical numbering, and every time we now discuss *trans*-zeatin or similar cytokinin phytohormones we have to alert readers in a footnote to avoid confusion, especially that the two key imidazole N atoms, N7 and N9, have swapped numbers. As another example, in a structure of a St John's wort protein in complex with ANS (8-anilinonaphthalene-1-sulfonate; Sliwiak *et al.*, in preparation) we had to use an incorrect numbering (Fig. 1*d*) of the ANS molecule (identified in the PDB as 2AN, with ANS being reserved for dansyl acid) because of a similar precedent. Here, the naphthalene numbering is even inconsistent with the officially used chemical name. In this case, the confusion could partly be owing to fluid IUPAC recommendations (Moss, 1998). The current rules require letters to mark the fusion atoms, as in Fig. 1(*c*), a system that is indeed used by the PDB for dansyl acid (in the old IUPAC system these naphthalene atoms were numbered 9 and 10). With the various validation campaigns and remediation of the PDB archives in progress, it is difficult to buy the argument that incorrect (or illogical) atom numbering of a ligand molecule should be perpetuated forever just because there had been an error (or negligence) before. The PDB should certainly make an effort to put the ligand atom-numbering schemes on a par with the IUPAC standards and the accepted customary rules.

Even with complex organic molecules composed of several functional groups (such as in the two examples above) it should be possible to identify the principal moiety (adenine, naphthalene) and assign it with the primary numbering scheme. Atoms in the remaining residues (substituents) could be numbered consecutively or using

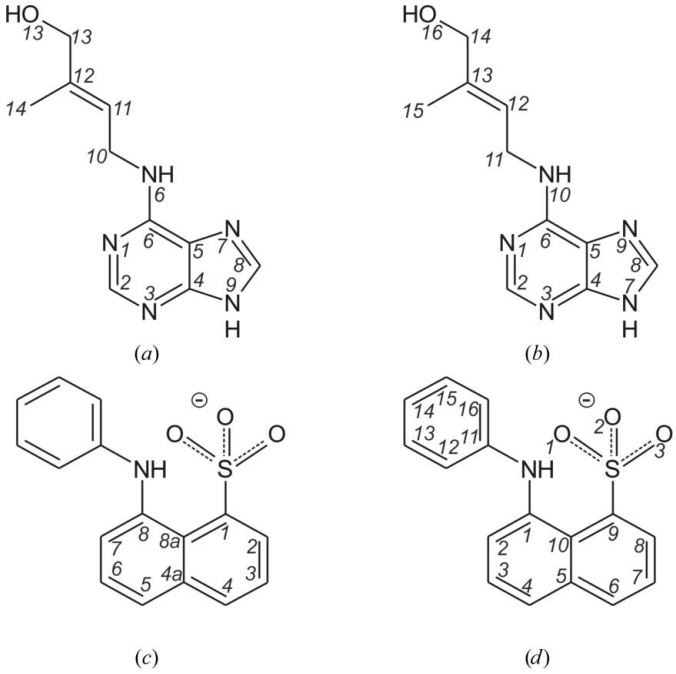

Figure 1
Different variants of atom numbering. *trans*-Zeatin numbered according to the adenine system (*a*) and according to a PDB template (*b*). In (*a*) the main moiety (6-aminopurine) is numbered according to the adenine convention, while the atoms of the secondary hydroxyisopentenyl substituent are numbered consecutively by analogy to the atom-labelling schemes of small molecules deposited in the CSD. Numbers may be repeated if they refer to different atom types (C13, O13). 8-Anilinonaphthalene-1-sulfonate (ANS; 2AN in the PDB) numbered according to IUPAC recommendations (*c*) and according to the PDB (*d*). The difference is in the numbering of the naphthalene skeleton, which is the main moiety.

non-numerical characters (as is practiced, for example, with nucleosides). Such problems are common in the small-molecule community and it could be beneficial if the PDB and CSD (Cambridge Structural Database; Allen, 2002) worked out suitable guidelines together.

As a related issue, the tautomeric or protonation state of ligand molecules is often not correct, or at least not obvious. An example of this is *trans*-zeatin, where the purine N—H atom can be found at N7 or N9 (in purine numbering; Fig. 1*a*) and the site of additional protonation (at low pH) is not obvious at all. As another example, in the 2AN ligand entry of the PDB the acid and anionic forms are happily confused. A more serious problem of this sort was encountered when we screened the PDB Z-DNA structures (Drozdzal *et al.*, 2013) in complex with the sperminium tetracation [Spk; spermine^{4+}; $H_3^+N(CH_2)_3NH_2^+(CH_2)_4NH_2^+(CH_2)_3NH_3^+$]. The spermine molecule (Spm) exists in the fully protonated form (Spk) at a wide range of pH values. Despite this, the sperminium tetracation has only been specified as a ligand in four structures deposited in the PDB (PDB entries 1se6, 1y0q, 1mg9 and 1kgk; Zhao *et al.*, 2006; Golden *et al.*, 2005; Zeth *et al.*, 2002; Wilds *et al.*, 2002), in contrast to 86 PDB entries (including the ultrahigh-resolution Z-DNA structure 3p4j at 0.55 Å; Brzezinski *et al.*, 2011) that indicate an unrealistic neutral form of the ligand (Spm). It should be noted that the bond lengths and angles of Spk and Spm differ appreciably and the distortions are particularly important at high resolution.

A separate and a more serious problem regards the generation of restraint dictionaries for ligand molecules. While the stereochemical libraries for macromolecules, especially for proteins (Engh & Huber, 1991, 2001), are pretty standard, exhaustively tested and generally free of errors (Jaskolski *et al.*, 2007), the numerous ligand libraries that are floating around are not. In our own experience, we have seen incorrect or unlikely ligand geometries, including a case of *trans*-zeatin with an erroneous assignment of atom hybridization by *eLBOW* (Moriarty *et al.*, 2009). If a ligand molecule, which is often of key interest in a macromolecular complex, is refined against a set of flawed stereochemical targets, the results could be quite lamentable. This problem is only touched on here, as its proper analysis and hopefully solution requires a very thorough study.

In summary, several lines of action can be recommended. (i) In ligand molecules, the principal moiety should be numbered by the PDB according to accepted conventions; (ii) secondary moieties (substituents) could be numbered using rules worked out jointly by the PDB and the CSD; (iii) ligands where protonation state or/and tautomeric form is an issue should be checked for the proper formula in conjunction with the reported chemical conditions, such as the pH of crystallization; and (iv) there is an urgent need for compilation of reliable standard restraint libraries for ligand molecules found in the PDB.

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.
Brzezinski, K., Brzuszkiewicz, A., Dauter, M., Kubicki, M., Jaskolski, M. & Dauter, Z. (2011). *Nucleic Acids Res.* **39**, 6238–6248.
Drozdzal, P., Gilski, M., Kierzek, R., Lomozik, L. & Jaskolski, M. (2013). *Acta Cryst.* D**69**, 1180–1190.
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, pp. 382–392. Dordrecht: Kluwer Academic Publishers.
Golden, B. L., Kim, H. & Chase, E. (2005). *Nature Struct. Mol. Biol.* **12**, 82–89.
Jaskolski, M., Gilski, M., Dauter, Z. & Wlodawer, A. (2007). *Acta Cryst.* D**63**, 611–620.
Moriarty, N. W., Grosse-Kunstleve, R. W. & Adams, P. D. (2009). *Acta Cryst.* D**65**, 1074–1080.
Moss, G. P. (1998). *Pure Appl. Chem.* **70**, 143–216.
IUPAC (2004). *Nomenclature of Organic Chemistry*, p. 1250. Research Triangle Park: IUPAC.
Pasternak, O., Bujacz, G. D., Fujimoto, Y., Hashimoto, Y., Jelen, F., Otlewski, J., Sikorski, M. M. & Jaskolski, M. (2006). *Plant Cell*, **18**, 2622–2634.
Pozharski, E., Weichenberger, C. X. & Rupp, B. (2013). *Acta Cryst.* D**69**, 150–167.
Weichenberger, C. X., Pozharski, E. & Rupp, B. (2013). *Acta Cryst.* F**69**, 195–200.
Wilds, C. J., Maier, M. A., Tereshko, V., Manoharan, M. & Egli, M. (2002). *Angew. Chem. Int. Ed. Engl.* **41**, 115–117.
Zeth, K., Ravelli, R. B., Paal, K., Cusack, S., Bukau, B. & Dougan, D. A. (2002). *Nature Struct. Biol.* **9**, 906–911.
Zhao, B. *et al.* (2005). *J. Biol. Chem.* **280**, 11599–11607.