BIOLOGICAL CRYSTALLOGRAPHY

ISSN 1399-0047

The rate of *cis–trans* conformation errors is increasing in low-resolution crystal structures

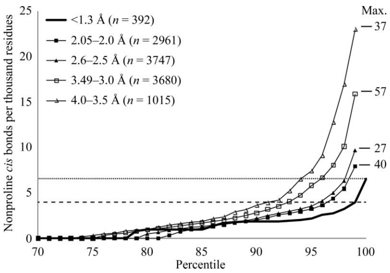
Tristan Ian Croll*

Institute of Health and Biomedical Innovation, Queensland University of Technology, 60 Musk Avenue, Kelvin Grove, QLD 4059, Australia. *Correspondence e-mail: tristan.croll@qut.edu.au

Received 4 October 2014
Accepted 15 January 2015

Keywords: *cis*-peptide bonds.

Cis-peptide bonds (with the exception of *X*-Pro) are exceedingly rare in native protein structures, yet a check for these is not currently included in the standard workflow for some common crystallography packages nor in the automated quality checks that are applied during submission to the Protein Data Bank. This appears to be leading to a growing rate of inclusion of spurious *cis*-peptide bonds in low-resolution structures both in absolute terms and as a fraction of solved residues. Most concerningly, it is possible for structures to contain very large numbers (>1%) of spurious *cis*-peptide bonds while still achieving excellent quality reports from *MolProbity*, leading to concerns that ignoring such errors is allowing software to overfit maps without producing telltale errors in, for example, the Ramachandran plot.



1. Introduction

It has been understood for decades that native proteins overwhelmingly favour the *trans* conformation for nonproline peptide bonds, with <0.05% found in the *cis* conformation (Stewart *et al.*, 1990). Those *cis*-peptide bonds that do occur are typically found in tightly restrained environments and play key structural and/or functional roles, and as such tend to be highly conserved (Craveur *et al.*, 2013). As such, a nonproline *cis*-peptide bond (nonPcis-pep) in a newly solved structure is a key feature of interest and should be carefully justified and remarked upon.

While investigating a recently published 3.05 Å resolution crystal structure (PDB entry 3puk) arising from an experienced crystallographic group (Hu *et al.*, 2011), I was surprised to note that 22 of 1204 residues were found with nonPcis-peps and that these went unremarked in the text. On closer inspection, I found that all could be returned to the *trans* conformation with an overall slight improvement in $R_{\rm work}$, $R_{\rm free}$ and Ramachandran statistics. Communication with the senior author and personal inspection revealed that while *Coot* has a menu option to manually check for *cis*-peptide bonds, neither *PHENIX* (Adams *et al.*, 2010), *MolProbity* (Chen *et al.*, 2010) nor the automated Protein Data Bank (PDB) validation tools (Read *et al.*, 2011) provide a report on *cis* bonds in their standard workflow, leading to these errors going unnoticed.

In order to evaluate the extent of this problem, I have performed an exhaustive check of all structures within key resolution bins to determine the rate of nonPcis-peps in solved structures as a function of time.

doi:10.1107/S1399004715000826

2. Methods

Reports (accession ID, deposition date, resolution, residue count, R_{work} and R_{free} values and the refinement software used) for all current structures with accession dates after 1985 with resolutions of ≤1.3, 2.0–2.05, 2.5–2.6, 3.0–3.49 and 3.5–4.0 Å were recovered from the PDB website. Each structure was downloaded in turn, and the number of protein C^{α} atoms and nonPcis-peps were counted using a simple script implemented in *VMD* (Humphrey *et al.*, 1996). Structures with fewer than 500 protein residues were excluded from per-structure analyses, while all structures with at least 100 protein residues were included in the calculation of aggregate rates of nonPcis-pep incorporation per year. Structural inspection and rearrangement was carried out using a haptic-guided interactive molecular-dynamics flexible fitting code developed in-house.

3. Results

The rate of nonPcis-pep incorporation into ≥500-residue structures in the chosen resolution bins is summarized in Fig. 1. In structures with resolution ≤1.3 Å, 99% of structures contain less than four nonPcis-peps per thousand residues. In comparison, 3.7% of 2.05–2.0 Å resolution structures exceed this rate, rising to 9.2% of 4.0–3.5 Å resolution structures. Extrapolation to the complete PDB set suggests that at least 2000 current structures have highly questionable rates of nonPcis-pep incorporation. The true rate of erroneous nonPcis-peps may, however, be substantially higher. Even in the ≤1.3 Å resolution set, three of the ten structures with the highest absolute counts of nonPcis-peps, PDB entries 3ncq (Helfmann *et al.*, 2010), 2gec (Jayaram *et al.*, 2006) and 2j82 (Schlicker *et al.*, 2008), show substantial evidence of error in these assignments, with inconsistent assignment between identical chains and/or poor fit to the local density.

While the average rate of nonPcis-pep incorporation over all high-resolution structures is 0.486 per 1000 residues, in keeping with prior analyses (Stewart *et al.*, 1990), since 2006 the rate of inclusion of nonPcis-peps has been steadily increasing in published ≥2.5 Å resolution structures (Fig. 2*a*). The increase appears to be dominated by a relatively small but growing population of structures with extremely high rates of such errors (Fig. 2*b*). No obvious correlation could be detected between high nonPcis-pep counts and the choice of refinement software.

The recent structure with PDB code 4q8j (Schäfer *et al.*, 2014; 3.8 Å resolution) is a useful case study to illustrate the problem. This structure arises from a highly prolific and experienced group (with the senior author having solved 102 structures since 1993) and was published in a top-ranking structural biology journal. According to the RCSB structural validation report, it ranks in the top quartile of structures in its resolution class for clashscore (13) and RSRZ outliers (0.4%), and in the top decile for Ramachandran (0.4%) and side-chain outliers (2.5%). Its *MolProbity* score of 2.28 puts it in the 99th percentile for 4.05–3.25 Å resolution structures. Yet of its 6206 protein residues, 86 (1.4%) appear with nonPcis-peps. Like the three chosen for closer inspection (Figs. 2*c* and 2*d*), the majority of these are found in environments atypical of nonPcis-peps and clearly appear to be erroneous. It therefore appears that a check for nonPcis-peps was not included at any stage of the refinement, peer-review or deposition process.

4. Discussion

The incorporation of an erroneous *cis*-peptide bond is, in many cases, a relatively minor structural problem. However, the unchecked incorporation of erroneous *cis*-peptide bonds provides extra, unnatural degrees of freedom to the crystallographer and the fitting software. For example, in the common circumstance of a helix or β-strand followed by a relatively weakly defined loop, it is not unusual for the initial assignment of register to be erroneous. In the presence of such an error, the inadvertent incorporation of one or more nonPcis-peps may allow the crystallographer to nevertheless achieve an apparently good fit to the loop density without the telltale bond, angle and/or Ramachandran outliers which typically accompany such errors. This may in fact be the case for the example pictured in Figs. 2(*c*) and 2(*d*), in which a 13-residue α-helix immediately N-terminal to the problem area appears to be out of register.

While it is of course impossible to determine the source of the errors in all of the identified cases, in the case of 3puk as mentioned in §1 it appears these were introduced during manual refinement in *Coot*, as poorly fitting loops were repeatedly stretched and relaxed back into the map (Martin, 2014). Investigation of 3puk and other homomultimeric structures suggests that inadvertent and unnoticed rather than deliberate incorporation is the norm, since it is common to see inconsistent nonPcis-pep incorporation between monomers.

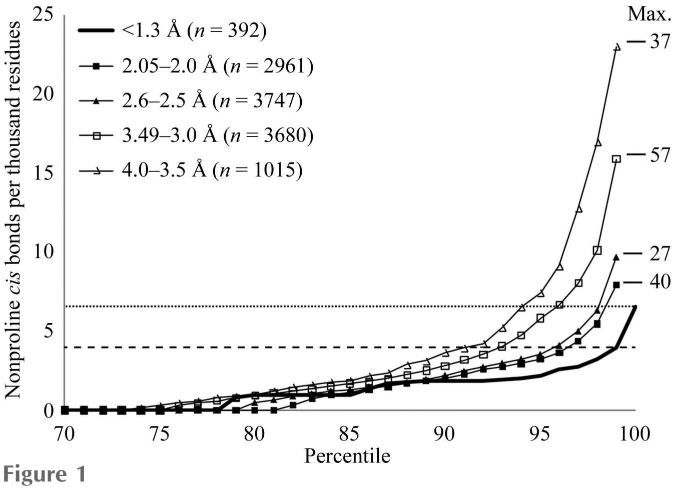


Figure 1
NonPcis-pep statistics for structures with ≥500 protein residues in selected resolution bins. The 99th and 100th percentile values for the <1.3 Å resolution set are denoted as dashed and dotted lines, respectively. The maximum value for each >1.3 Å resolution set is given on the right. Approximately 4% of 2.6–2.0 Å resolution structures, 7.3% of 3.49–3.0 Å resolution structures and 9.2% of 4.0–3.5 Å resolution structures have rates of nonPcis-pep incorporation above the <1.3 Å resolution 99th percentile rate.

3puk, a homodimer, contains 14 nonPcis-peps in chain *A* and eight in chain *B*, and the two chains have only one nonPcis-pep in common. Similarly, in the homotrimeric 3t6v (Ferraroni *et al.*, 2012) residues 160–164 have a different complement of nonPcis-peps in each chain. The homotetrameric chains *A*–*D* in 4kvm (Liszczak *et al.*, 2013) are inconsistent in their inclusion of nonPcis-peps at residues 59–60 and 357–358. I note that the most recent versions of *Coot* provide a warning when such manipulations introduce a *cis*–*trans* flip, but this is not highly prominent and may in some cases go unnoticed by the user. A preferable approach may be to modify the manual refinement tools to prevent such accidental flips, so that the only way that a *cis* bond may be incorporated is by the explicit choice of the user.

As the average size of protein structures solved by crystallography increases, it becomes ever more impractical to manually inspect structures for errors, and hence crystallographers are heavily reliant on the help of automated analysis software. Current packages such as *MolProbity* are indeed extraordinarily thorough in their reports on bond length, angle, Ramachandran and atomic distance outliers. I suspect that this very thoroughness is leading many practitioners to rely solely on these reports for refinement, allowing the undetected incorporation of substantial errors. To avoid this, it is imperative that the identification of *cis*-peptide bonds be included in the standard protein crystallography workflow. I suggest that this should be handled in a similar fashion to the current treatment of Ramachandran outliers. Like these, nonPcis-peps occur naturally at a very low frequency, and hence each should be carefully checked and justified by fit to the density, structural reasonability and/or reference to higher resolution homologues. I have communicated the existence of this problem to the developers of *MolProbity* and they have confirmed that a check for *cis*-peptide bonds will be included

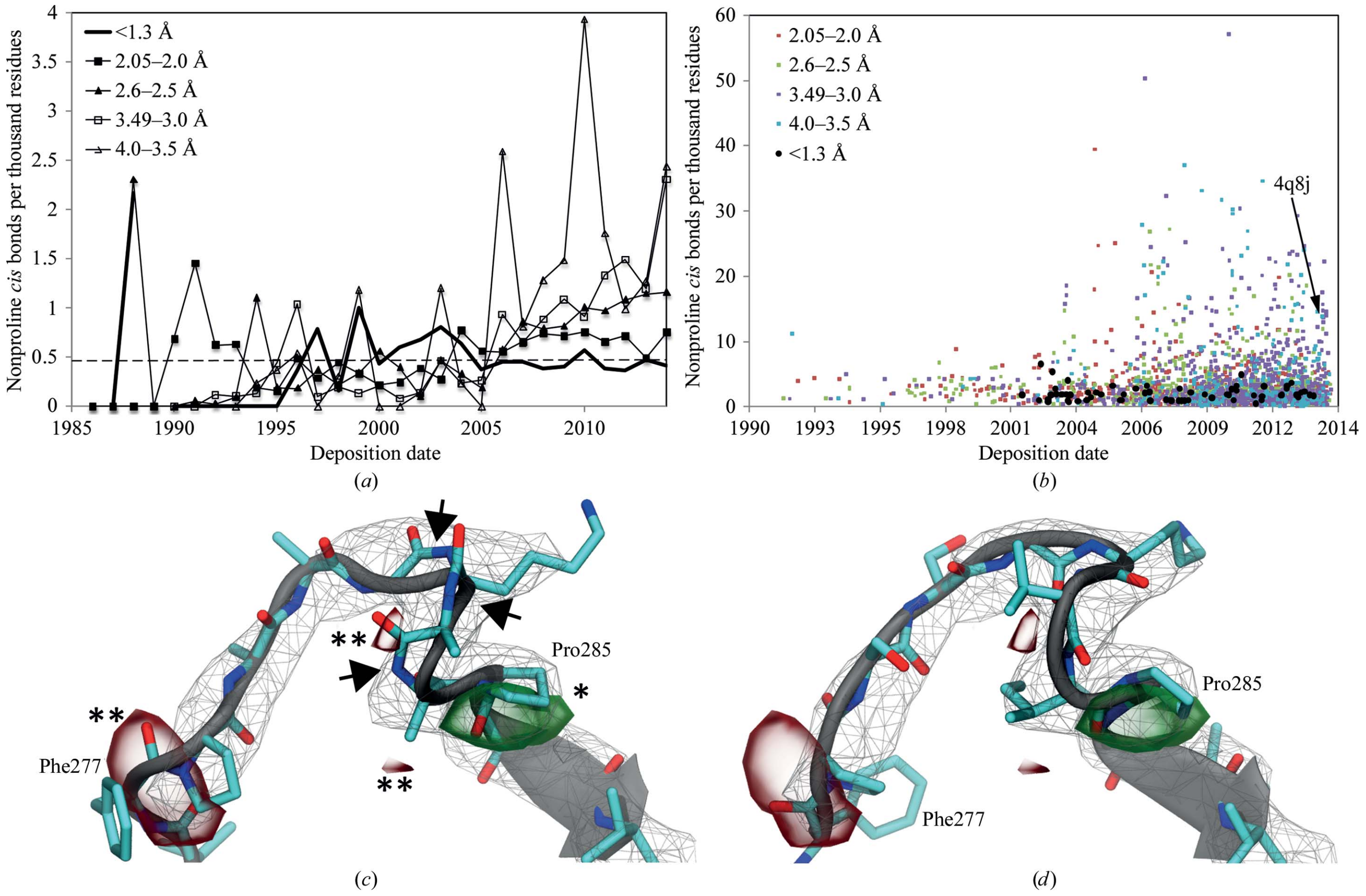


Figure 2

(*a*) Since 2006, the rate of appearance of nonproline *cis*-peptide bonds in low-resolution structures has increased steadily to 3–4 times the rate seen in very high resolution structures. Each data point was calculated by adding up all nonproline *cis* bonds appearing in >100-residue structures in a given resolution bin and year and dividing by the total number of solved protein residues in the same set. The average across all structures of resolution ≤1.3 Å is shown as a dashed line. (*b*) The excess *cis* bonds come from a relatively small population of highly aberrant structures. Note that only structures with ≥500 protein residues were considered for this panel. (*c*, *d*) Unnoticed nonPcis-peps may 'hide' more serious structural errors. (*c*) Structure 4q8j [indicated by an arrow in (*b*)], despite being given strong quality metrics from *MolProbity* and the RCSB, contains 86 nonPcis-peps. Three of these (residues 281–283 of chain *E*, indicated by arrows) form an exposed, weakly constrained turn. While some signs of error appear in the $\pm 2\sigma$ $F_o - F_c$ map (* and **, respectively), there are no Ramachandran outliers. The $2F_o - F_c$ map generated with *B*-factor sharpening of -60 Å^2 and contoured at 1σ is shown in wireframe. (*d*) Switching these bonds to *trans* resolves the density clashes and inspection points to a one-residue register error in the upstream 261–274 α-helix (not shown).

in a special low-resolution mode (LoRx) in the next official release of the package (Richardson & Richardson, 2014).

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Cryst.* D**66**, 12–21.

Craveur, P., Joseph, A., Poulain, P., de Brevern, A. & Rebehmed, J. (2013). *Amino Acids*, **45**, 279–289.

Ferraroni, M., Matera, I., Chernykh, A., Kolomytseva, M., Golovleva, L. A., Scozzafava, A. & Briganti, F. (2012). *J. Inorg. Biochem.* **111**, 203–209.

Helfmann, S., Lü, W., Litz, C. & Andrade, S. L. A. (2010). *J. Mol. Biol.* **402**, 165–177.

Hu, S.-H., Christie, M. P., Saez, N. J., Latham, C. F., Jarrott, R., Lua, L. H. L., Collins, B. M. & Martin, J. L. (2011). *Proc. Natl Acad. Sci. USA*, **108**, 1040–1045.

Humphrey, W., Dalke, A. & Schulten, K. (1996). *J. Mol. Graph.* **14**, 33–38.

Jayaram, H., Fan, H., Bowman, B. R., Ooi, A., Jayaram, J., Collisson, E. W., Lescar, J. & Prasad, B. V. V. (2006). *J. Virol.* **80**, 6612–6620.

Liszczak, G., Goldberg, J. M., Foyn, H., Petersson, E. J., Arnesen, T. & Marmorstein, R. (2013). *Nature Struct. Mol. Biol.* **20**, 1098–1105.

Martin, J. (2014). Personal communication.

Read, R. J. *et al.* (2011). *Structure*, **19**, 1395–1412.

Richardson, J. & Richardson, D. (2014). Personal communication.

Schäfer, I. B., Rode, M., Bonneau, F., Schüssler, S. & Conti, E. (2014). *Nature Struct. Mol. Biol.* **21**, 591–598.

Schlicker, C., Fokina, O., Kloft, N., Grüne, T., Becker, S., Sheldrick, G. M. & Forchhammer, K. (2008). *J. Mol. Biol.* **376**, 570–581.

Stewart, D. E., Sarkar, A. & Wampler, J. E. (1990). *J. Mol. Biol.* **214**, 253–260.