Storing diffraction data

SIR — *Nature* must consider structural biology to be of some interest to its readership, as almost every issue contains a new macromolecular structure. We therefore regret your publication of the letter from Hooft *et al.* on errors in protein structures[1]. The nonspecialist reader will immediately draw the conclusion that the crystallographic structures in the Protein Data Bank (PDB) are extremely unreliable. The quality of structures deposited at the PDB is indeed variable, but since the spate of high-profile errors that led to the Commentary by Brändén and Jones[2], the community has tried hard to develop new methods and protocols to make such errors less likely to occur or to escape attention.

We would like to make the following comments:

- *Nature*'s Correspondence section is not a suitable place to publish the results of a validation study of the PDB.
- Many listed items are simply not errors. From the wealth of structural knowledge in the PDB, we now know a lot about macromolecular preferences. For example, although proteins show clear preferences for certain main-chain conformations, an outlier in a Ramachandran plot is simply an outlier, it is not necessarily an error! In one enzyme family, the α/β hydrolases, the active-site nucleophile always has a so-called 'disallowed' conformation. Indeed, we would expect up to 5 per cent of all non-glycine residues to fall in 'disallowed' regions of this plot (G. J. K. & T. A. J., to be published).
- Stereochemical knowledge has been used in the refinement of most structures in the PDB, because the macromolecular crystallographer is usually limited by the resolution of the diffraction data. Improvements have been made in the dictionaries describing this information and in the refinement protocols. Using these newer dictionaries on older coordinate sets, refined with older protocols, is sure to produce many outliers. Even for models refined with the most modern dictionaries, stereochemical outliers can only indicate potential errors. On the other hand, it is perfectly possible to make a structure that has excellent stereochemistry and is totally wrong[3]. Good stereochemistry is not proof of a correct structure, although crystallographers usually provide these statistics as if it were.
- One cannot always use the coordinates without reading the remarks contained in the PDB entry or original publication. This is difficult for a computer program to do. For example, one structure is flagged as having no fewer than twelve D-amino acids that are, therefore, counted as errors. Even a cursory glance at the PDB file, however, reveals that this is the structure of gramicidin (an antibiotic made up of both L- and D-amino acids), which is one of the highest-resolution structures in the database. Other cases of D-amino acids may well be examples of poor stereochemistry. In another example, the entry 1GRH has been flagged as having a particularly large specific volume, V_m (954 Å^3 Da^{-1}, whereas one would expect something in the range 2 to 3), because the authors have deposited the coordinates of a single modified residue, as described in the file.
- Insufficient care has been taken in preparing the report. Some of the very close contacts that are flagged as errors arise because atoms sit on special positions in the unit cell. It is not uncommon for a water molecule or ion to sit on a crystallographic rotation axis so that the distance to one or more crystallographically related copies of itself will be zero. This is not the major reason for this class of error (the second most common class according to the authors), but we mention it to highlight the care that must be taken when preparing a report of alleged errors. Like crystallographers, Hooft *et al.* are human and their software has some shortcomings. For example, in the report on 2DRI, a ribose-binding protein from *Escherichia coli*

which has been refined to a resolution of 1.7 Å (ref. 4), most of the atoms itemized as lacking hydrogen bonds actually form hydrogen bonds to the bound ribose, while others involve residues that are external.

- Many book-keeping errors have been located. For example, some of the large V_m values are due to errors on the PDB CRYST1 card where the crystallographer has specified the number of noncrystallographic symmetry partners instead of the number of molecules in the cell. In other files, the temperature factors and occupancies may be in the wrong order. These are not errors in the structures but errors in the PDB files.
- Now that *Nature* requires the deposition of coordinates, the next step should be taken and the deposition of diffraction data should also be made mandatory, if necessary with an appropriate hold period. Many of the alleged errors listed by Hooft *et al.* could be properly evaluated if the experimental structure factor data were available. In many instances, these data may be lost to the community because of the rapidly changing world of computers where the optimal method of back-up is not constant. Deposition of diffraction data is probably the best way to improve the quality of the structural database and to ensure long-term archiving, but the process must be enforced. For example, a couple of months ago, two more-or-less identical structures were published in *Nature* and *Science*. The *Nature* coordinates have now been released, but the *Science* ones are still nowhere to be found (and we are still waiting for a response from the authors about their availability). Without the experimental data, one can suspect that there are problems in a particular structure but, except for trivial chirality errors that no-one would argue with, there is no proof. We have our own prejudices about how structures should be refined, and in particular we believe that low-resolution structures are being over-refined[3]. We have indicated some examples of this even at moderate resolution[5], but without the diffraction data there is little more that can be done.
- It would have been interesting to see how the 'errors' depend on the year of deposition. We think the quality of the deposited structures is improving, although the methodology being used is by no means uniform. The free R-value (R_{free}) has recently been introduced[6] as an aid in recognizing problems during crystallographic refinement. It can also help the crystallographer in designing the optimal refinement protocol. We have monitored its use in structures published between January and June 1996 (ref. 7). This shows a remarkable variation that depends on the journal in which the structure is published. In three journals (*Nature*, *Cell* and *Structure*), more than 80 per cent of the structures have R_{free} quoted. In *Nature Structural Biology* and in *Science*, about 65 per cent do so, while in *Acta Crystallographica D*, *Journal of Molecular Biology*, *Proteins*, *Protein Science* and *Biochemistry*, fewer than a third of the structures have a quoted R_{free}.

T. Alwyn Jones
Gerard J. Kleywegt
Department of Molecular Biology,
BMC Box 590,
S-751 24 Uppsala,Sweden
Axel T. Brünger
Howard Hughes Medical Institute
& Department of Molecular
Biophysics and Biochemistry,
Yale University,
New Haven,
Connecticut 06520-8114, USA

1. Hooft, R. W. W., Vriend, G., Sander, C. & Ebola, E. *Nature* **381**, 272 (1996).
2. Brändén, C. I. & Jones, T. A. *Nature* **343**, 687–689 (1990).
3. Kleywegt, G. J. & Jones, T. A. *Structure* **3**, 535–540 (1995).
4. Mowbray, S. L. & Cole, L. N. B. *J. Mol. Biol.* **225**, 155–175 (1992).
5. Kleywegt, G. J. *Acta Crystallogr. D* **52**, 842–857 (1996).
6. Brünger, A. T. *Nature* **355**, 472–475 (1992).
7. Kleywegt, G. J. & Brünger, A. T. *Structure* **4**, 897–904 (1996).

corres@nature.com

Letters submitted for Correspondence should be typed, double-spaced, on one side of the paper only, or e-mailed (with full mailing address) to corres@nature.com