Errors in protein structures

SIR — In 1990 Brandén and Jones[1] wrote in a Commentary to *Nature*: "Protein crystallography is an exacting trade, and the result may contain errors that are difficult to identify. It is the crystallographers' responsibility to make sure that incorrect protein structures do not reach the literature." Your recent leading article[2] about obligatory deposition of macromolecular coordinates and the underlying experimental data[3] revived this debate.

A FEW OF THE ERRORS IN THE LITERATURE...

Inconsistent symmetry information	19 files
Transformation matrix has determinant not equal to 1.0	5 cases
D amino acid	183 cases
Atom too close to symmetry axis leading to a clash	332 cases
Structure probably solved in wrong space group	24 files
Much too high Matthews' coefficient (V_m>7.0)	69 files
B-factors over-refined	533 files
Cell dimension off by more than 0.5%	1,914 files
Atomic occupancies negative or larger than 1.0	43,934 cases
Bond length deviates more than 4σ	61,051 cases
Bond angle deviates more than 4σ	309,186 cases
Atoms more than 0.4 Å too close to each other	265,290 cases
Side chain of His, Asn of Gln needs 180° flip	19,906 cases

The 1,159,804 outliers in Protein Data Bank data sets reflect discrepancies with conventions, statistical outliers and probable errors. Of the 76 classes of problems only 13 are listed in this table. The complete tables, full reports about every entry that we tested and detailed descriptions of all tests are available from http://www.sander.embl-heidelberg.de/pdbreport/

We have recently carried out a search for anomalies in 3,442 structures from the Brookhaven Protein Data Bank using our program WHAT_CHECK (which can be retrieved by anonymous FTP from swift.embl-heidelberg.de or from pdb.pdb.bnl.gov). We have analysed the deposited coordinates without using the underlying experimental data. We assume that the standard deviations in parameters such as bond lengths and bond angles that are observed in small molecules can be used as a standard of truth[4].

Our analyses reveal a much larger number of 4σ deviations than one would expect. Some of these outliers will reveal a novel feature (for example, a property typical for transmembrane proteins, of which there are now only a few in the database). Another fraction represents cases where the experimental data simply cannot be described by 'normal' coordinates (for example, two alternative conformations for a side chain, but insufficient data to resolve them). However, we also detected many straightforward errors. These range from wrong atom names or bond angles that are more than 10° off, via threonines with wrong C-β chirality and tryptophans with a 90° angle between the two rings, to major errors such as proteins that are misthreaded or solved in the wrong spacegroup (see table).

Many of these problems are not crucial to the casual browser. But for purposes of detailed biological studies or large surveys, it is important that these irregularities do not go undetected. The Protein Data Bank is aware of the importance of structure verification: it is increasingly making validation tools part of its production process, exchanging experience and software with European colleagues working in this field. (The BIOTECH structure validation server provides structure verification tools via the World Wide Web. WHAT_CHECK is part of this server. WWW addresses are: http://biotech.embl-heidelberg.de:8400/, http://biotech.embl-ebi.ac.uk:8400/ and http://biotech.pdb.bnl.gov:8400/.)

The current verification software can already solve some of the detected problems automatically. Others require manual electron density inspection. We hope that structure factor deposition will soon be as much an accepted practice as coordinate deposition, so that in the future more problems can be detected and more of the million or so found so far can be solved.

Rob W. W. Hooft
Gert Vriend
EMBL,
Meyerhofstrasse 1,
D-69117 Heidelberg, Germany
Chris Sander
EMBL-EBI,
Hinxton Hall, Hinxton,
Cambridge CB10 1RQ, UK
Enrique E. Abola
Protein Data Bank,
Brookhaven National Laboratory,
Biology Department, Building 463,
Upton, New York 11973, USA

[1] Brandén, C.-I. & Jones, T. A. *Nature* **343**, 687–689 (1990).
[2] *Nature* **379**, 191 (1996).
[3] Baker, E. N. *et al. Nature* **379**, 202 (1996).
[4] Engh, R. & Huber, R. *Acta crystallogr.* **A47**, 392–400 (1991).

Paradoxical error

SIR — Last year I published in News and Views, under the title "Multiple Kauzmann paradoxes" (*Nature* **373**, 475-476; 1995), an account of a paper, somewhat controversial as it proved subsequently, by two chemists, K. Kishore and H. K. Shobha (*J. chem. Phys.* **101**, 7037–7047; 1994). The paper discussed the extension of the well-known Kauzmann paradox, referring to an extrapolated temperature domain in which the entropy of a supercooled liquid is below that of the corresponding crystalline structure, to the relationship between a supercooled vapour and the corresponding liquid phase. Kishore and Shobha claimed to show that there is an upper limiting temperature above which a superheated liquid would indeed have a higher entropy than its vapour ('entropy crossing').

P. G. Debenedetti, M. M. Atakan and R. J. Speedy have now published (*J. chem. Phys.* **104**, 5349–5350; 1996) a rebuttal of Kishore and Shobha's treatment in terms of spinodal theory in the vicinity of the critical state, taking into account pressure, which Kishore and Shobha failed to do. They demonstrate that "the entropy of a superheated liquid and a vapor at the same temperature and pressure are never equal except at the critical point, where both phases are identical", and show how Kishore and Shobha arrived at what now appears to be their erroneous conclusion.

Robert W. Cahn
Department of Materials
Science and Metallurgy,
University of Cambridge,
Cambridge CB2 3QZ, UK

The last word

SIR — Webster and Erickson (*Nature* **380**, 386; 1996) call for a word to "designate the last person, animal, or other species in his/her/its lineage". They give several possibilities, preferring 'endling'.

In supporting instead 'ender', my colleague Ralph Elliott, a lexicographer in this university, points out that the *Oxford English Dictionary* attributes the word first to Chaucer and gives as one of its meanings 'He or that which puts an end or termination to anything'. The word seems not to be in use for any other purpose.

David Craig
Research School of Chemistry,
Australian National University,
Canberra 0200, Australia

SIR — 'Endling' has a somewhat pathetic feel to it, similar to 'foundling'. The suffix -*arch* means 'leader' and is used in the words matriarch and patriarch, meaning one who rules a family, clan or tribe. Terminal means 'forming an end or boundary'.

I suggest 'terminarch' to designate the last of lineage. It has a much stronger and more positive ring.

Elaine Andrews
RPR Gencell,
5301 Patrick Henry Drive,
Santa Clara, California 95054-1114, USA

SIR — The last remaining is a relict — a word with a decent Latin root, still used and understood.

There is no need invent a new one.

Mark Smith
Ballacurn, Ballaugh,
Isle of Man IM7 5EU, UK