Daresbury Laboratory

JOINT CCP4 AND ESF-EACBM NEWSLETTER ON PROTEIN CRYSTALLOGRAPHY

An informal Newsletter associated with the SERC Collaborative Computational Project No.4 on Protein Crystallography and the ESF Network of the European Association of the Crystallography of Biological Macromolecules.

Number 27 November 1992

Contents

CCP4 Suite News and Administrivia from Daresbury
D. Love ... 1

MIR with one derivative
M. Noble & R. Pauptit ... 3

Haloalkane dehalogenase
K.H.G. Verschueren, H.J. Rozeboom, K.H. Kalk & B.W. Dijkstra ... 7

Crystal Structure of Avidin-Biotin Complex at 2.7Åresolution
L. Pugliese & M. Bolognesi ... 11

SRS Specialist User Group for Laue Diffraction
J. Campbell & B. Kariuki ... 14

Notes on FEBS/EACBM Advanced Course "Techniques in Protein Crystallography ... 17

Status Report from ESRF
C.-I. Brändén ... 22

The Crystallography Workstation for X-Ray Structural Investigations of Single Crystals at the TNK Storage Ring
A.N. Popov, D.M. Kheiker & E.H. Harutyunyan ... 25

Availability of the MOSFLM program suite for processing image plate and film data
A.G.W. Leslie ... 30

On the Reliability of Atomic Models in the PDB: Stereochemical Point of View
A.G. Urzhumtsev ... 31

Editors: Wojciech M. Wolf

Keith S. Wilson

SERC Daresbury Laboratory
Daresbury, Warrington WA4 4AD UK

EMBL c/o DESY
Notkestrasse 85, D-2000 Hamburg 52
Federal Republic of Germany

ON THE RELIABILITY OF ATOMIC MODELS IN THE PDB : STEREOCHEMICAL POINT OF VIEW

A.G. Urzhumtsev

Institute of Mathematical Problems of Biology, Russian Acad.Sci., Pushchino, Moscow region, 142292, Russia
(current address: 15, rue Descartes, UPR de Biologie Structurale, IBMC, 67000 Strasbourg, France)

Abstract.

The reliability of a spatial structure is determined by the reliability of all its components. But usually only average characteristics are reported in the literature and they may "hide" large distortions of stereochemical parameters. An analysis of some files of the Protein Data Bank has confirmed this hypothesis.

The atomic model of protein is a complex system whose reliability depends on the reliability of all its components. In particular, the presence of only one strongly distorted bond length in the main chain is enough to raise doubts in reliability of the whole structure because usually local improvement of a model is hardly possible. This, in turn, makes doubtable the speculations concerning the mechanism of protein activity.

Unfortunately, the majority of published papers on protein structure determination only contain average information on stereochemical distortion, which is absolutely insufficient. For example, for a protein of 100 amino acid residues, the average error in bond lengths of about 0.01Å may be a result of errors in two polar situations:

a) all bond lengths contain erros of about 0.01Å;

b) all bond lengths are absolutely correct, except the one which has an error of about 3Å.

It is clear that the former type of models is quite good, which cannot be said of the latter one.

We have looked through a number of Protein Data Bank files (Bernstein et al., 1977) to analyse maximal distortions in stereochemical values. Of course, the quality of a model depends on many factors, of which the most important are the resolution, the refinement program used, the protein size, the date of structure determination, etc. Note that the proper geometry may be shown both by a well-refined structure and by a non-refined structure which is in bad agreement with experimental X-ray

data. Because of a very complex interrelation of these factors the models were analysed without classifying them into groups, we only separated refined structures from the unrefined ones.

The first 99 files were picked up from the available PDB version. Upon removal files with RNA structures, with incomplete models, with unrefined and partially refined models, etc, the errors $<\Delta d>$ of bond lengths were analysed for the rest of 50 files. Other stereochemical values were not analyses so thoroughly althouh it was noted, for example, that errors in bond angles were distributed roughly as in bond lengths.

Other 4 models were excluded intentionally from the analysus: two of them had very large average errors, $<\Delta d>$, of about 0.16Å, and for the other two the large maximal errors, Δd_{max}, exceeded 1Å.

An analysis has confirmed that the average errors $<\Delta d>$ are small enough both in refined and in majority of unrefined models (Fig.1). However, 18 refined models, almost 40% of the total, have main chain errors Δd_{max} larger than 0.1Å and 6 of them have errors larger than 0.2Å (Fig.2). When side chain errors are taken into account, the number of such models increased up to 30 and 9, respectively. Of course, it should be noted that some models including several unrefined ones have perfect characteristics, both average (Fig.1) and maximal (Figs.2,3).

The analysis which has been carried out shows that special efforts should be undertaken during refinemet to avoid large stereochemical distortions. For example, the refinement complex FROG (Urzhumtsev et al., 1989) provides a user with a histogram of distribution of different stereochemical characteristics but not with only average values.

Thus, the selective analysis of PDB files has revealed some PDB models with unrealistic stereochemical parameters. That is why one should be careful when dealing with macromolecular models. As a preventive step it will be very useful to ask the authors to publish not only the average but also the maximal distortions in macromolecular atomic models described.

Addition in proof : when the paper has been prepared (September 1990) for publication the author was informed that the similar work has been done by Dr. M.Basharov, Institute of Biological Physics, Pushchino.

References :

Bernstein, F.C., et al. (1977) *J.Molec.Biol.*, **112**, 535-542

Urzhumtsev, A.G., Lunin, V.Yu. & Vernoslova, E.A. (1989) *J.Appl.Cryst.*, **22**, 500-506

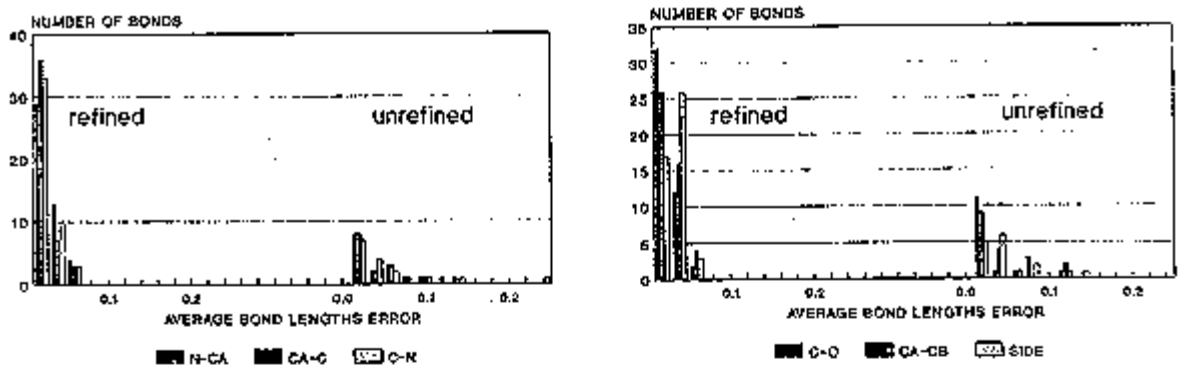

Fig. 1. The histogram of distribution of average bond-length errors, $<\Delta d>$, in refined and unrefined models : a) main chain bonds, b) other bonds.

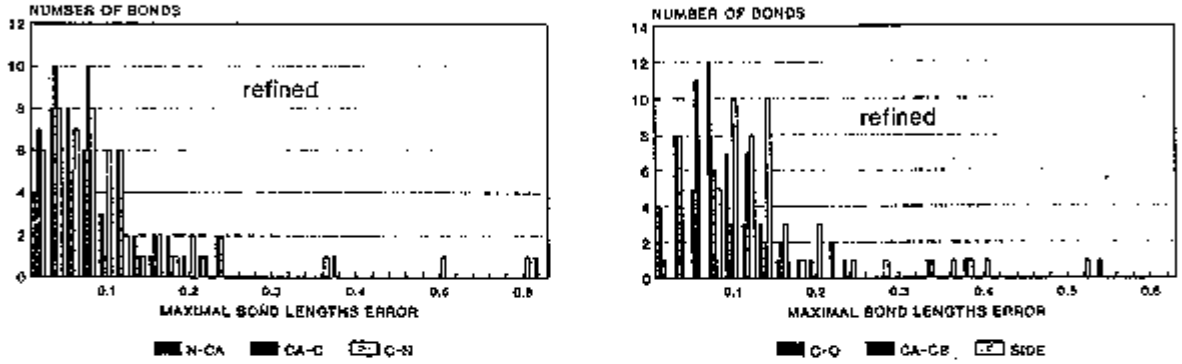

Fig. 2. The histogram of distribution of maximal bond-length errors, Δd_{max}, in refined models : a) main chain bonds, b) other bonds.

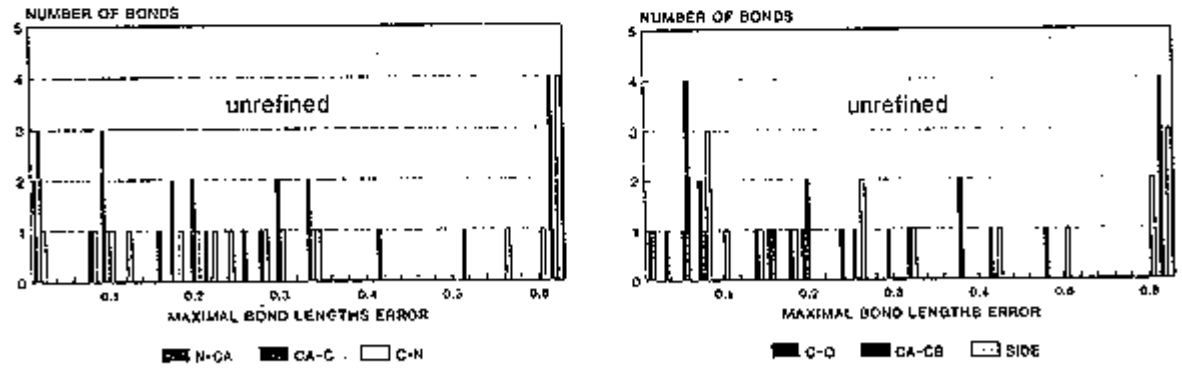

Fig. 3. The histogram of distribution of maximal bond-length errors, Δd_{max}, in unrefined models : a) main chain bonds, b) other bonds.