Journal Title: J. App. Cryst.
ISSN:
Article Author(s): Vriend, G, Hooft, W
Article Title: Verification of Protein Structures: Side Chain Planarity
Year: 1996
Volume: 29
Issue:
Pages: 714-716
PMID:

J. Appl. Cryst. (1996). **29**, 714–716

Verification of Protein Structures: Side-Chain Planarity

ROB W. W. HOOFT, CHRIS SANDER AND GERRIT VRIEND

European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany. E-mail: rob.hooft@embl-heidelberg.de

(*Received 23 April 1996; accepted 27 June 1996*)

Abstract

Nine of the 20 natural amino acids contain a planar group in their side chains. For these groups, normal deviations from planarity were derived by the study of similar fragments in accurately determined small-molecule structures. Comparison of these deviations with values found from a representative set of high-quality protein structures revealed that the planarity of the aromatic residues and arginine in protein structures is comparable to similar fragments in small molecules. For Asn, Gln, Asp and Glu, however, the deviations are up to twice as large as in comparable small-molecule structures, suggesting that adding an extra planarity restraint for these residue types could improve refinement procedures.

1. Introduction

Most three-dimensional protein structures that are currently known have been determined using X-ray crystallography or nuclear magnetic resonance techniques. Both of these techniques normally provide insufficient experimental data for unique determination of the complete structure. Consequently, additional information [like ideal bond lengths and angles (Engh & Huber, 1991)] is needed for a successful refinement process. Modern structure refinement procedures like *X-PLOR* (Brünger, Kuriyan & Karplus, 1987) obtain this extra information from a force field. As more information becomes available over the years, these force fields can be improved and consequently the results of the refinement can become more accurate. As part of a larger project directed towards the improvement of the quality of protein structures, an analysis was made of the planarity of groups in the side chains of Arg, Asp, Asn, Glu, Gln, His, Phe, Tyr and Trp residues. The results are available in the form of a verification procedure that can be used by the scientific community free of charge.

2. Methods

Protein structures were extracted from the Protein Data Bank (PDB) (Bernstein *et al.*, 1977). A representative set of high-quality structures was selected using the *WHAT IF* program (Vriend, 1990) as described below. Protein-like small molecules were selected from the Cambridge Structural Database (CSD) (Allen, Kennard & Taylor, 1983) with the CSD quest software.

2.1 *Construction of a database of reliable protein structures*

Using the *PDBFINDER* database (Hooft, Sander, Scharf & Vriend, 1997), a selection was made from the PDB using the following criteria:

(i) the keyword 'mutant' does not occur in the 'Compound' name;

(ii) the structure was solved using X-ray crystallography;

(iii) the resolution is better than 2.2 Å;

(iv) the R factor is lower than 0.22;

(v) there is at least one water molecule present in the PDB file.

The limits on R factor and resolution are chosen such that together with the quality and uniqueness criteria described below they lead to a final list of about 300 chains.

From the resulting set of PDB files, those for which a directional atomic contact analysis (Vriend & Sander, 1993) indicated abnormal features were excluded. The remaining structures were separated into the constituent amino-acid chains. A further selection on these chains was then performed:

(i) there should be no more than one chain break;

(ii) no more than two Cα-only residues should occur;

(iii) the number of amino acids in the chain should be between 32 and 749 (inclusive). The lower limit was set to avoid the inclusion of small peptides, the upper limit because of a software limitation.

This selection resulted in a set of approximately 1000 chains. These chains were sorted in order of increasing 'empirical unreliability factor' Q:

$$Q = \text{resolution} + 3.0 \times R \text{ factor}. \quad (1)$$

All chains were finally subjected to a sequence-based selection: for any pair of sequences where an alignment indicated that the longest sequence showed more than 30% sequence identity over its full length with the shorter one, the one with the higher Q value was removed from the list.

For the July 1995 release of the Protein Data Bank, this procedure resulted in 285 unique chains.

Table 1. *Number of hits found in the CSD for each fragment*

Fragment	Count
Asp/Glu	218
Phe	757
HisD/HisE/HisH	5 + 26 + 31
Asn/Gln	68
Arg	30
Trp	92
Try	91

Our selection algorithm is similar in spirit to the ones described by Hobohm, Scharf, Schneider & Sander (1992) for the generation of the *PDBSELECT* database. The main differences are that in our database there is no requirement for backward compatibility and that our focus is on optimizing the quality of the included structures rather than on maximizing the size of the resulting set.

The procedure described here has recently been introduced to update the general purpose *WHAT IF* relational database (Vriend, Sander & Stouten, 1994). The procedure is not specific to the planarity check.

2.2. *Planar side chains*

Nine of the 20 standard amino acids contain a planar group in their side chain (Fig. 1):

(i) Asp, Glu, Asn and Gln contain an sp2 hybridized group that has four non-H atoms in one plane;

(ii) the aromatic rings in Phe, His, Trp and Tyr are planar and the non-H atoms directly connected to them are almost coplanar;

(iii) the last five non-H atoms of the Arg side chain (the sp2 group and Cδ) form a plane.

We will be referring to the atoms in the planar groups (shaded dark in Fig. 1) as 'group 1 atoms' and to the non-H atoms connected to the aromatic systems in Phe, His, Trp and Tyr (shaded light in Fig. 1) as 'group 2 atoms'.

2.3. *Determination of normal deviations from planarity*

The Cambridge Structural Database was scanned for the complete structural fragments as shaded in Fig. 1, with the additional constraints that:

(i) coordinates should be present in the entry;

(ii) the crystallographic R factor should be below 7.5%;

(iii) the fragment must be connected to a C atom.

The number of fragments found for each type is listed in Table 1. For each of the fragments found, a least-squares plane was fitted through the group 1 atoms (shaded dark in Fig. 1). The root-mean-square (r.m.s.) distance of the constituent atoms to the plane was calculated, as well as the distance of each of the group 2 atoms (if any) to the plane. All of the calculated parameters were assumed to be normally distributed around 0.0 and the standard deviation of the population was calculated. Obvious outliers (deviating by more than 4σ) were ignored in this process. Results are given in Table 2.

Table 2. *Derived normal deviations from planarity from CSD structures (σ-CSD) and from 285 representative reliable PDB structures (σ-PDB)*

Also given is the number of cases where values over 4.0 times the σ-CSD were observed in the current PDB (N4σ). Estimated standard deviations are given in parentheses.

Parameter	σ-CSD	σ-PDB	N4σ
Asp	0.0063 (3)	0.0090	2176
Glu	0.0063 (3)	0.0091	2164
Phe	0.0089 (2)	0.0091	265
Phe-Cβ	0.064 (2)	0.059	333
His	0.0046 (4)	0.0082	1159
His-Cβ	0.050 (5)	0.066	722
Asn	0.0059 (5)	0.0104	2615
Gln	0.0059 (5)	0.0102	1565
Arg	0.037 (5)	0.034	1338
Trp	0.0135 (6)	0.0144	247
Trp-Cβ	0.074 (6)	0.074	100
Tyr	0.0082 (10)	0.0095	367
Tyr-Cβ	0.061 (5)	0.064	416
Tyr-OH	0.040 (3)	0.045	719

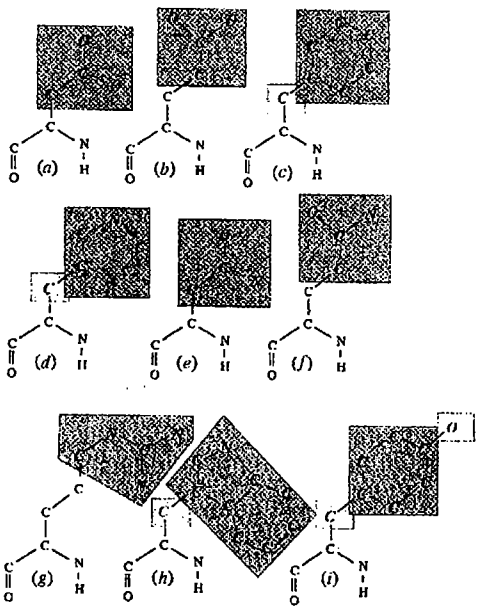

Fig. 1. Planar groups (dark shaded atoms, referred to as 'group 1 atoms') in the side chains of (*a*) aspartic acid, (*b*) glutamic acid, (*c*) phenylalanine, (*d*) histidine, (*e*) asparagine, (*f*) glutamine, (*g*) arginine, (*h*) tryptophane and (*i*) tyrosine. Atoms connected to aromatic systems (referred to as 'group 2 atoms') are shown by light shading; these atoms belong to the planar group, but are treated separately.

To be able to make an overall comparison of the planarity of the studied groups in the PDB and CSD databases, the calculations were repeated on the database of 285 reliable protein structures (Table 2).

3. Results and discussion

In the 285 representative protein structures, planarities for Phe, Tyr, Trp and Arg are comparable to the values obtained from the CSD. For Asp, Glu, Asn, Gln and His, the structures in the CSD are significantly closer to planarity than the structures in the PDB. It seems that planarity restraints used for the latter five residue types in refinement force fields may be too weak to reproduce structural preference of the side chains.

Determination of the side-chain planarities in all protein structures in the PDB reveals that 1916 out of 3866 structures contain one or more residues that deviate from planarity by more than 4.0 times the normal value. 1185 structures have at least two such residues. Deviations as high as 41 times the normal value are observed in refined structures and up to 100 times the normal value for unrefined structures or structures that were subjected to unrestrained refinement. Visual inspection of a number of residues deviating just over 4.0 times the normal value shows that this is a reasonable limit for the test.

The number of residues in a structure deviating more than 4.0 times the normal value from planarity is not significantly correlated with the publication date of the structure. Our results indicate that a difference can be seen between the different refinement programs, but all of the well known programs do produce extremely large deviations occasionally.

4. Availibility

The planarity verification is available as part of the *WHAT_CHECK* program (Hooft, Vriend, Sander & Abola, 1996). This program is available *via* anonymous FTP from swift.embl-heidelberg.de in directory /whatcheck/. The verification procedures in *WHAT_CHECK* are part of the Biotech protein structure verification suite at http://biotech.embl-heidelberg.de:8400/. Results for X-ray structures from the PDB are accessible as part of the *PDBREPORT* database, available at http://www.sander.embl-heidelberg.de/pdbreport/.

This work was carried out in the context of the PDB Verification Project funded by the European Commission (RWWH) and the RELIWE project funded by the BMFT (GV).

References

Allen, F. H., Kennard, O. & Taylor, R. (1983). *Acc. Chem. Res.* **16**, 146–153.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458–460.

Engh, R. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.

Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). *Prot. Sci.* **1**, 409–417.

Hooft, R. W. W., Sander, C., Scharf, M. & Vriend, G. (1997). In preparation.

Hooft, R. W. W., Vriend, G. Sander, C. & Abola, E. E. (1996). *Nature (London)*, **381**, 272.

Vriend, G. (1990). *J. Mol. Graph.* **8**, 52–56.

Vriend, G. & Sander, C. (1993). *J. Appl. Cryst.* **26**, 47–60.

Vriend, G. Sander, C. & Stouten, P. F. W. (1994). *Prot. Eng.* **7**, 23–29.