CABIOS

Vol. 13 no. 4 1997
Pages 425–430

Objectively judging the quality of a protein structure from a Ramachandran plot

Rob W.W.Hooft, Chris Sander and Gerrit Vriend

Abstract

Motivation: *Statistical methods that compare observed and expected distributions of experimental observables provide powerful tools for the quality control of protein structures. The distribution of backbone dihedral angles ('Ramachandran plot') has often been used for such quality control, but without a firm statistical foundation.*

Results: *A new and simple method is presented for judging the quality of a protein structure based on the distribution of backbone dihedral angles. Inputs to the method are 60 torsion angle distributions extracted from protein structures solved at high resolution; one for each combination of residue type and tri-state secondary structure. Output for a protein is a Ramachandran* Z-score, expressing the quality of the Ramachandran plot relative to current state-of-the-art structures.

Availability: *The Ramachandran test is available as part of the free WHAT_CHECK program. Information about this program can be obtained on the WWW from http://swift.embl-heidelberg.de/whatcheck/*

Contact: *E-mail: r.hooft@euromail.com*

Introduction

The three backbone torsion angles ϕ, ψ and ω are the main determinants of a protein fold. The allowed range of ω angles is very restrictive (MacArthur, 1996), so variations in this torsion angle do not give much conformational variety. Ramachandran *et al.* (1963) have created two-dimensional (2D) scatter plots of ϕ,ψ pairs, comparing them to a predicted distribution. These scatter plots are now commonly known as Ramachandran plots.

Simple polymer physics models can be used to make a predicted distribution of pairs of ϕ,ψ angles using a volume exclusion model: no two non-bonded atoms can overlap. Results of such calculations show a considerable conformational freedom in the two torsion angles, but with a number of clear restrictions. Deviations from the expected distribution for a new protein structure can now be used to judge the quality of that structure.

A distinct advantage of the ϕ,ψ distribution over many other diagnostics for structure quality is that it is very hard to improve the ϕ,ψ distribution using just structure refinement software (G.Kleywegt, personal communication); the Ramachandran plot is, therefore, an indicator of the intrinsic quality of the structure, and not an indicator of how well the responsible crystallographer is acquainted with the analysis tools.

Instead of volume exclusion models, many modern programs to make Ramachandran plots (e.g. PROCHECK; Laskowski *et al.*, 1993) use database statistics to create the reference distribution. A big advantage of these statistical techniques is that there are no simplifications involved, and the distributions thus represent the real conformational preference of a protein chain. However, these statistical techniques use a database of known structures, and the quality of the distributions is dependent on the quality of these known structures. Thus, it is very important that the reference database is kept up to date.

While it is possible to judge the quality of a plot by visual inspection or a simple cut-off criterion, an objective statistical analysis requires the exact definition of a reference distribution and a quantitative method to assess deviations from that distribution. The latter is the approach taken by our new procedure.

As in all statistical analyses, some deviations from normality are expected. Upon validating structures, great care should be taken not to call these expectations 'errors'. For example, ϕ,ψ torsion angles for active site residues often deviate from common values. For a normal distribution, ~5% of all observations are expected to be >2σ away from the average, and 0.01% >4σ. As the number of observables in a Ramachandran plot is approximately equal to the number of residues in the protein, a fair number of 2σ deviations are expected in each structure, and even a single 4σ deviation is not a rare event. Drawing conclusions from the individual residue scores is, therefore, difficult. PROCHECK addresses this problem by allowing 10% of the residues to be outside the most favoured areas. Our approach is to calculate a composite score in order to overcome the natural spread present in individual amino acid scores.

Algorithm

For each non-terminal residue in a protein chain, the two angles ϕ and ψ are given. The number of times (c) a similar pair of values appears in the database of reliable structures is

EMBL, Meyerhofstraße 1, D-69117 Heidelberg, Germany

taken as a measure of how 'normal' a certain ϕ,ψ combination is. Consider the following very simple procedure. A histogram of database occurrences is created with a ϕ,ψ grid size of $10° \times 10°$. Each residue k in a protein gets a score c_k from:

$$c_k = \text{number of database residues in the same } 10° \times 10° \text{ bin} \quad (1)$$

A large value of c_k indicates normality, as this local backbone conformation has frequently been observed in known structures. For example, values around $\phi = -60°$ and $\psi = -45°$ are commonly observed in α helices. However, by just counting the number of occurrences of a ϕ,ψ pair regardless of residue type or secondary structure element, information is lost. For example, not separating residue types will give anomalously low scores for Gly residues in reverse turns, and not separating secondary structure types will cause anomalously high scores for Pro residues in α helices. Furthermore, all-helical proteins like ROP (Banner *et al.*, 1987) will give higher scores than proteins like rubredoxin (Dauter *et al.*, 1992) that have almost no secondary structure, just because the ϕ,ψ distribution in the helical region is sharply peaked. To prevent these sources of bias, the data are subdivided by secondary structure and residue type, i.e. 3×20 histograms are created, one for each combination of three-state secondary structure [as determined by DSSP (Kabsch and Sander, 1983)] and amino acid type. The grid size for each of these histograms is $10° \times 10°$.

A straightforward application of these histograms in the sense of equation (1) would still be statistically unsatisfactory. This is because the counts for the different residues should not be compared directly: finding three Gly residues in a specific kind of loop will be much more common than finding three Trp residues in a β strand. Instead of using the count c_k for a residue k from the histogram directly, a normalized score z_k is calculated for each residue:

$$z_k = \frac{c_k^{ss,rt} - \langle c^{ss,rt} \rangle}{\sigma(c^{ss,rt})} \quad (2)$$

Here, $\langle c^{ss,rt} \rangle$is the database average of c for all residues with the same secondary structure type 'ss' and residue type 'rt' as residue k in the current protein, and $\sigma(c^{ss,rt})$ is the corresponding standard deviation. Rather than making a second pass over the database, the average and standard deviation can be efficiently calculated from the histograms:

$$\langle c^{ss,rt} \rangle = \frac{\sum_j (c_j^{ss,rt})^2}{\sum_j c_j^{ss,rt}} \quad (3)$$

$$\sigma^2(c^{ss,rt}) = \frac{\sum_j c_j^{ss,rt}(c_j^{ss,rt} - \langle c^{ss,rt} \rangle)^2}{(\sum_j c_j^{ss,rt}) - 1} \quad (4)$$

with summations (j) over all 36×36 ϕ,ψ bins.

Independent of residue type and secondary structure, the expected value of z_k in a normal protein structure is 0.0 with a standard deviation of 1.0 for each residue k. Since scores for all residues in a protein are on the same scale, a meaningful average score C for the entire protein can be calculated:

$$C = \frac{\sum_{k=1}^{K} z_k}{K} \quad (5)$$

In this equation, the summation (k) is over all K non-terminal amino acids in the protein.

To use C as a measure of quality, one needs to make reference to comparable values for all structures in the database (C_l for $1 < l < L$). This allows one to see how far a particular protein deviates from normality:

$$\langle C \rangle = \frac{\sum_{l=1}^{L} C_l}{L} \quad (6)$$

$$\sigma^2(C) = \frac{\sum_{l=1}^{L} (C_l - \langle C \rangle)^2}{L - 1} \quad (7)$$

$$Z = \frac{C - \langle C \rangle}{\sigma(C)} \quad (8)$$

When calculating C_l for database protein l, care is taken not to include protein l in the determination of the histograms. Failure to exclude the protein itself would result in inflated scores C_l for the database proteins, and consequently lower scores Z for new proteins under study.

In order to analyse a new structure, the scores for the residues in the protein are calculated using equation (1) and these are normalized using equation (2). Then the overall protein score is computed using equations (5) and (8).

A few modifications to this algorithm are used to make sure the results are well behaved in the general case:

- The resolving power is improved by using linear interpolation between adjacent bins in the histograms when looking up c_k.
- The number of usable residues in our database is ~60 000. This means that there are on average 1000 residues per histogram, or 1 per $10° \times 10°$ bin. For some histograms there are not enough data points to create a statistically significant number of counts ($\langle c^{ss,rt} \rangle < 2.0$). In such cases, a merged histogram for all amino acids except proline and glycine is used to score the residue instead.

Implementation

The Ramachandran Z-score procedure was implemented as a procedure in the WHAT IF program (Vriend, 1990). The 295 structures contained in our current (June 1996) non-redundant database (Hooft *et al.*, 1996a) were used for calibration. This database consists of ~60 000 non-terminal residues. Example distributions are shown in Figure 1.

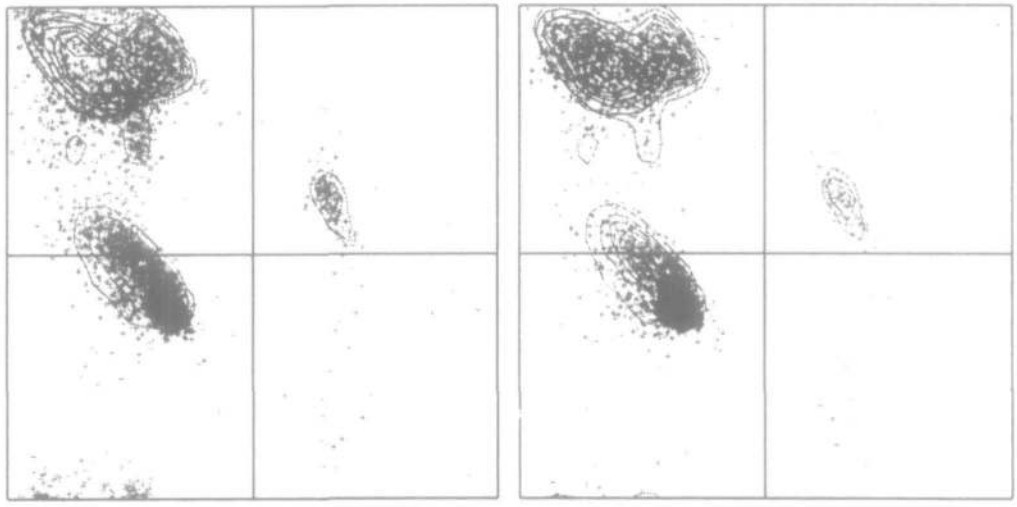

Fig. 1. Database densities of Asp (left) and Glu (right) residues shown with 'allowed areas' (averaged for non-Gly, non-Pro) for helical (blue), strand (red) and other (green) residues. Contours are drawn at 10, 20, 30, 40 and 50% of the maximum density for the three secondary structure types separately. Residues are colour coded by DSSP (Kabsch and Sander, 1983) secondary structure (same colouring as contour levels).

Discussion

The calibration process makes the average score for the database structures 0.0 with a standard deviation of 1.0. For 2897 protein X-ray structures with crystallographic *R*-factor below 25% and resolution less than 2.8 Å in the PDB (Bernstein *et al.*, 1977), the average Ramachandran *Z*-score is −0.7 with a standard deviation of 1.3 (Figure 2). A total of 162 of these structures have a score below −3.0, 55 score below −4.0 (more negative scores indicate lower quality). As expected, the average score is lower than that of the carefully selected dataset used for calibration. A study of Figure 3 shows that the resulting *Z*-scores correspond very well to an intuitive evaluation of the quality of the Ramachandran plot. It is our experience that a *Z*-value of −4.0 or lower indicates a serious problem with the structure.

A comparison of our *Z*-score with PROCHECK (Laskowski *et al.*, 1993) results is shown in Figure 4. It is clear that although there is a strong correlation between the two scores, pairs of structures can be located with quite different *z*-scores that both have 90% of their residues in the most favoured areas as defined by PROCHECK. Examples for these extremes are given in Figure 5. For these two structures, a clear difference can be seen between the distributions of the

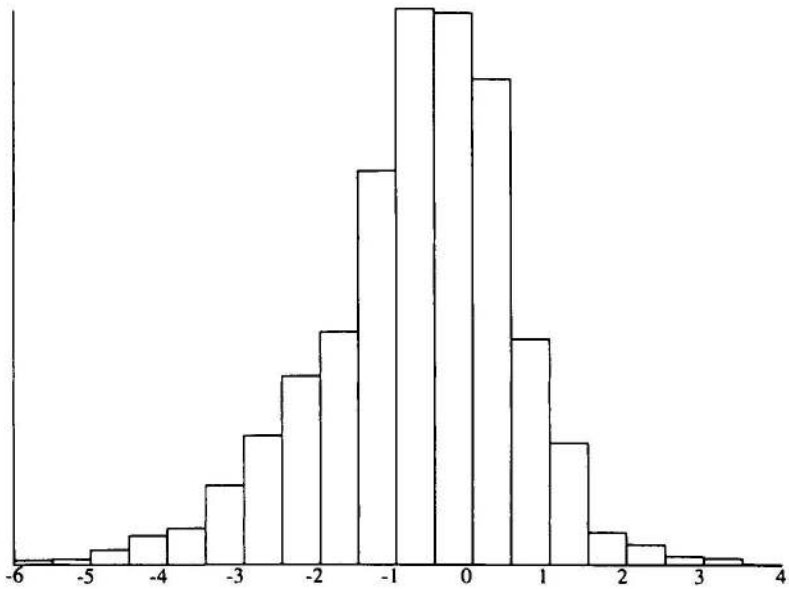

Fig. 2. Distribution of Ramachandran *Z*-scores for 2897 protein X-ray structures in the PDB.

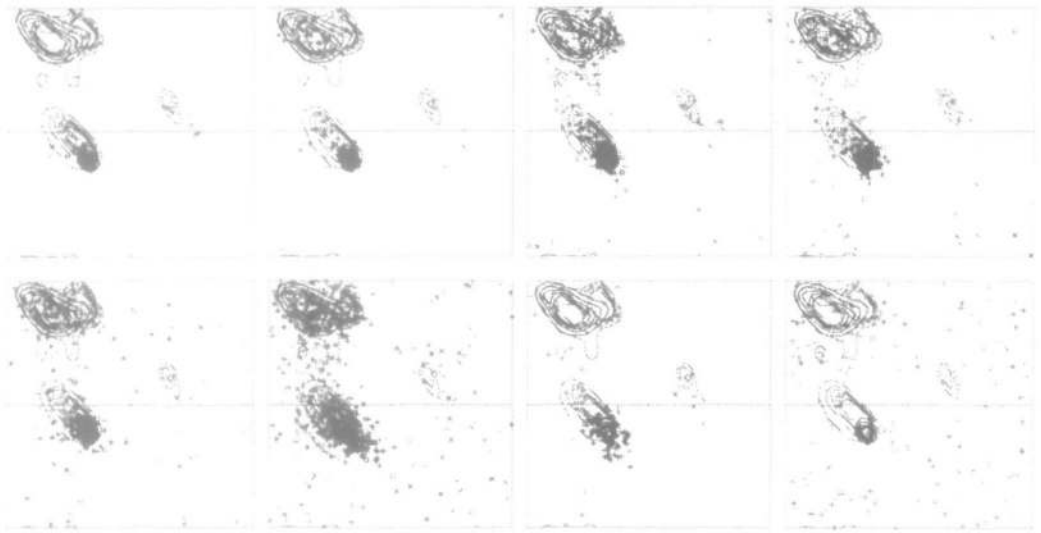

Fig. 3. Example Ramachandran plots for PDB structures resulting in Z-scores of 2.0, 0.0, −2.0, −3.0, −4.0, −5.0, −6.0 and −8.0 (left to right, top to bottom). Contouring and colouring are the same as in Figure 1; slightly different shades of green are used for turn and coil, and different shades of blue for α helix and 3_{10} helix.

residues within their allowed regions, a difference that is impossible to detect with a simple cut-off criterion. A visual inspection of the structures shows that the helices in the structure with low Z-score are much less regular than those in the structure with high Z-score: backbone oxygens are aligned less well with the helix axis, giving rise to much worse hydrogen bonding. The converse plot is shown in Figure 6: two structures both resulting in average Z-scores of 0.0, but with different percentages of residues in the most favoured areas. In this case, the structure with low PROCHECK score has a large percentage of loop residues, and a number of these are found near the edges of the contoured areas. The structure with high PROCHECK score does have more of its points inside the contoured areas. However, quite a number of residues have a helical hydrogen bonding pattern but ϕ, ψ angles representative for loops, and almost all strand residues are found at the edge of the strand area. The 96% PROCHECK score suggests that this structure is 'very regular'; the fact that it is not so regular can only be detected by our procedure because it evaluates the distribution of ϕ, ψ angles separately for each type of secondary structure.

The same statistical analysis of normality can be applied to

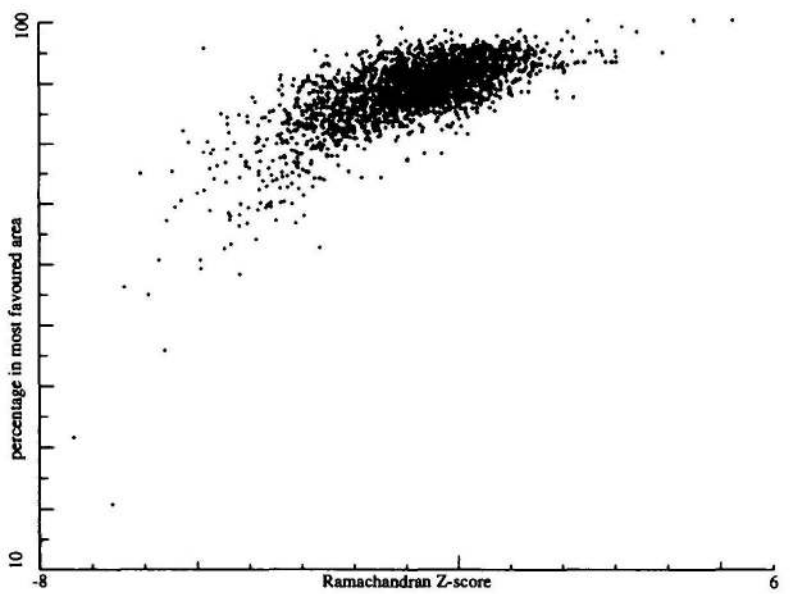

Fig. 4. Comparison of Ramachandran Z-scores with PROCHECK results.

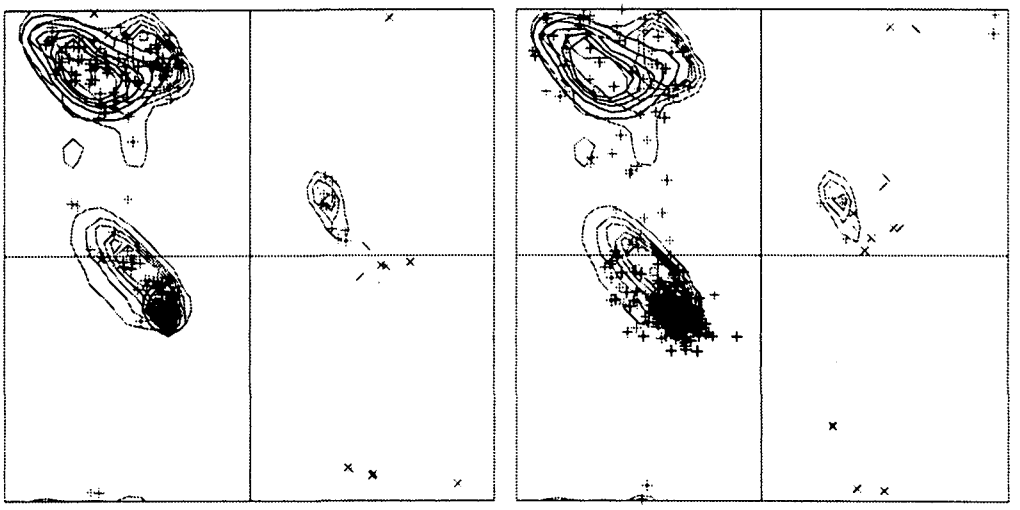

Fig. 5. Example Ramachandran plots for PDB structures both having 90% of their residues in PROCHECK's most favoured areas, but resulting in Z-scores of 1.4 (left) and −3.4 (right). Contouring and colouring are the same as in Figure 3.

other distributions in protein structures as well. We have implemented similar methods to assess $\chi 1/\chi 2$ distributions and five-residue backbone conformation normality.

Using automatic recalibration with the latest non-redundant dataset, no special efforts are required to ensure that new developments in the determination of the structure of proteins propagate into the Z-scores. The resulting Z-scores represent a 'current assessment' with respect to high-quality structures at a particular moment. They are dynamic entities that change when our understanding of what is 'perfect' improves, based on higher-quality X-ray data.

X-ray structures based on low-resolution data generally score worse than those based on high-resolution data. Other programs have special provisions to compensate for this effect, and their scores will thus indicate the quality of a particular structure as compared to other structures with similar resolutions. These scaled values can be used to find out whether it would pay off to put more efforts in the refinement of a new structure. We do not use resolution-dependent calibration because we prefer to indicate the quality of a structure as compared to current standards. Our unscaled values are more valuable, for example, in the selection of a good structure for modelling purposes.

Availability

The Ramachandran plot quality analysis is available as part of the WHAT_CHECK program (Hooft *et al.*, 1996b). This program is available via anonymous ftp from ftp://swift.embl-heidelberg.de/whatcheck/

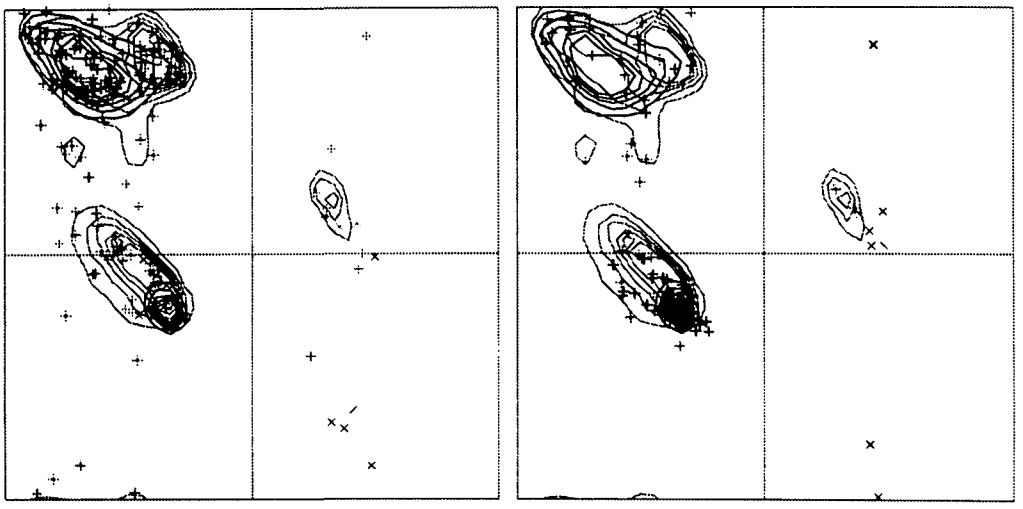

Fig. 6. Example Ramachandran plots for PDB structures both having a Z-score of 0.0, but having 81.6% (left) and 96.0% (right) of their residues in PROCHECK's most favoured areas. Contouring and colouring are the same as in Figure 3.

Many of 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 via http://www.sander.embl-heidelberg.de/pdbreport/

Acknowledgements

This work was carried out in the context of the PDB verification project funded by the European Commission (R.W.W.H.) and the RELIWE project funded by the German BMFT (G.V). We wish to thank Brigitte Altenberg, Karina Krmoian, and the EMBL Computer Group for their technical support. Discussions with Gerard Kleywegt and Roman Laskowski inspired us to work on Ramachandran plots.

References

Banner,D., Kokkinidis,M. and Tsernoglou,D. (1987) Structure of the ColE1 Rop protein at 1.7Å resolution. *J. Mol. Biol.*, **196,** 657–675.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The protein data bank: a computer-based archival file for macro-molecular structures. *J. Mol. Biol.*, **112,** 535–542.

Dauter,Z., Sieker,L.C. and Wilson,K.S. (1992) Refinement of rubredoxin from *Desulfovibrio vulgaris* at 1.0Å with and without restraints. *Acta Crystallogr.*, **B48,** 42–59.

Hooft,R.W.W., Sander,C. and Vriend,G. (1996a) Verification of protein structures: Side-chain planarity. *J. Appl. Crystallogr.*, **29,** 714–716.

Hooft,R.W.W., Vriend,G., Sander,C. and Abola,E.E. (1996b) Errors in protein structures. *Nature*, **381,** 272.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen bond and geometrical features. *Biopolymers*, **22,** 2577–2637.

Laskowski,R.A., MacArthur,M.W., Moss,D.S. and Thornton,J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.*, **26,** 283–291.

Ramachandran,G.N., Ramakrishnan,C. and Sasisekharan,V. (1963) Stereochemistry of polypeptide chain conformations. *J. Mol. Biol.*, **7,** 95–99.

MacArthur,M.W. and Thornton,J.M. (1996) Deviations from planarity of the peptide bond in peptides and proteins. *J. Mol. Biol.*, **264,** 1180–1195.

Vriend,G. (1990) WHAT IF: a molecular modelling and drug design program. *J. Mol. Graph.*, **8,** 52–56.

Received on December 6, 1996; accepted on March 27, 1997