Phi/Psi-chology: Ramachandran revisited

Gerard J Kleywegt and T Alwyn Jones*

Address: Department of Molecular Biology, Biomedical Centre, Uppsala University, Box 590, S-751 24 Uppsala, Sweden.

*Corresponding author.
E-mail: alwyn@xray.bmc.uu.se

Structure 15 December 1996, 4:1395–1400

 ISSN 0969-2126

The errors that can be introduced into a protein model during model building and refinement vary tremendously in their importance and severity [1,2]. At one extreme, the mainchain may be totally incorrectly traced in the experimental map, or the molecular replacement solution may be wrong. Minor errors may include an incorrect peptide orientation, or misplaced or excessive water molecules. The reasons why errors creep into models are many, but for a structure built into an experimental map, the main ones are limited resolution and poorly phased diffraction data. Other things being equal, the resolution of the diffraction data should be the ultimate variable that determines the accuracy of a structural investigation. Inevitably, life is more complicated, and a successful structural investigation is often a learning experience for the people involved.

To ensure the correctness of a study, the crystallographer has relied on the R factor which gives an overall measure of how well the final model fits the experimental diffraction data. The trust in this indicator for structures solved at medium and low resolution has been severely dented by a series of high profile studies where severe errors have been made and gone undetected. To supplement this single indicator, a number of new figures of merit have been suggested, of which the free R factor of Brünger [3] is particularly useful and is increasingly being used [4].

One of the more surprising results of high-resolution diffraction studies on proteins has been the observation that the conformational angles show preferences for (combinations of) values that are expected based on simple energy considerations. This has prompted us to rely on the use of sidechain rotamers during the initial map interpretation stage [5] and during refinement at low resolution. Deviations from the preferred conformations can then be used as indicators of potential error. It must be emphasized, however, that these are merely potential error indicators and they must be carefully evaluated with the experimental information that is available to the crystallographer. Due to steric hindrance, the mainchain of a polypeptide usually assumes preferred, energetically favourable conformations [6]. For each residue, these conformations can be characterized by the value of two torsion angles, ϕ and ψ (the third angle, ω, is largely restricted to values of 180° for *trans*-peptides, and 0° for *cis*-peptides). The φ angle of residue *i* is defined by the torsion C_{i-1}-N_i-$C\alpha_i$-C_i, and ψ by the torsion N_i-$C\alpha_i$-C_i-N_{i+1}. The distribution of ϕ and ψ is usually called the Ramachandran plot. More than ten years ago, such plots were used to remove two structures from a database of high-resolution structures that had been created for model building in experimental maps [7] (TAJ, unpublished results). Both structures have since been shown to contain severe errors.

The Ramachandran plot will clearly show how well the ϕ and ψ angles cluster and will reveal other oddities that may be the result of errors made during refinement. Unfortunately, many scientific magazines consider such a plot to be too technical for their readership, who are more interested in biological relevance and beautiful pictures. In our experience, the Ramachandran plot is one of the simplest and most sensitive means for assessing the quality of a protein model in the absence of experimental data. The major reason for this is probably the fact that the ϕ, ψ angles (or combinations of these) are not usually restrained during X-ray refinement, as opposed to bond lengths and bond angles, for instance. Therefore, an indicator that shows how much a particular structure deviates from the preferred areas of a Ramachandran plot is, we believe, a requirement for assessing the quality of a protein model.

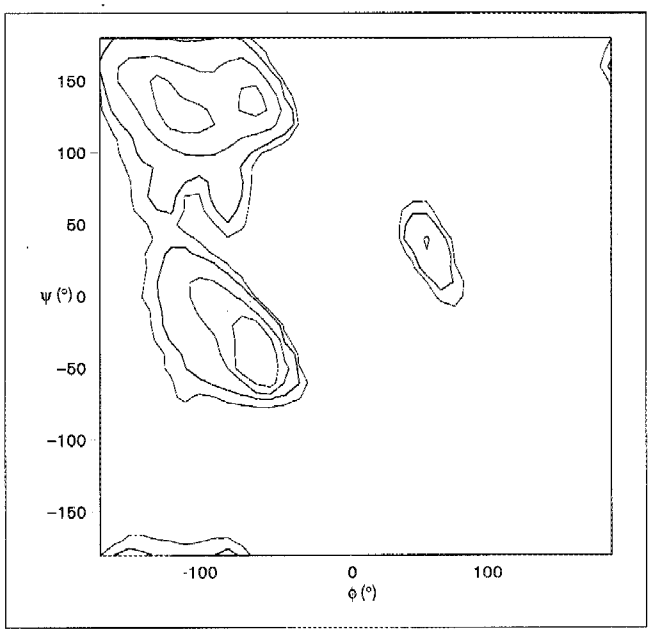
With the advent of the program ProCheck [8], Ramachandran plots have gained somewhat in popularity. ProCheck divides the Ramachandran plot into four types of area: most favoured, additional allowed, generously allowed and disallowed. A typical good model should not only have very few residues within the disallowed regions, but also very many in the most favoured regions. Unfortunately, the division into four regions has given rise to confusion when it comes to reporting the quality of the Ramachandran plot. Many authors only quote the number or percentage of residues in disallowed regions, others quote only those in the most favoured regions. Even more difficult to interpret is a phrase such as, '~80% of the residues were found to lie in allowed regions according to ProCheck'. This phrase may describe a high-quality model (where the authors meant to say 'most favoured') but can equally well be used to describe a very poor model (if the authors meant to say '~20% in disallowed regions'). For instance, the backwards-traced model of cellular retinoic acid-binding protein which we described earlier [1], has 8.9% of its residues in disallowed regions, and only 42.7% in the most favoured regions. Nevertheless, an unscrupulous crystallographer could report this as '91% of the residues lie in allowed

regions of the Ramachandran plot'. This problem was also recently noted by Karplus who, in an independent study, found that 'much of conformational space designated as allowed and generously allowed, and even some of the core region is very rarely (or not at all) observed' [9].

In order to remedy this problem, we have carried out an analysis of high-resolution protein structures (see the Methods section). This has resulted in a division of the Ramachandran plot into two areas: core and non-core. The core regions consist of the most populated 10° by 10° areas which together account for 98% of all non-glycine residues in our sample (Fig. 1); together they occupy only 19.7% of the entire plot area. By having a binary classification scheme, ambiguities concerning allowed regions are avoided. Also, we have chosen to include proline residues in the analysis, as the most populated areas of the Ramachandran plot for these residues are not outside the areas found for all other non-glycine residues (data not shown).

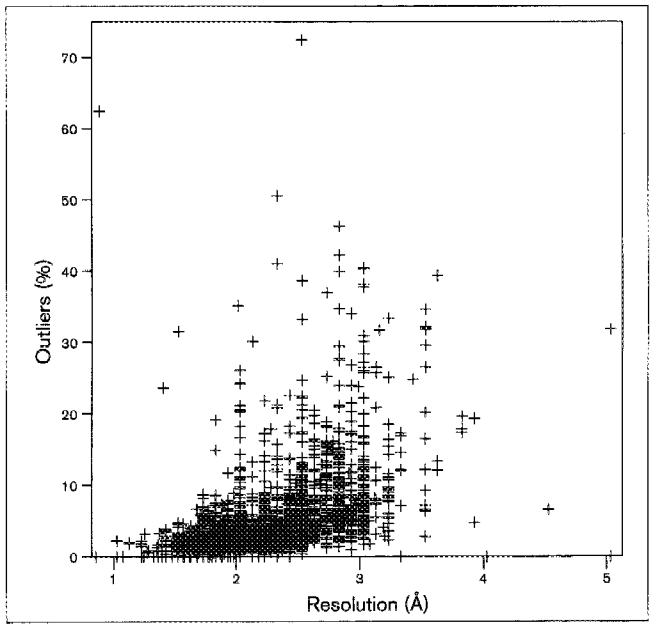
Figure 2 shows the relationship between resolution and the percentage of residues in non-core regions (outliers) for more than 3000 protein structures from the Protein Data Bank (PDB). Table 1 shows the distribution of the percentage of outliers for all protein X-ray structures (with full coordinates and at least 20 residues) that were in the PDB

Figure 1



Distribution of angle combinations for high-resolution structures. The outer (magenta) contour level encloses the most populated regions which together account for 98% of all non-glycine residues. Subsequent contour levels are at 95, 90, 80 and 50% (purple, yellow, green and red), respectively. (See the Methods section for details.)

Figure 2



Distribution of the percentage of Ramachandran outliers as a function of resolution for 3076 protein structures in the February 1996 release of the PDB.

in February 1996. This shows that ~91% of all structures have 10% or fewer outliers. Only ~4% have more than 15% outliers, and ~1.5% have more than 25% outliers. However, these numbers vary a great deal as a function of time: for structures deposited between 1973–1980, 66.7% have no more than 15% outliers; for the period 1981–1985, this number is 85.8%; for 1986–1990 it is 94.2%; and for 1991–1995 it is 97.1%. Indeed, there is a weak negative correlation between the year of deposition and the percentage of outliers (correlation coefficient –0.16).

In Figure 2, it is the high-resolution structure of gramicidin [10] which is responsible for the noticeable outlier. Figure 3 shows its Ramachandran plot, which has more than 60% outliers. However, this molecule (refined to 0.86 Å resolution) is a small peptide, which contains both D- and L-amino acids. The Ramachandran plot shows that the preferred ϕ, ψ angles of the D-amino acids are positioned symmetrically around the diagonal of the plot from those of the L-amino acids. This case clearly demonstrates that outliers in a Ramachandran plot are not necessarily errors. However, it is the responsibility of the crystallographer to investigate if outliers are due to errors in the model, or if they represent unusual features of the structure.

We have also looked at Ramachandran plots for protein models for which a free R value [3] is quoted in the PDB entry. We identified 127 such entries and find that the fraction of outliers is slightly more strongly correlated with the

Table 1

Distribution of the percentage of Ramachandran outliers in protein models in the PDB.

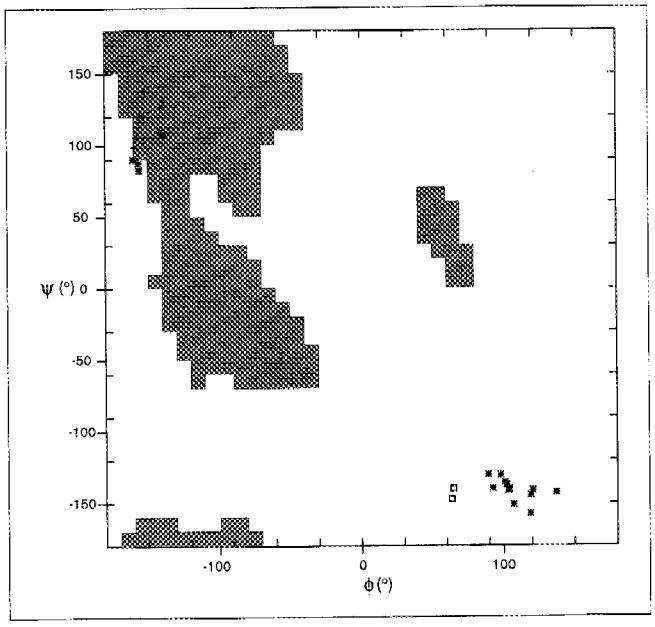
Ramachandran outliers (%)	Number of PDB entries	Fraction* (%)
0–5	2354	76.5
5–10	448	14.6
10–15	143	4.7
15–20	60	2.0
20–25	26	0.9
25–30	16	0.5
30–35	14	0.5
35–40	7	0.2
40–45	4	0.1
45–50	1	0.03
>50	3	0.1

*Fraction of total number of PDB entries.

free R value (correlation coefficient +0.57) than with the conventional R value (+0.49); however, resolution is the highest correlated factor (+0.64).

Of course, one could introduce ϕ, ψ restraints during refinement in order to cosmetically improve the model. We have tried this with X-PLOR [11] using the 3.2 Å model of the complex of the Fc fragment of human IgG with the C2 domain of protein G [12], which has 16% outliers. By introducing restraints for those residues which lie near any of the core regions, the fraction of outliers can be reduced to 11% even with a very low force constant (10 kcal mol^{-1} Å^{-2}. However, even with a force constant greater than 5000 kcal mol^{-1} Å^{-2} the fraction of outliers gets only as low as 9%, and this is then at the expense of an increase of both the conventional and the free R values. It therefore appears to be rather difficult to 'fudge' the indicator.

Figure 3



Ramachandran plot for gramicidin [10]. In this and subsequent Ramachandran plots, glycine residues are shown as squares. Non-glycine residues are shown as plus signs if they fall inside core regions, and as an asterisk if they lie outside the core regions; core regions are shaded in green.

Ramachandran plots can be used to monitor the progress of refinement and rebuilding of a protein model. For instance, Figure 4 shows the Ramachandran plots for various models of cellobiohydrolase I as the structure was built and refined [13]. Table 2 shows how the Ramachandran outlier indicator improved as the model improved (C Divne, personal communication).

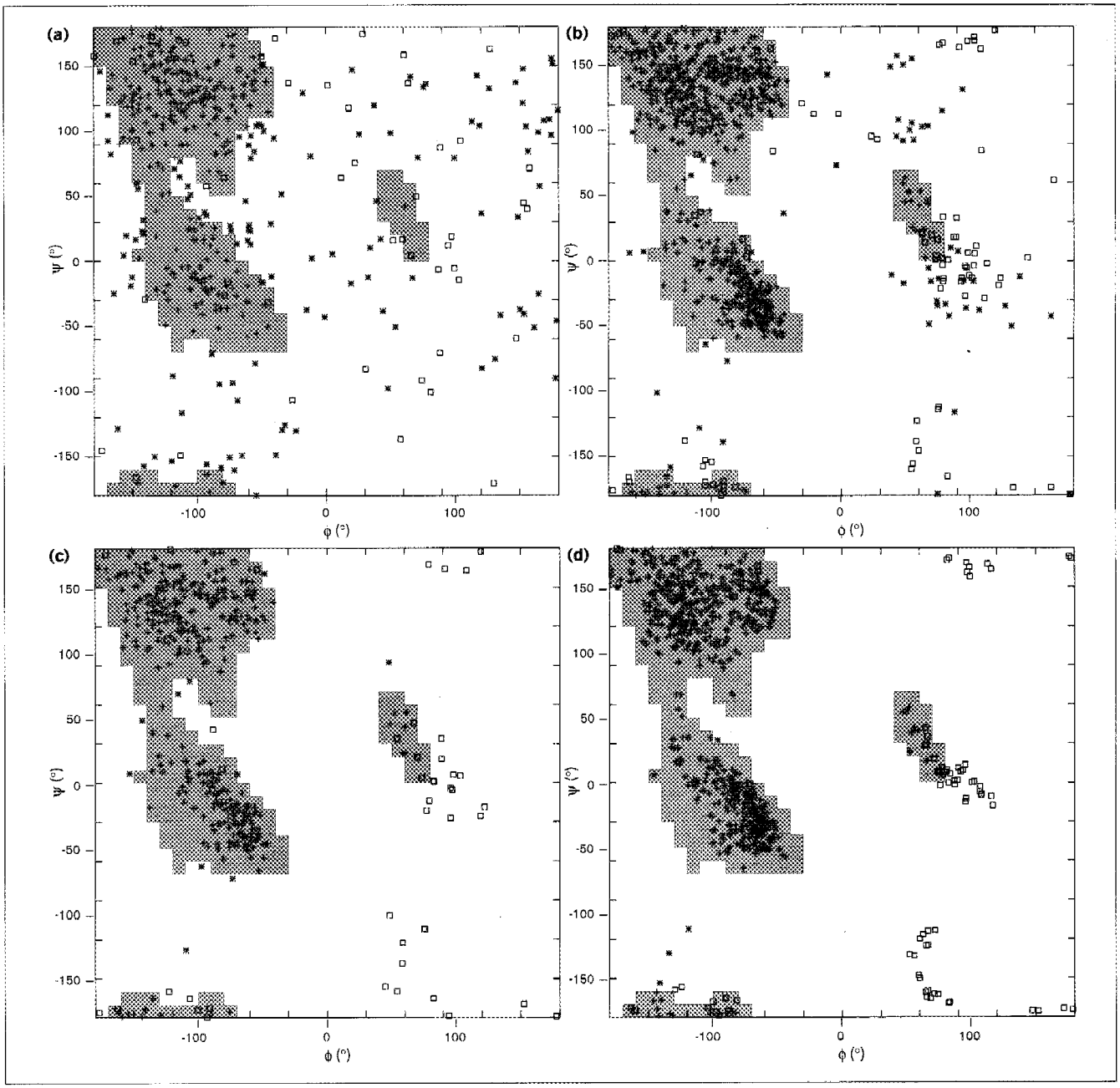
The Ramachandran outlier indicator described here is, we believe, a very useful measure as to how well the structure fits the expected mainchain torsion angle distribution. Nevertheless, a plot still contains a lot more information than a single number. This is clearly illustrated by the gramicidin example above, and by other examples (e.g. one case where the crystallographers appear to have gone to some extremes to reduce the number of residues with positive ϕ angles, is shown in Fig. 5).

Non-crystallographic symmetry

Sometimes macromolecules crystallize with more than one independent copy of the molecule inside the crystallographic asymmetric unit. In such a case of non-crystallographic symmetry (NCS), a simple modification of the Ramachandran plot instantly turns it into a means to visualize the difference in the backbone torsion angles for corresponding residues in the various NCS-related molecules [14]. The modification entails a simple calculation of the centroid ϕ, ψ angles for each set of NCS-related residues, and connecting the points in the Ramachandran plot to this centroid. Normally, one would expect most residues to cluster fairly tightly, although some clusters with larger spread may occur, for example in hinge regions [14]. However, if one finds that most or all clusters show severe scatter, one might want to introduce NCS-restraints in further refinement to avoid artefactual differences between the NCS-related molecules.

There are a number of cases known of structures containing NCS that have been deposited and show large numbers of Ramachandran outliers as well as large differences between the mainchain torsion angles of NCS-related residues. Models with more than 15% outliers should be regarded with caution; the depositing authors should probably try to correct these. If NCS is involved, we suggest that more care should be taken during refinement to prevent the introduction of artefacts [1,14]. Some structures with NCS display genuine conformational differences, for example as global domain motions. This is illustrated by a

Figure 4



Ramachandran plots for **(a)** the initial MIR model (A1) of cellobiohydrolase I [13]; **(b)** model A2; **(c)** model A3; and **(d)** the near final model, A16 (also see Table 2). Residues are labelled as described in Figure 3.

new structure of ligand-free ribose-binding protein, in which there are two molecules in the asymmetric unit. They differ by a domain rotation of ~12° in one molecule relative to the other (SL Mowbray, personal communication); Figure 6 shows a Ramachandran plot of these structures. Overall, there are only 1% outliers and the vast majority of the mainchain torsion angles are similar in both models. However, particularly in the hinge region, there are some real differences that manifest themselves as longer connecting lines that together generate the domain rotation.

Methods

We used the list of Hobohm and Sander [15], of August 1995, and the PDB [16] release of October 1995, to create a set of 403 protein models. These models had no more than 95% sequence identity, contained more than 20 amino acid

Table 2

Concomitant improvement of model quality and Ramachandran plot during the refinement of cellobiohydrolase I.

Model number	R factor (%)	Ramachandran outliers (%)	*Rmsd to final model (Å)
A1	49.9	32.7	0.715
A2	29.3	6.9	0.314
A3	31.8	2.6	0.236
A4	25.0	1.7	0.165
A9	20.6	0.5	0.084
A16	18.1	0.7	0.002

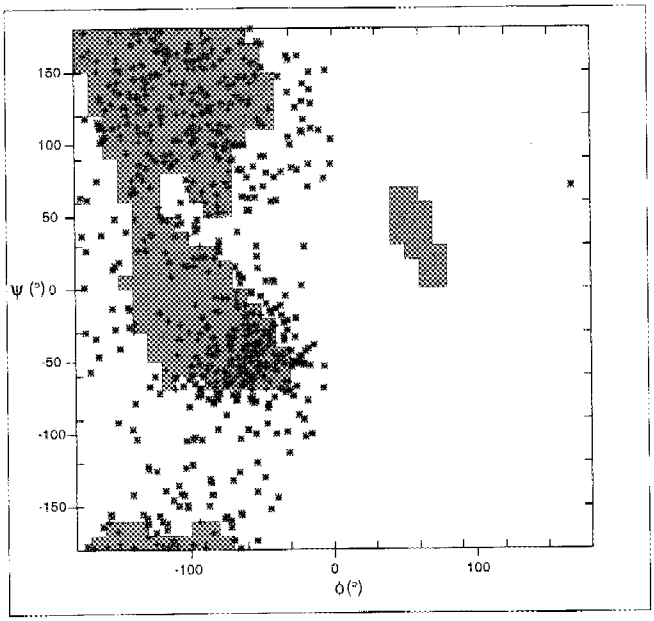
*Root mean square deviation.

residues, and had been solved by X-ray crystallography at a resolution better than, or equal to, 2.0 Å. For each model, all atoms (and their associated torsion angles) whose temperature factor was higher than the average protein temperature factor plus two standard deviations were discarded. This was done in order to exclude residues from the analysis whose conformation might have been determined more by the restraints or force field used in the refinement than by actual experimental data.

For the Ramachandran analysis, the plot was divided into squares of 10° by 10°, and the ϕ, ψ combinations in each square were tallied for 81 782 residues. Although the distributions are different for different residue types, here we only discuss the statistics pertaining to all (74 893) non-glycine residues. The distribution of ϕ, ψ values for these residues is shown in Figure 1. Note that the area commonly associated with β structure actually contains two maxima. The outer contour line delineates the most populated areas which together account for 98% of all non-glycine residues. For an average X-ray model determined at a resolution of 2.0 Å or better, one would expect ~0–5% of the non-glycine residues to lie outside the shaded areas (an estimate determined by analyzing all protein models in the PDB solved at a resolution of 2.0 Å or better). The average fraction of outliers for all structures (i.e. at all resolutions) is 4% (σ 5%).

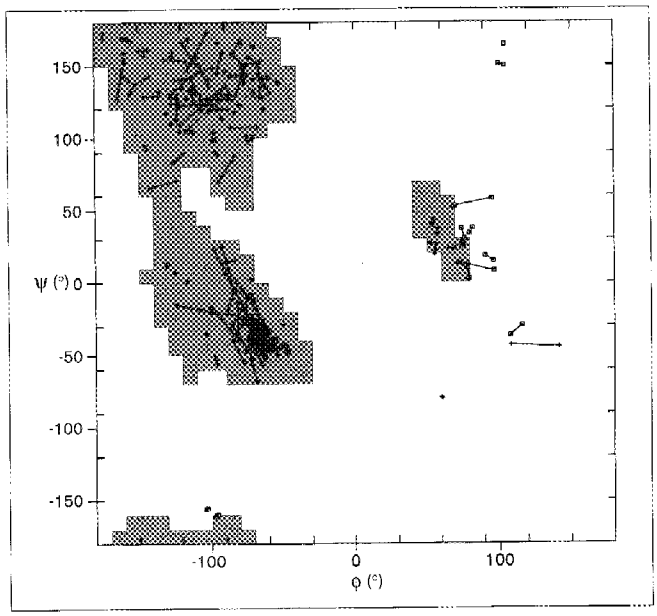
We have implemented this new definition of core regions in all our programs that use or produce Ramachandran plots, including O [5], OOPS [17], LSQMAN [14] and MOLEMAN2 (GJK, unpublished program). A list of outlier percentages of more than 3000 proteins from the February 1996 release of the PDB is available on the World Wide Web (URL: http://alpha2.bmc.uu.se/~gerard/rama/rama.html). This site also contains the 37 by 37 matrix of residue counts, as well as a Fortran subroutine that implements our core region definition.

Figure 5



An unusual Ramachandran plot for one enzyme in the PDB. Note that there are no glycine residues apparent in this plot because the sequence of the enzyme had not been assigned and all residues are listed as 'UNK' (unknown) in the PDB entry. Residues are labeled as in Figure 3.

Figure 6



Multiple-model Ramachandran plot [14] of two NCS-related molecules of ligand-free ribose-binding protein (SL Mowbray, personal communication). In this type of plot, the symbols for corresponding residues in NCS-related molecules are connected.

Acknowledgements

We would like to thank Dr C Divne for providing us with her CBH I models, and Dr SL Mowbray for allowing access to the ribose-binding protein model prior to publication. This work was supported by the Swedish Natural Science Research Council, Uppsala University and the European Union (grant number BIO4-CT96-0189 to TAJ).

References

1. Kleywegt, G.J. & Jones, T.A. (1995). Where freedom is given, liberties are taken. *Structure* **3**, 535–540.
2. Brändén, C.I. & Jones, T.A. (1990). Between objectivity and subjectivity. *Nature* **343**, 687–689.
3. Brünger, A.T. (1992). Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–475.
4. Kleywegt, G.J. & Brünger, A.T. (1996). Checking your imagination: applications of the free R value. *Structure* **4**, 897–904.
5. Jones, T.A., Zou, J.Y., Cowan, S.W. & Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Cryst. A* **47**, 110–119.
6. Ramakrishnan, C. & Ramachandran, G.N. (1965). Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units. *Biophys. J.* **5**, 909–933.
7. Jones, T.A. & Thirup, S. (1986). Using known substructures in protein model building and crystallography. *EMBO J.* **5**, 819–822.
8. Laskowski, R.A., MacArthur, M.W., Moss, D.S. & Thornton, J.M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26**, 283–291.
9. Karplus, P.A. (1996). Experimentally observed conformation-dependent geometry and hidden strain in proteins. *Protein Sci.* **5**, 1406–1420.
10. Langs, D.A. (1988). Three-dimensional structure at 0.86 Å of the uncomplexed form of the transmembrane ion channel peptide gramicidin A. *Science* **241**, 188–191.
11. Brünger, A.T. (1992). *X-PLOR, Version 3.1. A System for X-ray Crystallography and NMR.* Yale University Press, New Haven, CT, USA.
12. Sauer-Eriksson, A.E., Kleywegt, G.J., Uhlén, M. & Jones, T.A. (1995). Crystal structure of the C2 fragment of streptococcal protein G in complex with the Fc domain of human IgG. *Structure* **3**, 265–278.
13. Divne, C., *et al.*, & Jones, T.A. (1994). The three-dimensional crystal structure of the catalytic core of cellobiohydrolase I from *Trichoderma reesei*. *Science* **265**, 524–528.
14. Kleywegt, G.J. (1996). Use of non-crystallographic symmetry in protein structure refinement. *Acta Cryst. D* **52**, 842–857.
15. Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522–524.
16. Bernstein, F.C., *et al.*, & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.
17. Kleywegt, G.J. & Jones, T.A. (1996). Efficient rebuilding of protein structures. *Acta Cryst. D* **52**, 829–832.