Structure Validation by Cα Geometry: φ, ψ and Cβ Deviation


1Department of Biochemistry, Duke University, Durham, North Carolina
2Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom
3GlaxoSmithKline, Research Triangle Park, North Carolina

ABSTRACT Geometrical validation around the Cα is described, with a new Cβ measure and updated Ramachandran plot. Deviation of the observed Cβ atom from ideal position provides a single measure encapsulating the major structure-validation information contained in bond angle distortions. Cβ deviation is sensitive to incompatibilities between sidechain and backbone caused by misfit conformations or inappropriate refinement restraints. A new φ,ψ plot using density-dependent smoothing for 81,234 non-Gly, non-Pro, and non-prePro residues with $B < 30$ from 500 high-resolution proteins shows sharp boundaries at critical edges and clear delineation between large empty areas and regions that are allowed but disfavored. One such region is the γ-turn conformation near $+75°$, $-60°$, counted as forbidden by common structure-validation programs; however, it occurs in well-ordered parts of good structures, it is overrepresented near functional sites, and strain is partly compensated by the γ-turn H-bond. Favorable and allowed φ,ψ regions are also defined for Pro, pre-Pro, and Gly (important because Gly φ,ψ angles are more permissive but less accurately determined). Details of these accurate empirical distributions are poorly predicted by previous theoretical calculations, including a region left of α-helix, which rates as favorable in energy yet rarely occurs. A proposed factor explaining this discrepancy is that crowding of the two-peptide NHs permits donating only a single H-bond. New calculations by Hu et al. [Proteins 2002 (this issue)] for Ala and Gly dipeptides, using mixed quantum mechanics and molecular mechanics, fit our nonrepetitive data in excellent detail. To run our geometrical evaluations on a user-uploaded file, see MolPROBITY (http://kinemage.biochem.duke.edu) or RAMPAge (http://www-cryst.bioc.cam.ac.uk/rampage). Proteins 2003;50:437–450.

INTRODUCTION

The Cα is the most important locus for evaluating distortion of covalent geometry in protein structures, because it joins sidechain with backbone and responds to both and especially to their compatibility. If either side of that junction (the backbone φ,ψ or the sidechain rotamer) happens to be misfit into the wrong local minimum, the process of model refinement is forced to compromise by distorting Cα geometry. Therefore, the three major components of geometrical structure validation are backbone conformation (primarily φ,ψ angles), sidechain conformation (primarily sidechain rotamers), and Cα geometry, which signals backbone-sidechain compatibility.

Because bond lengths are very tightly restrained, geometrical distortion around the Cα ends up primarily in the bond angles; the τ (N-Cα-C) angle is primarily affected just by backbone, whereas the N-Cα-Cβ and C-Cα-Cβ angles are affected by sidechain-backbone compatibility. Ideal bond angle values are known from highly accurate small-molecule structures, and traditional structure validation reports flag outliers that deviate by more than a few standard deviations. Those lists can be very revealing, but the sheer number of entries discourages their routine use. At the Cα, it is also hard to evaluate the significance of deviations that often are split between two of the tetrahedral angles. A single, summary criterion would be highly desirable, and this article proposes the Cβ deviation for that role.

Torsion angles for the backbone and sidechain are the other main geometrical components around the Cα, both reporting their respective conformations and interacting with each other and with the Cα bond angles. When all-atom contact analysis showed severe atomic clashes for some members of all previously published sidechain rotamer libraries, we used B-factor filtering and careful analysis of other quality factors to compile the “penultimate” sidechain rotamer library. Those rotamer distributions were satisfyingly sharper and narrower than found before, highlighting the surprising degree to which sidechain conformations are relaxed and favorable even in protein interiors. That study also identified several types of systematic fitting errors that occur fairly often for Leu, Met, and...
branched-Cβ sidechains; those errors are of interest here, because they are most reliably diagnosed by considering Cβ deviations as well as all-atom clashes and nonrotameric torsion angles. The sidechain rotamers are not updated further here, because the database has not grown by a large enough factor to change the conclusions. The MOLPROBITY web server described here uses the rotamers from Lovell et al.5 to accompany the new Cβ and Cγ criteria described here.

The Ramachandran diagram,6 which plots Cα versus Cβ backbone conformational angles for each residue in a protein, has been with us nearly as long as macromolecular crystal structures. This same coordinate system is used to show either empirical scatter plots of the conformations observed in the database of known 3D structures, or else contours of calculated energies or steric criteria as a function of Cα and Cβ for a dipeptide (as in Fig. 1). Especially in recent years, Cφ, Cψ plots for individual proteins have also become central for structure validation, because Cφ, Cψ values are not optimized in the refinement process and, therefore, provide a sensitive indicator of local problem areas.7 To the approximation of ideal covalent geometry and trans peptides, the Cφ, Cψ plot encapsulates all information about backbone conformation in a remarkably concise and intuitive form; therefore, it has been a key abstraction central to our growing understanding of protein structure, energetics, and folding.

Our knowledge of the overall empirical Cφ, Cψ distribution has gradually improved in accuracy, both because a much larger number of structures are available and to an even greater extent because many of the recent structures are at extremely high resolution. The approximate location and shape of the most favorable, low-energy regions (consisting of α-310, β-polyp), and Lα, as labeled on Figure 1(a), plus some areas bridging α and β) became clear very early8,9; their limits are produced by collision among main-chain and Cβ atoms and are, therefore, largely sidechain-independent except for Gly and Pro. Those particularly favorable regions have now converged to agreement in all treatments, except for a slight remaining tendency of the boundaries to shrink inward as accuracy continues increasing.10

These Cφ, Cψ criteria have become accepted as a central aspect of protein structure validation. They are routinely applied during deposition of structures to the Protein Data Bank11 and have been made conveniently available on web sites such as the Biotech Validation Suite (http://biotech.embl-ebi.ac.uk:8400). Figure 1(b) shows the Cφ, Cψ regions scored as “strictly allowed” for WhatCheck’s single definition12 and the three “core,” “allowed,” and “generously allowed” regions for ProCheck.3,7 Inclusion of low-resolution, and especially of high-B, data in ProCheck gave noise throughout the plot, producing unrealistic outlines for the two outer regions. In reaction, WhatCheck chooses not to try distinguishing possible from forbidden conformations outside their 98% contour. However, several analyses 13–15 have specifically studied a number of individual low-B, high-resolution residues with Cφ, Cψ in the outer regions [especially in the γ-turn, II’ turn, and below-α plateau regions labeled in Fig. 1(a)]. They concluded the...
The γ-turn conformation (near \( \phi, \psi = 75^\circ, -60^\circ \), with CO(i-1) to NH(i+1) H-bond) was first described by Huggins in 1943,\(^{16} \) although he envisioned a repeating series of residues with these torsion angles, producing a shallow helix with two residues per turn. Such a series has not been observed, although the mirror-image γ' conformation (near \( -75^\circ, 60^\circ \)) can sometimes repeat as a highly pleated β-strand; the γ' region forms a small extension below the β region [Fig. 1(a)]. Némethy and Printz\(^{17} \) suggested, on the basis of modeling studies, that the γ-turn conformation may exist as a three-residue chain reversal, and it was first described in a protein by Matthews\(^{18} \) in thermolysin. Rose et al.\(^{19} \) and Milner-White\(^{20} \) published reviews of the occurrence of γ and γ' turns.

The γ-turn and the γ'-turn are also known as C\( ^\gamma \) and C\( ^\gamma' \), respectively, because the H-bond completes a seven-membered ring and the β-carbon is either axial or equatorial to the ring. Energy calculations for a dipeptide in vacuo\(^{21–24} \) find these two conformations as the overwhelmingly global energy minima in Ramachandran space, because the lack of water drives formation of the backbone H-bond. For proteins in aqueous solution, the γ' conformation is favorable but not optimal, and the γ-turn is accessible but rare because of a minor steric overlap of Cβ and the carbonyl O.

In type II' tight turns, the second of four residues adopts \( \phi, \psi \) angles near 50°, −125°. Gly is the most common residue in this position, although other residues are observed.\(^{25} \) As also seen in the γ-turn conformation, there is some steric overlap between the Cβ and carbonyl oxygen. The serine of the catalytic triad of all α/β hydrolases and lipases has this conformation.\(^{26} \)

The analysis and proper treatment of these “outlier” conformations is still controversial, because it is very difficult to distinguish rare but genuine features of the molecules from errors in the models. Strained conformations should be expected, although the expectation is that they should be observed only rarely. Genuinely strained conformations are often useful indicators of biological significance, and certainly no attempt should be made to “fix” them. However, conformational outliers are to be treated with suspicion, because they may merely reflect deficiencies of the structure determination. This vital distinction between strain and error can, we propose, be approached statistically by behavior as a function of data quality and in individual cases both by the combination of atomic clashes and geometrical distortion and also by determining whether a more “normal” conformation could explain the experimental data equally well.

The pattern of relative frequencies within the \( \phi, \psi \) distribution varies significantly among the 18 non-Gly, non-Pro amino acids,\(^{27–29} \) but the outlines of those regions are all nearly the same. In contrast, glycine and proline each have very different \( \phi, \psi \) distribution outlines than the other amino acids, with conformational constraints either significantly less (Gly, with no Cβ) or significantly more (Pro, with its pyrrolidine ring). However, their outliers have seldom been explicitly flagged and are not included in overall summary validation measures,\(^{3,4} \) because so much less data are available than for the general case. This is particularly unfortunate for glycine, because its lack of an observable Cβ atom makes its experimental \( \phi \) and \( \psi \) values especially error-prone.\(^{30} \) Pre-Pro residues (those that precede prolines) also have a very distinctive \( \phi, \psi \) distribution,\(^{28} \) and we treat pre-Pro as a separate fourth case here.

The current study revisits the Ramachandran plot for Gly, Pro, pre-Pro, the general case of the 18 other amino acids, and the residues in nonrepetitive structure, using both new data and new techniques to resolve the above problems. From a database of 500 structures selected by resolution, homology, and other criteria of quality, the residues with high crystallographic B-factors for the backbone are omitted. Hydrogens are added and optimized,\(^{31} \) and all-atom contact analysis\(^{32} \) is used to judge the reliability both of overall structures and of local regions. The validity of sparsely populated regions is determined by analyzing their occurrence frequency versus resolution or B-factor and by examining a sample of cases for degree of bond-angle distortion, ambiguity of the electron density, and the presence of compensating interactions. The result is a set of remarkably well-defined empirical \( \phi, \psi \) distributions that cleanly distinguish the truly disallowed regions from the dis favored but allowed regions.

Overall, this study presents a set of simple, accessible, and definitive tools for evaluating protein backbone conformation, sidechain rotamers, and distortions of Cu-Cβ geometry. Collectively, they complement all-atom contact analysis, and they are effective for structure validation purposes, for comparison with theoretical calculations, and for better understanding of the factors that control protein conformation.

**MATERIALS AND METHODS**

Selection of the data set of protein structures for this study started from our previous database of 240 structures at 1.7 Å resolution or better,\(^{5} \) augmented by the 30% homology cutoff PDB Select list (Hobohm and Sander\(^{33} \)) from February 2000 and new structures of 1.5 Å resolution or better released from February to May 2000. Structures were rejected from the PDB Select list if they were solved to a resolution of lower than 1.8 Å. The three sources were then reconciled by using all structures found both in the PDB Select list and in our previous database and, for closely related pairs of structures, by using the one with the best combination of clashscore (number of van der Waals overlaps \( \geq 0.4 \) Å per 1000 atoms\(^{35} \)) and resolution. If multiple identical chains were present, the first chain was chosen, unless the file header indicated that another was better ordered. The newer structures of high resolution (\( \geq 1.5 \) Å) were added if not already in the database and replaced structures if they were solved to higher resolution and had better clash score. Files with multiple, nonhomologous chains were split only if each formed a separate compact unit.
A number of filters were applied, in addition to those of resolution described above. Specifically, structures were rejected if they had a clash score $> 22$ for those atoms with crystallographic $B$-factors less than 30. A nonnormalized cutoff on $B$-factor was used, for reasons previously discussed. The final $\phi, \psi$ data set contained 97,368 total residues. Dihedral angles were calculated with DANG, and secondary structure assignments were made by using the DSSP algorithm as modified in ProCheck. Methods for the sidechain rotamer analysis are described in Lovell et al. and included use of a $B$-factor cutoff of 40 for sidechain atoms. $C_\beta$ deviations and directions were calculated by Prekin for all 500 structures. Many methods of calculating an ideal $C_\beta$ position are equivalent if the backbone geometry is ideal, but the answer is algorithm-dependent if $\tau$ (the $N$-Ca-C angle) is distorted. Our method compromises the difference by calculating $C_\beta$ twice, from both the $C$-N-Ca-$C_\beta$ and the $N$-C-Ca-$C_\beta$ angles and dihedrals, averaging the two, and then idealizing the $C_\alpha$-$C_\beta$ bond length. (Note that the ideal values differ slightly for Ala, branched $C_\beta$, and other amino acids.) This is the same algorithm that Prekin uses to produce an ideal $C_\beta$ for residue mutation in the interactive MAGE/PROBE system. The $C_\beta$ deviation is defined as the $C_\beta$(obs)-$C_\beta$(ideal) distance, and the direction is defined as the $N$-Ca-$C_\beta$(ideal)-$C_\beta$(obs) torsion angle. Two graphical representations of the $C_\beta$ deviation are produced in kinemage format: either the magnitude is shown as the radius of a ball centered at each ideal $C_\beta$ position in the protein structure, or else all $C_\beta$ deviations and directions for a structure are plotted in polar coordinates around the $C_\beta$ of a single ideal geometry residue.

All relevant data items are stored in and queried from MySQL database tables at the molecule level (e.g., PDB code, resolution) or the residue level ($\phi, \psi, \tau$, backbone or sidechain $B$-factor, $C_\beta$ deviation). Two- and three-dimensional visualizations of protein structures or of data plots are done interactively in the MAGE graphics program using kinemage files copyrighted by the authors, with PostScript output for figures edited in Adobe Illustrator. Figures including electron density were rendered in PyMOL.

Smoothly contoured boundaries for the $\phi, \psi$ distributions are obtained by using a density-dependent smoothing function. This function treats areas of sparse data as continuous regions of low, relatively constant density, while preserving the sharp transitions in regions where high density falls off rapidly. The smoothed distributions are used both for contouring the database distributions and (on the MOLPROBE web site) for determining the $\phi, \psi$ quality values for each residue of a user-submitted structure.

The smoothed, normalized density of points, $\rho$, is expressed in general as a sum of the contributions of all $N$ data points:

$$\rho(\phi, \psi) = \sum_{i=1}^{N} \sigma_i(\phi, \psi)$$

Each contribution $\sigma_i$ is computed as a normalized cosine mask that depends on both $\phi$ and $\psi$, with a specified radius, $\alpha_i$, according to:

$$\sigma_i(\phi, \psi) = \frac{\pi}{\alpha_i^2(\pi^2 - 4)} \left( \cos\left(\frac{\pi x_i}{\alpha_i}\right) + 1 \right), \quad x_i < \alpha_i$$

$$\sigma_i(\phi, \psi) = 0, \quad \text{otherwise}$$

where $x_i = \sqrt{(\phi - \phi_i)^2 + (\psi - \psi_i)^2}$

$\phi_i$ and $\psi_i$ are the values for the $i$-th data point, and $\alpha_i = k \rho_i(\phi, \psi)$ is the coefficient to normalize the mask volume to 1.0. Note the distinction between the many $\sigma$ values, each of which simply spreads out the density of one data point, and the single $\rho$, which is the overall density function describing the distribution of populated regions on the Ramachandran plot.

The final density function $\rho_2$ is calculated in two iterations. In the first iteration, a density $\rho_1$ is calculated by using the above equations, with identical $\alpha_i = \alpha_0$ for all $i$. The second iteration uses the same equations, but now the radius of each cosine mask varies according to:

$$\alpha_i = k \rho_i(\phi, \psi)$$

Because the objective is to smooth between the data points, we calculate the mask width from function (5), which approximates the average distance between points in $n$ dimensions, modified by a constant factor $k$ and an exponential factor $\lambda$. For this work, the number of dimensions in the data set, $n$, was always 2, and $\lambda$ was fixed at 0.5, which produced contours that stayed suitably close to the steep transitions. The value of $k$ was varied slightly,
depending on the data density in the sparse regions. As implemented, $r(\phi, \psi)$ is approximated by summing all the $\sigma_i$ on an evenly spaced ($2^\circ$) grid of sample points to give $r(j,k)$. Any desired density $r(\phi, \psi)$ is found by linear interpolation from the four nearest $r(j,k)$. The end result is a good approximation of $r(\phi, \psi)$.

Contour levels are specified here as percentages. For example, a contour at 85% means that the data have been contoured at the largest density value not greater than $p_2(\phi_i, \psi_i)$ for 85% of data points $i$ that is, the contour encloses 85% of the data points and excludes 15%. Density values for contouring are determined after sorting all data points $i$ by the value of $p_2(\phi_i, \psi_i)$; for example, the 85% value is the value of $p_2$ for the 0.85$\%$-th point in the sorted list. Contours are calculated with the KIN2DCONT program.\textsuperscript{34}

For smoothing the general distribution, we used $\alpha_0 = 10^\circ$ and $k = 13$. The final density ($p_2$) was then contoured at 99.95% (allowed) and 98% (favored) levels. The other distributions contained only 5–10% as much data, and so required a larger value of $k$ for optimal smoothing. For the glycine, proline, and proline proline distributions, we used $\alpha_0 = 10^\circ$ and $k = 16$. The final density ($p_2$) was then contoured at 99.8% (allowed) and 98% (favored) levels. For the nonsecondary structure distribution in Figure 7, we again used $\alpha_0 = 10^\circ$ and $k = 13$; $p_2$ was contoured at 99.9% (allowed) and 95% (favored) to match the general case contours closely, given the much smaller percentage of data in the helical region.

 Ninety-eight-percent contours outlining the “favored” regions were defined for all four of the inclusive cases shown in Figure 4, allowing a summary statistic to be calculated for all residues. For the general residue case, the 99.95% contour dividing the “allowed” from “outlier” regions is well behaved, but a 99.8% level had to be used for the single-residue distributions to avoid artifacts from small numbers. Residues in an individual structure are first assigned to Gly, Pro, pre-Pro, or general case and are then evaluated as favored, allowed, or outlier by comparing the interpolated density value for their $\phi, \psi$ to the relevant contour values. This procedure makes the use of smooth boundaries no harder than assignment by rectangular bins.

### RESULTS

#### C$\beta$ Deviation

As explained in the Introduction, distortion around the C$\alpha$ is an especially sensitive way of locating potentially serious problems in the model for a protein structure. However, the geometry of that tetrahedral center is usually described by three bond lengths, three bond angles, and the $\tau$, or $\psi$ handedness. Even if the handedness is evaluated as a torsion angle or a chiral volume and each measure is expressed in SDs from its ideal value, as in the extensive and useful lists provided by ProCheck,\textsuperscript{3} it is still not obvious how to combine these numerous, incommensurate, and not entirely independent measures into an overall evaluation.

As a new approach to simplifying this problem, we calculate an ideal-geometry C$\beta$ position from the backbone atoms, defining the “C$\beta$ deviation” as the distance of the observed C$\beta$ from the ideal one and using C$\beta$ deviation as a single, simple measure of geometrical nonideality around the C$\alpha$. The ideal C$\beta$ construction process (see Materials and Methods) is designed to respond conservatively to backbone distortions (e.g., splitting the difference between the N-C$\alpha$-C$\beta$ and C$\alpha$-C$\beta$ angles if the $\tau$ angle is non-ideal). For some purposes, it is also useful to know the direction of the C$\beta$ deviation relative to the local backbone, which we measure as the torsion angle N-C$\alpha$-C$\beta$(ideal)-C$\beta$(obs).

As a graphical representation for display on the 3D protein structure, we use a ball with radius equal to the magnitude of the C$\beta$ deviation, centered at the ideal C$\beta$ position (and thus tangent to the observed C$\beta$), as shown in Fig. 2(c). This particular case is one where the $\beta$-branched sidechain was fit backward [see Fig. 2(d)] into ambiguous electron density, with refinement then distorting bond angles and shifting the C$\beta$ position by nearly half an Å in the attempt to optimize overall fit. On the display for an entire protein, the large C$\beta$ deviations can easily be spotted, as in the example of Figure 2(b). For larger structures, it is useful to turn off the small C$\beta$ deviations and then zoom in on a big one for examination with the full model turned back on. Either on the MolPROBITY web service (http://kinemage.biochem.duke.edu) or running our PREKIN program by itself, the user can obtain a numerical listing of the C$\beta$ deviation and direction values, a kinemage display of C$\beta$ deviation balls on the 3D structure [Figs. 2(b) and (c)], and/or a scatterplot of C$\beta$ deviation and direction relative to an ideal geometry residue [Fig. 2(a)]. In practice, a C$\beta$ deviation of 0.25 Å is the approximate threshold of significant distortion: a value $< 0.2$ is not diagnostic, whereas a value $> 0.3$ indicates some sort of problem, either generic or local. Badly misfit sidechains often have C$\beta$ deviations above that threshold, especially for $\beta$-branched residues.

The C$\beta$ deviation versus direction plot for the full data set (excluding Pro) is shown in Figure 3(a). It has a significant asymmetry, corresponding to more and larger deviations in the directions perpendicular to the local backbone direction (vertical in the figure). Several factors, at least, contribute to this asymmetry. First, in the perpendicular directions C$\beta$ distortion is split between the N-C$\alpha$-C$\beta$ and C$\alpha$-C$\beta$ angles so that neither is highly deviant, therefore being more acceptable to the refinement program. Second, those are the directions in which valid C$\beta$ motions can be produced by small anticorrelated changes of $\phi$ and $\psi$; if C$\beta$ is placed well but with backbone error that includes $\phi$ and $\psi$, the resulting deviation will be perpendicular to the backbone. Our data suggest that both those mechanisms are probably involved, because the distribution shapes correlate strongly with refinement program [Fig. 3(b)], but all show elongation in the perpendicular direction.

There is also some tendency for the C$\beta$ deviation-direction plots to show a threefold elongation along the directions of the three backbone bonds, visible especially for CNS/XPlor in Figure 3(b). One factor contributing to
that shape is explained by Figure 3(c), which shows the highly threefold distribution for the branched Cβ residues with eclipsed χ₁ angles. These sidechains are all presumably misfit, like the example in Figure 2, and the resulting distortions are predictably large and in the observed three directions. A further asymmetry visible in Figure 3(a) and (b) is that deviations are somewhat larger along the CαH direction (upward) than away from it (downward). This asymmetry is at least partly due to the residues with αφ values between 0° and +100°, whose Cβ deviations are plotted in Figure 3(d); they have a mild clash between Cβ and O(n-1), which produces a genuine shift of the Cβ position upward.

Many issues go into evaluating the suitability of this new validation measure. Incorrect amino-acid handedness is rare except in the case of free-atom refinements (there are no cases in our database), but detecting it when present is important. Cβ deviation detects either reversal or distortion of Cα handedness as sensitively as chiral volume and more reliably than an improper-dihedral criterion. Cβ deviation is quite insensitive to bond length distortions, but that is an advantage: bond lengths are so tightly constrained that their variation responds primarily to refinement parameters and not to local strain in the structure. Bond angle deviations, on the other hand, are in practice the place where refinement allocates most of the strain when data and model requirements cannot be reconciled. If a sidechain is fitted backward, then usually the Cβ position ends up moved out of place to fit the sidechain bulk approximately into the electron density. This results in distorted bond angles around the Cα. If the distortion is all in one angle (either N-Cα-Cβ or C-Cα-Cβ), then the problem is easy to interpret from traditional criteria. If the distortion is split between the two angles (as

Fig. 2. Cβ deviation as displayed on the 3D protein structure: a: Radial plot of magnitude and direction for all Cβ deviations in 1bi5, showing a tight, round central distribution under 0.1Å, plus a few outliers up to 0.4Å. b: Cβ deviation balls for the entire 1bi5 structure, showing that the few problem residues stand out very clearly from a background with near-ideal geometry. c: Closeup of large Cβ deviation (radius of gold ball, centered on ideal Cβ position) for a Val sidechain fit backward with eclipsed χ₁ (PDB file 1bi5, Val342). d: All-atom contacts for Val342, including a serious clash (red spikes); an ideal rotamer (not shown) fits with good contacts.
Fig. 3 shows is very often the case), then neither change may look significant; however, the Cβ deviation from ideal position gives a clear and equivalent measure for either case.

**φ,ψ for the General Case**

As explained in Materials and Methods, we have developed a database of nearly 100,000 residues from 500 structures at \( \leq 1.8 \) Å resolution to determine which regions of Ramachandran space are populated for the best data: at very high resolution and low crystallographic B-factors. Plotting the \( \phi \) and \( \psi \) of these data points for the general case of non-Gly, non-Pro, and non-prePro [Fig. 4(a)], we find the usual primary peaks in \( \alpha \), \( \beta \), and \( \omega \) conformations, which have been known since Ramachandran.\(^6\) However, there are also significant, reproducible observations elsewhere on the plot, which not only persist but even increase slightly in percentage as B-factors decrease, all the way down to \( B < 10 \). The pattern does not change across the low-\( B \) data, but adding residues with backbone \( B \geq 30 \) leads to a dramatic increase in the amount of scatter, sparsely populating the entire \( \phi,\psi \) plot including areas near \( \phi = 0^\circ \), which force large steric overlaps of main-chain atoms and are clearly not physically possible. The most important difference from earlier empirical studies of \( \phi,\psi \) space, then, is the greatly improved signal-to-noise ratio [e.g., compare with Fig. 5(a) of Morris et al.\(^7\)], which allows a clear distinction between truly disallowed “outlier” conformations and those that are rare but allowed.

Previous definitions of \( \phi,\psi \) regions [Fig. 1(b)] were done by counting data points within 10° angle bins. Although a bit coarse-grained, that system works quite well at high data density such as the ProCheck “core” or the What-Check “strictly allowed” regions. However, at lower data density, that method is unduly sensitive to statistical fluctuations, giving jagged edges that would presumably alter with the use of more or different data. To produce more robust and even edges, we choose to smooth and then contour the distribution. A second problem is produced by the contrast between very shallow and very steep edges, especially the extreme example of the diagonal edge to the right of \( \alpha \)-helix, which goes from zero density to the global maximum in just over 20°. Data binning, uniform smoothing, or inclusion of low-accuracy data all tend to make this
edge appear to spread outward into truly forbidden areas with large steric clashes. Therefore, we have used density-dependent smoothing (see Materials and Methods) to help the contours suitably hug the sharp edges while still smoothing the sparse, gradual edges.

The final choice is the level at which contours are drawn, defined by the percentage of the high-quality data they enclose. Our “favored” region includes 98% of the data and agrees almost exactly with the “strictly allowed” region of Kleywegt and Jones\textsuperscript{12} as reproduced in Figure 1(b), except for the absence of the left “leg” extending down from the β-region; the correspondence would be complete if we had included pre-Pro residues in the general case distribution. However, we agree with the ProCheck authors\textsuperscript{3,7} that in addition it is important to define an outer region that encompasses nearly all high-quality data, and we also believe that aim can finally now be successfully achieved for the general case; therefore, we have defined an “allowed” region that includes 99.95% of the data. The two contours enclosing favored and allowed regions are shown in Figure 4(a), along with the 81,234 data points of the general case distribution.

The favored region comprises 17% of the area of the plot; both favored and allowed regions together cover only 41.5%. However, our allowed region is significantly different in shape from either the “allowed” or “generously allowed” regions of Morris et al.\textsuperscript{7} Our “outlier” region is both much larger and contains less data than their “outside” region. We agree that the plateau region below α should be considered allowed. We differ, however, on the

Fig. 4. \(\phi,\psi\) angle distributions for 97,368 residues with backbone B-factor < 30 from the 500-structure high-resolution database, along with validation contours for favored and allowed regions. a: The general case of 81,234 non-Gly, non-Pro, non-prePro residues. b: The 7705 Gly residues, shown with twofold symmetrized contours. c: The 4415 Pro residues with contours. d: The 4014 pre-Pro residues (excluding those that are Gly or Pro) with contours.
possible cases: there are a few data points above present data, however, are not adequate for settling all random error spreads data points everywhere. Even the 4(a), which contains the controversial frequency up to points (x) outside the allowed region, which occur at a near-zero Ramachandran space with positive /H9278 acceptability of the sinuous, sparsely populated stripe of noise in the highest range of /H9253 Gly and Pro residues, as a function of /H9251 B-factors.

Fig. 5. Occurrence of \( \phi, \psi \) values in certain regions (for general case residues), as a function of B-factor range. The lowest line shows outlier points (x) outside the allowed region, which occur at a near-zero frequency up to \( B = 30 \) but are quite prevalent at high \( B \). In contrast, the solid straight line shows the inverse slope fit to the allowed but disfavored \( \gamma \)-turn (\( \gamma \)), II' (\( \gamma \)), and plateau (\( \gamma \)) regions, until they increase as well from noise in the highest range of B-factors.

The empirical distribution of \( \phi, \psi \) for Gly is approximatley twofold-symmetric around the central 0°,0° point, because the lack of a C\( \beta \) produces a mirror symmetry in the steric constraints for Gly. However, the data obey that symmetry only inexactly: the right-handed \( \alpha \)-helix produces a small, intense peak around -60°, -40°, but the \( \alpha \) to \( \beta \) region is even better populated because it is a useful conformation accessible without any strain only for Gly. The steric constraints defining the outer limits of accessible \( \phi, \psi \) regions should be symmetrical for Gly, and so we calculate the 98% and the 99.8% contours for Gly from a twofold symmetrized version of the Gly \( \phi, \psi \) data. The data points plotted in Figure 4(b) are unsymmetrized, to show that they fit well within the outlines defined by the symmetrized contours. Near \( \phi = 180° \) there is not enough data to define the outer contour robustly, but toward \( \phi = 0° \) the steric clashes are strong and the boundaries are clear. Therefore, although Gly conformation is more permisive than the general case, the outlier region covers 37% of the \( \phi, \psi \) plot.

To the first approximation, Pro is very simple, because the ring closure restricts \( \phi \) near -70°. However, as seen in Figure 4(C), there is a rich detail in the distribution, with a \( \phi \) width from about -50° to -100° and three distinct peaks in \( \psi \). The two major peaks correspond to \( \alpha \) and polyproline II conformations, whereas the small central peak is in the \( \gamma \) region. \( \gamma \) is a slightly strained conformation stabilized by an NH(i-1) to CO(i+1) hydrogen bond. Figure 4(c) includes prolines preceded by either trans or cis peptide bonds; the cis examples have relatively more negative \( \phi \) values and do not occur in the \( \gamma \) peak, because they cannot form the \( \gamma \) hydrogen bond.

Pre-Pro \( \phi, \psi \)

As mentioned in the description of the general case, residues that precede proline are treated separately here because they have a distinctively different \( \phi, \psi \) distribution, shown in Figure 4(d). As originally pointed out by Karplus,\textsuperscript{28} pre-Pro residues preferentially populate a region near -130°, 80° [marked "pre-Pro" on Fig. 1(a)] below the left side of the broad \( \beta \)-region. The pre-Pro distribution in Figure 4(d), indeed, shows more than twice as many data points in that region than the general distribution of Figure 4(a), despite the fact that there are only 5% as many total pre-Pro residues; their preference for that region is thus 40-fold higher.

Gly and Pro \( \phi, \psi \)

Glycine and proline are significantly different from the other amino acids in their backbone stereochemistry. The
The only other regions populated by pre-Pro residues are \( \beta \), poly-Pro, and two small, round areas at the very tip of \( \alpha \) and of \( \alpha' \). The constraints on pre-Pro conformation are produced by the Pro C\( \delta \), which not only prevents H-bonding of the Pro NH but clashes with other backbone or C\( \beta \) atoms in many conformations. \( \gamma \) and \( \gamma' \) conformations are impossible for pre-Pro residues because the clash of CO(pre-Pro-1) with C\( \delta \)(Pro) takes a large bite out of the pre-Pro \( \phi, \psi \) distribution, producing the convex curve that forms the lower right edge of the \( \beta \)-region for pre-Pro. As is often true, occurrence is enhanced just inside that sharp boundary, presumably because there is then a favorable van der Waals contact for the same atom pair that clashes just outside the boundary.

By looking at the shape of the pre-Pro distribution in Figure 4(d), it is hard to deny its significant difference from any of the patterns in Figure 4(a)–(c); therefore, we treat pre-Pro \( \phi, \psi \) separately as a fourth case for validation purposes. Examination of \( \phi, \psi \) distributions for the other amino acids, either individually or in related groups, shows substantial differences in peak heights of various regions, especially for \( \alpha; \) however, the contours enclosing favored and allowed regions do not differ enough to justify their separate treatment at the current database size.

**Web Servers**

Starting from a user-uploaded coordinate file or from a PDB file selected by ID code, the Ramachandran plot evaluations described above can be run on the MolPROBITY server at http://kinemage.biochem.duke.edu or the RAMPAGE server at http://www-cryst.bioc.cam.ac.uk/rampage. MolPROBITY also provides C\( \beta \) deviations, sidechain rotamer evaluations, hydrogen addition, and all-atom contacts. Results are displayed as 3D kinemage graphics in JAVA\( \text{M} \)\( \text{AGE} \), for viewing directly online, plus tables or lists as appropriate. The output graphics and coordinate files can also be downloaded. The software used in this work (PREKIN, MAGE, REDUCE, PROBE, DANG, etc.) is free and open-source, written and copyrighted by various of the authors, and available at the kinemage web site along with the \( \phi, \psi \) distributions and the 500 database files with hydrogens added.

**DISCUSSION**

The process known as structure validation serves several distinct purposes. The most traditional is for the benefit of journal and grant reviewers, lab directors, database entry, and so forth to certify whether a structure meets generally accepted current standards of good practice in the field. In macromolecular crystallography there is a reasonable consensus, for a given resolution range, about respectable values of residual and free \( R \).\( ^{39, 40} \) (see also real-space fit of model to density at Kleywegt and Jones 2002 url, http://portray.bmc.uu.se/eds), and for ideality of bond lengths, bond angles, and \( \phi, \psi \) values.\( ^{3, 12} \) Such standards are extremely important, and we hope that the criteria developed here will become accepted by the community as additions to, or improvements on, current standards.

However, the present work is primarily addressed to strengthen two other important aspects of structure validation. The first is to provide the end users of 3D data with convenient but critical assessments of probable accuracy that apply to local regions as well as to overall structures. The second aspect is to provide crystallographers themselves with a suite of tools for locating and fixing local problems during the process of fitting and refinement. The major centerpiece of this strategy is all-atom contact analysis,\( ^{11, 32, 37} \) which has the advantage of using information (the hydrogen contacts) that is independent both of the usual refinement targets and of traditional validation criteria. That contact information, however, is most powerful if used in conjunction with suitable measures of geometric ideality, because choice of refinement strategy can to some extent trade off nonideality between those two types of criteria. Therefore, there is a need for geometrical validation tools that are updated with large and quality-filtered data sets, which are tuned to complement all-atom contact analysis and which are collectively optimized for sensitivity to backbone or sidechain conformations trapped in the wrong local minimum.

The present study has treated Ramachandran plot criteria for backbone conformation with special emphasis on distinguishing rare from erroneous \( \phi, \psi \) values, touched briefly on backbone rotamer criteria,\( ^{9} \) and presented C\( \beta \) deviation as the single number most sensitive to geometrical distortion at the C\( \alpha \) where backbone and sidechain requirements must be reconciled. These validation criteria capture the three major structural aspects of geometry around the critical C\( \alpha \) locus in proteins.

A residue with good fit to density, low \( B \)-factor, favored \( \phi, \psi \) values, a rotameric sidechain, no atomic clashes, and ideal covalent geometry is almost certain to be modeled correctly. Whenever several of those factors are far from optimal, however, an error should be suspected unless there are mitigating circumstances such as compensating favorable interactions, tight packing constraints, or functional requirements for a locally strained conformation. Examples where the combined validation criteria diagnose a clear error include the backward-fit sidechain in Figure 2(a) (with a high C\( \beta \) deviation, a bad rotamer, all-atom clashes, and an ambiguous fit to the electron density) and the Ramachandran-outlier residue of Figure 6(a), with \( \phi, \psi \) values of \( +44^\circ, -29^\circ \) well inside the truly forbidden area near \( \phi = 0^\circ \), a bad all-atom clash, and two successive backbone bond angles opened up by \( >10^\circ \). For both these cases, it is an important argument that there are normal, favorable conformations that could occupy nearly the same position in space and connect well with the continuing chain on either side. That also means these two examples are both correctable, which is certainly not always the case but is made more likely by good rotamer libraries and multiple, independent validation criteria that are local and direction-specific.

These criteria are equally valuable for positively validating the correctness of well-placed residues with disfavored but allowed, or even outlier, conformations. Figure 6(b) shows the classic \( \gamma \)-turn residue from thermolysin, with
of $79^\circ$, $63^\circ$ classed as a serious outlier by both ProCheck and WhatCheck [see Fig. 1(b)]. The new Ramachandran criteria class it not as an outlier but as allowed (although disfavored). It forms the $\gamma$-turn H-bond with no atomic clashes and only small bond angle distortions, at the end of a well H-bonded $\beta$-hairpin, and it has moderate $B$-factors (near 20) and clear, well-fit electron density. Figure 6(c) shows a serine with $\phi, \psi$ of $+85^\circ, +171^\circ$, which is a Ramachandran outlier even by the new criteria but is only just outside an allowed region. The other bond and torsion angles are reasonable, there are no bad clashes, the electron density is clear and well fit, the $B$-factors are low (about 7), and the residue makes two backbone and two sidechain H-bonds (one to help bind the adenosine-3'-5' diphosphate product). In both of these examples, the model is validated as clearly correct, because a single worrisome feature is outweighed by the total evidence of the other favorable indicators.

We are suggesting, on the basis of their reproducible but relatively rare occurrence patterns, that conformations outside the favored but inside the allowed, modestly strained, with a significant but not huge energetic penalty relative to the favored, and conformations. Experimental evidence relevant to that claim can come from stability measurements of Ala versus Gly mutants for residues that start out in (and are likely to stay in) conformations favored for Gly but merely allowed for the general case. Several studies fulfill those conditions for the $\alpha'$-turn and the below-$\alpha$ “plateau” and are discussed below. We have found no similar mutation studies for the $\gamma$-turn conformation, but in any case their interpretation would be less clear because the $\gamma$-turn region is not very well populated even for Gly.

Stites et al.41 mutated Gly79 of Staphylococcal nuclease (which has plateau $\phi, \psi$ of $-102^\circ, -145^\circ$ in 1SNc) to Ala and found a decrease in stability of 1.3 kcal mol$^{-1}$. For
that winds through L ...

Aside from greatly lowering the enormous peak at general case without Gly, Pro, or pre-Pro (shown in Fig. 7).

/H9278

most nearly appropriate comparison is with the empirical ...

tions such as those involved in secondary structures, the ...

accessible to the dipeptide, done with a water model or ...

tions for the energy or free energy of conformations ...

lowered by a conformational change. (Note that an appar-

uncertainty about whether the energy difference might be lowered by a conformational change. (Note that an apparently anomalous mutation result, showing a 0.3 kcal mol⁻¹ stabilization for Ala over Gly50 at φ,ψ 97°, –154° in 1SNC, is not applicable to this issue: the loop containing Gly50 has very poor electron density and very high B-factors, so that Ala50 is likely to adopt an entirely different conformation.)

Mutational results and empirical occurrence frequencies agree, therefore, that the II’ and plateau conformations are allowed but disfavored for residues with β-carbons, by a significant but modest energy penalty on the order of 1–2 kT.

Comparison is also relevant with theoretical calculations for the energy or free energy of conformations accessible to the dipeptide, done with a water model or with intermediate values for the dielectric constant. Because the dipeptide cannot include longer-range interactions such as those involved in secondary structures, the most nearly appropriate comparison is with the empirical φ,ψ distribution found for nonrepetitive protein structure, compiled either just for Ala (without pre-Pro) or for the general case without Gly, Pro, or pre-Pro (shown in Fig. 7).

Aside from greatly lowering the enormous peak at α-helical φ,ψ, which is 10 times the density of any other region, the nonrepetitive distribution differs very little from the general case distribution: favored and allowed contours for the nonrepetitive (dark lines) and the general (thin lines) cases in Figure 7 coincide almost perfectly.

The original Ramachandran calculations [Fig. 1(a)] were purely steric, based on a hard sphere model, with an outer contour from minor relaxation of bond angles. They show the three major areas clearly and form the basis for all later work. Our calculations of all-atom contact scores⁴² are also primarily steric but have soft sphere van der Waals repulsions and include an H-bonding term. All-atom contact scores as a function of φ,ψ calculated for Ala in ideal geometry are dominated by two deep elliptical troughs (negative scores are unfavorable), which cover the central region around φ = 0°. These troughs involve truly dire atomic clashes of the O(n – 1) atom deserving the term “forbidden,” whereas the unfavorable regions with positive φ are less extreme. In particular, there is a “shoal” that winds through θα and includes conformations such as γ-turn and II'-turn which are only slightly disfavored and which do occur in the empirical distribution.

Many energy calculations have been performed for the Ala dipeptide in water; some of the excellent samples have used quantum mechanics,⁴³ a variety of molecular mechanics force fields,⁴⁴ and estimates of free energy.⁴⁵ All show the α, β, and Lα regions, and all show the “shoal” in positive φ, but the shapes and positions are not accurate (e.g., the strong diagonal shape of the α and Lα regions is usually not evident). In particular, both steric and energetic calculations show as quite favorable a region left of α, near –150°,–60° on the φ,ψ plot, which is almost unpopulated in the empirical distributions. The two peptide NH groups are close in this conformation (see Fig. 8), but the problem cannot just be electrostatic because those effects are included in all of the energy calculations. Our conjecture to explain this discrepancy between theory and observation is that the angle and crowding of the two NH groups near –150°,–60° permits H-bond donation to only a single acceptor, rather than the two H-bonds that are normally possible for two successive NHs.

Fortunately, there is now a new set of theoretical calculations, reported by Hu et al.¹ in the accompanying article (this issue), which matches the empirical distribution (Fig. 7) in very good detail. They performed dynamics simulations in which the Ala or Gly dipeptide portion was calculated quantum mechanically, whereas the solvent and solvent-peptide energies were calculated with a molecular mechanics force field. Their theoretical distribution shows the forbidden troughs around φ = 0°, the diagonal edges of α and Lα regions, and only a sparse population left of α. Even the Gly distribution matches the empirical data satisfactorily. This achievement of closer agreement between theoretical and empirical φ,ψ distributions provides...
CONCLUSION

One unifying theme of these proposed geometrical validation criteria is that bond angles and torsion angles are much more effective when analyzed in the appropriate local combinations than they are if treated individually. It has been evident from the first Ramachandran plot\(^6\) onward to the current update that \(\phi\) and \(\psi\) are not even approximately independent and must be analyzed together. The original insight of Ponder and Richards\(^46\) in defining sidechain rotamers was that the sidechain angles are much more powerful if analyzed in combination rather than individually, and our recent rotamer analyses\(^5\) confirm that principle even more strongly. The new definition of C\(\beta\) deviation as an especially revealing index of distortion around the C\(\alpha\) provides a final example of this principle, where the suitable combination of multiple bond angles yields a criterion far superior to the individual angle deviations.

The geometrical structure validation criteria described here are a revision and extrapolation of previous standards. The defined \(\phi,\psi\) regions make a significant improvement by virtue of providing more stringent limits where that is needed but also by validating rare but quite acceptable conformations that most crystallographers have encountered at least once in an active site. It is both appropriate, and perhaps even overdue, to use modern high-resolution, low-\(B\) data to define the standards which structures in general should aspire to approximate; these updated standards are especially compelling, because they have converged to agreement and remain constant from 1.8 Å down to 0.5 Å resolution and across all \(B\)-factor ranges <30. An acceptable structure at a given resolution can be defined by having suitably high overall percentages of \(\phi,\psi\) values, sidechain rotamers, and C\(\beta\) deviations within allowed ranges. However, the truly important claim is that by examining the outliers for each of those criteria in conjunction with all-atom contacts and electron density, and by correcting them when appropriate, essentially any protein structure can feasibly be rendered significantly more accurate than without these tools.

ACKNOWLEDGMENTS

We thank Jan Hermans for a productive collaboration on the theoretical issues and Joe Patel for sharing the well-fit 1KA1 “outlier” before publication. Ian W. Davis is a Howard Hughes Medical Institute predoctoral fellow.

REFERENCES