Structure
Article

Conformation Dependence of Backbone Geometry in Proteins

Donald S. Berkholz,[1] Maxim V. Shapovalov,[2] Roland L. Dunbrack, Jr.,[2] and P. Andrew Karplus[1,*]
[1]Department of Biochemistry and Biophysics, Oregon State University, 2011 ALS, Corvallis, OR 97331, USA
[2]Institute for Cancer Research, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111, USA
*Correspondence: karplusp@science.oregonstate.edu
DOI 10.1016/j.str.2009.08.012

SUMMARY

Protein structure determination and predictive modeling have long been guided by the paradigm that the peptide backbone has a single, context-independent ideal geometry. Both quantum-mechanics calculations and empirical analyses have shown this is an incorrect simplification in that backbone covalent geometry actually varies systematically as a function of the Φ and Ψ backbone dihedral angles. Here, we use a nonredundant set of ultrahigh-resolution protein structures to define these conformation-dependent variations. The trends have a rational, structural basis that can be explained by avoidance of atomic clashes or optimization of favorable electrostatic interactions. To facilitate adoption of this paradigm, we have created a conformation-dependent library of covalent bond lengths and bond angles and shown that it has improved accuracy over existing methods without any additional variables to optimize. Protein structures derived from crystallographic refinement and predictive modeling both stand to benefit from incorporation of the paradigm.

INTRODUCTION

Structural details at the 0.1 Å scale guide our understanding of enzyme catalysis, how mutations cause disease, and what makes a good inhibitor and potential drug. Since the work of Pauling et al. (1951), protein model building at all levels has been guided by the assumption that the peptide backbone has a certain ideal geometry independent of context (Figure 1). This paradigm underlies the restraints used to guide protein structure refinement (e.g., Evans, 2007) and is also the basis of the rigid-geometry approximation used to simplify model building in the most successful structure-prediction packages such as Rosetta and I-TASSER (Rohl et al., 2004; Zhang, 2009). The rigid-geometry approximation uses fixed bond lengths and angles, leaving torsion angles as the only variables needed to define the structure. Ideal target values for the peptide backbone have varied little over the years, and a set of values most recently updated in 1999 (EH; Engh and Huber, 1991; Engh and Huber, 2001) is commonly used (Figure 1).

Experimentally derived crystal structures at all but the highest resolutions reflect the influence of the single-value ideal-geometry paradigm that is applied in the form of geometric restraints. However, strong evidence exists that this paradigm is flawed. Quantum-mechanics calculations and empirical analyses of high-resolution protein structures from over a decade ago suggested that the concept of a single, context-independent ideal value for backbone bond angles and lengths was wrong (Schäfer and Cao, 1995; Karplus, 1996). Instead, both approaches showed that backbone covalent geometry varies systematically as a function of the conformation of the backbone torsion angles. The systematic conformation dependence of ideal geometry was most notable for the N-C_α-C bond angle ($\angle NC_\alpha C$) that varied by 8.8°, from 105.7° to 114.5° (Karplus, 1996). Similarly, systematic distortions of geometry are known to occur for classically disallowed but experimentally observed conformations (e.g., Gunasekaran et al., 1996; Ramakrishnan et al., 2007). And finally, particularly intriguing has been the observation that at increasingly higher resolution, protein structures are in progressively worse agreement with the supposedly "ideal" values (e.g., Longhi et al., 1998). This observation resulted in a recent literature debate about how to adjust the target values used for geometric restraints and how heavily to weight them (Jaskolski et al., 2007a; Tickle, 2007; Jaskolski et al., 2007b; Stec, 2007). We contributed to this debate with the suggestion that the root of the problem is not simply a matter of incorrect ideal target values or weights, but instead is a matter of an incorrect paradigm in that ideal geometry should be a function, not a single value (Karplus et al., 2008).

With the explosion of protein structures solved at 1.0 Å resolution or better, the time is ripe to extend the earlier analysis (Karplus, 1996) and more accurately determine the nature and extent of the systematic variations of peptide geometry with conformation. To accomplish this, we created a nonredundant database of atomic-resolution structures that has nearly 20,000 residues. Here, we use this database to analyze conformation-dependent trends in backbone geometry in all bond angles and lengths. We also show that accounting for these trends has the potential to improve both crystallographic refinement and homology modeling.

RESULTS AND DISCUSSION

Data Source and Analysis Strategy

To accurately characterize the nature and extent of conformation-dependent variations in geometry, we used a data set of 16,682 well-defined 3-residue segments from 108 diverse protein chains determined at 1.0 Å resolution or better (see Experimental Procedures). A 3-residue segment includes all of

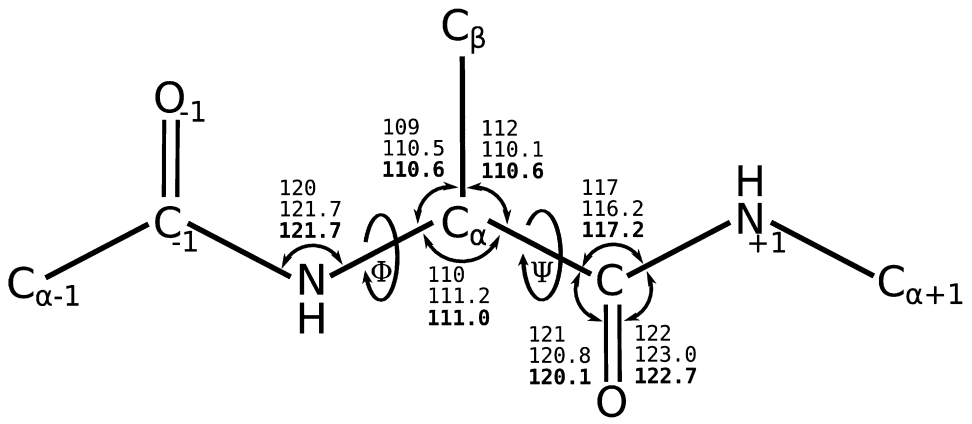


Figure 1. Evolution of the Ideal Values for Backbone Geometry Used in the Single-Value Paradigm

A central residue (residue 0) is shown with atoms from residues −1 and +1 that contribute to its two adjacent peptide units. For each of the seven bond angles associated with residue 0, three ideal values from earlier work are shown from oldest (top) to most recent (bottom). They are from Corey and Donohue (1950), Engh and Huber (1991), and Engh and Huber (2001). Most refinement and modeling programs use one of the Engh and Huber sets or a slight variation on them. Rotatable bonds defining the backbone torsion angles Φ and Ψ are indicated. Figure created with Inkscape.

the atoms in two complete peptide units, and the data set included the bond lengths and bond angles for the peptide units uniquely identified by whether they mostly involve atoms from residue −1, 0, or +1 in the 3-residue segment (Figure 1). Based on previous work (Karplus, 1996) indicating distinct geometric behavior of Gly, Pro, the β-branched residues Ile and Val (Thr behaves more like a general residue because of stabilizing sidechain-backbone hydrogen bonds), and residues preceding proline (pre-Pro), we carried out separate statistical analyses for those five groups. The data set used here included 1,379 Gly, 639 Pro, 511 general pre-Pro (644 before exclusion of Gly/Pro/Ile/Val), 1,822 Ile/Val, and 10,921 general residues (the 16 other residue types taken together). All pre-Pro residues are excluded from the other classes. As seen in Figure 2, these residues were distributed in Φ,Ψ as has been seen for many well-filtered data sets (Karplus, 1996; Kleywegt and Jones, 1996, Lovell et al., 2003). Figure 2 also provides the shorthand nomenclature we will use for certain regions of the Ramachandran plot.

We analyzed these results to visualize and to document the Φ,Ψ-dependent variations in bond lengths and angles. Our approach was to use kernel-regression methods to smooth the data and to produce continuously variable functions for each parameter (see Experimental Procedures). The figures and tables in this paper are based on the kernel-regression analysis and only include regions of the Ramachandran plot having an observation density of at least 0.03 residues/degree^2 (i.e., 3 residues in a 10° × 10° area) and a finite standard error of the mean.

Ubiquitous, Systematic, Φ,Ψ-Dependent Variations Exist in Peptide Geometry

Bond Angles

The data reveal that for general residues, all 15 bond angles in the two peptides adjacent to the central residue vary systematically with Φ and Ψ (Figure 3 and Table 1). The most prominent observation is that the variations do not occur only in rare outlier conformations, but they occur throughout even the most populated areas of the plot for all residue types (Figure 3; see Figures S1–S4 available online). Consistent with the lower-resolution analysis (Karplus, 1996), ∠NC$_\alpha$C varies the most (6.5°), but four other angles also vary by ≥ 5°. An important difference from the results of the earlier study is that the conformation-dependent standard deviations of the bond angles are about half what was seen previously (Karplus, 1996), ranging from 1.3°–1.8° (Table 1). These are also substantially smaller than the standard deviations of ∼2.5° used for the single ideal values defined by Engh and Huber (1991) based on small-molecule structures. It is notable that ultrahigh-resolution crystal structures are generally refined using geometric restraints that do not match the local averages, so the narrow (small σ) distributions cannot be an artifact of the restraints used. Interestingly, the variations in the averages are 2–4 times the standard deviations (Table 1), implying that current modeling restraints would work to wrongly pull angles away from their actual optimal values in many regions. Dramatically, the distributions at the extremes can even be completely nonoverlapping because of the small standard deviations (Figure 4). The standard errors of the Φ,Ψ-dependent means (i.e., σ/√N) for bond angles are less than 0.5° in nearly all regions and typically less than 0.2° in the highly populated regions (Figures S5–S9)—however, errors should be considered when examining averages for the lowest-populated edges and other regions, such as the pre-Pro region for general residues. In comparison, the 2°–7° ranges seen for the expected values are 10–50 times greater than their uncertainties. This shows that the variations are well-determined and backbone geometry in no way obeys the single ideal value paradigm.

Bond Lengths

In the 1996 study, the resolution of the data did not allow reliable visualization of bond-length variations. Here at atomic resolution, systematic Φ,Ψ-dependent trends are now visible in bond lengths (Figure 5) but the variation ranges (0.01–0.02 Å) are only on par with the standard deviations (0.012–0.016 Å), meaning the distributions are highly overlapping. The standard errors of the mean are smaller (∼0.002 Å), so the variations in the means seen are nevertheless significant (∼10-fold larger). Given that the standard deviations are on par with the expected coordinate accuracy, we hypothesize that the true underlying

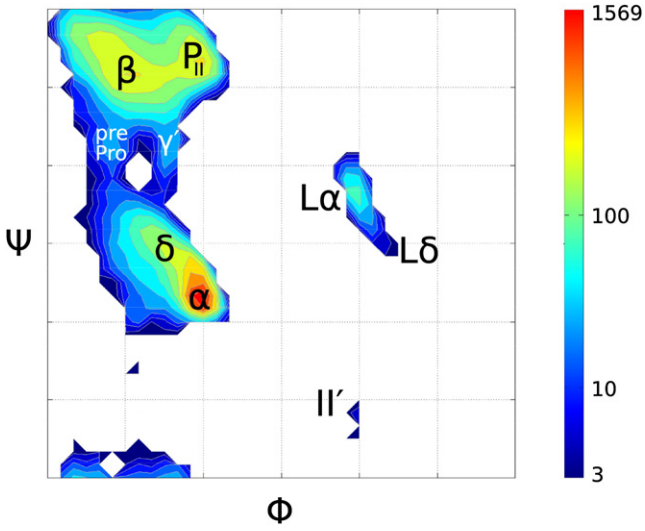


Figure 2. Protein Backbone Conformations of Non-Gly Residues
This Ramachandran plot is colored by empirical observation counts in atomic-resolution proteins. Labels indicate regions of particular interest (Karplus, 1996; Lovell et al., 2003; Hollingsworth et al., 2009). Coloring uses a logarithmic function to allow lower- and higher-population regions to be seen simultaneously. Observation density was calculated using kernel regressions (see Experimental Procedures). Unlabeled versions of this plot and another for only Gly residues are available as supplementary material (Figures S12 and S13). Figure generated with Matlab and edited with Inkscape.

bond lengths are distributed more narrowly and thus will require still higher resolution analyses to determine accurately. Because of this limitation and the expectation that, because of the very small distances involved, the bond-length variations will have little impact on modeling accuracy, we will not further describe the bond-length trends here. Nevertheless, we suspect the variations involved will be chemically informative (e.g., Esposito et al., 2000; Figure 5).

Variations Are Correlated with Local Interactions

Comparison of conformation-dependent trends across the two sequential peptide units reveals that the trends are largely locally influenced. For each of the seven angles associated with the central residue, the range is larger than the range for the same angle associated with the previous or subsequent residue (Table 1). For instance, $\angle N_{-1}C_{\alpha-1}C_{-1}$ and $\angle N_{+1}C_{\alpha+1}C_{+1}$ have ranges of 5.5° and 3.0°, whereas $\angle NC_{\alpha}C$ has a range of 6.5°. This implies that the angles in Table 1 associated with residues −1 and +1 show highly local effects, being more influenced by the Φ,Ψ values of their respective residues than the Φ,Ψ values of residue 0 (the central residue). For modeling purposes, it makes sense to assign the "ideal" target values for all seven of these angles based on Φ,Ψ of the central residue.

Furthermore, among these seven angles, additional evidence of the dominance of local effects is seen as each angle is influenced mostly by the single closest torsion angle, whether it is Φ or Ψ (Figure 3). Starting at the N-terminal end, $\angle C_{-1}NC_{\alpha}$ is heavily Φ-dependent as is seen in the vertical pattern of variation, then the C_{α}-centered angles are a mixture, displaying diagonal patterning, and the angles at the C-terminal end, such as $\angle C_{\alpha}CN_{+1}$, have Ψ-dependent horizontal patterning. Even among the C_{α}-centered angles, $\angle NC_{\alpha}C_{\beta}$ shows enhanced dependence on Φ and $\angle C_{\beta}C_{\alpha}C$ shows enhanced dependence on Ψ. This extreme locality agrees with much prior work noting that local steric interactions are critical factors in determining observed conformational and secondary-structure preferences (e.g., Dunbrack and Karplus, 1994; Baldwin and Rose, 1999).

Comparison of Trends with Quantum Mechanics

As noted in the introduction, quantum-mechanical (QM) calculations of isolated alanine peptides (Jiang et al., 1997; Yu et al., 2001) also produce conformation-dependent trends in bond angles and bond lengths. The QM calculations are computationally intensive and they have only been carried out at 30° resolution in Φ,Ψ (Jiang et al., 1997; Yu et al., 2001), making detailed features of the trends unavailable. Beyond a certain level, the method and basis set used in QM calculations is unimportant to this analysis because they produce trends on the same scale with a nearly constant offset (Yu et al., 2001). As was reported by Karplus (1996), the QM results have similar trends, but now it is apparent that QM results show larger deviations, ranging farther both positively and negatively than experimental protein structures. For example, the empirical deviations from the central value for $\angle NC_{\alpha}C$ are roughly 70% of the calculated deviations. Additionally, QM calculations show serious discrepancies in some less populated regions, such as a false global maximum for $\angle O_{-1}C_{-1}N$ in Lδ (Figures 2 and 3). The mis-scaling seen in QM-calculated angles has been suggested by others to be caused by a lack of long-distance structural effects (Jiang et al., 1997; Yu et al., 2001; Feig, 2008). However, if that were the case, comparison of residues in secondary structure versus those in loops should show this same difference, but Karplus (1996) did not see a difference, and here we confirm that observation (Figures S10 and S11). One potential underlying cause is the difference between a protein environment and vacuum rather than a long-distance effect caused by repeating secondary structure, but the reason that calculations in small peptides fail to predict the correct details of conformation-dependent geometry for proteins is uncertain.

Local Variations Make Structural Sense

The bond-angle trends for five classes of residues for all Φ,Ψ possibilities comprise a massive amount of information that cannot be exhaustively described in the context of this article. A survey of the results, however, reveals a general principle that the observed trends in geometry make structural sense in terms of accommodating local steric and electrostatic interactions, extending the rationale for observed conformations proposed by Ho et al. (2003). In Karplus (1996), the behavior of $\angle NC_{\alpha}C$ in the well-populated α, β, and δ regions (Figure 2) was rationalized in these terms, including the proposal of a π-peptide interaction in the δ region optimized by the opening of $\angle NC_{\alpha}C$ (see Figure 8 of Karplus, 1996). Instead of rehashing those observations, here we present four illustrative examples of Φ,Ψ regions with significant distortions. The conformations are shown in Figure 2, the relevant bond-angle values can be seen in Figure 3, and the specific collisions being ameliorated are illustrated in Figure 6.

In the Lα/Lδ region, non-Gly residues are disfavored because when using single ideal values for bond angles and lengths, there

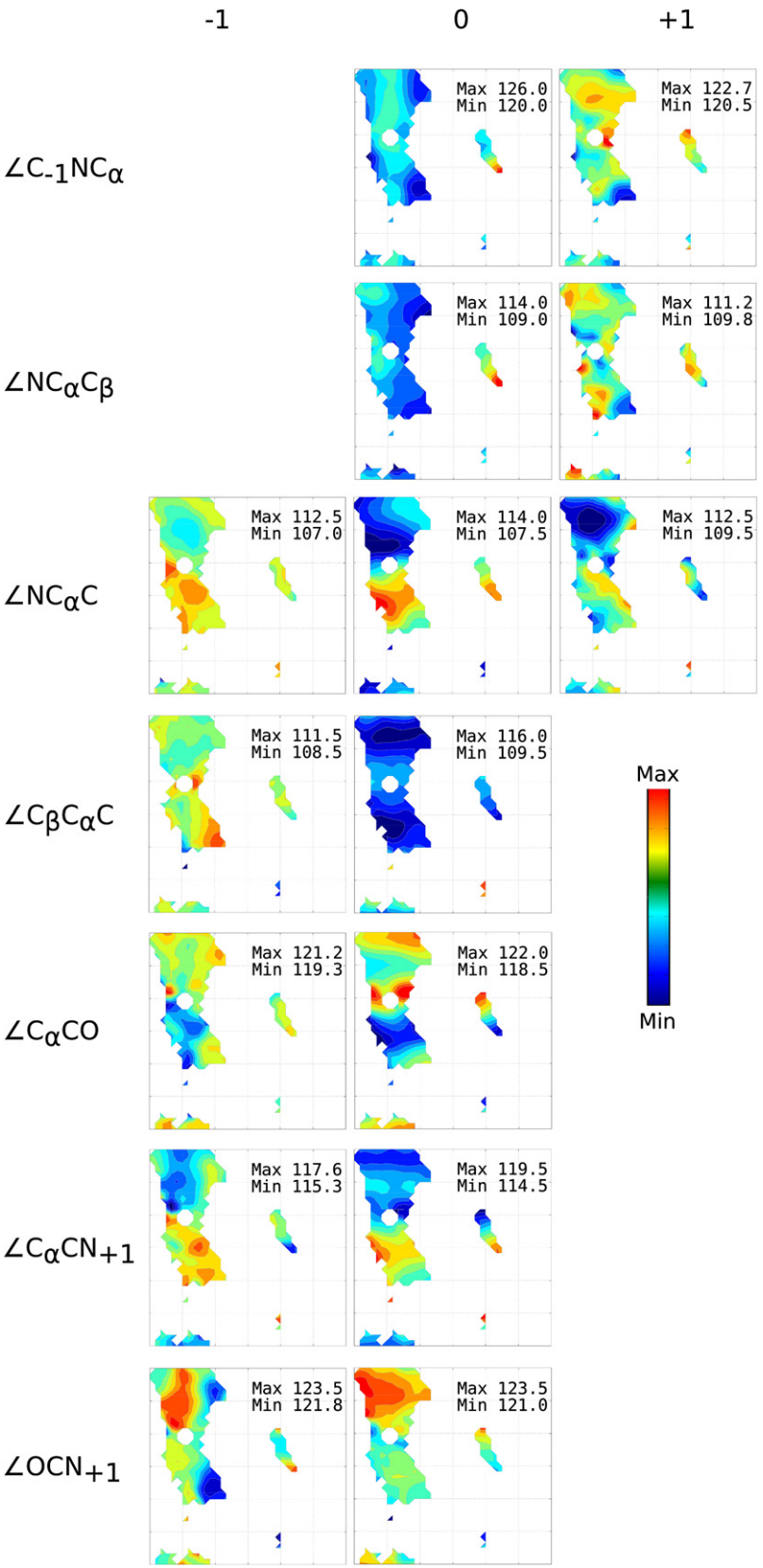


Figure 3. Conformation-Dependent Variation in Bond Angles of General Residues as a Function of the Φ,Ψ of the Central Residue

A Ramachandran plot is shown for each backbone bond angle in the two peptide units surrounding the central residue of the tripeptide. The seven unique peptide bond angles are associated with either residue −1, 0, or +1 based on which residue contributes at least two atoms to the angle. Φ and Ψ in each plot, however, refer to the central residue, 0. Within each plot, colors indicate averages ranging from the global minimum (blue) to the global maximum (red) as calculated using kernel regressions (see Experimental Procedures). The global minima and maxima are provided in each plot. Figure created with Matlab.

Table 1. Observed Ranges for Peptide Bond Angles

Residue	Angle	EH[a]	Min(CDL)	Max(CDL)	Range	σ(EH)	σ(CDL)[b]
−1	$\angle NC_\alpha C$	111.0	107.0	112.5	5.5		
	$\angle C_\beta C_\alpha C$	110.6	108.5	111.5	3.0		
	$\angle C_\alpha CO$	120.1	119.3	121.2	1.9		
	$\angle C_\alpha CN_{+1}$	117.2	115.3	117.6	2.3		
	$\angle OCN_{+1}$	122.7	121.8	123.5	1.7		
0	$\angle C_{-1}NC_\alpha$	121.7	120.0	126.0	6.0	1.8	1.7
	$\angle NC_\alpha C_\beta$	110.6	109.0	114.0	5.0	1.7	1.6
	$\angle NC_\alpha C$	111.0	107.5	114.0	6.5	2.8	1.5
	$\angle C_\beta C_\alpha C$	110.6	109.5	116.0	6.5	1.9	1.8
	$\angle C_\alpha CO$	120.1	118.5	122.0	3.5	1.7	1.3
	$\angle C_\alpha CN_{+1}$	117.2	114.5	119.5	5.0	2.0	1.3
	$\angle OCN_{+1}$	122.7	121.0	123.5	2.5	1.6	1.3
+1	$\angle C_{-1}NC_\alpha$	121.7	120.5	122.7	2.2		
	$\angle NC_\alpha C_\beta$	110.6	109.8	111.2	1.4		
	$\angle NC_\alpha C$	111.0	109.5	112.5	3.0		

All values are in degrees. CDL indicates the conformation-dependent kernel regressions from this work.

[a] Values are from Engh and Huber (2001).

[b] Values are typical for the majority of the plot, although they are greater in the least populated regions. See Figure S5 for details.

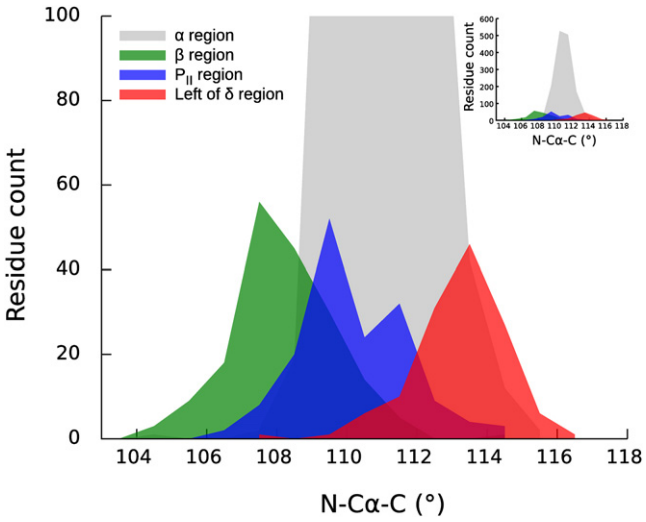


Figure 4. $\angle NC_\alpha C$ Distributions Are Well-Defined and Distinct

Shown are observations from selected 10° × 10° bins in each of four conformations: α (gray), β (green), P_{II} (blue), and a region left of δ at (−125°, −5°) (red). The *y* axis range is set to visualize the distributions in non-α bins. Histograms were calculated using 1° bins and plotted as frequency polygons. Distributions are visibly separate and thus conformation dependent. Inset: The same plot, with the *y* axis range set to display the full height of the α distribution. If not broken out by conformation, the non-α bins would be indistinguishable from tails of the α distribution. Figure created with gnuplot and Inkscape.

is a close-contact collision between O_{-1} and $C_\beta H$. As Φ increases, this collision becomes worse. The conformation-dependent trends show that these conformations become accessible by a systematic increase in $\angle O_{-1}C_{-1}N$, $\angle C_{-1}NC_\alpha$, and $\angle NC_\alpha C_\beta$ that opens the ring between O_{-1} and C_β. At the extreme tip of the region near (+90°, 0°), these angles open compared with the EH values (Figure 1) by 0.4°, 4.3°, and 2.8°, respectively, to increase the $O_{-1}\ldots C_\beta$ distance from 2.59 Å to 2.85 Å. Although this change in distance is small, as are others described in this section, they can make large energetic differences by transforming unfavorable atomic clashes to close contacts.

The II′ region is adopted by the i+1 residue of type II′ turns—a tight turn with a hydrogen bond between O_{-1} and $N_{+2}H$. In this conformation, C_β is quite close to both O_{-1} and N_{+1}, which results in this region being unfavorable for nonglycine residues. Under the rigid-geometry approximation, the entire region should be disallowed because of this clash, but distortions in covalent geometry allow it to be accessible. The variations seen in Figure 3 show that the distortions relative to EH values (Figure 1) include a large opening in $\angle C_\beta C_\alpha C$ (5.9°) as well as opening of $\angle C_\alpha CN_{+1}$ (3.3°) to reduce the $C_\beta\ldots N_{+1}$ clash. This also reduces the $O_{-1}\ldots C_\beta$ clash, where the $\angle C_\beta C_\alpha C$ distortion acts like opening jaws to move C_β away from O_{-1}. The extreme bond openings are enabled by a closing of $\angle NC_\alpha C$ (2.5°), $\angle C_\alpha CO$ (1.8°), and $\angle OCN_{+1}$ (2.0°). The $C_\beta\ldots N_{+1}$ distance increases from 2.65 Å to 2.71 Å, and the $O_{-1}\ldots C_\beta$ distance increases from 3.06 Å to 3.09 Å.

Left of the δ region is a Ramachandran-allowed but sparsely populated region. The primary clash is between HN and HN_{+1}. This clash is prevented through a combination of distortions relative to EH values: the dominant increases are in $\angle NC_\alpha C$ (3.5°) and $\angle C_\alpha CN_{+1}$ (2.8°) that both exhibit their extreme values (Figure 3), coupled with a decrease in $\angle C_\alpha CO$ (2.0°). The combined effect is to open and twist a nearly planar ring between NH and $N_{+1}H$ to prevent a van der Waals overlap by increasing the $HN\ldots HN_{+1}$ distance from 1.78 Å to 1.92 Å and the $N\ldots N_{+1}$ distance from 2.66 Å to 2.76 Å.

As a final example, we illustrate the importance of treating pre-Pro as a special residue type. Preproline residues are classically disallowed in the α region, yet they are experimentally observed with low populations (Hurley et al., 1992). The primary clash occurs between N and $C_{\delta+1}$ with a secondary clash between $C_\beta H$ and $C_{\delta+1}H$ (Figure 6). To alleviate this clash, the Pro ring bends away from the pre-Pro residue through increases in $\angle NC_\alpha C$ (2.0°), $\angle C_\beta C_\alpha C$ (2.4°), and $\angle C_\alpha CN_{+1}$ (3.3°), enabled by decreases in $\angle C_\alpha CO$ (2.3°), $\angle OCN_{+1}$ (2.6°), and $\angle CN_{+1}C_{\alpha+1}$ (3.8°). In comparison to calculations by Hurley et al. (1992) that suggested a single, very large deviation of 8.5° in $\angle C_\beta C_\alpha C$, here we observe that the distortions have diffused across all of the angles between the sterically hindered atoms. These distortions increase the $N\ldots C_{\delta+1}$ distance from 2.65 Å to 2.85 Å and the $C_\beta H\ldots C_{\delta+1}H$ distance from 1.86 Å to 1.90 Å to reduce the van der Waals overlap. $\angle CN_{+1}C_{\delta+1}$ was not included in the database, but we expect it also opens to further alleviate the collision.

A 10° Resolution Conformation-Dependent Library

With the knowledge of these systematic trends comes the possibility of leveraging them to improve the accuracy of crystallographic refinement and homology modeling. To provide a convenient form in which the documented systematic variations can be used in modeling applications, we created a binned conformation-dependent library (CDL) for distribution. Similar to the approach taken by Karplus (1996), we divided Φ,Ψ space

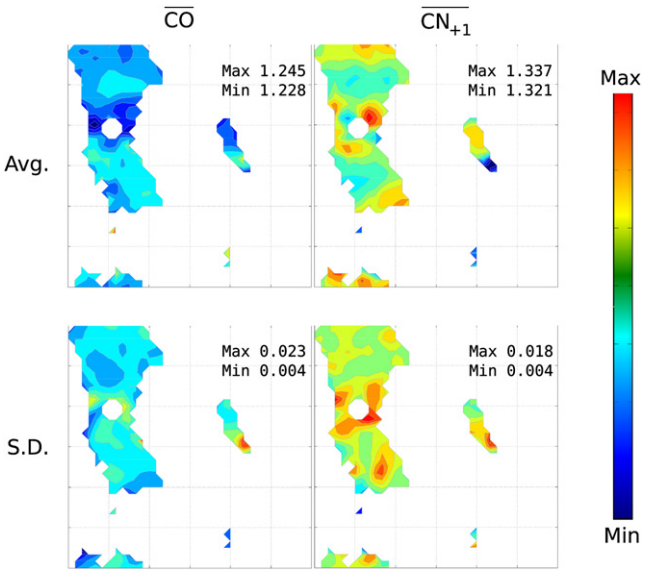


Figure 5. Conformation-Dependent Variation in Bond Lengths Is Partially Masked by Experimental Uncertainty

Ramachandran plots are shown for average lengths and standard deviations of the C = O bond (left panels) and the C-N bond (right panels) using colors as in Figure 3. These bonds are involved in the partial double-bond character of the peptide bond, so the expectation is for them to be anticorrelated as electron density shifts between them. Some such anticorrelation is visible as a Ψ-dependent effect in averages (shown in the top panels) but it is not as clear as trends seen in bond angles, possibly because the standard deviations (shown in the bottom panels) are near the level of experimental uncertainty.

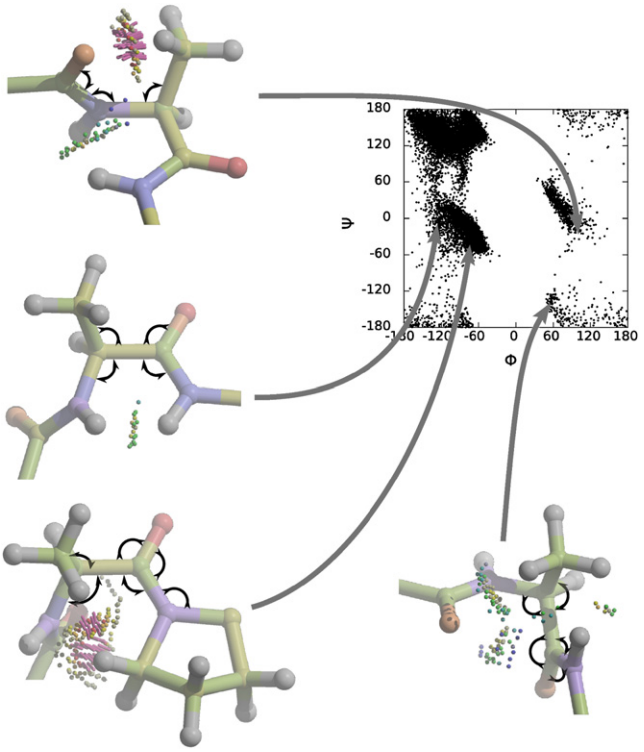


Figure 6. Structural Basis for Geometry Variations of Selected Conformations

Four Ala residues with adjacent peptides are shown, built using EH values and with dots showing van der Waals overlap between atoms: blue (wide contact), green (close contact), yellow (small overlap), and red (bad overlap). Clockwise from top left: tip of the Lα/Lδ region; left of the δ region; a pre-Pro–Pro dipeptide in the α region; and the II′ region. Arrows indicate angles that open or close substantially relative to EH values. Note that all of these distortions serve to alleviate atomic clashes. The overlaps were calculated by MolProbity (Davis et al., 2007) and are shown in Coot (Emsley and Cowtan, 2004).

into 1296 10° × 10° bins and calculated the averages and standard deviations for each bin for each of the five residue-type categories (Gly, Pro, pre-Pro, Ile/Val, general). This first-generation CDL (v1.0), available from the authors or at http://proteingeometry.sourceforge.net/, uses a simple precalculated lookup table that accepts conformations and returns the appropriate target value for each bond angle and length. When considering crystallographic refinement and homology modeling, it is important to note that using more accurate CDL values in place of EH values does not increase the number of variable parameters used in the modeling.

Conformation-Dependent Angles Are More Accurate

A variety of simple control calculations can be carried out to show that the CDL is an improvement over the single-value paradigm (EH values) and even context-dependent values derived from molecular mechanics (MM) force fields. Because an MM force field allows bond angles and lengths to vary with conformation, it could in theory provide better conformation-dependent values than the empirical approach.

As one simple assessment, we compared how well the $\angle NC_\alpha C$ values in a 1.15 Å ribonuclease structure (Protein Data Bank [PDB] code 1rge; Sevcik et al., 1996) matched with EH values, the CDL, and bond-angle values from the structure after minimization using a MM force field (see Experimental Procedures). As seen in Figure 7, the conformation-dependent library outperforms both the single ideal value and MM. Importantly, the CDL produces more angles with very close (<1°) agreement with the reference crystal structure as well as fewer extremely large deviations. In terms of modeling accuracy, there appears to be no downside to using the CDL.

For a more thorough comparison of the CDL with EH values, we compared how well each matched the $\angle NC_\alpha C$ values for the set of protein structures used to generate the CDL, with each protein jackknifed out during its comparison. Averaged over the whole data set, the median deviation from the native bond angles for the EH single-value paradigm was 1.5°/residue and the median deviation for the CDL dropped to 1.1°/residue. This amounts to an improvement of ~25% in $\angle NC_\alpha C$ accuracy relative to the old paradigm.

To understand the impact this difference could have upon protein modeling, coordinates for each jackknifed structure were rebuilt from torsion and bond angles using EH or CDL values. Holmes and Tsai (2004) have shown that the replacement of experimental bond angles with ideal ones while holding Φ and Ψ fixed shifts coordinates by an average of 6 Å (unnormalized by protein length), and this limits model-building accuracy. Here, applying the same approach, we find that the median C_α $rmsd_{100}$ (root-mean-square deviation; normalized to the length of a 100-residue protein) from the native structure for EH values

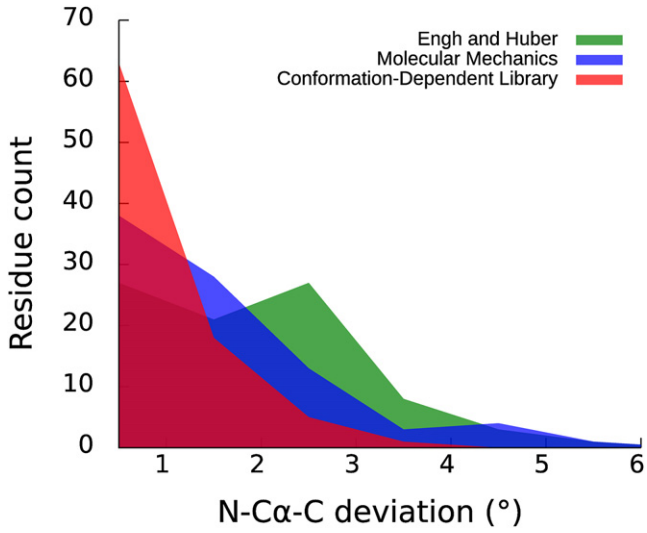


Figure 7. CDL ∠NC$_\alpha$C Values Match Ultrahigh-Resolution Structures Best

Deviations of predicted angles from the experimental ones for atomic-resolution ribonuclease (PDB code 1rge; Sevcik et al., 1996) with various methods are shown: EH single ideal values (green), molecular mechanics (blue), and the CDL (red). Results are shown in a histogram-like manner using 1° bins and frequency polygons. Of these three methods, the CDL matches best, followed by molecular mechanics, then single ideal values. Figure created with gnuplot.

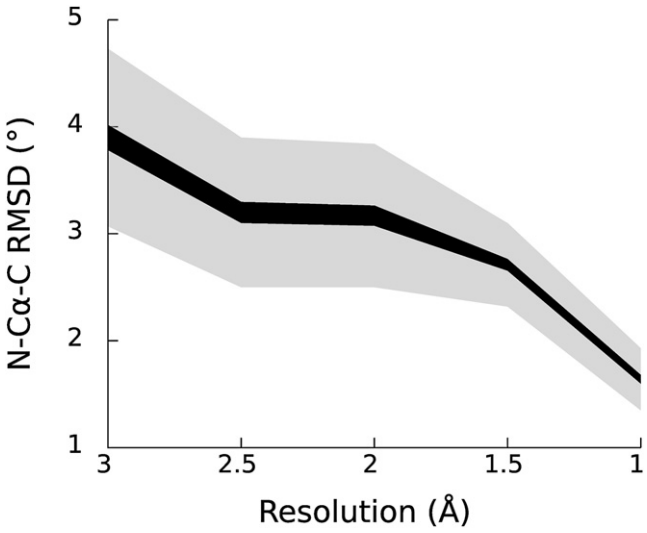


Figure 8. ∠NC$_\alpha$C Deviation of the CDL Values from Crystal Structures as a Function of Resolution of the Analysis

At each of five resolutions ranging from 1.0-3.0 Å, the ∠NC$_\alpha$C rmsds from the CDL were calculated for 50 nonredundant structures. The distributions of rmsds at each resolution are shown. The thickness of the black line indicates the standard error of the mean, and the thickness of the gray line indicates the standard deviation. Figure created with gnuplot.

was 3.23 Å, and for CDL values it was 2.85 Å. The CDL produced a significant improvement in the C$_\alpha$ rmsd$_{100}$ of ~0.4 Å over the old single-value paradigm.

Potential Applications: Crystallographic Refinement and Homology Modeling

To assess the potential impact of accounting for Φ,Ψ-dependent variations upon X-ray crystal structures at various resolutions, we evaluated how much the experimental ∠NC$_\alpha$C values deviated from those in the CDL as a function of resolution (Figure 8). To avoid bias, none of the structures used in the survey were used in the generation of the CDL. Analysis of the data shows that for structures solved at near 1 Å resolution, the rmsd of ∠NC$_\alpha$C from the CDL is ~1.6°. This matches well with the standard deviation seen in the CDL for this angle and serves as an effective validation of the CDL. Additionally, the small standard deviation of the rmsds at this resolution shows that each of the individual high-resolution structures is well-described by the CDL. Already at a resolution of 1.5 Å, normally considered very high resolution, the match of ∠NC$_\alpha$C values to the CDL is nearly twice as poor as for the 1.0 Å resolution structures. This loss of accuracy became steadily more pronounced in lower-resolution structures, rising to nearly 4° at 3.0 Å resolution. We conclude that by using the CDL, high-, medium-, and low-resolution structures could all be improved. We suspect that at resolutions worse than 3 Å, the CDL would have less impact because dihedral angles would be less reliable.

To understand the potential benefit of accounting for Φ,Ψ-dependent geometry variations in predictive modeling of protein structure, we carried out a test using the Rosetta modeling program (Rohl et al., 2004). A standard control calculation for homology modeling is to ask how far a crystal structure moves from the experimental structure when minimized by the force field. This provides a lower limit on how accurately a structure can be predicted (e.g., Bradley et al., 2005). For our test, we performed a series of 100 Monte Carlo energy minimizations starting with different random seeds using both native and "ideal" bond geometries for two ultrahigh-resolution protein structures: ribonuclease chain A at 1.15 Å resolution (PDB code 1rge; Sevcik et al., 1996; Figure 9) and the PDZ domain of syntenin at 0.73 Å (PDB code 1r6j; Kang et al., 2004; data not shown). "Native" geometry refers to the bond lengths and angles as seen in the crystal structure. As seen in Figure 9A, minimizations using the "native" bond geometry instead of the idealized geometry resulted in better convergence (tighter grouping) and allowed the minimized structure to be about 30% closer to the true structure (~0.6 Å versus ~0.9 Å). One notable feature is that the improved behavior occurs despite the force field's optimization based on the traditional "ideal" geometry values. We conclude from this that the use of the rigid-geometry approximation with standard single ideal values limits modeling accuracy substantially. Thus, it is worthwhile to adapt modeling programs to account for the new conformation-dependent geometry paradigm.

To pinpoint exactly where in the structure the improvements occurred, we calculated the deviations between the crystal structure and the energy-minimized structures using native versus ideal geometry (Figure 9B). As an indication of the variation that can occur for this protein in two environments, the deviations with chain B from the same structure are also shown. The largest differences between EH and experimental geometry occur in loops rather than regular secondary structure

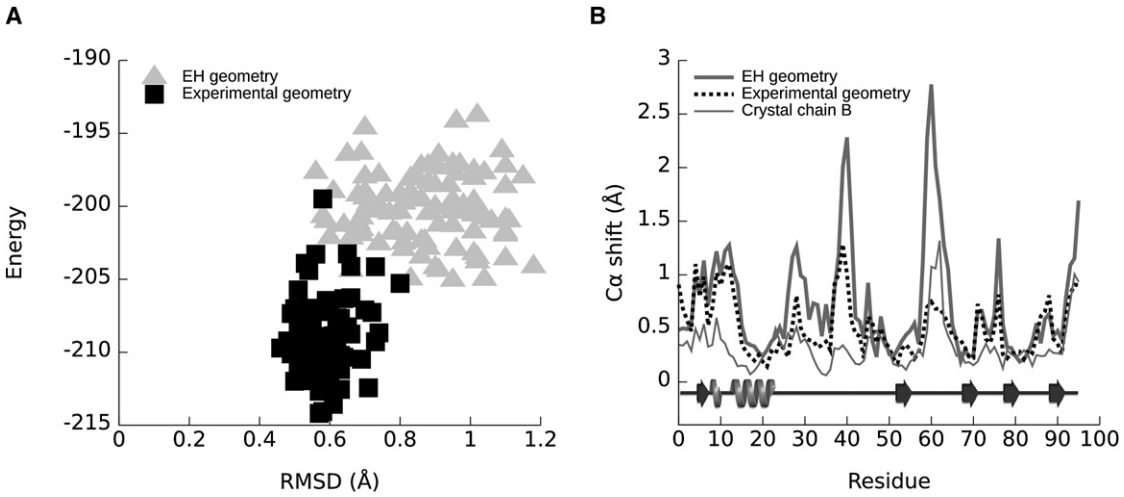


Figure 9. Energy Minimization Behaves Better Using Experimental Geometry as Opposed to the Rigid-Geometry Approximation

(A) Shown are 100 trials minimized with experimental (squares) and with EH (triangles) geometries. They are plotted as Rosetta energy versus the C_α rmsd from the crystal structure (as calculated by Rosetta). Figure created with gnuplot.

(B) Shown are C_α shifts between the crystal structure chain A and a structure selected from the cluster center using experimental (dashed line) or EH (solid thick line) geometry, C_α shifts for chain B versus chain A from the same crystal (solid thin line), and a schematic of the secondary structure (using spirals for helices and arrows for β strands). The crystal structure chain B reflects the differences in the same protein in two environments. Overlays were created using the McLachlan algorithm as implemented in ProFit by iteratively overlaying structures using a subset of C_α atoms with a maximum per-atom rmsd of 0.1 Å until convergence was reached. The secondary structure is taken from PDBsum (Laskowski et al., 2005). Figure created with gnuplot and Inkscape.

(Figure 9B). This meets the expectation that the largest systematic deviations from single ideal values should occur in parts of the protein with less observed, more diverse Φ,Ψ values. This result was expected because the most highly populated regions dominate the global averages, resulting in the illusion of single ideal values assumed in EH, whereas more conformationally diverse, less populated regions contribute less to the global average. Importantly, the two loops that were highly improved by using experimental geometry are at the active site of the protein, so the accuracy with which they are modeled would significantly impact the ability of this mock homology model to provide insight.

Outlook

The studies here show that the dependence of backbone geometry on conformation is unmistakably real, significant, and systematic, and it has a rational structural basis. These systematic distortions in covalent geometry additionally explain how some conformations are accessible to amino-acid residues despite being theoretically disallowed by modeling based on single ideal values for backbone geometry. Extending these studies to the conformation dependence of the ω and χ_1 torsion angles will be described elsewhere. The conformation-dependent library we derived from the database represents the first step toward implementing the new paradigm of "ideal-geometry functions." With its much-improved agreement to ultrahigh-resolution crystal structures, the ideal-geometry functions provide an intellectually satisfying resolution to the debate among crystallographers as to what ideal values should be used during refinement. Also, because the ideal-geometry functions captured in the CDL are simply a highly enlarged set of immutable ideal values, their use in the place of single ideal values represents no increase in algorithmic complexity. Use of the CDL thus offers the potential for improved modeling accuracy in a wide variety of experimentally based and predictive modeling applications without increasing the risk of overfitting.

EXPERIMENTAL PROCEDURES

Data Set Construction

A Protein Geometry Database being developed in our laboratory (http://pdf.science.oregonstate.edu/) was used to generate our data set of atomic-resolution geometry information. To optimally analyze Φ,Ψ-dependent geometry trends, the data set must be large but also have independent and accurate information about geometry. The plethora of new atomic-resolution protein structures allowed us to use stringent criteria for independence and accuracy, yet still have sufficient observations for reasonable statistics. To ensure independence, we used the PDBSelect (Hobohm and Sander, 1994) list from March 2006 to choose protein chains with less than 90% sequence identity to any other chain in the data set. To ensure high accuracy, we only used structures determined at 1.0 Å or better. At this resolution, we estimate Φ and Ψ dihedral angle accuracy to be better than 3° (see next paragraph). Also, as in Karplus (1996), to ensure that individual residues used were well-resolved, we required that all residues in a five-residue segment were all well-ordered, having B-factors < 25 $Å^2$ for the main-chain average, the side-chain average, and C_γ, and alternative conformations were discarded.

To estimate the experimental uncertainty in Φ and Ψ for 1 Å resolution structures, we chose to use a straightforward, empirical approach—randomize and re-refine a test structure multiple times and then examine the spread of the dihedral angles among the structures. Specifically, we applied 10 coordinate randomizations with a mean shift of 0.2 Å using phenix.pdbtools (Adams et al., 2002) to the coordinates of glutathione reductase at 0.95 Å resolution (PDB ID 3dk9; Berkholz et al., 2008) and re-refined each in SHELXL (Sheldrick, 2008). Dihedral rmsds for the vast majority of residues were between 1° and 2°. The 90th percentile of the per-residue rmsds in both Φ and Ψ was 2.2°, and the rmsd values of the per-residue rmsds for Φ and Ψ were 1.7° and 2.4°, respectively.

Kernel Regression for the Bond Lengths and Bond Angles

The data value of any structural parameter *a* of residue *i* (or of the left or right neighbor of residue *i*) can be expressed:

$$a_i = m(\phi_i, \psi_i) + v^{\frac{1}{2}}(\phi_i, \psi_i)\varepsilon_i$$

where *m* is a regression function, and ε are random Gaussian-distributed errors with mean 0 and σ = 1:

$$m(x,y) = E(a|\phi = x, \psi = y)$$
$$v(x,y) = \mathrm{Var}(a|\phi = x, \psi = y).$$

In these expressions, *E* is the expectation value of *a* and Var is the variance of *a*.

To obtain an estimate of *m* and ν, we use a zeroth-order or Nadaraya-Watson kernel regression (Naradaya, 1964) by summing over *N* data points:

$$\widehat{m}(\phi,\psi) = \frac{\sum_{i=1,N} K(\phi_i - \phi, \psi_i - \psi)a_i}{\sum_{i=1,N} K(\phi_i - \phi, \psi_i - \psi)}$$

$$\widehat{v}(\phi,\psi) = \frac{\sum_{i=1,N} K(\phi_i - \phi, \psi_i - \psi)(a_i - \widehat{m}(\phi_i,\psi_i))^2}{\sum_{i=1,N} K(\phi_i - \phi, \psi_i - \psi)}.$$

The latter is $\mathrm{Var}(a|\phi,\psi)$, an estimate of the heteroscedastic data variance as a function of ϕ and ψ.

The functions *K* are kernels that weight the data points based on how far away they are from the query ϕ,ψ value. Because ϕ and ψ are angles, we use the product of two von Mises kernel functions (Mardia and Zemrock, 1975)

$$K(\phi - \phi_i, \psi - \psi_i) = \frac{1}{4\pi^2}\frac{1}{(I_0(\kappa))^2}\exp(\kappa(\cos(\phi_i - \phi) + \cos(\psi_i - \psi))).$$

At large values of κ, these functions behave very similarly to Gaussian distributions, except that they are periodic. We investigated several values of κ and plotted the resulting regressions as a function of ϕ and ψ. We empirically chose a value of κ = 50 to produce distributions that varied smoothly with ϕ and ψ in a reasonable way.

The ϕ,ψ map is not uniformly populated by data points, each of them representing a single residue backbone conformation. Therefore, for the unpopulated regions of the map, the kernel regression analysis generates nonlocal estimates of *m* and ν. A query point (ϕ,ψ) in which we estimate expectation and variance values of *a*, can be surrounded by an effective radius *r*, equal to half of a bandwidth, *b* of the kernel function, *K*. We can count the effective number of data points, N_{eff} within the radius, *r*, around any query point. These points will have an impact on the estimated local values of *m* and ν.

We define the bandwidth, $b(\kappa)$ as a diameter of the circle centered on the query point (ϕ_0,ψ_0) within which the von Mises kernel function integrates to 68.2% (the value of integral of the normal distribution probability density function within one standard deviation from its center):

$$\int_{\sqrt{\phi^2+\psi^2}<b(\kappa)} K(\phi - \phi_0, \psi - \psi_0)d\varphi d\psi = 0.682.$$

The bandwidth of the von Mises kernel at κ = 50 is approximately 16°.

In order to depict the trends of $\widehat{m}(\phi,\psi)$ and $\widehat{v}(\phi,\psi)$, we only plot their estimates at ϕ,ψ grid points where $N_{eff}(\phi,\psi) \geq 3$ within a circle with a diameter equal to the bandwidth $b(\kappa = 50) = 16°$.

In the sparsely populated areas of the ϕ,ψ map the threshold of at least 3 data points within the effective bandwidth may lead to estimates with high standard errors of mean (SEM). We calculated an estimate of SEM, as

$$SEM(a|\phi,\psi) = \sqrt{\frac{v(\phi,\psi)}{N_{eff}(\phi,\psi)}}.$$

It is very important to analyze the trends of *m* and ν as a function of ϕ,ψ together with $SEM(a|\phi,\psi)$. The values of SEM will indicate the significance of the trend in the more sparsely populated areas.

Creation of the Binned Conformation-Dependent Library

To create a binned CDL for each residue class, averages and standard deviations were calculated in 10° × 10° bins in Φ,Ψ. . The results were stored in a set of files, one per residue class. Python scripts provide an interface to the CDL, allowing easy retrieval of the conformation-dependent values when given a residue name and conformation. Additional tools building upon this simple interface are also part of the distributed code, including a tool that will compare the bond angles and lengths in any PDB coordinate file with CDL values, EH values, or another PDB coordinate file. The CDL and accessory tools are available under an open-source license from http://proteingeometry.sourceforge.net/. The CDL is also available at http://dunbrack.fccc.edu/nmhrcm/.

Molecular Mechanics Calculations

MM-derived context-dependent values for bond angles for two test cases (PDB codes 1rge [Sevcik et al., 1996] and 1r6j [Kang et al., 2004]) were generated using the following protocol: the structures were minimized in CHARMM (Brooks et al., 1983) using the parm_all22_prot force field with the CMAP correction (Mackerell, 2004) using the GBMV implicit solvent model (Lee et al., 2003). The protocol used cycles of 100 steps of steepest-descent minimization with heavy-atom restraints of 5, 3, 1, and 0 × atomic mass kcal/mol/Å^2. Following the last cycle (which had no restraints), 1000 steps of adopted basis Newton-Raphson minimization were performed, and the typical gradient root mean square was about 0.05 kcal/mol/Å.

CDL Assessments

Building Ideal Models and Analysis of Nonbonded Interactions

Ideal peptides with EH or CDL backbone geometry were built using PyRosetta (http://pyrosetta.org/), Python bindings to Rosetta (Rohl et al., 2004). To account for the length dependence of rmsd calculations (e.g., Holmes and Tsai, 2004), we linearly rescaled all rmsds to those of 100-residue proteins using the EH rmsds and the assumption that rmsds intersect the origin. Based on the linear fit of EH rmsds versus length produced, we calculated a scaling factor of (0.0332519/100) / (0.0332519/length). To understand the structural basis of variations between these theoretical peptides, van der Waals clashes were visually analyzed using the Coot (Emsley and Cowtan, 2004) interface to MolProbity (Davis et al., 2007).

Crystal Structure ∠NC$_\alpha$C Angles

Nonredundant structures with a 25% sequence-identity threshold were taken from PDBSelect (Hobohm and Sander, 1994). Among these, 50 structures were selected from each of five resolution ranges: 1.0–1.1 Å, 1.5–1.6 Å, 2.0–2.1 Å, 2.5–2.6 Å, 3.0–3.1 Å. For each residue in these structures, we then calculated the difference in the observed ∠NC$_\alpha$C and the CDL value. These were used to calculate the per-structure rmsds, which were then used to calculate averages, standard deviations, and standard errors of the mean for each of the five resolution shells.

SUPPLEMENTAL DATA

Supplemental Data include thirteen figures and can be found with this article online at http://www.cell.com/structure/supplemental/S0969-2126(09)00335-9.

ACKNOWLEDGMENTS

We thank Charles L. Brooks III (University of Michigan) for performing the MM minimizations used in this study. We additionally thank the David Baker lab (University of Washington at Seattle), in particular Srivatsan Raman, James Thompson, and Elizabeth Kellogg, for their help with Rosetta. We thank Jeffrey Gray (Johns Hopkins University) for providing PyRosetta, the Python bindings to Rosetta. We thank Lothar Schäfer (University of Arkansas) for providing a database of QM-calculated dipeptides and an extrapolation program to obtain values for conformation-dependent bond angles and lengths. This work was supported in part by National Institutes of Health (NIH) grant R01-GM083136 (to P.A.K.), National Science Foundation grant MCB-9982727 (to P.A.K.), and NIH grant P20-GM76222 (to R.L.D.).

Received: July 20, 2009
Revised: July 20, 2009
Accepted: August 20, 2009
Published: October 13, 2009

REFERENCES

Adams, P.D., Grosse-Kunstleve, R.W., Hung, L.W., Ioerger, T.R., McCoy, A.J., Moriarty, N.W., Read, R.J., Sacchettini, J.C., Sauter, N.K., and Terwilliger, T.C. (2002). PHENIX: building new software for automated crystallographic structure determination. Acta Crystallogr. D Biol. Crystallogr. *58*, 1948–1954.

Baldwin, R.L., and Rose, G.D. (1999). Is protein folding hierarchic? I. Local structure and peptide folding. Trends Biochem. Sci. *24*, 26–33.

Berkholz, D.S., Faber, H.R., Savvides, S.N., and Karplus, P.A. (2008). Catalytic cycle of human glutathione reductase near 1 A resolution. J. Mol. Biol. *382*, 371–384.

Bradley, P., Misura, K.M.S., and Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. Science *309*, 1868–1871.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. *4*, 187–217.

Corey, R.B., and Donohue, J. (1950). Interatomic distances and bond angles in the polypeptide chain of proteins. J. Am. Chem. Soc. *72*, 2899–2900.

Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., Wang, X., Murray, L.W., Arendall, W.B., 3rd, Snoeyink, J., Richardson, J.S., and Richardson, D.C. (2007). MolProbity:all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res. *35*, W375–W383.

Dunbrack, R.L., and Karplus, M. (1994). Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nat. Struct. Biol. *1*, 334–340.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. *60*, 2126–2132.

Engh, R.A., and Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structure refinement. Acta Crystallogr. A Found. Crystallogr. *47*, 392–400.

Engh, R.A., and Huber, R. (2001). International Tables for Crystallography. In International Tables for Crystallography, M.G. Rossmann and E. Arnold, eds. (Dordrecht, The Netherlands: Kluwer Academic Publishers), pp. 382–392.

Esposito, L., Vitagliano, L., Zagari, A., and Mazzarella, L. (2000). Experimental evidence for the correlation of bond distances in peptide groups detected in ultrahigh-resolution protein structures. Protein Eng. *13*, 825–828.

Evans, P.R. (2007). An introduction to stereochemical restraints. Acta Crystallogr. D Biol. Crystallogr. *63*, 58–61.

Feig, M. (2008). Is alanine dipeptide a good model for representing the torsional preferences of protein backbones? J. Chem. Theory Comput. *4*, 1555–1564.

Gunasekaran, K., Ramakrishnan, C., and Balaram, P. (1996). Disallowed Ramachandran conformations of amino acid residues in protein structures. J. Mol. Biol. *264*, 191–198.

Ho, B.K., Thomas, A., and Brasseur, R. (2003). Revisiting the Ramachandran plot: Hard-sphere repulsion, electrostatics, and H-bonding in the α-helix. Protein Sci. *12*, 2508–2522.

Hobohm, U., and Sander, C. (1994). Enlarged representative set of protein structures. Protein Sci. *3*, 522–524.

Hollingsworth, S.A., Berkholz, D.S., and Karplus, P.A. (2009). On the occurrence of linear groups in proteins. Protein Sci. *18*, 1321–1325.

Holmes, J.B., and Tsai, J. (2004). Some fundamental aspects of building protein structures from fragment libraries. Protein Sci. *13*, 1636–1650.

Hurley, J.H., Mason, D.A., and Matthews, B.W. (1992). Flexible-geometry conformational energy maps for the amino acid residue preceding a proline. Biopolymers *32*, 1443–1446.

Jaskolski, M., Gilski, M., Dauter, Z., and Wlodawer, A. (2007a). Numerology versus reality: a voice in a recent dispute. Acta Crystallogr. D Biol. Crystallogr. *63*, 1282–1283.

Jaskolski, M., Gilski, M., Dauter, Z., and Wlodawer, A. (2007b). Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them? Acta Crystallogr. D Biol. Crystallogr. *63*, 611–620.

Jiang, X., Yu, C., Cao, M., Newton, S.Q., Paulus, E.F., and Schäfer, L. (1997). ø/ψ-Torsional dependence of peptide backbone bond-lengths and bond-angles: comparison of crystallographic and calculated parameters. J. Mol. Struct. *403*, 83–93.

Kang, B.S., Devedjiev, Y., Derewenda, U., and Derewenda, Z.S. (2004). The PDZ2 domain of syntenin at ultra-high resolution: bridging the gap between macromolecular and small molecule crystallography. J. Mol. Biol. *338*, 483–493.

Karplus, P.A. (1996). Experimentally observed conformation-dependent geometry and hidden strain in proteins. Protein Sci. *5*, 1406–1420.

Karplus, P.A., Shapovalov, M.V., Dunbrack, R., Jr., and Berkholz, D.S. (2008). A forward-looking suggestion for resolving the stereochemical restraints debate: ideal geometry functions. Acta Crystallogr. D Biol. Crystallogr. *64*, 335–336.

Kleywegt, G.J., and Jones, T.A. (1996). Phi/psi-chology: Ramachandran revisited. Structure *4*, 1395–1400.

Laskowski, R.A., Chistyakov, V.V., and Thornton, J.M. (2005). PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. Nucleic Acids Res. *33*, D266–D268.

Lee, M.S., Feig, M., Salsbury, F.R., and Brooks, C.L. (2003). New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. J. Comput. Chem. *24*, 1348–1356.

Longhi, S., Czjzek, M., and Cambillau, C. (1998). Messages from ultrahigh resolution crystal structures. Curr. Opin. Struct. Biol. *8*, 730–737.

Lovell, S.C., Davis, I.W., Arendall, W.B., III, de Bakker, P.I.W., Word, J.M., Prisant, M.G., Richardson, J.S., and Richardson, D.C. (2003). Structure validation by Ca geometry: φ, ψ and Cβ deviation. Proteins *50*, 437–450.

Mackerell, A.D. (2004). Empirical force fields for biological macromolecules: overview and issues. J. Comput. Chem. *25*, 1584–1604.

Mardia, K.V., and Zemrock, P.J. (1975). Algorithm AS 86: The Von Mises distribution function. Appl. Stat. *24*, 268–272.

Naradaya, E. (1964). On estimating regression. Theory Probab. Appl. *9*, 141–142.

Pauling, L., Corey, R.B., and Branson, H.R. (1951). The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl. Acad. Sci. USA *37*, 205–211.

Ramakrishnan, C., Lakshmi, B., Kurien, A., Devipriya, D., and Srinivasan, N. (2007). Structural compromise of disallowed conformations in peptide and protein structures. Protein Pept. Lett. *14*, 672–682.

Rohl, C.A., Strauss, C.E.M., Misura, K.M.S., and Baker, D. (2004). Protein structure prediction using Rosetta. Methods Enzymol. *383*, 66–93.

Schäfer, L., and Cao, M. (1995). Predictions of protein backbone bond distances and angles from first principles. J. Mol. Struct. *333*, 201–208.

Sevcik, J., Dauter, Z., Lamzin, V.S., and Wilson, K.S. (1996). Ribonuclease from Streptomyces aureofaciens at Atomic Resolution. Acta Crystallogr. D Biol. Crystallogr. *52*, 327–344.

Sheldrick, G.M. (2008). A short history of SHELX. Acta Crystallogr. A. Found. Crystallogr. *64*, 112–122.

Stec, B. (2007). Comment on Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them? by Jaskolski, Gilski, Dauter and Wlodawer (2007). Acta Crystallogr. D Biol. Crystallogr. *63*, 1113–1114.

Tickle, I.J. (2007). Experimental determination of optimal root-mean-square deviations of macromolecular bond lengths and angles from their restrained ideal values. Acta Crystallogr. D Biol. Crystallogr. *63*, 1274–1281.

Yu, C.H., Norman, M.A., Schäfer, L., Ramek, M., Peeters, A., and van Alsenoy, C. (2001). Ab initio conformational analysis of N-formyl L-alanine amide including electron correlation. J. Mol. Struct. *567*, 361–374.

Zhang, Y. (2009). Protein structure prediction: when is it useful? Curr. Opin. Struct. Biol. *19*, 145–155.

Supplemental Data

Conformation Dependence

of Backbone Geometry in Proteins

Donald S. Berkholz, Maxim V. Shapovalov, Roland L. Dunbrack, Jr., and P. Andrew Karplus

Figure S1.
Conformation-dependent variation in bond angles for Ile/Val residues as a function of the Φ,Ψ of the central residue. Otherwise, all is as in Figure 3.

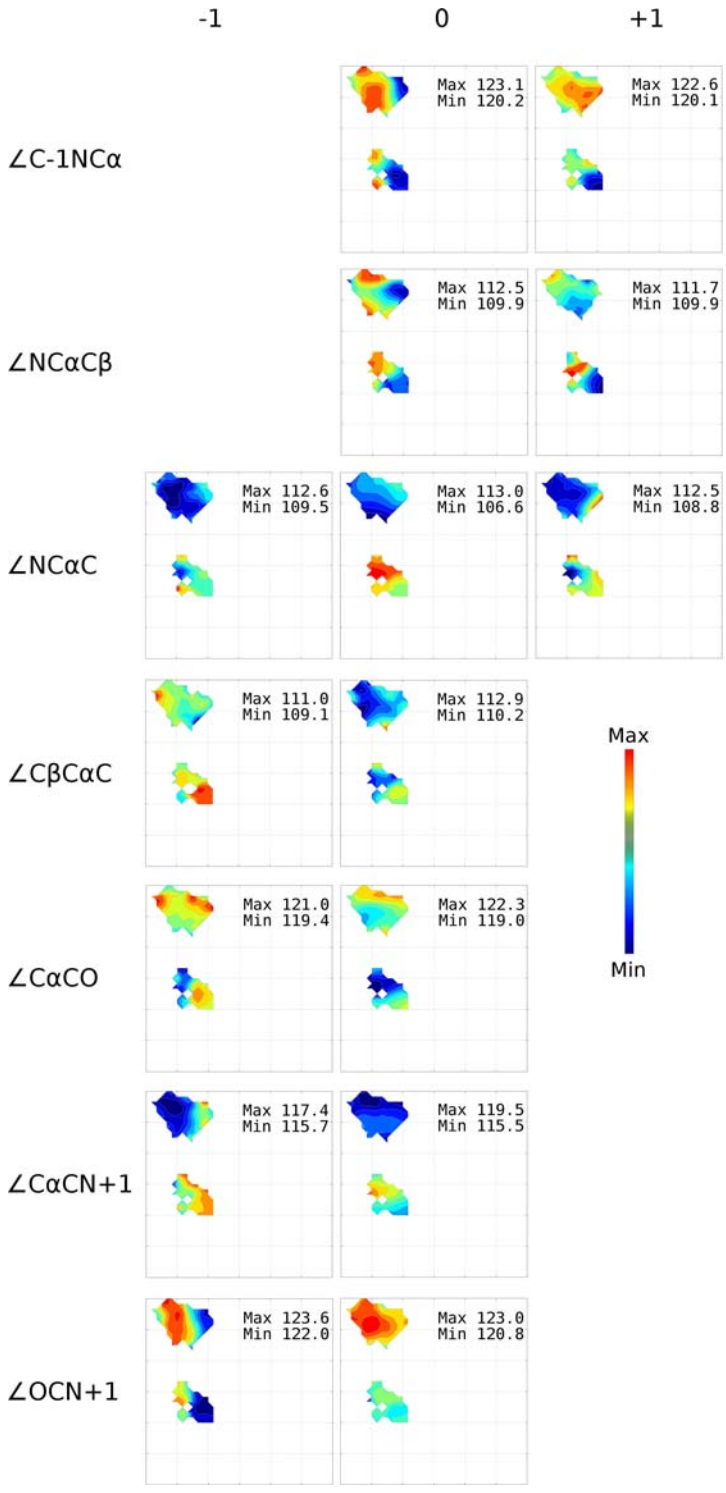


Figure S2. Conformation-dependent variation in bond angles for Pro residues as a function of the Φ,Ψ of the central residue. Otherwise, all is as in Figure 3.

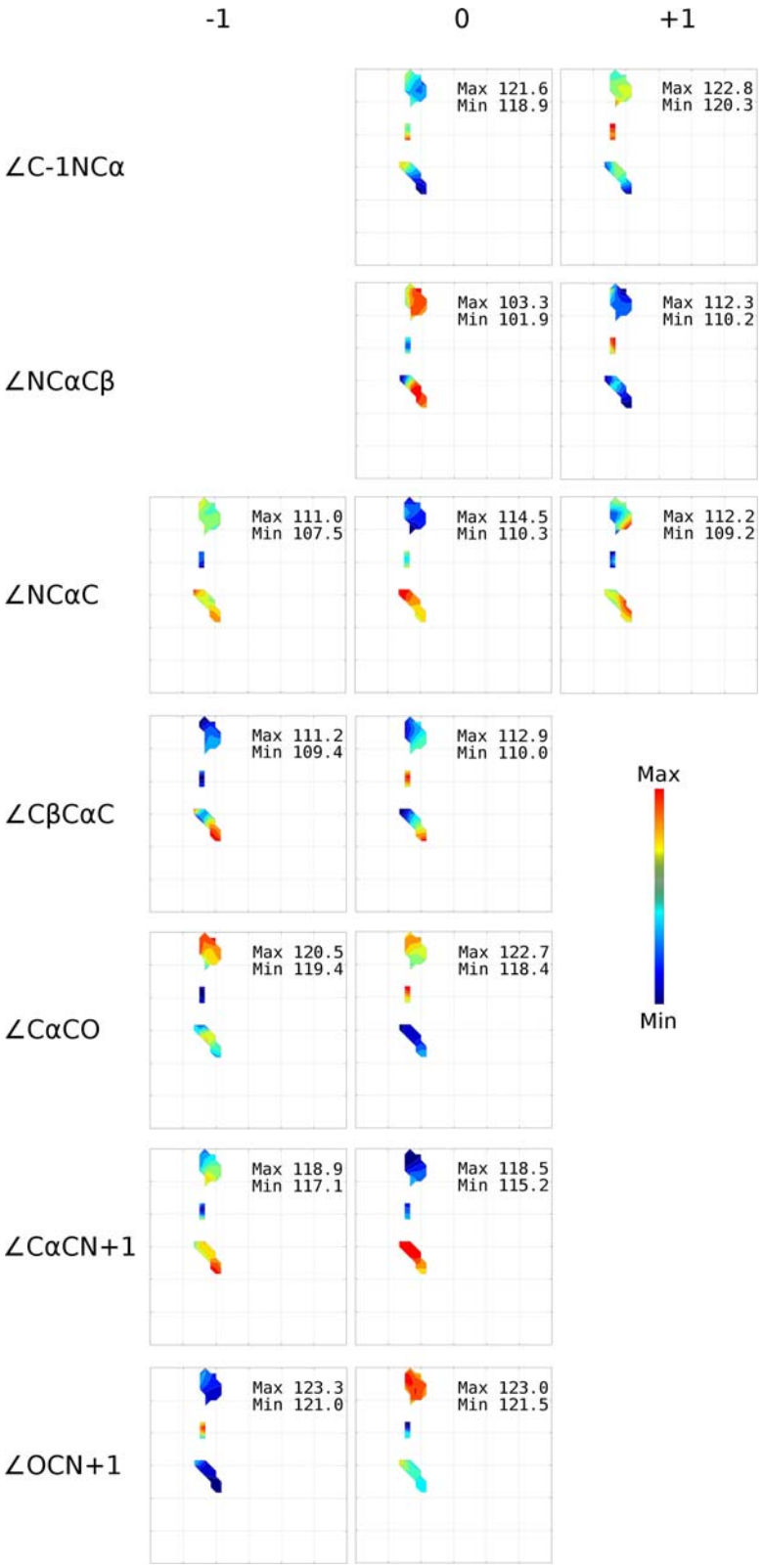


Figure S3. Conformation-dependent variation in bond angles for Gly residues as a function of the Φ,Ψ of the central residue. Otherwise, all is as in Figure 3.

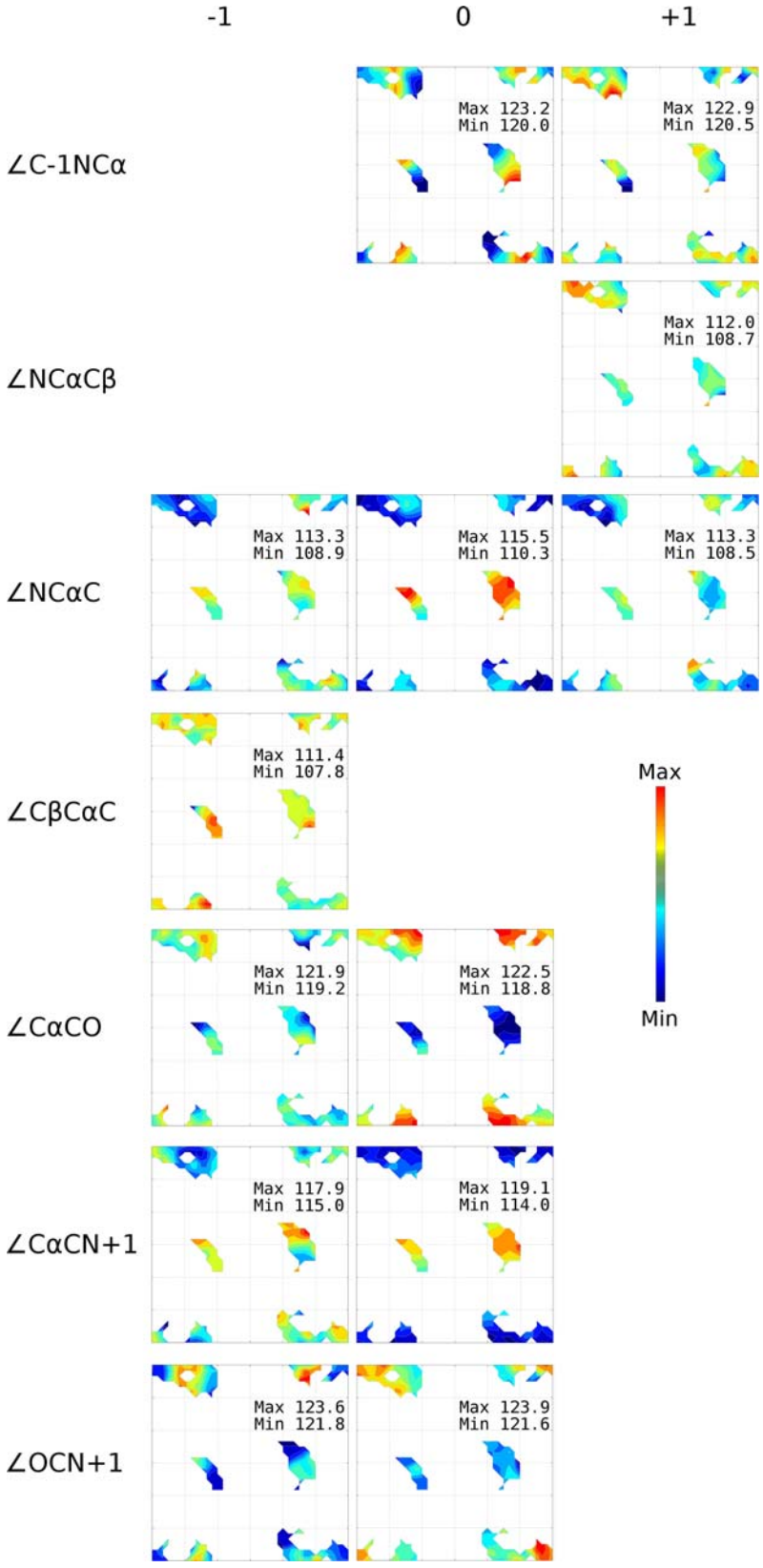


Figure S4. Conformation-dependent variation in bond angles for general prePro (i.e., nonGly, nonPro, nonIle/Val residues preceding proline) residues as a function of the Φ,Ψ of the central residue. Otherwise, all is as in Figure 3.

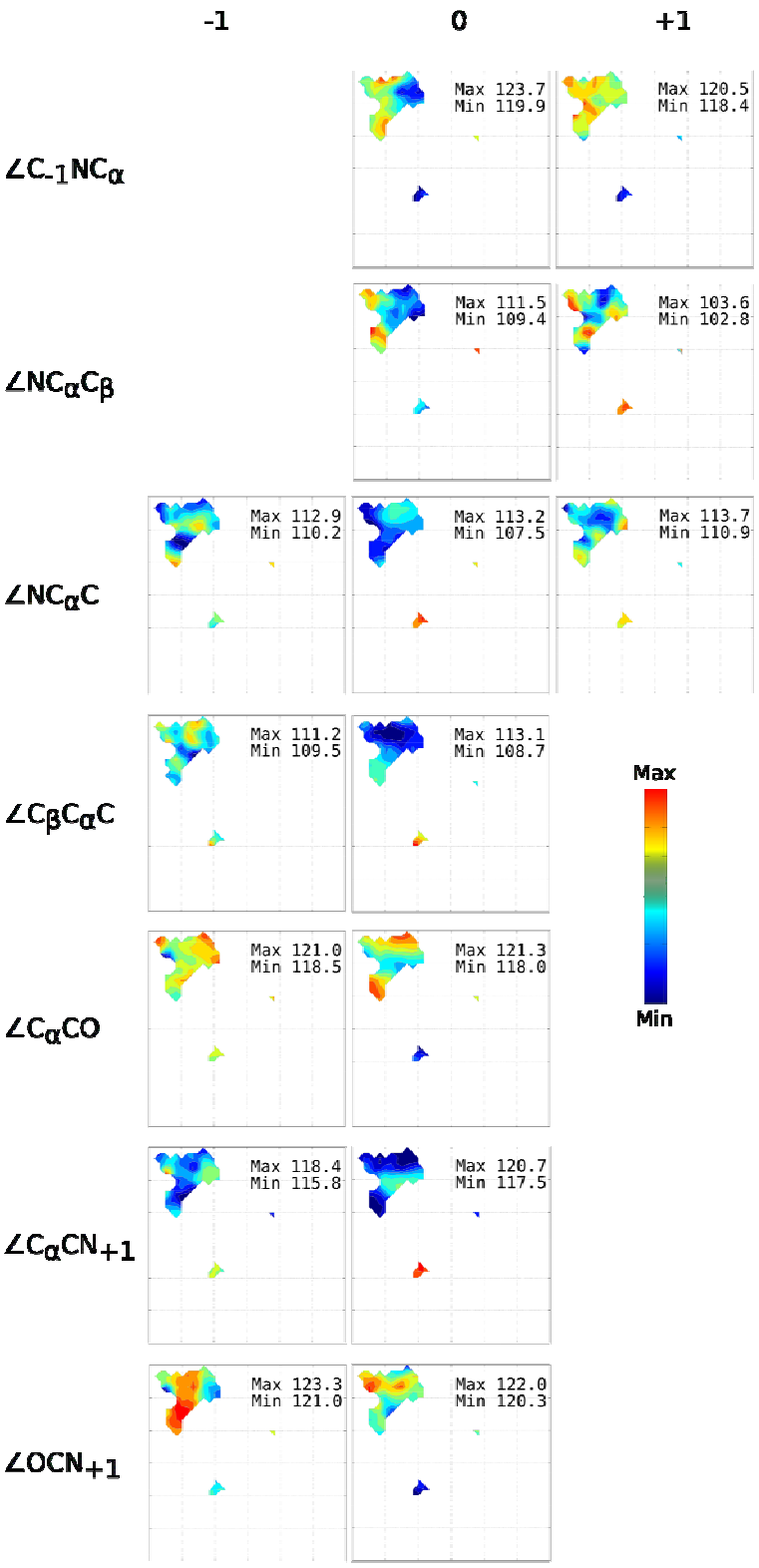


Figure S5. Conformation-dependent variation in the standard errors of the means of bond angles for general residues as a function of the Φ,Ψ of the central residue. Otherwise, all is as in Figure 3.

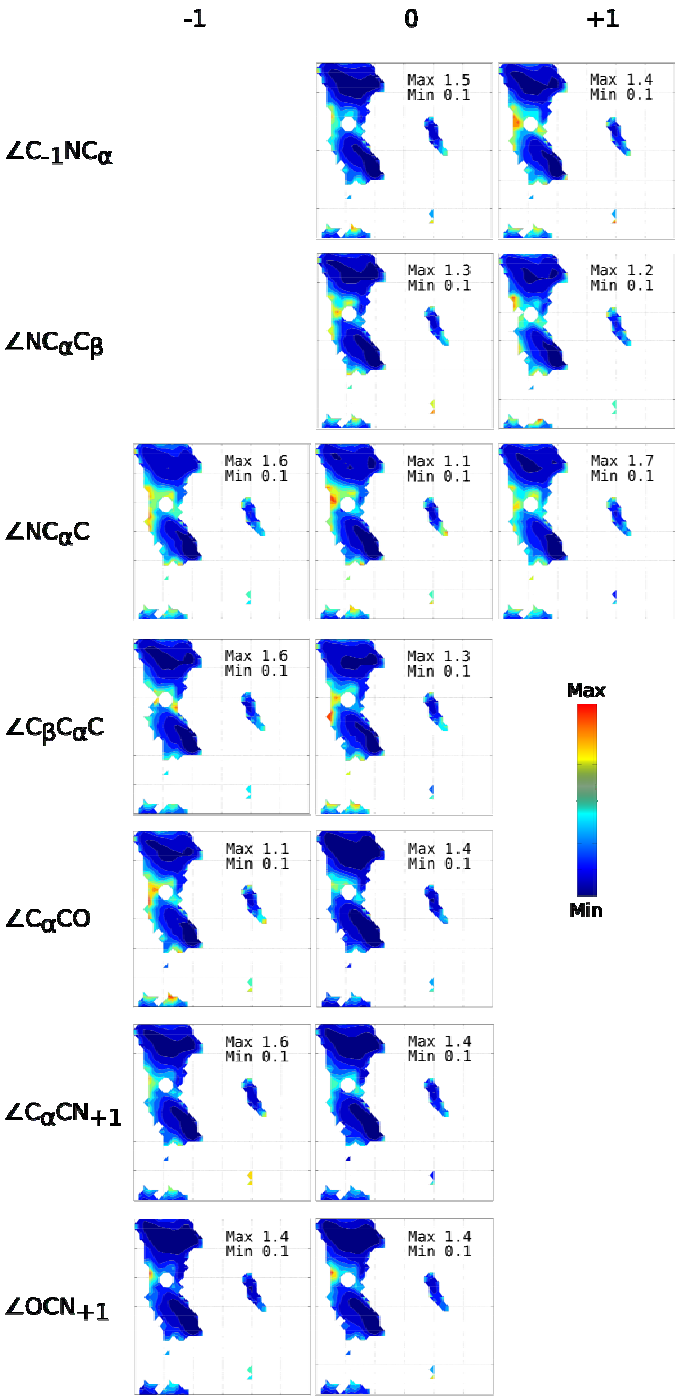


Figure S6. Conformation-dependent variation in the standard errors of the means of bond angles for Ile/Val residues as a function of the Φ,Ψ of the central residue. Otherwise, all is as in Figure 3.

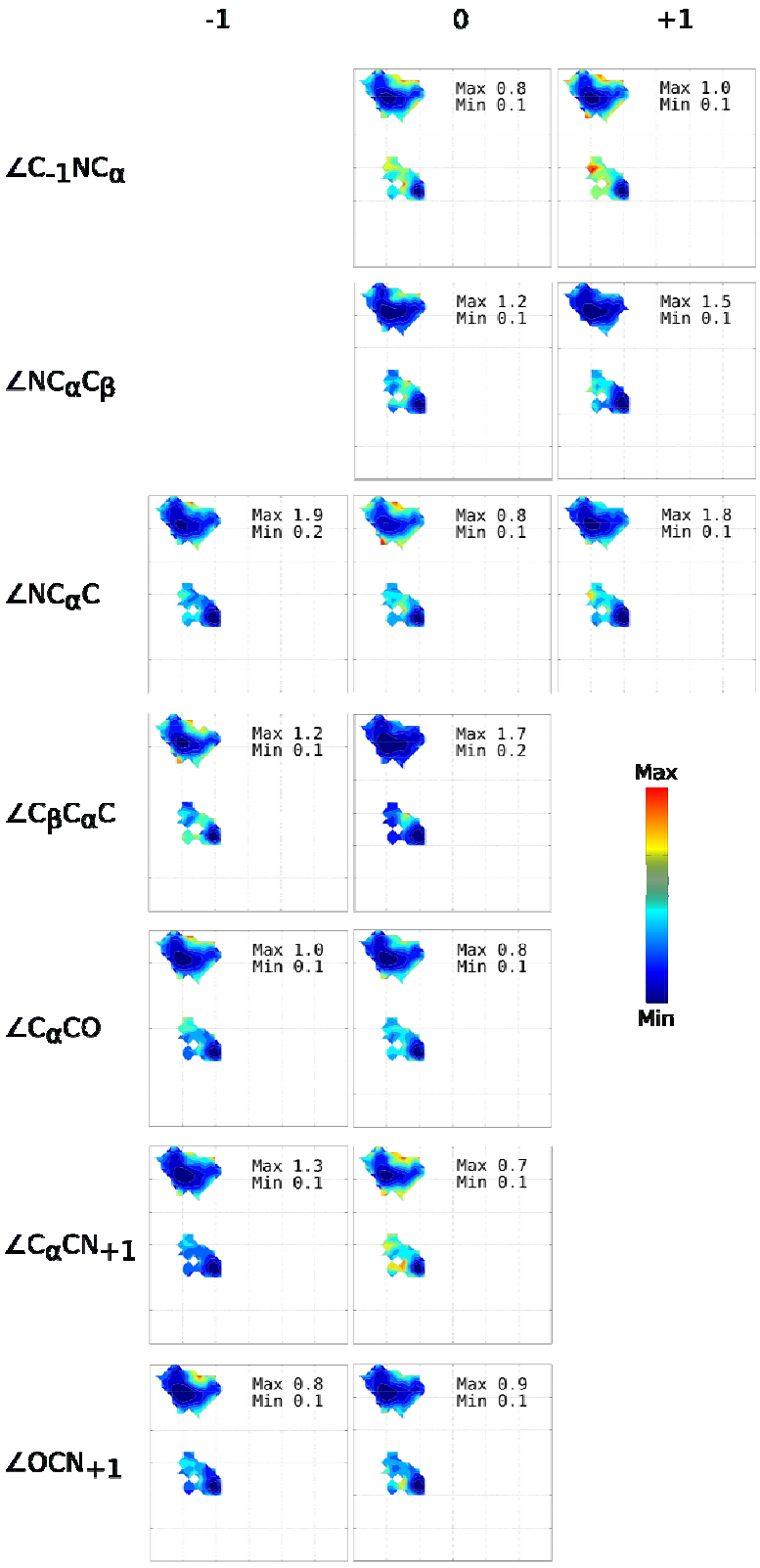


Figure S7. Conformation-dependent variation in the standard errors of the means of bond angles for Pro residues as a function of the Φ,Ψ of the central residue. Otherwise, all is as in Figure 3.

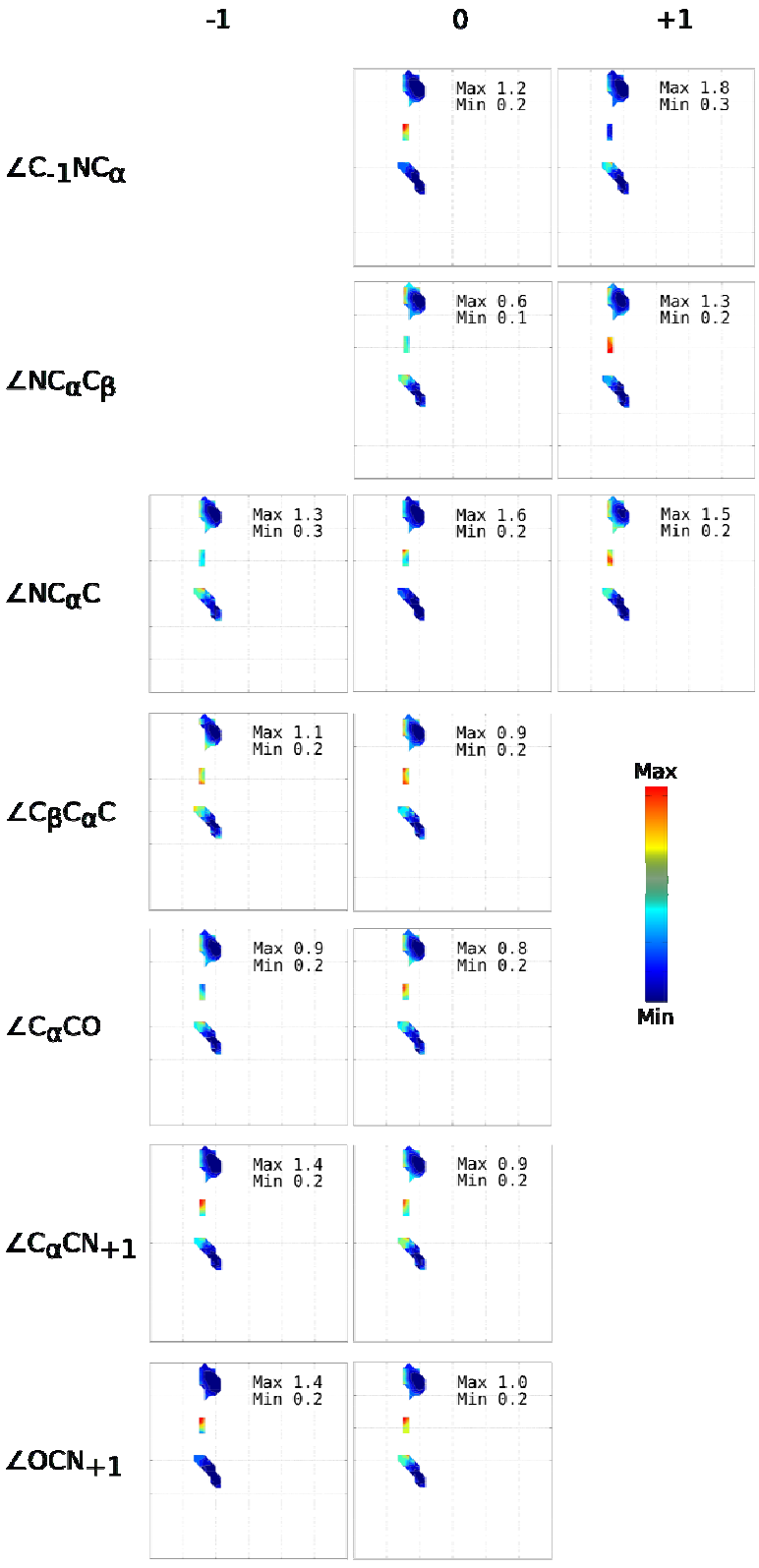


Figure S8. Conformation-dependent variation in the standard errors of the means of bond angles for Gly residues as a function of the Φ,Ψ of the central residue. Otherwise, all is as in Figure 3.

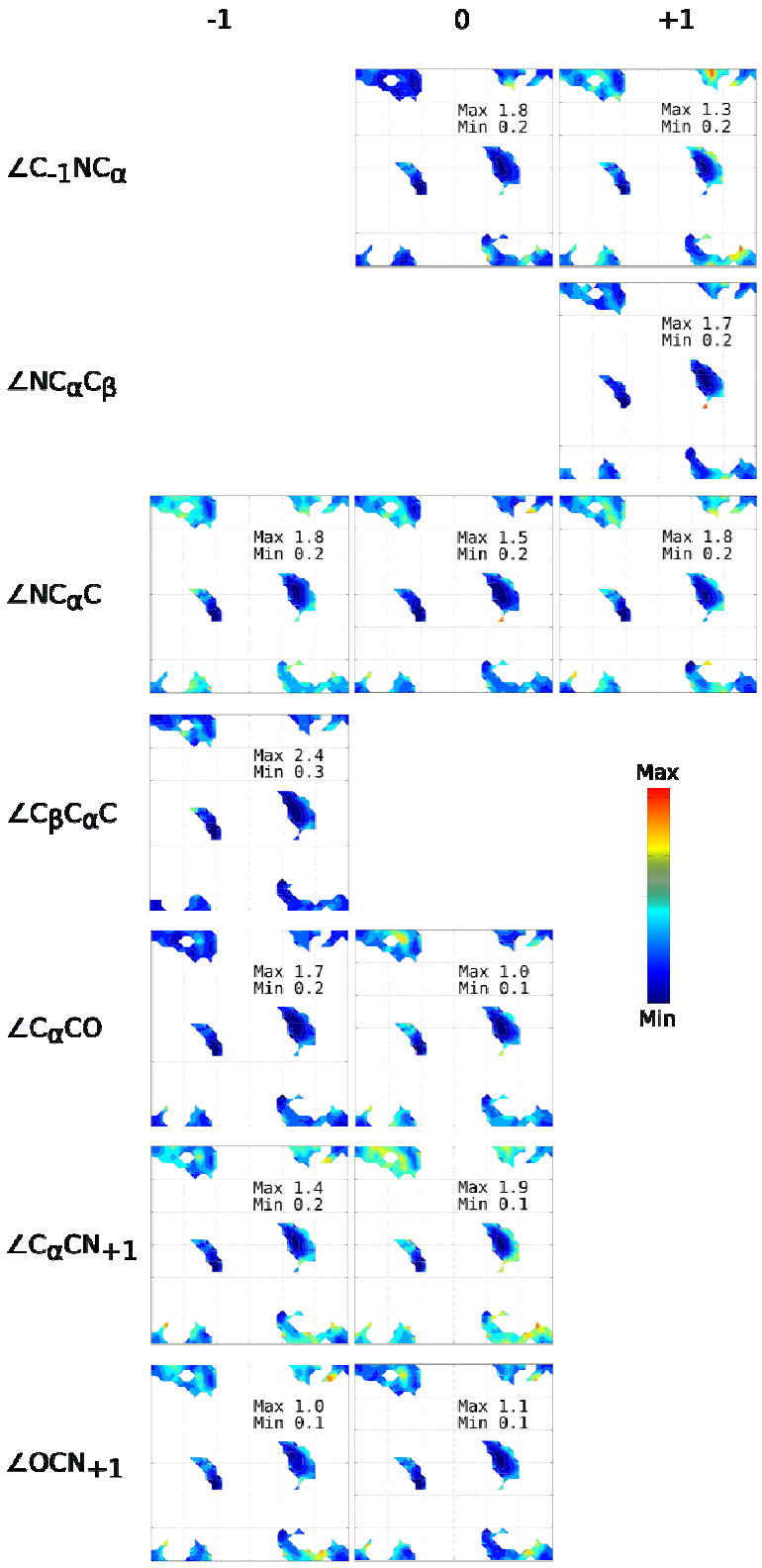
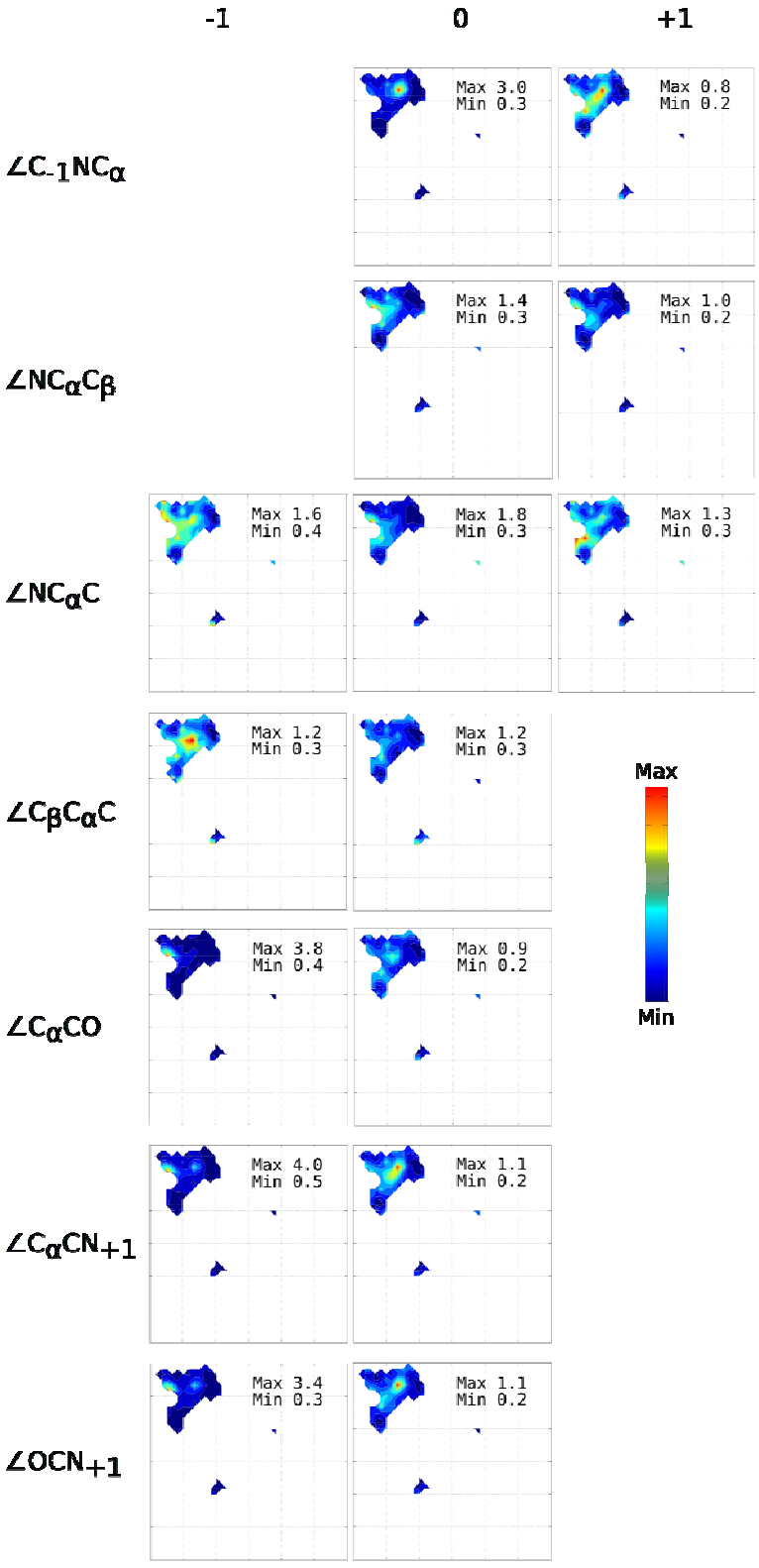


Figure S9. Conformation-dependent variation in the standard errors of the means of bond angles for general prePro (i.e., nonGly, nonPro, nonIle/Val residues preceding proline) residues as a function of the Φ,Ψ of the central residue. Otherwise, all is as in Figure 3.



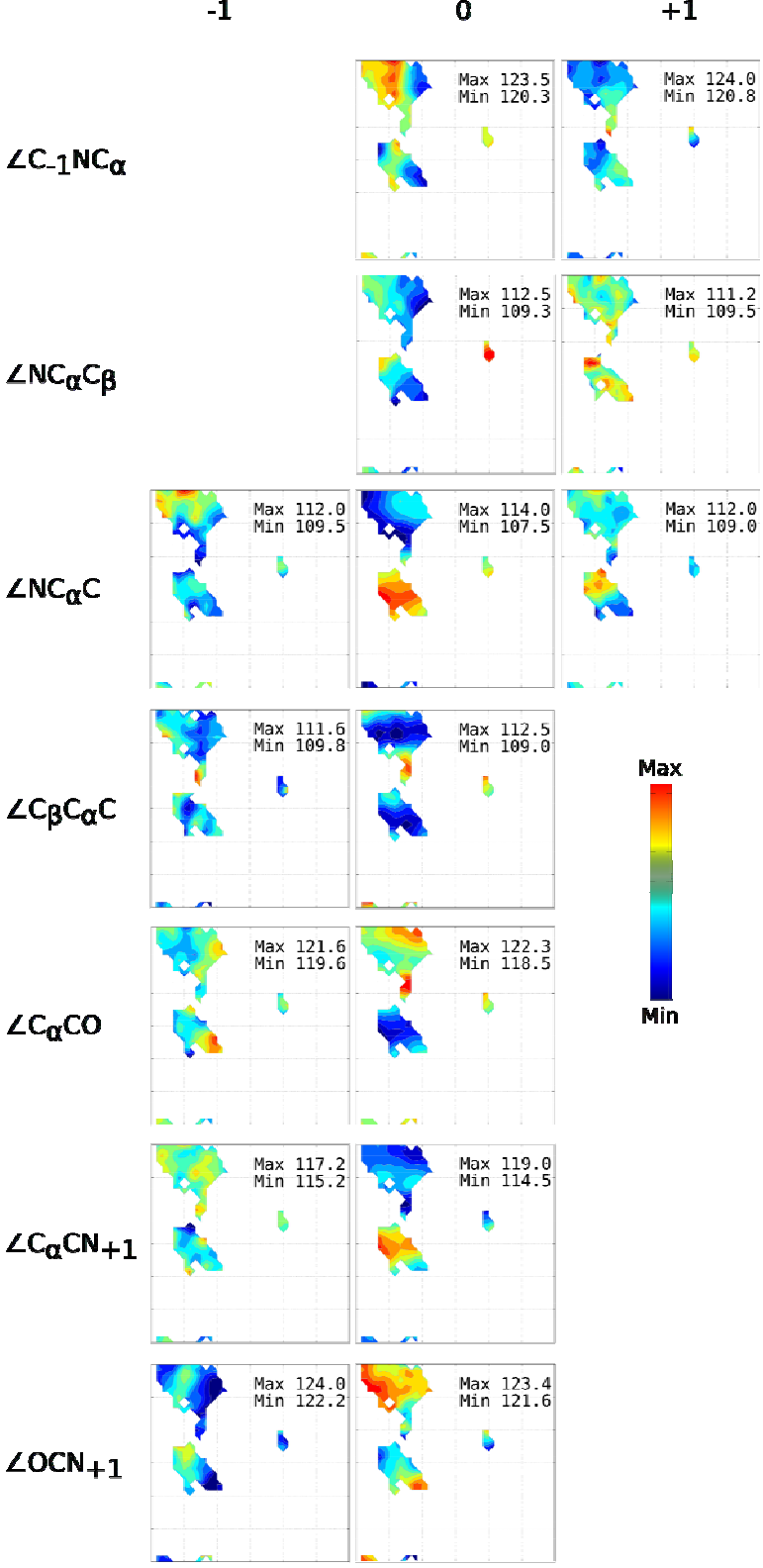


Figure S10. Conformation-dependent variation in bond angles for general residues without defined secondary structure as a function of the Φ,Ψ of the central residue. The lack of secondary structure is defined by DSSP codes 'S' or '-'. The calculations used 1,342 residues. Otherwise, all is as in Figure 3.

Figure S11. Conformation-dependent variation in the standard errors of the means of bond angles for general residues without defined secondary structure as a function of the Φ,Ψ of the central residue. The lack of secondary structure is defined by DSSP codes 'S' or '-'. The calculations used 1,342 residues. Otherwise, all is as in Figure 3.

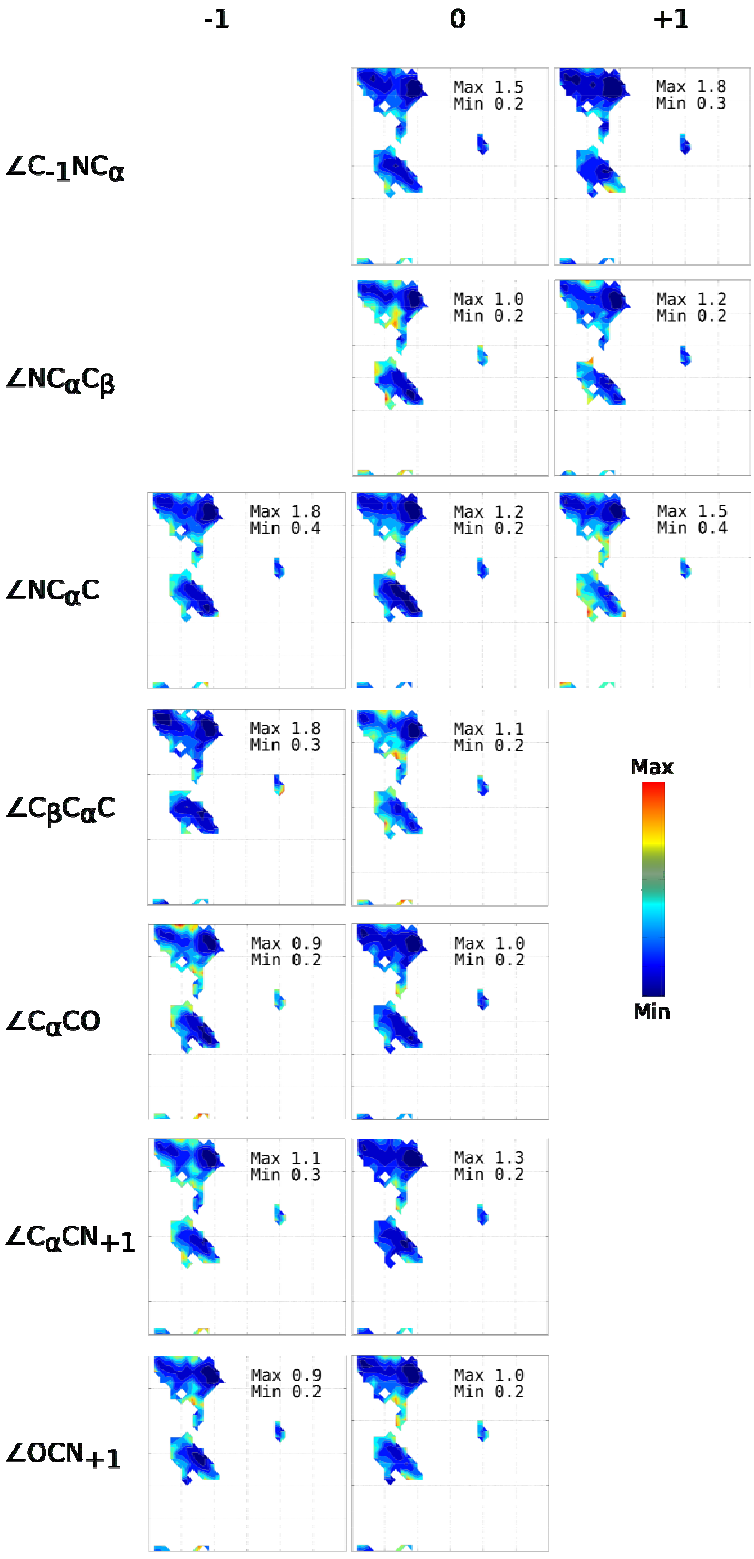


Figure S12. Protein backbone conformations of non-Gly residues. This is identical to Figure 2 but unlabeled.

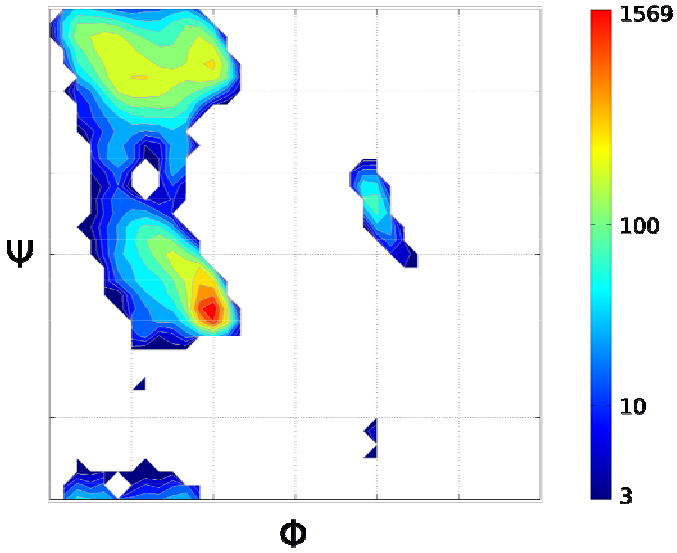


Figure S13. Protein backbone conformations of Gly residues. This is otherwise identical to Figure S12.