BioMol Concepts, Vol. 1 (2010), pp. 271–283 • Copyright © by Walter de Gruyter • Berlin • New York. DOI 10.1515/BMC.2010.022

Review

A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins

Scott A. Hollingsworth and P. Andrew Karplus*

Department of Biochemistry and Biophysics, Oregon State University, Corvallis, OR 97331, USA

*Corresponding author
e-mail: karplusp@science.oregonstate.edu

Abstract

The Ramachandran plot is among the most central concepts in structural biology, seen in publications and textbooks alike. However, with the increasing numbers of known protein structures and greater accuracy of ultra-high resolution protein structures, we are still learning more about the basic principles of protein structure. Here, we use high-fidelity conformational information to explore novel ways, such as geo-style and wrapped Ramachandran plots, to convey some of the basic aspects of the Ramachandran plot and of protein conformation. We point out the pressing need for a standard nomenclature for peptide conformation and propose such a nomenclature. Finally, we summarize some recent conceptual advances related to the building blocks of protein structure. The results for linear groups imply the need for substantive revisions in how the basics of protein structure are handled.

Keywords: linear groups; polypeptide conformation; Ramachandran plot, secondary structure; torsion angles.

Introduction

The use of torsion angles to describe polypeptide and protein conformation was developed by Sasisekharan as part of his studies of the structure of collagen chains during his work as a graduate student in the research group of G.N. Ramachandran (1). The power of this approach was readily apparent and its use quickly became widespread (2, 3). Using revised definitions, this so-called Ramachandran plot or φ, ψ-plot has remained nearly unchanged in the ensuing 50 years and continues to be an integral tool for protein structure research and education (4).

It is from the origins of this unmistakable plot, based on treating atoms as simple impenetrable spheres (2, 4), that we get the widespread conceptual framework that alanine-like residues (those residues that are not glycine nor proline) can occupy one of three major 'allowed' regions (Figure 1): two larger regions known as the alpha-region and the beta-region (that include conformations giving rise to α-helices and β-strands, respectively), and a much smaller $alpha_L$-region representing backbone conformations that are mirror images of those in the alpha-region. These regions are 'allowed' in the sense that when the peptide atoms are given standard radii they do not collide. An additional region, sometimes called the bridge region (because it bridges the alpha- and beta-regions) becomes allowed if the atoms are given smaller radii that represent the smallest values that could be considered plausible (Figure 1).

Overlaid on the classically allowed regions shown in Figure 1 is a set of approximately 60 000 observations of alanine-like residues that are well defined in protein structures by X-ray crystallography determined at $\leq$1.2 Å resolution. The high-fidelity sets of residues used in Figure 1 and elsewhere in this paper were selected using the Protein Geometry Database [(7); http://pgd.science.oregonstate.edu]. What is readily apparent in Figure 1 is that the majority of observations do fall into three major clusters closely associated with the alpha-, beta- and $alpha_L$-regions, but with features and limits that do not completely match with the classically defined boundaries of the 'allowed' regions. Given this discrepancy, now instead of using the originally predicted allowed and disallowed regions, Ramachandran plots with regions based on empirical distributions are used to assess the stereochemical quality of solved crystal structures as a part of validation tools such as PROCHECK (8), MOLEMAN2 (9), and more recently MOLPROBITY (10).

In addition, Biochemistry and Biophysics textbooks use Ramachandran plots to introduce students to the fundamentals of protein structure (5, 6). As the amount and accuracy of structural data has grown over the years, there has also been real growth in understanding the basics of protein conformation and some concepts that were once widely held have been displaced. Especially now, with the explosive growth of well-defined high-resolution crystal structures, there is the potential for greater understanding as one can study φ, ψ-distributions that are minimally influenced by experimental error. Here, we present some such high-fidelity plots derived from a dataset of approximately 72 000 well-ordered residues from a set of diverse protein structures determined at $\leq$1.2 Å resolution. We also present a few informative ways of viewing Ramachandran plots and make a proposal for a standard nomenclature to describe the most populated regions of the Ramachandran plot. Then, we

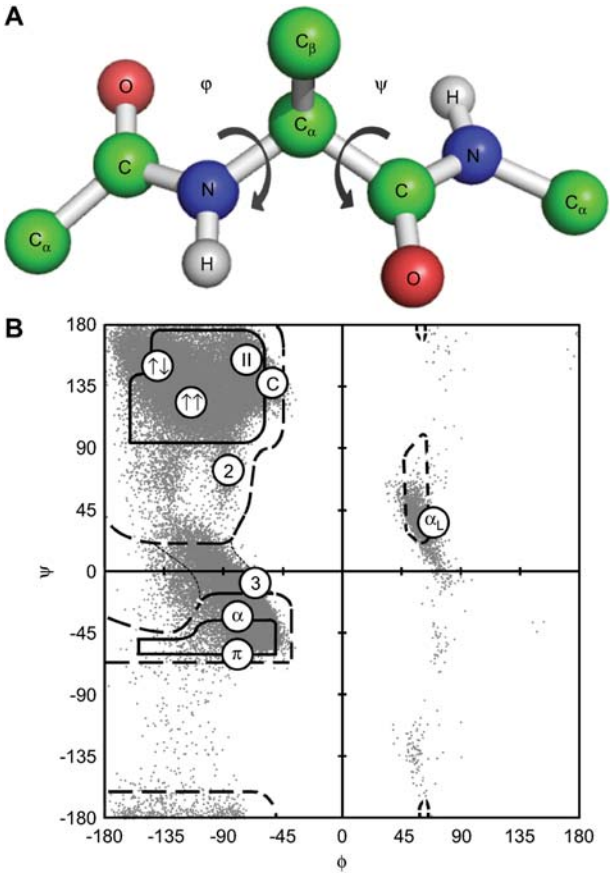

Figure 1 The classical Ramachandran or φ, ψ-plot.
(A) Ball and stick model of a dipeptide with a central Ala residue indicating the rotations defined by the torsion angles of φ and ψ. φ is defined by the torsion angle created by C_{i-1}-N_i-C_{α_i}-C_i and ψ is that defined by N_i-C_{α_i}-C_i-N_{i+1}. Figure created with PyMol (Schrödinger LLC, New York, USA). (B) The canonical Ramachandran plot from Ramachandran and Sasisekharan (4) with outlines defining the classically allowed (dashed lines), core allowed (solid lines) and extreme-limit allowed (dotted lines) regions for an Ala dipeptide. Also shown is a scatter plot of 63 149 Ala-like (non-Gly, non-Pro) residues from a diverse set of crystal structures determined at ≤1.2 Å resolution. The locations of linear groups as commonly presented in textbooks (5, 6) are also shown for the α-helix (α), 3_{10}-helix (3), π-helix (π), left-handed α-helix (α_L), 2.2_7 ribbon (2), polyproline-II (II), collagen (C), parallel β-sheet (↑↑) and anti-parallel β-sheet (↑↓).

summarize some recent developments in the understanding of standard structures such as linear groups in proteins.

Ramachandran plot visualization and nomenclature

A three-dimensional (3D) plot

To help visualize the features of high-fidelity Ramachandran plots, we have found it helpful to look beyond the common two-dimensional φ, ψ-plot, which for a large dataset does not serve to convey well the true nature of the distribution. In particular, when a large subset of the observations is found very narrowly distributed within one small region (such as occurs in the α-helical region), this is not well seen in the simple plot because the data points occlude one another. 3D versions of the Ramachandran plot, with the third dimension representing observations, have been used previously (11), but we show here that when it is colored as is familiar from maps of the physical features of the earth, it gives a much more compelling impression of the proportions of residues in the different parts of the Ramachandran plot (Figure 2). Using our high-fidelity dataset and including all 20 amino acid types, this 3D plot allows several interesting observations. Most obvious is the titanic and sharp peak resulting from residues in α-helices. This very sharp peak towers over every other portion of the Ramachandran plot, including the other portions of the classically defined alpha-region. This salient observation suggests that the classically defined alpha-region does not behave as a unit and so would be better defined as two regions, one corresponding to the α-helix and the other representing mostly what was originally called the bridge region. Focusing on the classically defined beta-region, this plot reveals a natural break into four regions: two main peaks that encompass the β-strand and what is known as the P_{II} (see below) conformation, and two low-

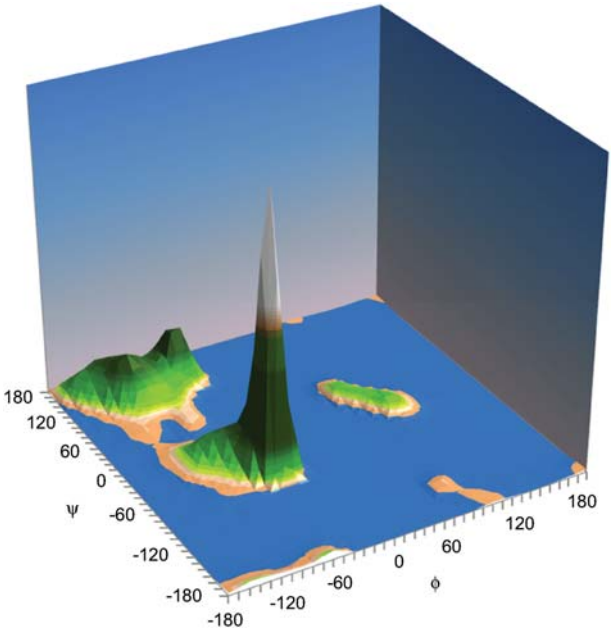

Figure 2 A geo-style 3D-Ramachandran plot.
A Ramachandran plot is shown with a third dimension representing the number of observations. For this plot, the 72 376 residue high-fidelity dataset (see text) was used to generate total numbers of observations in each 20°×20° bin centered every 10° in φ and ψ. The figure was then created using Excel 2007 (Microsoft, Redmond, WA, USA) defining every 50 observations a new vertical step. The coloring of the plot has been chosen to show off the natural patterns in the data and to provide a pleasing and admittedly fun, geographic-like representation. Major contour levels in observations are: blue ocean (0–49), sandy beach (50–149), vegetated region (150–4599), snowy peaks (4600 and higher).

lying peninsulas extending towards the alpha-region. Moving to the alpha$_L$-region, this 3D plot clearly shows that its distribution is not a mirror image of the alpha-region with a prominent peak and a lower plateau, but is simply a relatively evenly populated island. Finally, the plot reveals a lowly populated region located around φ=+60, ψ=-150 to -180. This latter region happens to be the mirror image conformation of the P_{II} region, just as the alpha$_L$-region is the mirror of the alpha-region. Indeed, any pair of points on the Ramachandran plot related by inversion symmetry through (0, 0) [i.e. (φ, ψ) and (-φ, -ψ) pairs] are mirror image conformations.

Ramachandran plot nomenclature

It has long been recognized that there are notable regions of the Ramachandran plot beyond the broadly defined alpha-, beta- and alpha$_L$-regions and over the years many different naming strategies have attempted to capture various important aspects of the plot. Figure 3 surveys a few of the varied nomenclatures found in the literature. Each paper had a rationale for its nomenclature, but the lack of adoption of a single nomenclature not only reflects the complexity of the plot but has also contributed to confusion even among structural biologists. This is especially confusing because the same designator has been used for completely different regions. Here, we present a proposal for a standard nomenclature that takes into account previous strategies, with the better-known names adopted for regions in the plot corresponding to the natural groupings seen in Figure 2, and also takes into account the mirror image aspects of certain conformations. As was initiated for turn types (3), we will use a prime symbol (′) to represent a mirror image conformation.

The five natural clusters of the plot that are already associated with a well-established name are residues forming α-helices, β-strands, P_{II}-spirals, γ-turns and γ'-turns. We propose the standard labels α, β, P_{II}, γ and γ' be used for these regions (Figure 4). In this way, the α-region encompasses a relatively small area centered around the canonical α-helix peak of φ, ψ=(-63, -43). This small region contains approximately one-third of residues in diverse proteins, as is easily visible in Figure 2. The β-region, not to be confused with the classical beta-region that made up the entire northwestern quadrant of the original Ramachandran plot, is the actual natural grouping encompassing the region that includes most residues forming β-strands. The P_{II}-region constitutes the right-hand portion of the classical beta-region and a natural break separates it from the population in the β-region (Figure 2). The P_{II} name as used by Perskie et al. (16) is better than the earlier used β_P descriptor (Figure 3A and B) because these residues are not generally found in β-sheets, but when repeated make what has been called the polyproline-II conformation (17) that we have suggested be renamed the polypeptide-II conformation (18). The P_{II} conformation being lumped together with β-structures is simply a historic mistake related to the under appreciation of the P_{II} conformation owing to its lack of regular backbone hydrogen bonding (19). The γ-region around φ, ψ=(+80, -80) is relatively rare, but was named early for the γ-turn proposed by Némethy and Printz (20) that has a specific NH_{i+2} to O_i backbone hydrogen bond. Its mirror image region centered at (φ, ψ)=(-80, +80), called γ', has the same single residue hydrogen bond turn geometry as described by Matthews (21) and, as is seen here (Figure 4), has been noted to be much more common than the γ-turn. It is ironic that in this case the more common conformation was given the 'mirrored' name γ'.

In addition to the five major regions described above, there is one additional region that has a consistently used name in the literature. This is the ε- or e-region (Figure 3A–C). This region, located in the sparsely populated upper and lower parts of the left-handed quadrant of the Ramachandran plot, reflects a fairly extended chain and is largely populated by glycine, the only amino acid that can readily adopt the conformations required for the region. The two apparent parts of the ε-region are actually continuous as is better shown by a Ramachandran plot is wrapped around new axes (see below). In our proposed nomenclature the ε-region encompasses a subregion that we suggest be designated P_{II}'.

With these regions named, the most populated unnamed region is within the classically defined alpha-region, for which the tail extends from the α-region and continues up and to the left at a 45° angle. As noted in the introduction section, this region has been called the bridge region as it bridges the broad alpha- and beta-regions (Figure 1). Karplus (22) showed this region is not only allowed but is even strongly favored in real protein structures because it is characterized by a $NH_{i+1} \rightarrow N_i$ 'π-peptide' interaction [see figure 6 in Karplus (22)] and an opening up of the N-C_α-C bond angle to relieve the collision. Berkholz et al. (23) have followed up that work to show that in all regions of the Ramachandran plot, the expected 'ideal' bond angles vary in a way that makes sense to optimize non-covalent interactions. This means that according to ultra-high resolution protein structures the paradigm of a single ideal geometry for the peptide unit is not correct. Instead, one must think in terms of an 'ideal geometry function', with the expected values for bond angles changing systematically as a function of φ and ψ (23).

In terms of their role in protein structure, residues in this bridge region are involved in a wide variety of turns including classical type I, type III and type II′ turns. Using the already present pattern of Greek letters, we follow the lead of Karplus (22) to choose the next available Greek letter δ for naming this populated cluster.

Given the conventions mentioned above, it would be sensible to adopt α', β', P_{II}' and δ' for naming the regions that have conformations that are the mirror images of α, β, P_{II} and δ, respectively. Looking at the natural distribution, however, several observations can be made. The first is that there are no large clusters of observations in what would be the α'- and β'-regions. In the α'-region, there are residues present but there is not a sharp peak that mirrors the α-region. Instead, there appears to be a smooth extension of the distribution of residues that are in the δ'-region (Figure 2). Therefore, our proposed nomenclature (Figure 4) does not label any region α' as that would be misleading. The β'-

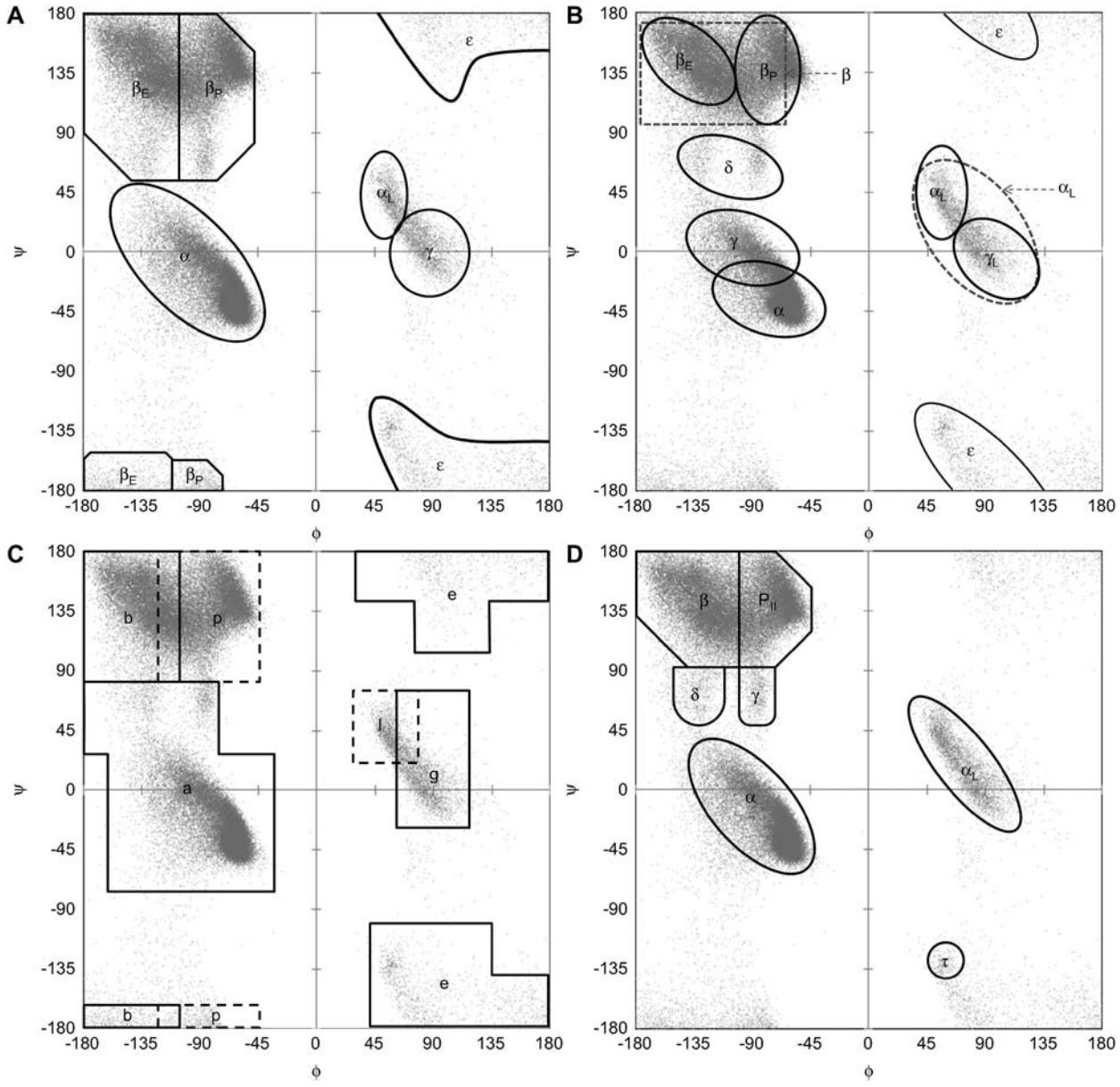

Figure 3 Representative published Ramachandran plot nomenclatures.
(A) Nomenclature from Wilmot and Thornton's 1990 landmark paper on β-turn conformations (12). (B) Nomenclature from Efimov's (13) work on common structures in proteins. (C) One example taken from Oliva et al. (14) of nomenclature used in an automated machine learning study; this is also similar to the nomenclature of structural states Rooman et al. (15) defined in an approach to simplify protein structure prediction. (D) Nomenclature by Perskie et al. (16) investigating the coil library. In each panel, the scatter plot included is based on all 72 376 residues in the high-fidelity dataset.

region overlays with the ε-region (Figure 5) and is occupied diffusely by Gly residues, although they do not cluster in the center of the region and thus β′ does not seem the best designator and is not proposed.

In contrast, the δ′- and P_{II}′-regions do correspond to natural clusters of residues and thus these given names are indeed appropriate to describe the mirrored regions. We extend the δ′-region for the reasons given above to include what in terms of φ, ψ angles is the mirror to the α conformation (what would be the α′-region). This also emphasizes (consistent with what is known) that even the α conformation itself will contain residues that are not present in α-helices and in that sense conceptually belong to an extended δ-region. The P_{II}′-region has not generally been discussed as a mirror image of the P_{II}-region but it can be seen that the shape of the distributions suggest it truly is a mirrored state and would have similar backbone energetics for Gly residues (for which no C_β collision occurs). Consistent with this point, Gly residues cluster as a clearly defined population in that region rather well (Figure 6B and discussion section). This novel designation of P_{II}′ as a region decreases the scope of the unique portion of the ε-region from its com-

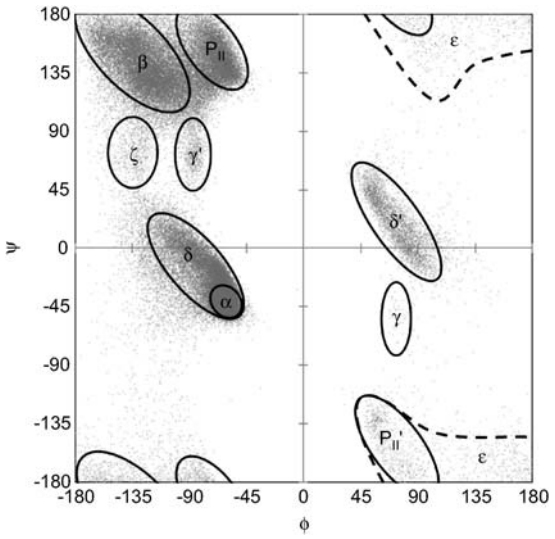

Figure 4 Proposed Ramachandran nomenclature.
Ramachandran plot as in Figure 3 but defining our proposed nomenclature for commonly populated regions of the plot (see text).

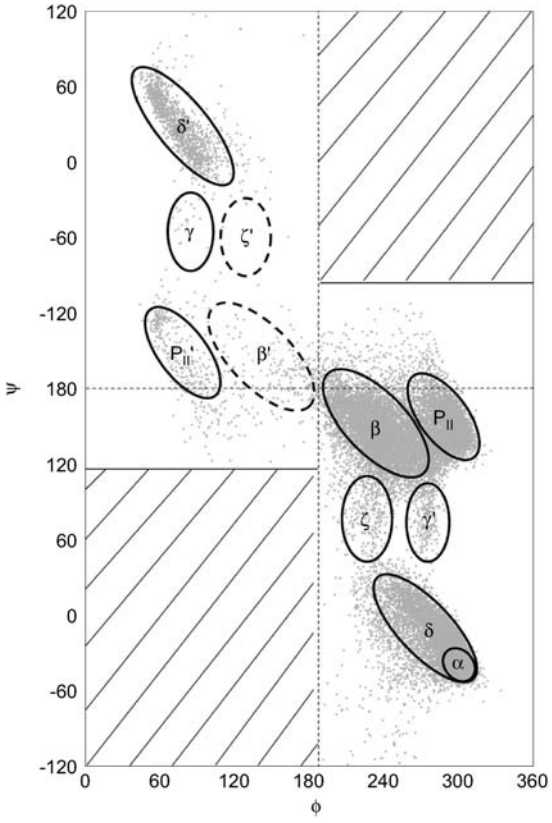

Figure 5 Mirrored Ramachandran plot.
A mirrored Ramachandran plot is shown with an extended ψ axis and mirrored redundant ψ-values but showing each point in the 72 376 residue dataset only once by not showing residues in the regions of the plot that are shaded. This highlights the mirror qualities of certain conformations. The β′- and ζ′-regions, shown with dashed lines, are not true occupied and are not recommended for use (see text).

monly assigned extent, so that it is only needed to account for the Gly residues in conformations closer to the (180, 180) corner of the plot. We have not done it here, but to account for the natural grouping, it seems that for Gly residues the P_{II}- and P_{II}'-regions need to be extended further across the $\psi=180^\circ$ line (Figure 6B).

At this point, inspection of Figure 2 reveals that the only major natural cluster of observations for non-glycine residues that is not yet named is the peninsula extending from the β-region near φ, ψ=(+130, +80). Based on using the next available Greek letter after ε, this region was first dubbed the ζ (zeta) region by Karplus (22) who noted that the region was dominated by residues occurring before proline (i.e., pre-Pro or XPr residues). For this reason, the region has also been called the pre-Pro region (24) but we prefer ζ because it is shorter and avoids the misconception that only residues before Pro can be found in this region. In the 1.2 Å high-fidelity dataset, only 51% of the residues that fall into the ζ-region precede proline. There is, of course, no corresponding ζ′-region cluster because in that region C_β atoms experience major collisions and because there are no mirror image Pro residues for these residues to precede.

Mirrored Ramachandran plot

One aspect of the Ramachandran plot that is often not well perceived is the mirror qualities of the various conformations. To better convey this, we have produced a version of the plot with each point only shown once but that emphasizes the mirror image relationships of various populations (Figure 5). What results from this action is that the main recognizable populations of the Ramachandran plot are inverted through an alternate origin of (φ, ψ)=(180, 180) point making it easier to identify the opposite handed mirror conformations. Highlighted much more clearly are, for example, the P_{II}'-region and the absence of clusters in what would be the β′- and ζ′-regions.

Wrapped Ramachandran plot

Another useful way to represent the data in a Ramachandran plot is to rewrap the entire plot by using different limits for each axis. The common Ramachandran plot has been centered on φ, ψ=(0, 0), but as well known as this plot is, it unfortunately separates portions of several major regions of the plot itself such as the upper portions of the β-, P_{II}- and ε-regions. An inspection of the φ, ψ distribution of residues in proteins (Figure 4) shows that in the φ direction the line at φ=0 is completely unpopulated, and in the ψ direction the line at ψ=-120 is minimally populated. Thus, a minimal disruption of natural groupings can be obtained by rewrapping the plot by setting the lower limit edges at φ=0 and ψ=-120 (and adding 360° to any value less than the new lower limits). The resulting plot leaves all major populations intact allowing a much better perception of their true extents (Figure 6). An additional value of such a plot is its suitability for the use of machine learning tools that look for common

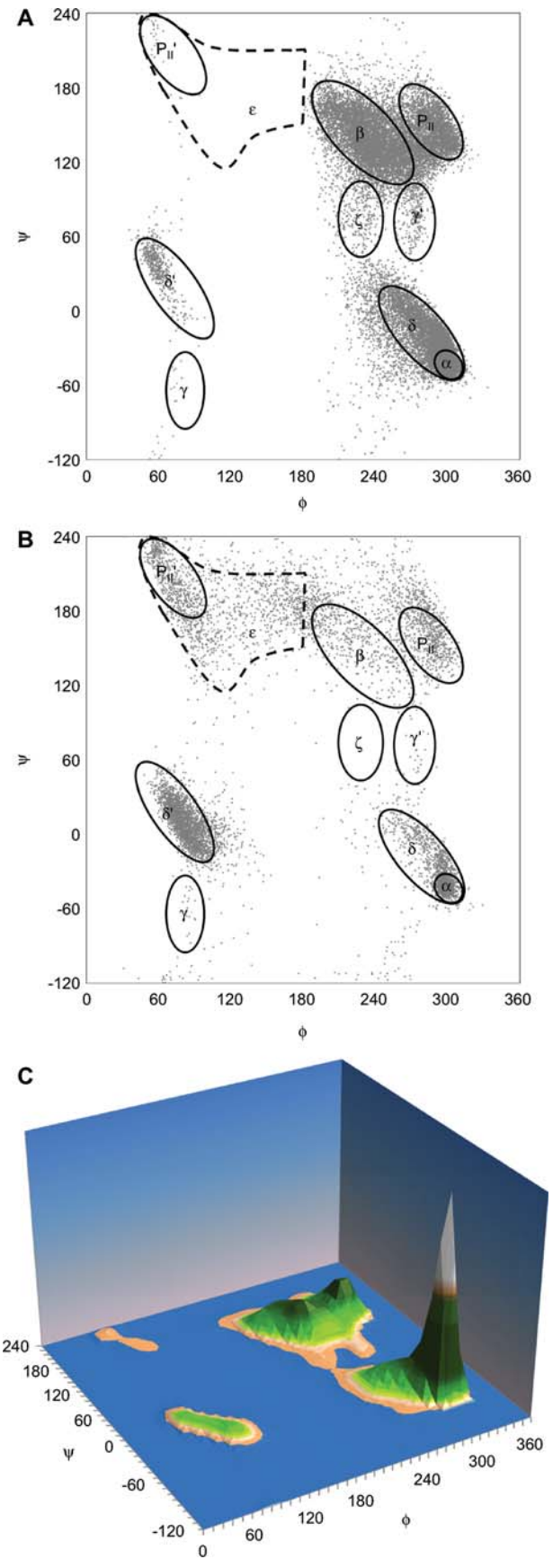

Figure 6 Wrapped Ramachandran plot.
(A) A Ramachandran plot showing the 63 149 Ala-like residues only (non-Gly, non-Pro), rewrapped around new axis origins (φ=0 and ψ=-120, see text) so that no region is cut unnaturally. To create this new plot, any φ<0 was reset to φ+360 and any residue with ψ<-120 was reset to ψ+360. (B) Wrapped Ramachandran plot with only the 6046 Gly residues. (C) Wrapped geo-style 3D Ramachandran plot. The plot shown is the same one as is shown in Figure 2, but using the rewrapped limits which allows all 'land masses' to appear as single intact entities.

motifs, because such programs are rarely capable of handling circular datasets such as the data displayed by the standard Ramachandran plot.

High-fidelity distributions for the individual amino acid residues

Figure 7 provides a set of separate Ramachandran plots for each amino acid plus Xpr (i.e., pre-Pro) residues based on the observed conformations of 72 376 well-ordered residues in protein structures determined at 1.2 Å resolution or better. These plots are in good agreement with previously published sets of plots (11, 22, 25) but are higher fidelity having estimated φ, ψ uncertainties of only around $\pm 3°$ (23). The plots in Figure 7 illustrate one important principle that we have noted before (22) but is still not widely appreciated. It is tempting to think that the distribution of each residue reflects the stability of that residue at various conformations, for instance with Gly being more stable in δ′ than in δ and Ala being more stable in the upper left-hand portion of the β-region than it is in the central β-region. But this is not the case, as can easily be shown by noting that Gly is physically a symmetric amino acid and thus a dipeptide with Gly in it must have equivalent energetics in the δ′- and δ-regions.

The principle that can explain this asymmetry of the Gly distribution, and that must similarly influence all distributions, is that the observed φ, ψ distributions for any amino acid residue type do not reflect the conformational energetics of the residue itself but how the energetics of that residue compares with the energetics of other residues. This can be understood by considering a simplified model for protein evolution that captures a few of its major aspects. In this simplified view, once a protein fold is established, the protein fold is considered to not change over time and the selection pressure at most positions in the protein has to do with the stability of the protein fold. This is a reasonable approximation as it is well established that sequences change much, much faster than folds (26). Given this model, mutation events over time will allow each of the 20 residue types to be tried out at each position, whereas the φ, ψ angles at that position for the most part do not change. The selective pressure will then lead to a preference for residues at each position in the structure based on their relative ability to adopt the conformation (i.e., those φ, ψ angles) present at that position.

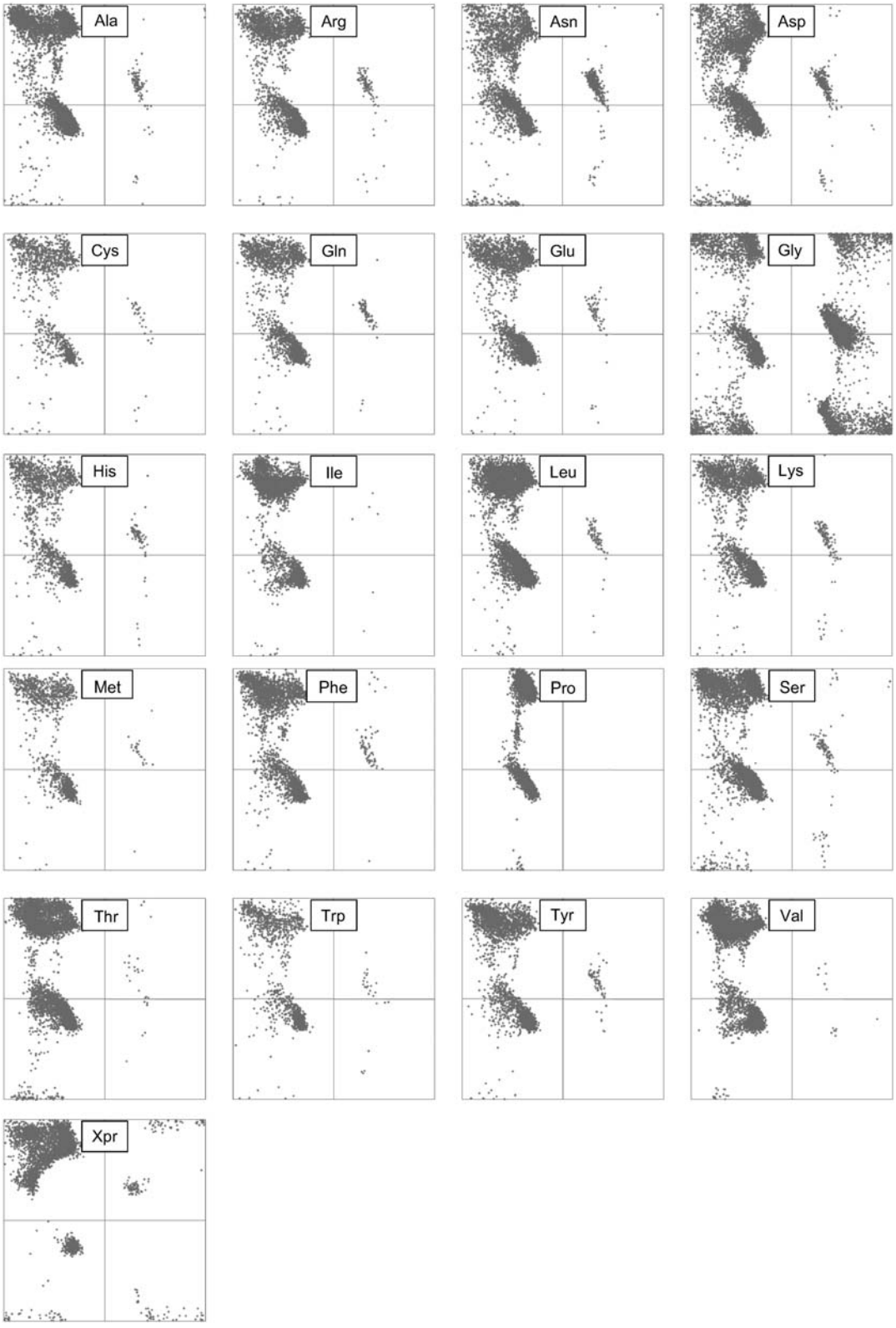

Figure 7 Individual residue distributions.
Shown are individual Ramachandran plots of each of the 20 residue types as well as Xpr (i.e., pre-Pro). Glycine, by far the most flexible amino acid, is the only amino acid that truly populates the β′-portion of the ε-region. The number of residues for each plot are as follows (based on the 72 376 residue plot): Ala: 6781 residues; Arg: 3208; Asn: 3267; Asp: 4300; Cys: 1167; Gln: 2540; Glu: 3819; Gly: 6046; His: 1748; Ile: 4128; Leu: 6334; Lys: 3287; Met: 1342; Phe: 2904; Pro: 3185; Ser: 4340; Thr: 4545; Trp: 1197; Tyr: 2764; Val: 5474; Xpr: 3185.

Applying this to the case of Gly, we can consider that the lower part of the δ′-region near φ, ψ=(+90, 0) is uniquely accessible for Gly residues compared to all other residue types, but the δ-region φ, ψ=(-90, 0) can be adopted relatively easily by all residues except Pro. This phenomenon explains the asymmetry of the Gly distribution, because Gly will be the dominant residue in the δ′-region (as it is far more stable than all other types), but it will only account for ~1/19th of all residues in the δ-region (because it has similar energetics to 18 other residue types in this region). Similarly, the upper left-hand portion of the β-region is uniquely accessible to Ala because the γ atoms of larger side chains experience collisions with the backbone. Thus, in that conformation Ala is a high proportion of all residues, whereas it is only approximately 1/19th of all residues in the central part of the β-region (as all but Pro can adopt this conformation). Of course protein evolution and energetics is more complex than this simple model accounts for, but these simplified arguments make clear that the φ, ψ-distribution of each residue is influenced by a competition with other residue types during evolution and thus it should not be taken to be an indicator of the φ, ψ energetics of that residue.

Common structures in proteins

The Ramachandran plot is a foundational concept used in biochemistry courses to describe the basic elements of protein structure, but in most cases the approach is based on a decades old view of secondary structure types summarized in the International Union of Pure and Applied Chemistry (IUPAC) nomenclature from 1970 (27). Much has been learned in the ensuing 40 years and recently high-accuracy distributions have been used to better define the basic parts list that builds proteins structures (28). One study focused on the simple repeating conformations (linear groups) and others have looked at the variety of simple non-repeating conformations that include many types of turns.

Linear groups in proteins

A dominant class of structures in proteins is the so-called linear group. The defining characteristic of linear groups is that they are made up of a series of residues having a repeating single conformation (29). Each linear group forms a unique helical structure. For instance, a segment having the repeating φ, ψ angles of near (-60, -40) for several consecutive residues leads to the very common α-helix. As part of their early research, Ramachandran and Sasisekharan (4) published an analysis that documented the helical parameters of residues per turn and the rise per residue for a linear group based on all possible φ, ψ pairs (Figure 8A). Over the years, many such linear group structures have been predicted and a subset of these has been observed in protein structures. Some of the supposed linear groups included in basic textbook presentations of protein structure are shown in Figure 1A. Until last year (18), modern, accurate protein structures were not systematically used to assess the fundamental question of what linear groups exist in proteins. Unique to the Hollingsworth et al. study (18) was the use of primary analysis criteria that focused exclusively on φ, ψ-angles rather than on regular hydrogen bonding. Their empirically minimal definition for a linear group was simply three residues in a row with φ, ψ-angles within ±10° of a certain value. That study revealed a limited number of real structures and failed to find any examples of several theorized linear groups (18). The study led to very clear conclusions that only three broad types of linear groups occur in proteins; these are a group of conformations that include the α- and 3_{10}-helices, a group that are largely β-strands and a group that adopt a P_{II} conformation. These results are visualized in Figure 8B and the groups are each described in the following paragraphs.

The most common linear group motif in proteins is the well-known α-helix, first hypothesized by Pauling et al. in 1951 (30). Conceived while confined in bed, Pauling stumbled upon the canonical conformation the α-helix adopts to fulfill the internal hydrogen bonds of the peptide backbone (31). Almost immediately after his original publication, the evidence began to pile up in favor of its existence (32). With the Hollingsworth et al. criteria, it is the most linear of the linear groups with φ, ψ-angles clustered very tightly around the most populated φ, ψ-value of (-63, -43) (Figure 2). Although the α-helix does have some variability in conformation, the local hydrogen bonding pattern that defines it keeps this variation minimal, with nearly all residues within ±15° of the peak φ, ψ-values.

Interestingly, in terms of φ, ψ-angles, the single cluster of linear groups in this region of the Ramachandran plot includes not only the α-helix but also the 3_{10}-helix and segments that combine both 3_{10}- and α-helical hydrogen bonding (18). The 3_{10}-helix, known from the early days of protein crystallography (33) and from more recent analyses (34–36) to be a minor component of protein structure, is more tightly wound than the α-helix and does not occur in long segments. The subset of the linear group peak with 3_{10}-helical hydrogen bonding is but a shoulder with a distinct peak position centered near φ, ψ=(-60, -25). It includes nearly 100 times fewer residues than the α-helical occurrences. In the study by Hollingsworth et al., many examples existed in the database of α-helical linear groups of 10 residues and longer, but the longest linear group with a 3_{10}-helical hydrogen bonding pattern with repeating φ, ψ-angles was only five residues, implying that the 3_{10}-helicies might not in general be stable enough to be true linear groups that could extend indefinitely.

The second most common structural motif, the β-strand, was originally also anticipated by Pauling and Corey in 1951 (37). Based on their modeling studies, it came in two distinctive conformations depending on the orientation of the strands to which it was hydrogen bonded, i.e., whether it had a parallel or anti-parallel alignment (N to C) compared with its neighbor. Whereas the concept that parallel and anti-parallel β-strands had distinct conformations has been maintained over the years (5, 6, 27, 38), Hollingsworth et al. (18) showed that, in agreement with one previous study (39), the linear groups representing both the parallel and anti-parallel β-strands cover equivalent ranges of φ, ψ-space. Although

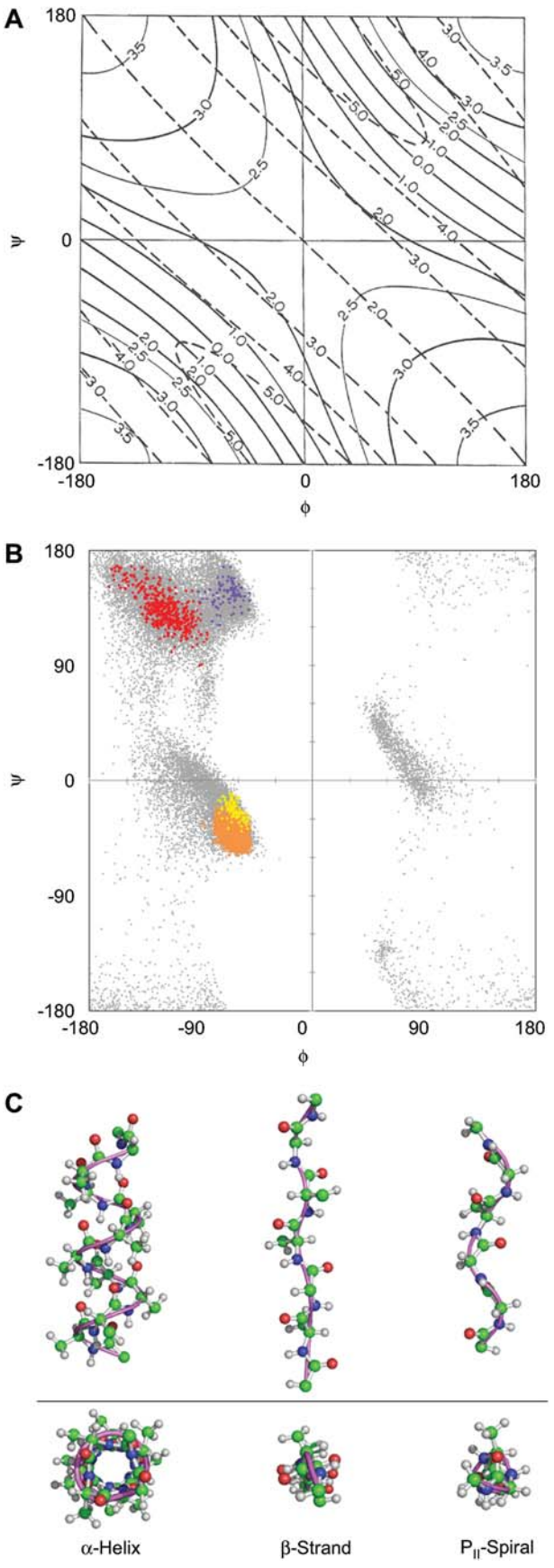

Figure 8 Location and distribution of the three major linear groups.
(A) The helical parameters of rise per residue (solid lines) and residues per turn (dashed lines) are shown as a function of linear group φ, ψ-values assuming 'ideal' geometry. Adapted from Ramachandran and Sasisekharan (4). (B) Shown are the locations and distributions of the three major linear groups. A natural break separates the two linear groups in the classical beta-region, the β-strand (red) and P_{II}-spiral (purple). The α-helices (orange) and 3_{10}-helices (yellow) are overlapped. Figure reprinted with permission from Hollingsworth et al. (18). (C) Examples of each of the three major linear groups, a 10-residue α-helix from PDBID 2OB3 [residues 182–191, all falling within φ=(63±4°, 43±4°)]; a five-residue β-strand from PDBID 1LC0 [residues 212–216, all falling within (φ, ψ)=(121±10°, 130±10°)]; a five-residue P_{II}-spiral from PDBID 3BOG [residues 47–51, all falling within (φ, ψ)=(-65±10°, 145±10°)]. Atoms are colored in the same manner as Figure 1A. Edge-on and end-on views are shown with a violet ribbon highlighting the backbone.

the anti-parallel arrangement is more common in the database population, both types of strands as a linear group have a maximal population near centered around φ, ψ=(-120, +130) (Figures 2 and 8B). Also notable is that the populations are broadly distributed over a range of approximately 80° in φ and 80° in ψ. The spread is in stark contrast to the tight distribution of the α-helix population.

The third and only other conformational region that builds linear groups in proteins is the P_{II}-region. This reinforces other studies noting that even though P_{II} is often overlooked, it is an important regular structure in folded proteins (19). When considering linear groups (Figure 8B) rather than individual residues (Figures 1 and 2), this P_{II}-region splits even more distinctly from the β-region emphasizing its uniqueness. As for the β linear group, the P_{II} linear group is broad; it is centered around φ, ψ=(-65, +145) and extends over approximately 50° in both φ and ψ. As documented in the early days of protein structure analysis (40–42) and recently reviewed by Woody (17), the P_{II} conformation is very common for all types of polypeptide chains in water being adopted among others by poly-L-proline, poly-glycine and unfolded poly-L-alanine. Furthermore, studies have also now shown that unfolded proteins in general have a tendency to take on the P_{II} conformation (43). The stability of P_{II} has been ascribed to maximizing chain entropy while exposing all hydrogen bond capable backbone atoms to the water in a way that minimally disrupts the natural water organization (28). Given the context of this popular conformation, it is of interest to note that because Gly is achiral, the poly-glycine II conformation would be expected to include both the P_{II}- and P_{II}'-regions, emphasizing the relevance and appropriateness of the designation P_{II}'. This broad role for the P_{II} conformation is what led Hollingsworth et al. (18) to recommend that the full name of the P_{II} structure itself be changed from poly-proline II to polypeptide II allowing the maintenance of the traditional short-hand designator of P_{II} while removing the completely misleading focus on type of amino acid residues involved (18).

Aside from the three linear groups described above ($\alpha/3_{10}$, β and P_{II}) no other linear groups exist in proteins. Among the putative linear groups shown in Figure 1B, the 2.2_7 Ribbon (44) simply does not exist; however, the π-helices and left-handed α- and 3_{10}-helices require some special comment. The π-helix has been a controversial structure. First proposed in 1952 (45), short segments with π-helical hydrogen bonding patterns do exist in proteins (46–48). We have recently discovered that the π-helix, although not a true conformational linear group, occurs in approximately 1 in 7 proteins as a perturbation of an α-helix owing to the insertion of a single residue. The insertion can be accommodated as a bulge that creates the i+5 hydrogen bonding pattern that defines the π-helix (R. Cooley et al., submitted).

The left-handed α-helix, incidentally the helix that was actually pictured in Pauling et al. (30) when they proposed the α-helix structure (49), does not exist in long segments, although a few examples of three-residue long segments with this conformation do exist (18). Interestingly, although the left-handed 3_{10}-helix also does not exist as an extended linear group, it does occur in proteins as short segments that tend to be missed because their annotation by secondary structure algorithms does not indicate that they are left-handed (18, 34).

This new research reveals a remarkably simple view of the basics of protein structure, with just three major linear groups: the $\alpha/3_{10}$-helix, the β-strand (either parallel or antiparallel) and the P_{II}-spiral. Whereas all three are technically 'helices' in the mathematical sense, we suggest based on their appearance that the common names α-helix and β-strand be continued for the first two, but that the third be called P_{II}-spirals instead of P_{II}-helices to capture its more extended structure (Figure 8C). Also, it should be noted that each of these groups have an inherent amount of variability, with the β-strands and P_{II}-spirals covering a broader diversity of conformations than the $\alpha/3_{10}$-helices. For this reason, any figure that shows only the average location of the linear groups (i.e., Figure 1B) would miss out on communicating the variability of these groups. We suggest something like Figure 8B showing both the location of the major linear groups and their variation would serve as a much more effective introduction to the common regular structures in proteins.

Less regular structures

Taken together, the linear groups described above do account for the bulk of residues in most protein structures, but in complete protein folds there are many other conformations that occur. The next level of complexity can be considered two adjacent residues adopting distinct φ, ψ-values (2, 3). These structures have also been rather well studied. Although such structures are more complex to represent on a Ramachandran plot, they can be designated by a short-hand that assigns to each residue the designator for the region of the Ramachandran plot it occupies. For example, a classic type II β-turn would be described as a $P_{II}\delta'$ using our proposed nomenclature. Perskie et al. (16) showed that if one creates a library of residues in proteins that are not in β-strands or α-helices, which they termed the 'coil library,' that 75% of those residues adopt one of five simple motifs: P_{II}-spirals, β-turns (50, 12), inverse γ-turns (γ') and two motifs involving prolines ($\zeta\alpha$ and ζP_{II}). If somewhat relaxed criteria are used for the β-turns and β-strands, the total fraction of all residues accounted for by α-helices, β-strands, P_{II}-spirals and the other four above motifs reaches 90% [see table 4 in Perskie et al. (16)]. We presume that the remaining 10% of residues participate in less common turn and loop structures and encompass a great diversity of conformations.

For segments of chain three residues and longer, the inherent complexity grows dramatically and categorization becomes increasingly challenging. Given that there has not been agreement on the nomenclature for the conformations of single residues, it should come as no surprise that there is not an agreement on categorization of these more complex segments. Many of the most fruitful efforts have involved the use of automated programs rather than manual inspection for codifying structures of two-residue segments or longer (51–56, 14, 57–59), but even so none of these approaches resulted in a compelling, clear categorization for such segments that led to its widespread recognition and use. In principle, however, such a categorization should be possible and it is our expectation that the application of such studies to the increased amount of high-fidelity data now available will be the path to a complete and clear classification of the 'parts list' of proteins (28).

Expert opinion

How the basics of protein structure are often taught about and thought about are too heavily influenced by outdated concepts. Part of the challenge is the very real complexity of protein structure, but another part is simply the level to which our thinking about structure has been dictated by and limited by the concepts developed during the early days of protein structure determination when there were few structures available and those that were available were not highly accurate. Now that tens of thousands of protein structures are available, many with coordinate accuracy approaching that of small molecule crystal structures, new insights are being developed that need to be translated into an updated framing of the basic concepts of protein structure. Here in the context of highlighting some of the recent insights about linear groups in proteins, we present some high-fidelity Ramachandran plots and explore nontraditional ways to present the φ, ψ-plot data so as to allow the features of the natural distributions to come across more clearly. We also note the lack of consensus about nomenclature for conformations and make a concrete suggestion for a step towards adopting a standard that better reflects the patterns seen in the natural distributions.

In terms of dipeptide conformation, the most important concept we present is the idea of a 'wrapped' Ramachandran plot that uses different limits on the φ and ψ axes so that all natural groupings of residues can be viewed as intact clusters. The mainstream adoption of this wrapped plot could be

a challenging adjustment for those of us who have already been trained using the current plot, but newcomers would find it makes sense and the change would be beneficial in the long run. The novel geo-style 3D Ramachandran plot emphasizes how the wrapped plot allows the 'continents' of conformation to remain intact (compare Figure 6C with Figure 2). Extending the 'geo' analogy further, we note that when creating maps of the world, cartographers do not force splits at exactly 0° or 90° latitude or longitude, but generally split maps following natural divisions in the middle of oceans so that land masses remain intact. Visual impressions make a big impact in how people perceive information.

Some highlighted concepts regarding dipeptide conformations in the paper are the distinction of the α- from the δ-region and the β- from the P_{II}-regions (brought home by a novel geo-style 3D Ramachandran plot), the lack of true α′- or β′-regions, and the recognition of the conformation near φ, ψ=(+60, -150) as the P_{II}'-region, i.e., a mirror image to the P_{II} conformation. We have no delusion that the nomenclature we propose will be a panacea that all will agree on and expect the need for further discussion by the community about how to handle certain complexities, such as what to do about the nomenclature of the sparsely populated regions, whether to extend the size of the P_{II}- and P_{II}'-regions to include extremes only adopted by Gly, and whether residues in the α conformation, but not in an α-helix should be described as being in the α-region or the δ-region.

With regard to conformations adopted by a segment of residues, the result that α-/3_{10}-helices, β-strands and P_{II}-spirals are the only linear groups in proteins is no surprise to those who work on protein structure, but the explicit documentation of the types of linear groups and their scopes should clarify the conformations making up the 'protein parts kit'.

Two additional concepts have been the focus of some attention in this review. The first is a brief mention of recent research showing that the ideal bond angles of the peptide unit vary systematically with φ, ψ-values and that explains how certain classically 'disallowed' regions such as the δ′-region can occur. The second is a discussion making the point that the observed φ, ψ distribution for a given residue type is not only influenced by the energetics of the residue itself but also its relative energetics compared with those of the other residue types.

Outlook

With high-fidelity Ramachandran plot availability, the time is ripe for the completion of a definitive analysis of the 'protein parts list' conceptualized by Fitzkee et al. (28). The study of linear groups by Hollingsworth begins this process and extension to the analysis of all of the most common structures (51–56, 14, 57–59) will be the next step. Within a few years we should have such a complete and accurate list of all protein parts and this will mostly impact our understanding and pave the way for fundamental, first principles understanding of the energetics behind the conformational preferences. It will also stimulate great improvements in both validation tools and the accuracy with which we can model and predict protein tertiary structure either *de novo* from its sequence or by template-based modeling.

Highlights

- Ultra-high resolution crystal structures, conveniently searchable in the Protein Geometry Database, provide a wealth of high-fidelity information for studying protein conformation.
- Ultra-high resolution structures show that bond angles depend systematically on conformation and this impacts what conformations occur.
- The continued description of the Ramachandran plot as having three main regions – alpha, beta and $alpha_L$ – is inadequate and misleading. To accurately describe the dominant regions of conformational space, five regions are minimally required: α, δ, β, P_{II} and δ′. An additional five regions capture less populated, but still discretely defined regions. These are ζ, γ′, P_{II}' and γ, and the region ε that is for the most part uniquely occupied by Gly.
- The complexity of protein structure makes this difficult, but a proposal is made for a much needed agreement on nomenclature for the natural groupings of conformations.
- Novel informative presentations such as a 3D, mirrored and wrapped Ramachandran plots are introduced.
- Owing to evolutionary history, φ, ψ distributions for individual residues are influenced by the relative energetics of the 20 residue types.
- α/3_{10}-helices, β-strands and P_{II}-spirals are the only three structures of repeating φ, ψ-angles that occur in proteins.
- A next step is to complete a protein parts list by using high-fidelity distributions to more precisely define common structures that have varying φ, ψ-angles.

Acknowledgments

This research was supported in part by National Institutes of Health grant GM083136 (to P.A.K.) and Howard Hughes Medical Institute grant 52005883 (supporting S.A.H.). We thank the many colleagues with whom we have discussed fundamental aspects of protein structure. The authors have no financial conflicts of interest related to this work.

References

1. Sasisekharan V. Stereochemical criteria for polypeptide and protein structures. In: Ramanathan N, ed., Collagen, New York: Wiley, 1962: 39–78.
2. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol 1963; 7: 95–9.
3. Venkatachalam CM. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. Biopolymers 1968; 6: 1425–36.

4. Ramachandran GN, Sasisekharan V. Conformation of polypeptides and proteins. Adv Protein Chem 1968; 23: 283–438.
5. Voet D, Voet JG. Biochemistry. Hoboken, NJ: J. Wiley & Sons, 2004.
6. Lehninger AL, Nelson DL, Cox MM. Lehninger principles of biochemistry. New York, NY: W.H. Freeman, 2008.
7. Berkholz DS, Krenesky PB, Davidson JR, Karplus PA. Protein geometry database: a flexible engine to explore backbone conformations and their relationships to covalent geometry. Nucleic Acids Res 2010; 38: D320–5.
8. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK – a program to check the stereochemical quality of protein structure. J Appl Cryst 1993; 26: 283–91.
9. Kleywegt GJ, Jones TA. Phi/psi-chology: Ramachandran revisited. Structure 1996; 4: 1395–400.
10. Davis IW, Murray LW, Richardson JS, Richardson DC. MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. Nucleic Acids Res 2004; 32: W615–9.
11. Beck DA, Alonso DO, Inoyama D, Daggett V. The intrinsic conformational propensities of the 20 naturally occurring amino acids and reflection of these propensities in proteins. Proc Natl Acad Sci USA 2008; 105: 12259–64.
12. Wilmot CM, Thornton JM. Beta-turns and their distortions: a proposed new nomenclature. Protein Eng 1990; 3: 479–93.
13. Efimov AV. Standard structures in proteins. Prog Biophys Mol Biol 1993; 60: 201–39.
14. Oliva B, Bates PA, Querol E, Aviles FX, Sternberg MJ. An automated classification of the structure of protein loops. J Mol Biol 1997; 266: 814–30.
15. Rooman MJ, Kocher JP, Wodak SJ. Prediction of protein backbone conformation based on seven structure assignments. Influence of local interactions. J Mol Biol 1991; 221: 961–79.
16. Perskie LL, Street TO, Rose GD. Structures, basins, and energies: a deconstruction of the Protein Coil Library. Protein Sci 2008; 17: 1151–61.
17. Woody RW. Circular dichroism spectrum of peptides in the poly(Pro)II conformation. J Am Chem Soc 2009; 131: 8234–45.
18. Hollingsworth SA, Berkholz DS, Karplus PA. On the occurrence of linear groups in proteins. Protein Sci 2009; 18: 1321–5.
19. Adzhubei AA, Sternberg MJ. Left-handed polyproline II helices commonly occur in globular proteins. J Mol Biol 1993; 229: 472–93.
20. Némethy G, Printz MP. The gamma turn, a possible folded conformation of the peptide chain. Comparison with the beta-turn. Macromolecules 1972; 5: 755–8.
21. Matthews B. The gamma turn. Evidence for a new folded conformation in proteins. Macromolecules 1972; 5: 818–9.
22. Karplus PA. Experimentally observed conformation-dependent geometry and hidden strain in proteins. Protein Sci 1996; 5: 1406–20.
23. Berkholz DS, Shapovalov MV, Dunbrack RL Jr, Karplus PA. Conformation dependence of backbone geometry in proteins. Structure 2009; 17: 1316–25.
24. Lovell SC, Davis IW, Arendall WB 3rd, de Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by C alpha geometry: phi, psi and C beta deviation. Proteins 2003; 50: 437–50.
25. Hovmoller S, Zhou T, Ohlson T. Conformations of amino acids in proteins. Acta Crystallogr D Biol Crystallogr 2002; 58: 768–76.
26. Orengo CA, Thornton JM. Protein families and their evolution – a structural perspective. Annu Rev Biochem 2005; 74: 867–900.
27. International Union of Pure and Applied Chemistry (IUPAC). IUPAC-IUB Commission on biochemical nomenclature. Abbreviations and symbols for the description of the conformation of polypeptide chains. Biochemistry 1970; 9: 3471–9.
28. Fitzkee NC, Fleming PJ, Gong H, Panasik N Jr, Street TO, Rose GD. Are proteins made from a limited parts list? Trends Biochem Sci 2005; 30: 73–80.
29. Schulz G, Schirmer R. Principles of protein structure. New York: Springer-Verlag, 1979.
30. Pauling L, Corey RB, Branson HR. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci USA 1951; 37: 205–11.
31. Pauling L. How my interest in proteins developed. Protein Sci 1993; 2: 1060–3.
32. Perutz MF. New X-ray evidence on the configuration of polypeptide chains. Nature 1951; 167: 1053–4.
33. Davies DR. X-ray diffraction studies on polypeptide conformations. Prog Biophys Mol Biol 1965; 15: 189–222.
34. Pal L, Basu G, Chakrabarti P. Variants of 3_{10}-helices in proteins. Proteins 2002; 48: 571–9.
35. Enkhbayar P, Hikichi K, Osaki M, Kretsinger RH, Matsushima N. 3(10)-helices in proteins are parahelices. Proteins 2006; 64: 691–9.
36. Pal L, Basu G. Novel protein structural motifs containing two-turn and longer 3_{10}-helices. Protein Eng 1999; 12: 811–4.
37. Pauling L, Corey RB. The pleated sheet, a new layer configuration of polypeptide chains. Proc Natl Acad Sci USA 1951; 37: 251–6.
38. Richardson JS. The anatomy and taxonomy of protein structure. Adv Protein Chem 1981; 34: 167–339.
39. Nagano K. Logical analysis of the mechanism of protein folding. IV. Super-secondary structures. J Mol Biol 1977; 109: 235–50.
40. Sasisekharan V. Structure of poly-L-proline II. Acta Crystallogr 1959; 12: 897–903.
41. Arnott S, Dover SD. The structure of poly-L-proline II. Acta Crystallogr B 1968; 24: 599–601.
42. Scott RA, Scheraga HA. Conformational analysis of macromolecules. III. Helical structures of polyglycine and poly-L-alanine. J Chem Phys 1966; 45: 2091–101.
43. Shi Z, Woody RW, Kallenbach NR. Is polyproline II a major backbone conformation in unfolded proteins? Adv Protein Chem 2002; 62: 163–240.
44. Donohue J. Hydrogen bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci USA 1953; 39: 470–8.
45. Low B, Baybutt R. The pi-helix – a hydrogen bonded configuration of the polypeptide chain. J Am Chem Soc 1952; 74: 5806–7.
46. Weaver TM. The pi-helix translates structure into function. Protein Sci 2000; 9: 201–6.
47. Dasgupta B, Chakrabarti P. pi-Turns: types, systematics and the context of their occurrence in protein structures. BMC Struct Biol 2008; 8: 39.
48. Fodje MN, Al-Karadaghi S. Occurrence, conformational features and amino acid propensities for the π-helix. Protein Eng 2002; 15: 353–8.
49. Dunitz J. Pauling's left-handed α-helix. Angew Chem Int Ed 2001; 40: 4167–73.
50. Hutchinson EG, Thornton JM. A revised set of potentials for β-turn formation in proteins. Protein Sci 1994; 3: 2207–16.

51. Adzhubei AA, Eisenmenger F, Tumanyan VG, Zinke M, Brodzinski S, Esipova NG. Approaching a complete classification of protein secondary structure. J Biomol Struct Dyn 1987; 5: 689–704.
52. Rooman MJ, Rodriguez J, Wodak SJ. Automatic definition of recurrent local structure motifs in proteins. J Mol Biol 1990; 213: 327–36.
53. Zhang X, Fetrow JS, Rennie WA, Waltz DL, Berg, G. Automatic derivation of substructures yields novel structural building blocks in globular proteins. Proc Int Conf Intell Syst Mol Biol 1993; 1: 438–46.
54. Schuchhardt J, Schneider G, Reichelt J, Schomburg D, Wrede P. Local structural motifs of protein backbones are classified by self-organizing neural networks. Protein Eng 1996; 9: 833–42.
55. Wintjens RT, Rooman MJ, Wodak SJ. Automatic classification and analysis of α α-turn motifs in proteins. J Mol Biol 1996; 255: 235–53.
56. Fetrow JS, Palumbo MJ, Berg G. Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme. Proteins 1997; 27: 249–71.
57. Wintjens R, Wodak SJ, Rooman M. Typical interaction patterns in αβ and βα turn motifs. Protein Eng 1998; 11: 505–22.
58. Anishetty S, Pennathur G, Anishetty R. Tripeptide analysis of protein structures. BMC Struct Biol 2002; 2: 9.
59. Ikeda K, Tomii K, Yokomizo T, Mitomo D, Maruyama K, Suzuki S, Higo J. Visualization of conformational distribution of short to medium size segments in globular proteins and identification of local structural motifs. Protein Sci 2005; 14: 1253–65.