Amino Acids (2013) 44:781–790
DOI 10.1007/s00726-012-1402-z

ORIGINAL ARTICLE

A proteomic Ramachandran plot (PRplot)

Oliviero Carugo · Kristina Djinović-Carugo

Received: 17 March 2012 / Accepted: 10 September 2012 / Published online: 25 September 2012

Abstract Each protein structure can be characterized by the average values of the main chain torsion angles ϕ and ψ and, as a consequence, be plotted on a bidimensional diagram, which resembles the Ramachandran plot. Here, we describe a proteomic ϕ–ψ plot (PRplot) where each protein structure is associated with one point, allowing in this way to represent the entire protein structure universe. It was verified that the PRplot is a robust tool since it does not depend on the dimension of the proteins, on the crystallographic resolution of the structures, nor on the biological source; moreover, it is little affected by disordered and structurally uncharacterized residues. The proteins mapped on the PRplot tend to cluster in three regions that correspond to the structures rich in alpha-helices, in beta-strands, and in both helices and strands, and are distributed along a sigmoidal curve that connect these three highly populated regions. PRplots are a unique instrument to project all protein structures on a single bidimensional plane where the entire structural complexity is reduced to a striking simplicity, with the sigmoid curve clearly delineating the space fraction accessible to a stable protein.

O. Carugo (✉)
Department of Chemistry, University of Pavia,
Viale Taramelli 12, 27100 Pavia, Italy
e-mail: oliviero.carugo@univie.ac.at

K. Djinović-Carugo (✉)
Max F. Perutz Laboratories, Department of Structural and Computational Biology, Vienna University,
Campus Vienna Biocenter 5, 1030 Vienna, Austria
e-mail: kristina.djinovic@univie.ac.at

K. Djinović-Carugo
Department of Biochemistry, Faculty of Chemistry and Chemical Technology, University of Ljubljana,
Aškerčeva 5, 1000 Ljubljana, Slovenia

Keywords Protein backbone conformation · Protein fold · Protein structure · Proteome · Ramachandran plot · Structural bioinformatics

Introduction

When G.N.I. Ramachandran ideated the homonymous plot, where each amino acid of a protein is characterized by its backbone ϕ (on the x-axis) and ψ (on the y-axis) torsion angles, he probably did not imagine the future success of his invention (Ramachandran et al. 1963). During the past half a century, the Ramachandran plot has been used (Lesk 2001), revisited (Walther and Cohen 1999; Kleywegt and Jones 1996), and commented (Zhou et al. 2011; Berkjolz et al. 2009) countless times.

Ramachandran plots are routinely used to check the stereochemical quality of protein three-dimensional structures (Laskowski et al. 1996; Chen et al. 2010). There are in fact regions of the plot, i.e., combinations of ϕ and ψ values, that are energetically more favorable than others and the quality of a protein structure is considered to be acceptable if its amino acids fall in regions of the Ramachandran plot that are energetically allowed.

Here, we present an expansion of the Ramachandran's original idea by using the same bidimensional framework to describe not a single amino acid, but an ensemble of amino acids. This is the concept described here: the position of a protein is defined by the average values of its ϕ and ψ torsion angles. This results in a proteomic Ramachandran plot, named PRplot, which can host several proteins and allow a global description of all protein structures.

Three clusters of protein structures can be identified on the PRplot, which correspond to the structures rich in helices, in strands, and in both helices and strands, and

which are connected by a sigmoid curve that delineates the fraction of the PRplot that can be populated by folded proteins. A striking reduction of the extraordinary complexity of protein structure is then achieved.

Methods

Several unique datasets were created by using the Protein Data Bank (Bernstein et al. 1977; Berman et al. 2000), the Scop database (Andreeva et al. 2007), and the PISCES (Wang and Dunbrack 2003) and PDBselect (Griep and Hobohm 2009) web servers. All structures that were incomplete were removed from each of them. A structure was considered to be incomplete if its PDB file contained one of the following lines: the REMARK 465 lines, which list the residues that lack completely the positional coordinates; the REMARK 470 lines, which list the non-hydrogen atoms of the amino acids that do not have positional coordinates; and the REMARK 475 lines, which enumerate the residues modeled with zero occupancy.

The values of the ϕ and ψ torsion angles were computed with the program DSSP (http://swift.cmbi.ru.nl/gv/dssp/). Average values were calculated with locally written software.

The comparisons between PRplots were made by means of a contingency table analysis (Dowdy et al. 2004). The conformational space, which is defined by the ϕ and ψ torsion angles, was divided into 324 squares of $20° \times 20°$. Each PRplot was then described by the number of proteins in each of the 324 squares and the comparison between two PRplots performed by comparing two, ordered strings of 324 numerical values. Empty squares were ignored since they are obviously very numerous and would bias the comparisons, by causing the systematic underestimation of the differences between the plots.

Results and discussion

PRplot generation

With the exception of the first and last amino acid, each protein residue is associated with a pair of ϕ and ψ torsion values. A protein of N residues has therefore $N - 2$ pairs of ϕ and ψ values, which can be averaged with circular statistics methods (Batschelet 1981) as

$$A_{\text{ave}} = \arctan\left(\frac{\sum \sin A_i}{\sum \cos A_i}\right),$$

where A_{ave} is the average value of the $N - 2$ A_i values ($A = \phi$ or ψ). The use of circular statistics is necessary since dihedral angles are periodical quantities that range from $-180°$ to $+180°$: for example, the average of two angles of $-179°$ and $+179°$ is $180°$ and not $0°$.

Three examples are presented in Fig. 1. The first protein is a truncated globin [PDB file 1dlw (Pesce et al. 2000)] rich in α-helices that does not contain strands. Many points in its Ramachandran plot are clustered around the region typical of the α-helices and the average ϕ and ψ values are right in the middle of the cluster of the helical residues. The second molecule is a bacterial outer membrane protein [PDB file 1qj8 (Vogt and Schulz 1999)] rich in β-strands. As a consequence, a dense cluster of points is observed in the region of the Ramachandran plot where β-strands are expected. The average ϕ and ψ values are in the middle of this cluster. The third protein is the archeal methenyltetrahydromethanopterin cyclohydrolase [PDB code 1qlw (Grabarse et al. 1999)] that contains both α-helices and β-strands. As a consequence, the Ramachandran plot contains several points in both the α- and the β-regions and the average ϕ and ψ values are in between the α and the β clusters.

PRplots of the protein universe

The PRplots are intended to show the distribution of proteins in the ϕ–ψ space. With this respect, the redundancy of the Protein Data Bank and of the databases that derive from the Protein Data Bank must be properly addressed. Redundancy can be removed in various ways, with different strategies and slightly different results (Sikic and Carugo 2010). For this reason we used three strategies, which resulted in the following three datasets:

(1) *Pisces-3.0* a precompiled list of 1,111 protein chains from the Pisces server (maximal sequence identity of 25 %), the structures of which were refined at 3.0 Å resolution (or better) with final R factor better than 0.25 (Wang and Dunbrack 2003).

(2) *PDBselect* a precompiled list of 695 protein structures from the PDBselect server (maximal sequence identity of 25 %) the structures of which were refined at a resolution of 3.0 Å or better with final R factor not worse than 0.30 (Griep and Hobohm 2009).

(3) *Scop-nr* an ensemble of 919 domains of the SCOP database, with only one example per fold type (Andreeva et al. 2007).

These three datasets are relatively small, since all the PDB entries with absent atoms/residues were removed (see "Methods" for pertinent details).

It is also necessary to remember that each of the three datasets contains only single chain protein structures and that one of the chains was randomly selected when the protein contains more than one polypeptide chain.

The three PRplots generated with these three datasets are shown in Fig. 2. It clearly appears that they are extremely similar, which is also confirmed by contingency table analysis ($P > 99.9$ %).

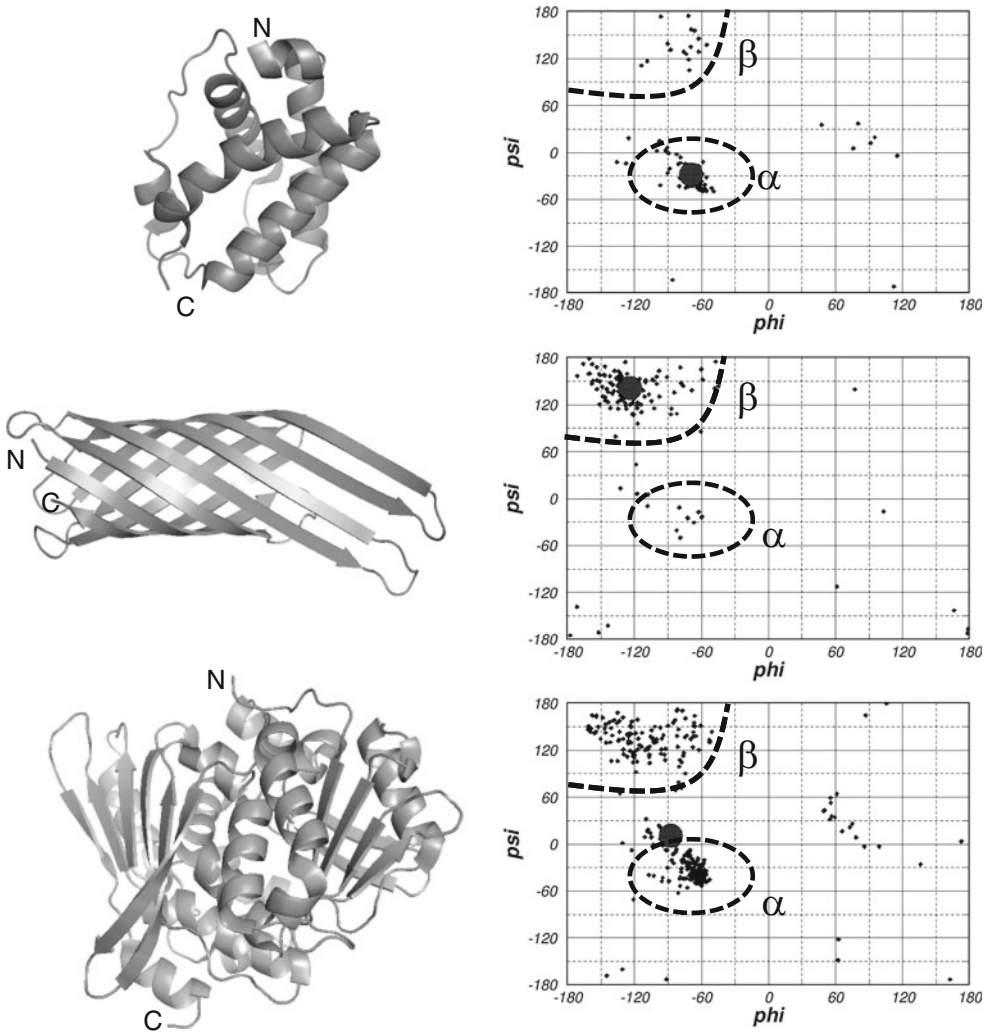

Fig. 1 Ribbon view and Ramachandran plot for the PDB files 1dlw (*top*), 1qj8 (*middle*) and 1qlm (*bottom*). The average ϕ/ψ is indicated with a *large circle* in the Ramachandran plots

It is important to observe that the similarity between the PRplots obtained with different ensembles of protein structures suggests that the trends shown in Fig. 2 are likely to be a genuine feature of the protein universe. In fact, two of the datasets (Pisces-3.0 and PDBselect) comprise protein molecules that might contain several domains while the third dataset (Scop-nr) includes only distinct structural domains. This is particularly important, since some of the protein that were used to produce Fig. 2 might not be native proteins but only smaller constructs, suitably optimized to allow expression and crystallization of the proteins.

In each of the PRplots of Fig. 2 it is possible to identify two regions populated much more than the rest of the plot. The first, centered at ϕ and ψ values close to $-75°$ and $-50°$, corresponds to proteins that contain many residues in helical conformation. The second region is located at ϕ and ψ values close to $-100°$ and 130° and corresponds to proteins that are rich in residues with an extended conformation of the backbone, as in the β strands. The proteins of the above two clusters are classified in the α and β classes, respectively, in the databases of protein domains Scop (Andreeva et al. 2007) and CATH (Orengo et al. 1997). In between these two regions there is a less populated zone, which includes proteins that have comparable amounts of α and β secondary structures and are classified in the α–β class of the CATH database and in the α–β and α/β classes of the Scop database. The few points that are considerably far from these regions systematically represent small proteins, with few (if any) secondary structural elements. In the Scop and CATH databases, they belong to classes termed "small proteins", "peptides" or "few secondary structures".

Distances between structures (dpp)

The possibility to cluster the data represented on the PRplots was examined by performing cluster analyses with

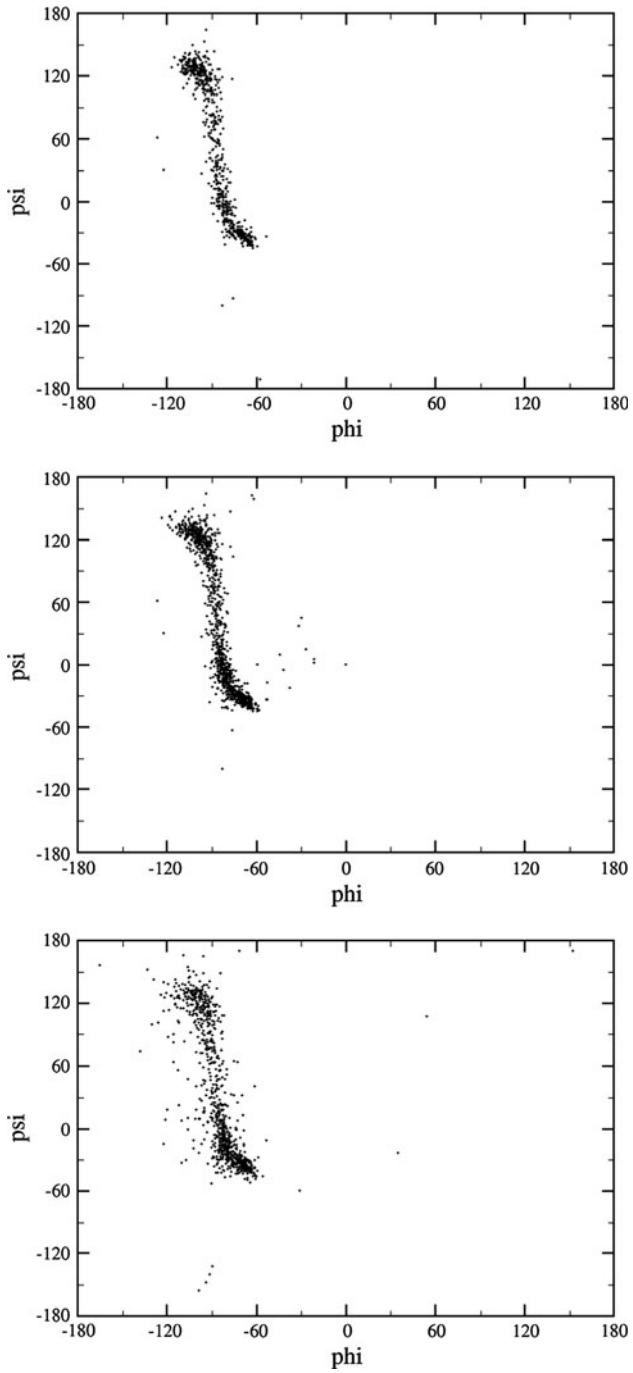

Fig. 2 PRplots computed with three different non-redundant sets of protein structures: PDBselect (*top*), Pisces-3.0 (*middle*), and Scop-nr (*bottom*) (see text for details)

the computer program MVSP (http://www.kovcomp.com/mvsp/downl2.html). For simplicity a small dataset was generated by retaining (randomly) only 31-folds of the SCOP database, which are plotted in Fig. 3a. A cluster analysis performed on a larger dataset containing all the entries of the Scop-nr list (see previous section for details)

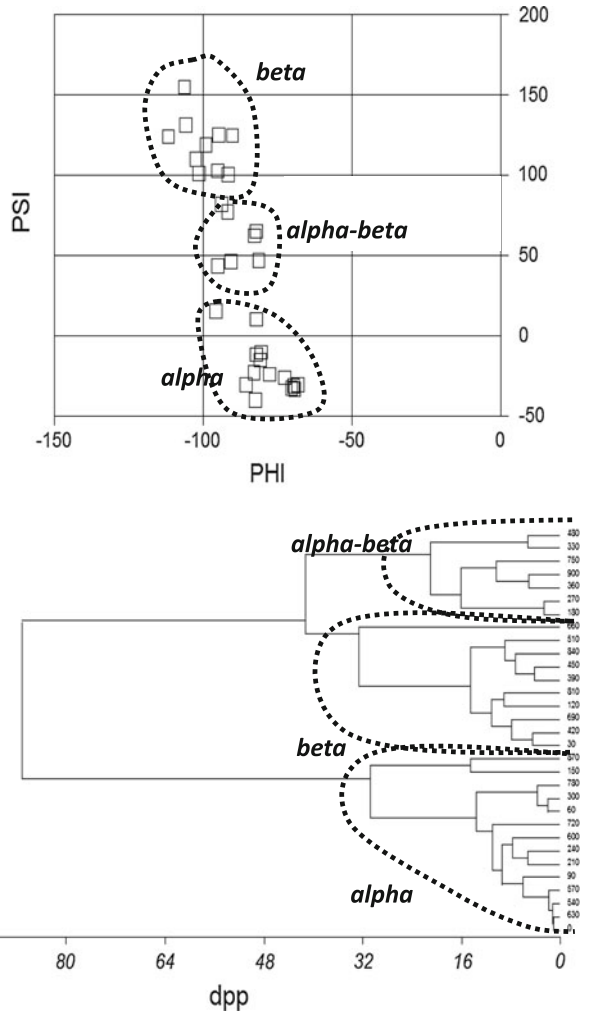

Fig. 3 **a** PRplots of the 31 folds that were randomly selected from the SCOP database. On the basis of the cluster analysis summarized by the dendrogram shown in **b**, they can be divided into the three clusters that are surrounded by *dotted lines*

produced essentially unchanged results and it is not reported here.

The distance (dpp) between two structures was measured with the Euclidean distance dpp defined as

$$\mathrm{dpp}_{ij} = \sqrt{\mathrm{d}x^2 + \mathrm{d}y^2}$$
$$\mathrm{d}x = |\phi_i - \phi_j|$$
$$\mathrm{d}y = |\psi_i - \psi_j|,$$

where i and j refer to the ith and jth proteins and where ϕ and ψ are the average values of the backbone torsion angles in each structure. Again both torsion angles were treated as periodical quantities with values that can range conventionally from $-180°$ to $180°$. Therefore, when necessary, the following transformations were applied:

$$\mathrm{if}(\mathrm{d}x > 180°) \rightarrow \mathrm{d}x = |360° - \mathrm{d}x|$$
$$\mathrm{if}(\mathrm{d}y > 180°) \rightarrow \mathrm{d}y = |360° - \mathrm{d}y|.$$

An agglomerative hierarchical clustering procedure and the centroid clustering criterion were adopted. Other clustering criteria, for example, UPGMA, farthest neighbor, median or WPGMA, produced closely similar results, also after randomizing the input order.

On the basis of the dendrogram shown in Fig. 3b the 31 protein structures were divided into three distinct clusters. The first contains 14 proteins that are rich in helical segments (α), the second includes 7 elements that contain both β strands and α helices (α–β), and the third comprises the remaining ten structures rich in β strands (β). Given that similar results were obtained by using also larger datasets (data not shown), it can be concluded that the clustering tendency (Carugo 2010a) in the PRplot reflects the same tendency observed in the protein fold space (Carugo 2010b).

The distributions of the dpp distances measured in the three protein datasets described above (Pisces-3.0, PDBselect, and Scop-nr) are depicted in Fig. 4. They are extremely similar, with two maxima, one at *ddp* close to 10° and the other, broader, at dpp close to 150°. The minimum that separates the two maxima is observed at dpp values close to 85° in all the three cases. Such a bimodal distribution is typical of the datasets that show an intrinsic clustering tendency (Carugo 2010a). The peak at low dpp values is due to the distances between members of the same cluster; the other peak, at higher dpp values, is due to the distances between structures that belong to different clusters; and the minimum at dpp values of about 85° suggests that this is the average distance between different clusters. Interestingly, the value of 85° is perfectly one half of the dpp distance (172°) that would be observed between a point corresponding to an ideal β conformation ($\phi = -110°$ and $\psi = 113°$) and a point corresponding to an ideal α conformation ($\phi = -57°$ and $\psi = -47°$) (Zhou and Faraggi 2010). In other words, half of the distance between α and β is the average distance between the clusters observed in the PRplots.

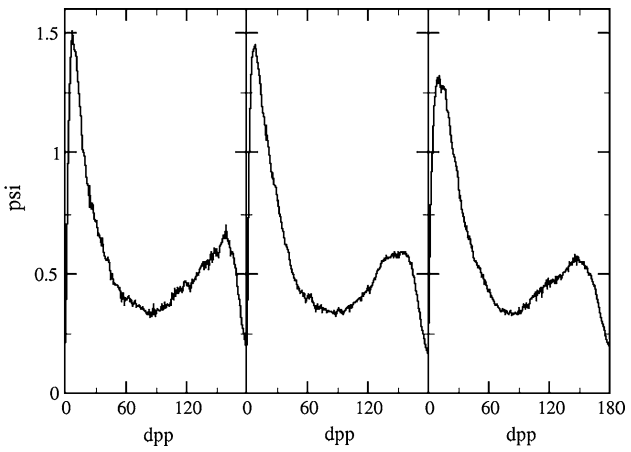

Fig. 4 Distribution of the dpp distances computed from the PRplots of non-redundant sets of proteins (PDBselect on the *left*, PISCES-3.0 in the *middle*, and Scop-nr on the *right*)

PRplot curve

In Fig. 2, the points on the PRplot are distributed along a sigmoid curve that links the regions that are populated more than the rest of the map, one at ϕ and ψ close to $-100°$ and 130° (β) to the other at ψ and ψ close to $-75°$ and $-50°$ (α). The points corresponding to the PDBselect dataset can be fitted by the function

$$\psi = \frac{a}{1 + b \cdot e^{c \cdot \phi}} + d,$$

where $a = 165.9$, $b = 2.03 \times 10^9$, $c = 0.248$, and $d = -36.6$ (correlation coefficient = 0.933). Similarly, the points associated with the Pisces-3.0 and with the Scop-nr datasets can be fitted by similar functions with correlation coefficients equal to 0.918 and 0.817, respectively.

Although the sigmoid relationship between ϕ and ψ lacks a precise physical interpretation, it is striking that a simple curve can describe most of the variability of the average backbone torsional angles of folded proteins. The sigmoid curve delineates the region of the PRplot that can be populated by protein structures and delineates the conformational space that connects α- and β-folds. In other words, this is not a folding trajectory that allows a protein to modify its structure, but on the contrary a path in the two-dimensional PRplot that allows to describe the structural variability of proteins.

Interestingly, while "wrong" structures are mapped on the sigmoidal curve, disordered proteins tend to lie a bit outside the curve.

Examples of wrong structures can be found in the Decoys 'R' Us database (Samudrala and Levitt 2000), which contains, e.g., the list of wrong experimental protein structures compiled by Branden and Jones (1990) and the misfolded proteins obtained by placing the sequences on radically different folds (Holm and Sander 2000). The good mapping of wrong structures on the sigmoidal curve is not surprising, since they are acceptable from a physicochemical point of view, though they are wrong in the sense that they are different from the native structures. These wrong structures can thus be seen as possible relative energy minima, different from the native and more stable structures. Since their stereochemistry is reasonable, it is difficult to distinguish them from real protein structures on the PRplots.

On the contrary, the conformational behavior of unfolded proteins can be described as an ensemble of coils (Bernardo et al. 2005; Jha et al. 2005) where the average ϕ and ψ torsion angles tend to deviate from the values typically observed in folded and ordered proteins rich in

secondary structure elements. For example, in the highly disordered structure of the spinach thylakoid soluble phosphoprotein of 9 kDa [PDB file 2fft (Song et al. 2006)], the average ϕ values range from $-118°$ to $-96°$ and the ψ values range from $-9°$ to $+42°$ (20 models were deposited in the PDB), slightly outside the ϕ–ψ sigmoid curve described above.

PRplots of different organisms

In order to verify that the PRplots do not depend on the biological origin of the proteins, two datasets, one composed of 723 human proteins and the other of 269 *Escherichia coli* proteins were compared. The redundancy of both ensembles of proteins was reduced to 30 %. The two PRplots, shown in Fig. 5, were statistically identical, suggesting that PRplots prokaryotic and eukaryotic species have comparable populations of all α- all β- and α–β proteins.

Although it would have been surprising to observe differences between two groups of proteins because of their different biological origin (human and bacterial), this was explicitly verified. The result clearly suggests that the folding of a protein molecule obeys rules that are essentially chemical and depend on the physicochemical constraints of the amino acids and of the peptides. Although other factors might be species-dependent, for example, the frequency of specific post-translational modifications or metal uptakes, all these modifications are regulated by the same chemical rules.

Ensembles of NMR models

The determination of protein three-dimensional structures in solution by means of nuclear magnetic resonance (NMR) spectroscopy results typically in ensembles of models, each of which is compatible with the experimental data. The structural variability between the models may span a wide range of levels. In some cases, all the models are nearly identical, with the structural divergence limited to the N- and C-terminal segments. In other cases, several polypeptide segments, especially the loops, may have different conformations amongst the models, and just a small portion of the protein appears to have the same structure in all the models. The structural variability of these model ensembles was examined by means of the PRplots.

All the data were extracted from the Protein Data Bank (Bernstein et al. 1977; Berman et al. 2000). The selection was limited to the protein structures deposited as ensembles of 20 models (this is not a strict rule in the Protein Data Bank but it is a very common practice to deposit 20 models). The maximal identity between pairs of sequences was limited to 30 %, which resulted in 2,179 structures.

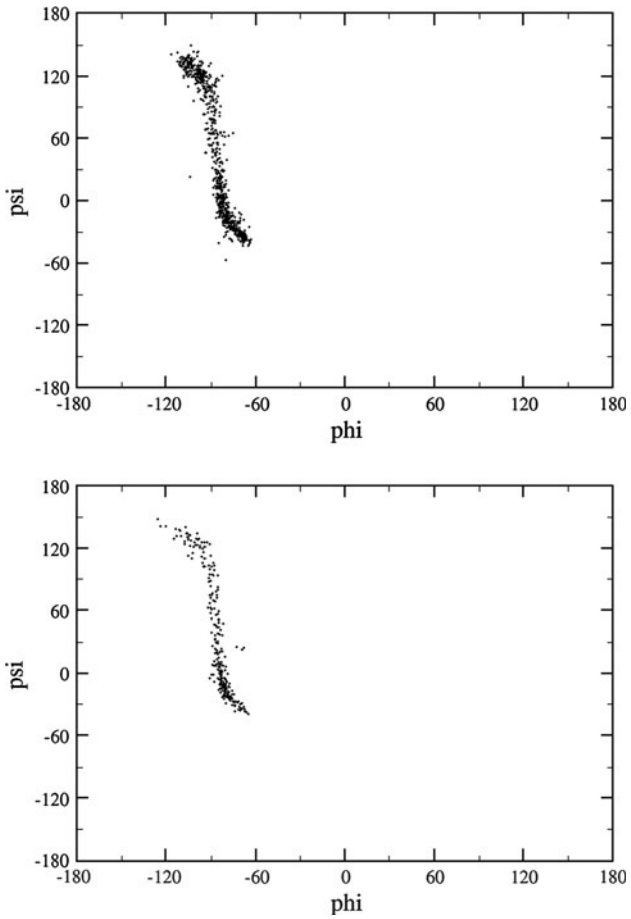

Fig. 5 PRplots computed with a non-redundant set of monomeric and globular proteins from *Homo sapiens* (*top*) and a non-redundant set of monomeric and globular proteins from *Escherichia coli* (*bottom*). In both cases there are more proteins in the α than in the β region, but in the *top image* there are many superposed points in the α region, which consequently appears less populated than it is in reality

The distribution of the dpp values is shown in Fig. 6. Not surprisingly, small values are much more frequently observed than large values, with the consequence that the distribution is unbalanced, with a long queue on the right side of the maximum, at high dpp values. The average dpp value is equal to 11.10° (SD 0.02°). An example is shown in Fig. 7. The chemokine domain of human fractalkine [PDB file 1b2t; (Mizoue et al. 1999)] contains 76 residues with a well-structured core and the two fluctuating termini. The average dpp distance is equal to 16.6° (SD 0.9°). Another example is shown in Fig. 8. The PKD domain of human polycystin-1 is well ordered from the N- to the C-terminus [PDB file 1b4r; (Bycroft et al. 1999)]. There is a very modest structural divergence only in some of the loops. The average dpp is consequently very small (2.30°; SD 0.1°).

It is necessary also to observe that, according to the Ramachandran plot, the stereochemical quality of the models shown in Fig. 7 is considerably worse than what is

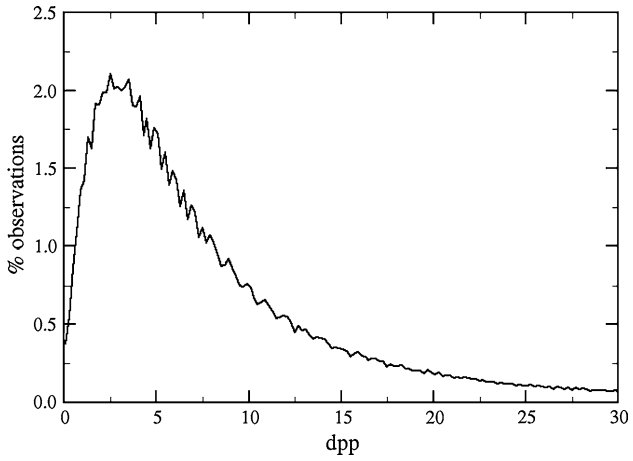

Fig. 6 Distribution of the dpp distances in the NMR protein structures. Only structures deposited as ensembles of 20 models in the Protein Data Bank were considered. All the models of the same ensemble were compared: 190 comparisons were made for each structure of the Protein Data Bank

usually seen in protein structures (Hovmoeller et al. 2002; Chen et al. 2010). To a minor extent, this is also true for the models of Fig. 8, the conformations of which are much less variable. A more detailed analysis of the Ramachandran plots of protein structures determined in solution by NMR spectroscopy is beyond the scope of the present manuscript, though it might enlighten the understanding of protein structure and flexibility.

Protein size and crystallographic resolution

In order to verify if the PRplots depend on the dimensions of the proteins, the Scop database (Andreeva et al. 2007) was divided into four parts, one containing protein domains that have less than 150 amino acids, the second with protein domains that have 150–300 residues, the third containing protein domains with 300–450 amino acids, and the fourth with protein domains that have more than 450 residues. Sequence redundancy was limited to 25 %. Four PRplots were constructed by using these four datasets (see Fig. 9). No statistically significant differences were observed between them, suggesting that the distribution of the proteins on the PRplot does not depend on their size. On the basis of this observation it can therefore be supposed that the three major classes of folds (all α, all β, α–β) are equally distributed amongst proteins independently of their size.

Analogously, in order to verify if different PRplots can be observed at different crystallographic resolutions, four non-redundant (maximal sequence identity = 25 %; R factor <0.3; number of residues between 40 and 10,000) ensembles of proteins were constructed with the PISCES web server (Wang and Dunbrack 2003). One contained protein structures refined at a resolution better (or equal to) 1.5 Å, the second structures refined at 1.5–2.0 Å resolution, the third proteins refined at 2.0–2.5 Å resolution, and the fourth structures refined at 2.5–3.0 Å resolution. The four PRplots corresponding to these four datasets (see Fig. 10) were statistically identical, suggesting that the crystallographic resolution does not affect the PRplots.

The independence of the PRplots appearance from protein size and crystallographic resolution is important since it makes the PRplot tool extremely robust, in the sense that it can be used to describe any protein structure without further (and often heuristic) parameterizations.

PRplots and missing residues

Frequently the conformation of residues cannot be determined experimentally, because some regions of the electron density maps calculated in crystallographic studies cannot be interpreted, leading to undetermined positions of some atoms and/or residues. This can be a problem when the average ϕ and ψ dihedral angles must be computed, because some of the ϕ and ψ values consequently cannot be calculated. It is therefore important to estimate the effects missing residues on the PRplot.

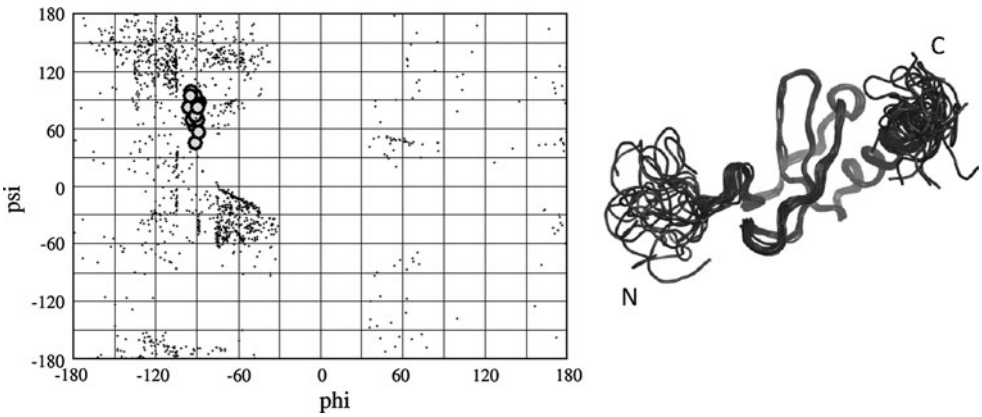

Fig. 7 *Left* Ramachandran plot of the 20 models of the PDB file 1b2t: the single residues are represented by *small black points* while the *larger circles* represent the average backbone torsion angles of each model. *Right* schematic view of the 20 models, which display considerable structural differences at the N- and C-termini

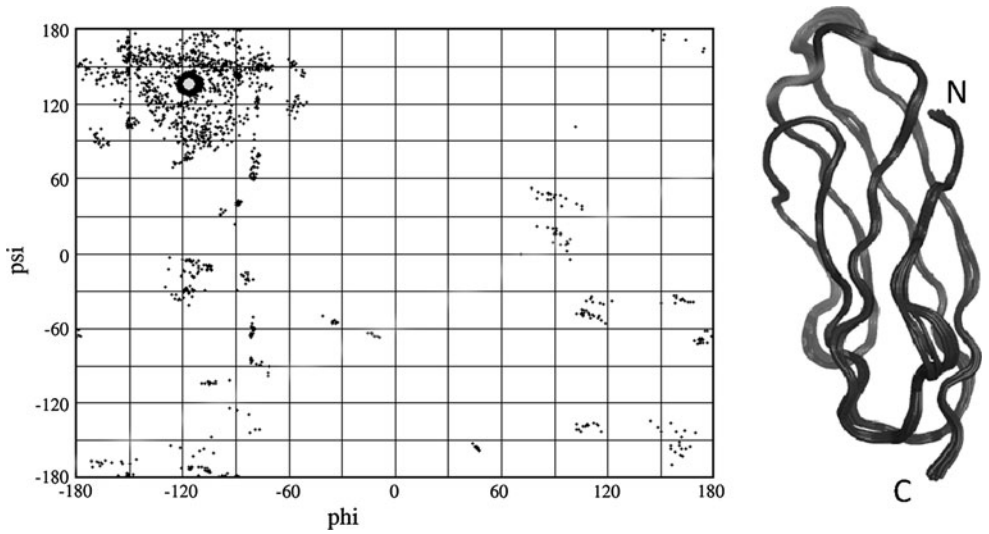

Fig. 8 *Left* Ramachandran plot of the 20 models of the PDB file 1b4r: the single residues are represented by *small black points* while the *larger circles* (nearly superposed to each other) represent the average backbone torsion angles of each model. *Right* schematic view of the 20 models, which differ slightly only in some loops

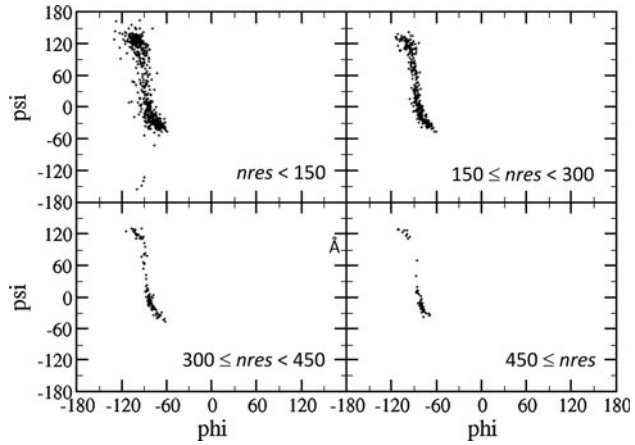

Fig. 9 PRplots for proteins of different dimensions (measured by the number of residues, *nres*)

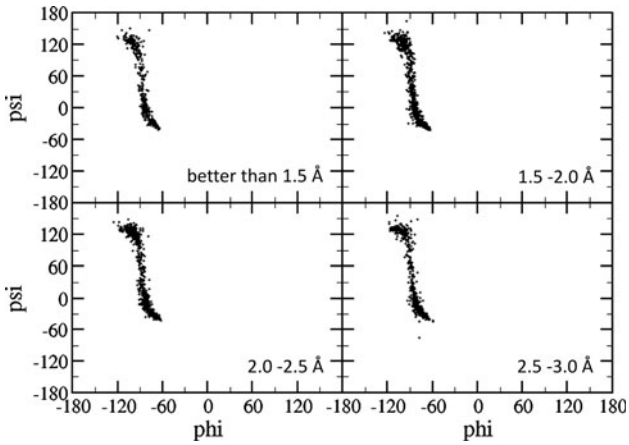

Fig. 10 PRplots for protein structures refined at different crystallographic resolutions

In principle, the best way to address this problem would be the comparison of a large number of protein structures where one or several residues are missing, but for which there a complete entry is available in the PDB. Unfortunately, the paucity of the data hampers this approach, which would be limited to a small number of diverse proteins (e.g., myoglobins and lysozymes). As a consequence, several artificial data were constructed. It is important to realize that these ensembles of artificial data allow only a semi-quantitative estimation of the effects of incomplete protein structure with some of its residues being structurally uncharacterized. In fact, it is reasonable to suppose that the missing residues are not randomly positioned within the protein structure, but often reside in loops at the protein surface. Furthermore, some residue types might be more prone than others to be experimentally undetectable.

Residues were intentionally deleted from the protein structures and the average ϕ and ψ values of the resulting, incomplete proteins were compared to the ϕ and ψ average values of the complete and intact proteins. The comparison was done by computing the average absolute difference between the average values of the ϕ and ψ angles of the intact structures and the average ϕ and ψ values of the proteins lacking some of their residues.

Three types of residue deletion protocols were used:

In the first, termed *Random*, residues selected randomly, were removed. Arbitrarily, we decided to remove 1, 2, 5, 10 or 20 residues in each protein, by using a pseudo-random number generator.

Table 1 Average distances dpp (°) observed by deleting selected residues in a protein structure

Deletion mode	Protein dimension	Number of deleted residues					
		1	2	5	10	20	All loops
Random	100–150	3.7	4.6	6.3	8.4	14.3	–
	300–350	3.4	4.2	5.5	7.8	12.9	–
String	100–15	–	–	5.9	8.5	12.2	–
	300–350	–	–	5.5	7.8	11.0	–
Loops	100–150	–	–	–	–	–	38.1
	300–350	–	–	–	–	–	35.0

See text for the description of the three strategies of residue deletion (random, string, and loops)

In the second deletion protocol, named *String*, a randomly positioned string of adjacent residues was removed in each protein, by using a pseudo-random number generator to select the position of the string to be removed. The length of the string was arbitrarily fixed to 5, 10 or 20 residues.

The difference between the *Random* and the *String* mode of deleting residues is that in the first case the residues that are removed from the protein structures are not necessarily close to each other, while in the *String* mode the residues that are eliminated constitute a short stretch of adjacent amino acids along the protein chain.

In the third deletion scheme termed *Loops*, all the loops, independently of their lengths, were removed. According to this protocol, all the residues that do not assume a helical or an extended secondary structure are assumed to be invisible in the crystal structures. This is of course an extreme hypothesis, since it is well known that loops, not necessarily short, can be well structured in protein crystal structures. However, one should remember that in most NMR solution structures of proteins the conformation of the loops cannot be determined uniquely because of their conformational plasticity.

The comparisons between ϕ and ψ torsions of intact proteins and of proteins lacking some of their residues were performed on two different sets of proteins. Both of them were extracted from the Scop-nr ensemble of protein domain structures; one of them contained smaller domains (100–150 residues) and the other larger domains (300–350 residues).

The variations of dpp distances relative to the native and intact protein are shown in Table 1. The dpp values are rather small, comparable to those observed amongst different NMR models of the same structure (see Figs. 6, 7, 8). They are larger if all the loops are deleted from each protein structure, though this is, as mentioned above, an extreme and probably unrealistic condition.

The fact that these dpp values are not too large is probably due to a compensation effect: some of the missing residues might modify the average ϕ and ψ values in a certain direction, while others might modify them in the opposite direction. Nevertheless, the fact that the position of a structure in a PRplot is little influenced by the absence of some amino acids is a positive feature of the PRplots themselves, highlighting their robustness and wide applicability.

Nevertheless, it is essential to underline that these estimations are only semi-quantitative and thus provide a limited indication of the effects of the incompleteness of protein structures on the position of the structures on the PRplot.

Conclusions

Each protein structure is associated with a point in the PRplot, according to the average values of its ϕ and ψ backbone torsion angles. The PRplot can therefore be used to map the protein structure universe in a simple and intuitive way that resembles closely the Ramachandran plot. Points tend to cluster in the three regions that correspond to the structures rich in α-helices, in β-strands, and in both types of secondary structure.

A sigmoid curve crosses three regions and we propose here that it demarcates the regions of the PRplot that can be populated by protein structures and delineates the structural variability of proteins.

We showed that PRplots are roughly independent of a series of factors (protein dimension, crystallographic resolution, biological source). This is a favorable feature since it makes PRplots highly robust and universal.

Moreover, PRplots appear little affected by the absence of a considerable fraction of residues, which can be "invisible" in the crystal structure. It is reasonable to suppose that this is due to the fact that different missing residues will move the protein in different directions on the PRplot plane, with the consequence that the average ϕ and ψ values are nearly invariant.

Acknowledgments This work was in part supported by the BIN-III project of the Austrian GEN-AU initiative.

References

Andreeva A, Howorth D, Chandonia J-M, Brenner SE, Hubbard TJP, Chothia C, Murzin AG (2007) Data growth and its impact on the SCOP database: new developments. Nucl Acids Res 36:D419–D425

Batschelet E (1981) Circular statistics. Academic Press, London

Berkjolz DS, Shapovalov MV, Dunbrack RLJ, Karplus PA (2009) Conformation dependence of backbone geometry in proteins. Structure 17:1316–1325

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucl Acids Res 28(1):235–242

Bernardo P, Blanchard L, Timmins P, Marion D, Ruigrok RWH, Blackledge M (2005) A structural model for unfolded proteins from residual dipolar coupling and small-angle X-ray scattering. Proc Natl Acad Sci USA 102:17002–17007

Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 112(3):535–542

Branden VI, Jones TA (1990) Between objectivity and subjectivity. Nature 343:687–689

Bycroft M, Bateman A, Clarke J, Hamill SJ, Sandford R, Thomas RL, Chothia C (1999) The structure of a PKD domain from polycystin-1: implications for polycystic kidney disease. EMBO J 16:297–305

Carugo O (2010a) Clustering criteria and algorithm. Methods Mol Biol 609:175–196

Carugo O (2010b) Clustering tendency in the protein fold space. Bioinformation 4:347–351

Chen VB, Arendall WB III, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Cryst D 66:12–21

Dowdy S, Wearden S, Chilko D (2004) Statistics for research. Wiley, Hoboken

Grabarse W, Vaupel M, Vorholt JA, Shima S, Thauer RK, Wittershagen A, Bourenkov G, Bartunik HD, Ermler U (1999) The crystal structure of methenyltetrahydromethanopterin cyclohydrolase from the hyperthermophilic archaeon *Methanopyrus kandleri*. Structure 7:1257–1266

Griep S, Hobohm U (2009) PDBselect 1992–2009 and PDBfilterselect. Nucl Acids Res 38:D318–D319

Holm L, Sander C (2000) Evaluation of protein models by atomic solvation preference. J Mol Biol 225:93–105

Hovmoeller S, Zhou T, Ohlson T (2002) Conformations of amino acids in proteins. Acta Cryst D 58:768–776

Jha AK, Colubri A, Freed K, Sosnick T (2005) Statistical coil model of the unfolded state: resolving the reconciliation problem. Proc Natl Acad Sci USA 102:13099–13104

Kleywegt GJ, Jones TA (1996) Phi/psi-chology: Ramachandran revisited. Structure 4:1395–1400

Laskowski RA, Hutchinson EG, Michie AD, Wallace AC, Jones ML, Thornton JM (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 8:477–486

Lesk AM (2001) Introduction to protein architecture. Oxford University Press, Oxford

Mizoue LS, Baxan JF, Johnson WC, Handel TM (1999) Solution structure and dynamics of the CX3C chemokine domain of fractalkine and its interaction with an N-terminal fragment of CX3CR1. Biochemistry 38:1402–1414

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH—a hierarchical classification of protein domain structures. Structure 5:1093–1108

Pesce A, Couture M, Dewilde S, Guertin M, Yamauchi K, Ascenzi P, Moens L, Bolognesi M (2000) A novel two-over-two alpha-helical sandwich fold is characteristic of the truncated hemoglobin family. EMBO J 19:2424–2434

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

Samudrala R, Levitt M (2000) Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction. Protein Sci 9:1399–1401

Sikic K, Carugo O (2010) Protein sequence redundancy: comparison of various methods. Bioinformation 5:234–239

Song J, Lee MS, Carlberg I, Vener AV, Markley JL (2006) Micelle-induced folding of spinach thylakoid soluble phosphoprotein of 9 kDa and its functional implications. Biochemistry 45:15633–15643

Vogt J, Schulz GE (1999) The structure of the outer membrane protein OmpX from *Escherichia coli* reveals possible mechanisms of virulence. Structure 7:1301–1309

Walther D, Cohen FE (1999) Conformational attractors on the Ramachandran map. Acta Cryst D 55:506–517

Wang G, Dunbrack RLJ (2003) PISCES: a protein sequence culling server. Bioinformatics 19:1589–1591

Zhou Y, Faraggi E (2010) Prediction of one-dimensional structural properties of proteins by integrated neural networks. In: Rangwala H, Karypis G (eds) Introduction to protein structure prediction. Wiley Series in Bioinformatics. Wiley, Hobokan, pp 45–65

Zhou AQ, O'Hern CS, Regan L (2011) Revisiting the Ramachandran plot from a new angle. Protein Sci 20:1166–1171