PROTEINS: Structure, Function, and Genetics 18:324–337 (1994)

Analysis of Cα Geometry in Protein Structures

T.J. Oldfield and R.E. Hubbard
Department of Chemistry, University of York, Heslington, York YO1 5DD, United Kingdom

ABSTRACT **The polypeptide of a protein molecule can be considered as a chain of Cα atoms linked by pseudobonds between the Cα atoms of successive amino acid residues. This paper presents an analysis of the angle and dihedral angles made by these pseudobonds in protein structures determined at high resolution by X-ray crystallography. This analysis reveals a strong correlation between Cα geometry and the protein fold. The regular features of protein secondary structure such as α-helix and β-sheet are very clearly defined. In addition, it is possible to identify with some confidence the discrete populations of particular conformations of β-turn. Comparison with the traditional Ramachandran type of plot demonstrates that an analysis of protein structure on the basis of Cα geometry provides a richer description of protein conformation. In addition, the characteristics of this geometry could be a useful guide in model building of protein structure.**

Key words: protein structure, secondary structure, peptide geometry, Ramachandran plot, β-turns

INTRODUCTION

For a particular amino acid residue n in a polypeptide chain, the angle ϕ is the dihedral angle C(n − 1), N(n), Cα(n), C(n) and the angle ψ is the dihedral angle N(n), Cα(n), C(n), N(n + 1) (Fig. 1a). Ramachandran and Sasisekharan[1] demonstrated that only certain combinations of ϕ and ψ were possible by considering steric overlap of the atoms in the amino acid residues $n - 1$, n and $n + 1$. The plot of the allowed regions for all combinations of ϕ and ψ (known as the Ramachandran plot) remains an essential method in analyzing the quality of protein structures. With some important exceptions, the conformations of all the amino acid residues of high resolution protein crystal structures are found to fall in these allowed regions.

Further analysis of the Ramachandran plot revealed that regions on the plot can be classified on the basis of the type of secondary structure observed experimentally in proteins. The largest of these is for those residues in β-strands, the second largest for the helices, and a third that is only partially allowed and is associated with left-handed helical forms of structure. Apart from the residue glycine (which with no Cβ atom has much greater conformational freedom), exceptions usually indicate poorly defined structure or a portion of the molecule that is sterically strained for folding or functional reasons. The Ramachandran plot can therefore be used to characterize the quality of a protein structure.

The secondary structure of proteins can be defined in terms of α-helices, β-strand (and β-sheet), β-turns composed of 4 residues, and undefined loop structure. Overall for the known protein structures, these categories are found in approximately equal proportions. α-Helical structure is usually very well defined, with the value of ϕ and ψ limited to within ±20°. In addition, the α-helix is stabilized by a hydrogen bond between the i and i + 4 residue. The β-strand is not as well defined by ϕ and ψ, and many residues that lie in β-sheet have a range of conformations that allows for twist and bulges in sheet structures.[2–4] The β-turn is normally defined as a sequence of four residues where the distance between the Cα(i) and Cα(i + 3) is less than 7 Å, and the residues are not in a helix.[5] Two types of classification have been described. The first by Venkatachalam[6] and Lewis et al.[5] is subdivided into the types I, I′, II, II′, III, III′, IV, V, VI, VII, though this is now truncated to I, I′, II, II′, III, III′, VIa, VIb, VII.[7] The second, and more recent, uses the Ramachandran angles ϕ and ψ for the residues (i + 1) and (i + 2), and defines the turn by the region occupied within the Ramachandran plot for these two residues.[8] This latter analysis shows that the possible number of 4 residue segment conformations is varied, but specific populations of turn type can be defined.

In this paper we present an analysis of the geometry of protein structure by considering the internal geometry for successive Cα atoms in a polypeptide chain. For each four residue segment we define three geometric terms. These are the internal angle θ_1 defined as Cα(i), Cα(i + 1), Cα(i + 2), the internal angle θ_2 defined as Cα(i + 1), Cα(i + 2), Cα(i + 3), and the dihedral angle τ defined by the atoms Cα(i),

Received August 23, 1993; revision accepted November 15, 1993.
Address reprint requests to R.E. Hubbard, Department of Chemistry, University of York, Heslington, York YO1 5DD, UK.

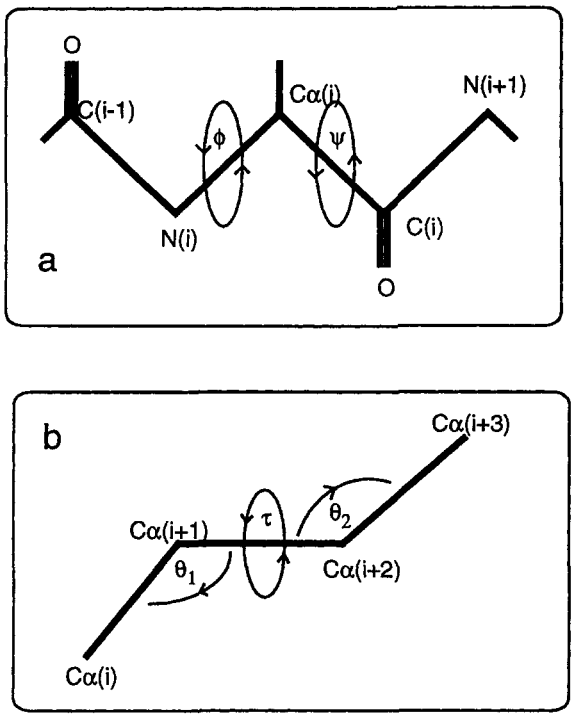


Fig. 1. (a) Definition of the Ramachandran torsion angles φ and ψ. (b) Definition of the pseudoangles and pseudotorsion angles θ_1, θ_2, and τ.

Cα(i + 1), Cα(i + 2), Cα(i + 3) (Fig. 1b). A number of previous studies have used these pseudotorsion angles and distance, but a detailed analysis of the correlation between these parameters has not been presented before. Levitt[9] used the reduced description of structure in terms of θ_1 and τ to generate simplified geometric energy calculations, and showed that

$$\tau \approx 180° + \phi(i + 2) + \psi(i + 1) + 20\,[\sin\{\phi(i + 1)\} + \sin\{\psi(i + 2)\}]$$
$$\theta_1 \approx 106° + 13\cos(\tau - 45)$$

The distribution of values of θ_1 as a function of τ has been used as a probability function to predict "native-like" random walk structures of proteins.[10] A series of articles by Rackovsky and Scheraga describes the derivation of two parameters which define the conformation of a 4 residue segment of Cα atoms and used this for protein analysis and classification.[11–16]

Our analysis shows that the Cα geometry within a protein falls into well-defined regions. We present here a preliminary analysis of these regions and demonstrate that a distinctive classification of protein structure is possible considering the Cα positions of four successive amino acids in a structure. This definition of protein geometry using only α carbon atoms has several potential advantages in the fields of protein crystallography and molecular modeling, by providing a rule base to guide model building. Details of the applications of Cα geometry will be presented elsewhere.

METHODS

For the Cα geometry analyzed here, it is not possible to calculate a simple energy function or hard sphere region equivalent to that used for a Ramachandran plot because there are many degrees of freedom over the four residues. It was therefore necessary to define an allowed region empirically from a database of proteins. The coordinates for 83 protein structures were selected from the October 1991 release of the protein data bank[17] on the basis of the following criteria:

1. The structures were determined by X-ray crystallographic analysis and refined by restrained least squares techniques to a resolution of 2 Å or better.
2. The highest resolution structure is used where multiple entries occur.
3. Where there are multiple entries of the same resolution the latest entry is taken.
4. Site-directed mutations of a structure are not considered different.

The selection of structures used in the analysis was made to minimize the number of closely related proteins. We realize that there remains a number of homologous proteins in the database, but using a high threshold for significance (50) ensures that we are sampling conformations from nonhomologous structures.

Table I lists brief details of these coordinate sets. These selection criteria give an approximately equal spread of β-sheet and α-helical structure within the proteins. The coordinates from the data bank files for these proteins were processed into a statistical data base and analyzed to generate the 18,503 values for each of the parameters θ_1, θ_2, and τ defined in Figure 1b. Atomic data were selected for each calculation of angle and torsion on the basis that all the atoms have nonzero occupancy and were from consecutive residues covalently bound to each other in the polypeptide chain.

Although the three parameters θ_1, θ_2, and τ could be represented in three dimensions as a contoured plot, it was found that such a detailed map was difficult to interpret. In addition, analysis of such a plot would be complicated by the interdependency of θ_1 and θ_2. For this reason, the data were analysed by considering two-dimensional contour plots. The program SQUID[18] was used to analyze this database of angles to produce contoured plots of $\tau(\theta_1)$, $\tau(\theta_2)$, $\tau(\theta_1)$–$\tau(\theta_2)$, and $\theta_2(\theta_1)$ where, for example, $\tau(\theta_1)$ is the two-dimensional distribution of τ plotted as a function of θ_1. The data were quantized in 10° steps for the torsion angles, and 5° steps for the angles. This results in 1296 bins of data, giving a 36 × 36 point

TABLE I. Proteins Used in This Study*

Number	Code	Resolution	Nres	% Helix	% Sheet	% Coil
1	1alc	1.7	122	39	18	43
2	1amt	1.5	60	88	0	12
3	1ccr	1.5	111	39	5	56
4	1cdp	1.6	108	47	1	52
5	1cho	1.8	291	11	34	55
6	1crn	1.5	46	43	17	39
7	1cse	1.2	336	28	27	45
8	1ctf	1.7	68	51	16	32
9	2cyp	1.7	293	47	10	43
10	1fd2	1.9	106	34	9	57
11	1fx1	2.0	147	29	24	46
12	1gcr	1.6	174	5	64	31
13	1gd1	1.8	572	30	27	43
14	1gox	2.0	350	42	18	40
15	1gp1	2.0	368	26	22	52
16	1hne	1.8	213	5	39	56
17	1hoe	2.0	74	0	51	49
18	1i1b	2.0	150	3	59	38
19	1mba	1.6	146	79	0	21
20	1ntp	1.8	219	9	39	53
21	1p01	2.0	178	4	56	39
22	1paz	1.5	120	17	44	39
23	1psg	1.6	322	22	37	41
24	1rdg	1.4	52	12	27	62
25	1s01	1.7	275	31	25	44
26	1sgc	1.8	168	7	61	32
27	1sgt	1.7	213	10	46	44
28	1snc	1.6	135	30	28	42
29	1thb	1.5	574	74	0	26
30	1tld	1.5	220	9	40	50
31	1tmn	1.9	318	38	22	40
32	1ton	1.8	220	10	40	50
33	1ubq	1.8	76	22	37	41
34	1utg	1.3	70	76	1	23
35	1xy1	1.0	14	0	0	100
36	1ypi	1.9	494	42	18	40
37	256b	1.4	212	82	0	18
38	2alp	1.7	172	5	66	30
39	2aza	1.8	258	16	41	43
40	2ca2	1.9	256	15	39	46
41	2cga	1.8	490	10	32	58
42	2ci2	2.0	65	18	31	51
43	2cro	2.0	63	65	0	35
44	2er7	1.6	323	10	40	50
45	2fb4	1.9	435	6	49	45
46	2fd2	1.9	106	32	16	52
47	2gbp	1.9	309	44	22	34
48	2ltn	1.7	458	2	34	65
49	2mlt	2.0	52	83	0	17
50	2ovo	1.5	56	20	23	57
51	2prk	1.5	279	29	28	43
52	2sec	1.8	337	28	27	45
53	2sga	1.5	167	8	58	34
54	2tmn	1.6	317	38	22	40
55	2wrp	1.6	105	74	0	26
56	3app	1.8	323	10	43	47
57	3apr	1.8	331	10	43	47
58	3b5c	1.5	86	35	14	51
59	3bcl	1.9	322	19	44	37
60	3cpp	1.9	405	46	10	43

(continued)

TABLE I. Proteins Used in This Study* (Continued)

Number	Code	Resolution	Nres	% Helix	% Sheet	% Coil
61	3ebx	1.4	51	0	53	47
62	3est	1.6	229	10	41	48
63	3grs	1.5	461	34	28	39
64	3ins	1.5	102	48	0	52
65	3lzm	1.7	164	60	8	32
66	3mcg	2.0	430	6	44	51
67	3rnt	1.8	104	15	29	56
68	3rp2	1.9	430	7	41	52
69	3sgb	1.8	226	9	48	43
70	4pep	1.8	326	12	42	46
71	4ptp	1.3	219	9	39	53
72	4tnc	2.0	160	60	4	36
73	5cha	1.7	474	8	35	57
74	5cpv	1.6	108	46	1	53
75	5cts	1.9	429	61	2	38
76	5cyt	1.5	103	38	6	56
77	6rxn	1.5	45	13	29	58
78	7gch	1.8	237	8	34	58
79	7pcy	1.8	93	9	47	44
80	7rsa	1.3	124	23	38	40
81	7wga	2.0	340	6	31	63
82	8dfr	1.7	186	20	37	44
83	9pap	1.6	212	29	25	47

*For each protein the table includes its code, quoted resolution, number of amino acid residues (Nres), percentage of residues that are in an α-helix, percentage of residues in β-sheet, and the remaining percentage of residues. In each case the proportion of α-helix and β-sheet was calculated using the rules of Kabsch and Sanders,[19] using the hydrogen bonding criteria of Baker and Hubbard.[23]

surface which when contoured forms the basis of the analysis presented here.

In some analyses, the data were selected on the basis of the residue type and secondary structure type at each of the four positions of the 4 residue segments. The secondary structure for each residue was assigned using the procedures of Kabsch and Sander.[19] In this way an assessment could be made of the effect on the parameters θ_1, θ_2, and τ of residue type or fold at a particular position in the segment.

RESULTS AND ANALYSIS

General Descriptions of the Distributions

Figure 2a shows a contoured plot of the observed ϕ and ψ values for all residues within the database of 83 proteins. One of the important features of this Ramachandran plot is that there is a clear definition of allowed and disallowed regions of conformational space, and this can be seen in this figure. As expected, there are 3 rather large but discrete regions that correspond to residues in helical conformations, β-strand conformations, and reverse turn conformations. There are some subpeaks particularly within the β-strand region, the main part of which is bilobal. However, it is difficult to associate these regions with particular subclasses of secondary structure. At lower contour levels, the distribution of glycine residues becomes apparent within the regions of the Ramachandran plot not normally accessible to L-α-amino acids (Fig. 2b).

Figures 3a–c and 4 show the contoured plots of $\tau(\theta_1)$, $\tau(\theta_2)$, $\theta_2(\theta_1)$, and $\tau(\theta_1)$–$\tau(\theta_2)$, for all the residues in the 83 proteins. There are three overall features of these plots which are of note. First, the values of θ_1 and θ_2 are confined almost exclusively to angles between 75° and 160°, and although all values of τ are observed they are not evenly distributed. The second striking feature of the plots of $\tau(\theta_1)$ and $\tau(\theta_2)$ is that there appear to be two main regions, each with a large peak for which the values of $\tau(\theta_1)$ or $\tau(\theta_2)$ are tightly defined at higher contour levels. This is surprising considering the number of degrees of freedom within the polypeptide chain that the functions $\tau(\theta_1)$ and $\tau(\theta_2)$ encompass. Third, there appears to be considerable fine structure in the contour plots.

Before discussing the results of the data within these plots in detail, we consider the effect of the choice of proteins on the distribution of Cα geometry.

Data Selection

The results presented here are based on an empirical calculation and therefore depend on the database of proteins used. The distributions $\tau(\theta_1)$, $\tau(\theta_2)$, $\tau(\theta_1)$–$\tau(\theta_2)$, and $\theta_2(\theta_1)$ have been generated from sev-

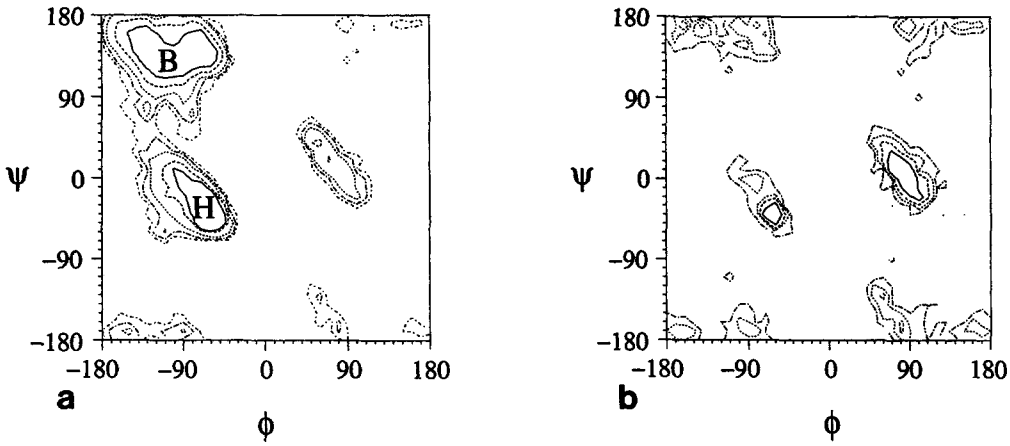


Fig. 2. (**a**) Distribution of 18,670 Ramachandran torsion angles ϕ and ψ from the database of 83 proteins. The distribution is contoured at the levels of 100, 50, 20, 10, and 5 examples per 10° × 10° bin. (**b**) Distribution of 1741 Ramachandran torsion angles ϕ and φ for the glycine residues from the database of 83 proteins. The distribution is contoured at the levels of 20, 10, 5, and 2 examples per 10° × 10° bin.

eral subsets of proteins from the database. The overall distributions have been found to be particularly robust, but weaker features tend to become more obvious with "better" quality proteins. To study the effect of resolution, two subsets of the 83 proteins were generated using a resolution range of 1.6 Å or better (38 proteins) and a resolution range between 1.6 and 2.0 Å (45 proteins), and their Cα geometry analyzed as described above (results not shown). The only statistically significant feature is a sharpening of the major peaks, which would be expected for better resolved structures. The changes as a function of resolution were calculated as the difference between the normalized distributions for the two sets of data. The remaining plot has variations of up to 15 values of occurrence for each bin of data with apparently no consistent bias. Limiting the number of proteins does influence the proportion of structure types (α, β, α + β, and α/β). This makes it difficult to separate detailed effects of resolution on θ_1, θ_2, and τ from changes in the distribution of such structure types.

The Major Peaks

Analysis of the conformation of the residues that correspond to the two major regions of $\tau(\theta_1)$ and $\tau(\theta_2)$ shows, not surprisingly, that they define α-helices and β-strands. The peaks are marked on Figure 3 (H, α-helix; B, β-strand). The largest peak at $\tau = 50° \pm 20°$ and $\theta_1 = \theta_2 = 95° \pm 10°$ is due to the residues that lie in α-helices. The peak is essentially the same height and extent in both of the plots $\tau(\theta_1)$ and $\tau(\theta_2)$ due to the internal symmetry of a helix. The peak is also very tightly defined occupying only about 1.5% of the total area of the $\tau(\theta_1)$ and $\tau(\theta_2)$ map.

The second largest peak at $\tau = -145° \pm 50°$ and $\theta_1 = \theta_2 = 125° \pm 20°$ is due to the residues that lie in a β-sheet conformation. This region, as expected, is more extensive and occupies about 6% of the map. The peak does not have the same extent in the two plots of $\tau(\theta_1)$ and $\tau(\theta_2)$, but has a skew in the opposite direction within the two functions. The skew pattern was found to be independent of the parallel/antiparallel nature of a β-sheet, but dependent on the twist within the β-sheet. As the torsion angle τ increases from −165°, the angle θ_1 increases, and the angle θ_2 decreases. In the case of parallel β-sheet this is due to the slight left-handed helical character preferred by residues[20] resulting in a twisted sheet.[3] For antiparallel sheet structure, there is the possibility for greater degrees of twist and it is common to find a coiled double antiparallel sheet[4] which would explain the more extreme values observed for differences between θ_1 and θ_2.

The plots of $\theta_2(\theta_1)$ and $\tau(\theta_1)$–$\tau(\theta_2)$

The distribution $\theta_2(\theta_1)$ explores the relationship between the two virtual Cα angles. If there was a complete correlation between these angles, then this plot would appear as a straight line. Figure 3c is neither a straight line nor symmetric showing that the two internal angles of a four residue segment are different and not directly related. So this plot reflects that for the virtual Cα description of a molecule, a direction can be associated with a polypeptide chain. This observation explains the differences between the plots $\tau(\theta_1)$ and $\tau(\theta_2)$, while the distribution $\tau(\theta_1)$–$\tau(\theta_2)$ shows where these major differences in θ_1 and θ_2 occur as a function of τ. The plot of $\tau(\theta_1)$–$\tau(\theta_2)$ can be used as an aid in determining interesting features from $\tau(\theta_1)$ and $\tau(\theta_2)$ plots as it removes by subtraction the large symmetric peaks (e.g., α-helix). The most striking feature of Figure 4 is the number of discrete regions that can be identified. The analysis of these features forms the remainder of this paper.

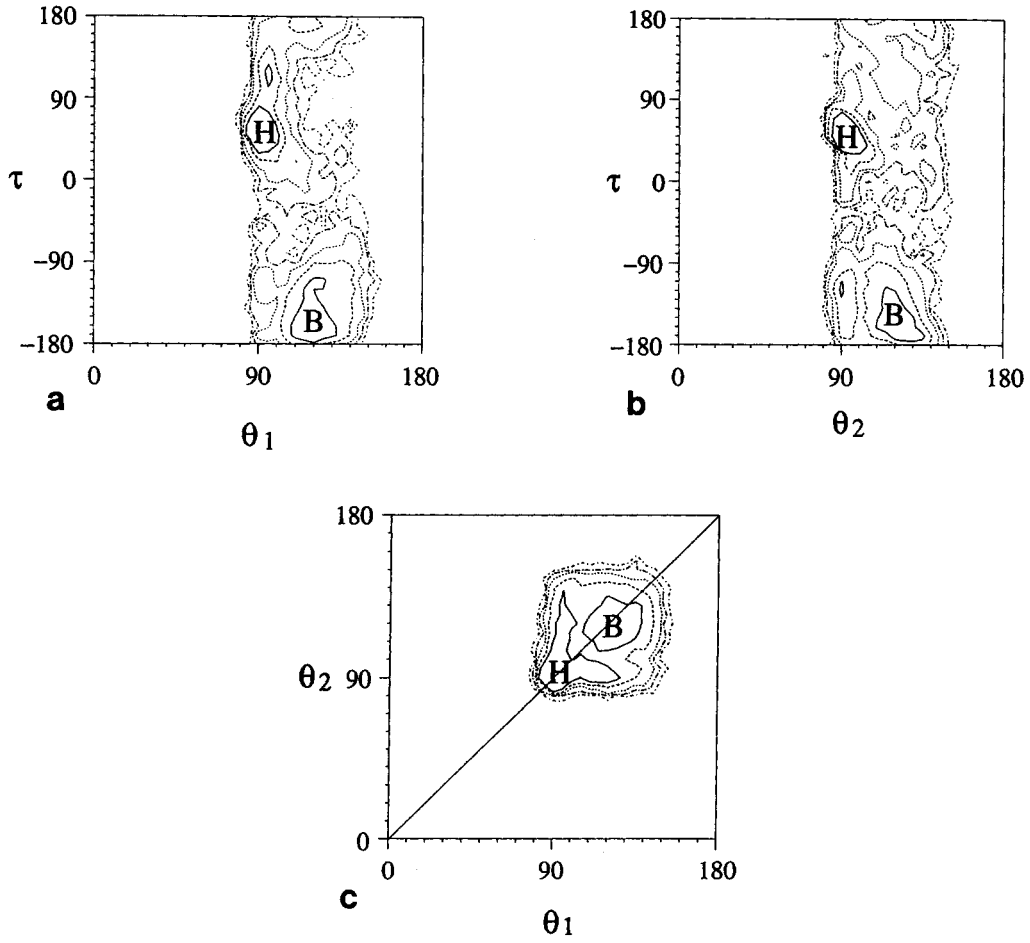


Fig. 3. Distribution of 18,503 values of the pseudotorsion τ and the pseudoangles θ_1 and θ_2. (**a**) Pseudotorsion τ plotted as a function of the pseudoangle θ_1. (**b**) Pseudotorsion τ plotted as a function of the pseudoangle θ_2. (**c**) Pseudoangle θ_2 plotted as a function of the pseudoangle θ_1. The distribution is contoured at the levels of 100, 50, 20, 10, and 5 examples per 5° × 10° bin for (**a**) and (**b**), and per 5° × 5° bin for (**c**). The region corresponding to residues in α-helices is marked H, that for residues in β-sheet conformations as B.

The Minor Peaks

For all the plots shown in Figures 3 and 4, there are many features at low contour levels that appear as discrete peaks within the distributions $\tau(\theta_1)$ and $\tau(\theta_2)$. About 25% of the area of the plots of functions $\tau(\theta_1)$ and $\tau(\theta_2)$ are taken up with these low probability regions which have no specific equivalent areas within a Ramachandran plot. These small features correspond to those residues that do not have a conformation that is either α-helical or extended β-strand. The median values for the nonzero bins of the distributions of $\psi(\phi)$, $\tau(\theta_1)$, and $\tau(\theta_2)$ presented here are 4, 12, and 14, respectively. These median values reflect the number of small but significant populations of data that are not in the two secondary structure types α-helix and β-strand. The values of the median for the plots of $\tau(\theta_1)$ and $\tau(\theta_2)$ also allow the estimation of a limit for the significance of a feature in the distribution. In general, in this paper we have analyzed only features for which there are at least 50 values in a data bin.

It is interesting that the diffuse region is markedly different for $\tau(\theta_1)$ and $\tau(\theta_2)$, and stronger features in one are not present in the other. It should also be noted that as τ is the same in both plots, any peak in one plot can only move horizontally in the other plot. This can be seen in the plot of $\tau(\theta_1)$–$\tau(\theta_2)$ shown as Figure 4.

Analysis of the minor peaks

We have looked in detail at the conformation and composition of the 4 amino acid segments that make up each region in the plot of $\tau(\theta_1)$–$\tau(\theta_2)$. We have investigated the effect on the distribution of the presence of glycine or proline at each of the 4 residue positions, on the effect of other amino acids at various positions, and attempted to correlate particular regions on the $\tau(\theta_1)$–$\tau(\theta_2)$ plot with particular combinations of φ and ψ on the Ramachandran diagram. Conformations defined by φ and ψ for residues $i + 1$ and $i + 2$ are referenced by the nomenclature of Wilmot and Thornton.[8] The Ramachandran plot is

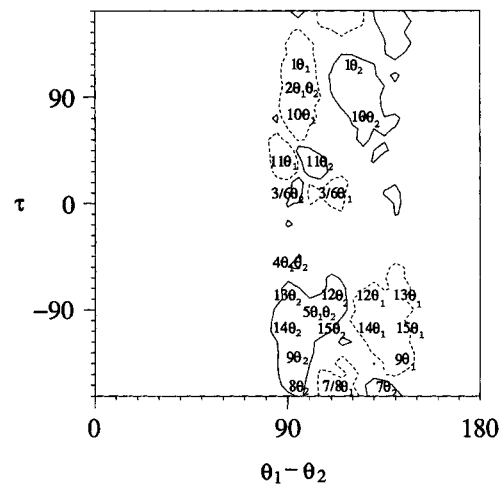


Fig. 4. Distribution calculated as the difference between the function $\tau(\theta_1)$ and $\tau(\theta_2)$. The regions that define the peaks 1 to 15 are marked on the difference plot where, for example, $1\theta_1$ is the region that corresponds to peak 1 for the pseudoangle θ_1, and $1\theta_2$ is the regions that corresponds to peak 1 for the pseudoangle θ_2. The label, for example, $4\theta_1\theta_2$ indicates that the pseudoangles θ_1 and θ_2 are the same, and hence cannot be observed as a significant peak on this difference diagram.

divided into regions αR, αL, β, βP, γL, and ε for right- and left-handed α-helical, β-sheet, β-proline, left-hand γ turns and extended conformations, respectively.

A four residue turn has been analyzed here if the conformation defined by θ_1, θ_2, and τ occurs more than 50 times out of the database of 18,503 values over the assigned area of the difference plot. The feature was also required to have a minimum area of $100^{\circ 2}$. This area restriction for deciding if a peak is significant is rather arbitrary, but is well above the noise level in the distributions. There are probably other turn types with lower frequency of occurrence to be identified from these plots. As the size of the protein databank increases, these structural features may become more clearly identifiable.

The results of this analysis are shown in Table II, and Figures 4 and 5a–g. Figure 4 shows the difference plot $\tau(\theta_1)$–$\tau(\theta_2)$ with the significant peaks identified. For example, peak $1\theta_1$ is the region that corresponds to peak 1 for the pseudoangle θ_1, and peak $1\theta_2$ is the region that corresponds to peak 1 for the pseudoangle θ_2. The major peaks in Figure 4 are labeled from peak 1 to peak 15 with θ_1 and/or θ_2 indicating the angle from which the peak arises in the difference plot. A subdivision of these regions was made after inspection of the Ramachandran diagrams. For example, for peak 6, inspection of the Ramachandran diagram for the residue fragments that contribute to this peak in Figure 4 clearly shows that there are two distinct populations of paired Ramachandran angles that contribute to the same peak (see Fig. 5a). These populations are labeled as conformations 6(i) and 6(ii).

It was found possible to associate some of the peaks with the presence of glycine or proline at particular positions. It was not possible to identify with any confidence any affect on the distributions attributable to the presence of other residue types. Table II shows that 7 of the peaks are highly correlated to the position of a glycine residue. Although glycine, without a β carbon, can lie in many different conformations, the conformations taken up by glycine tend to be those specifically dissallowed for L-α-amino acids. The Ramachandran plot for all glycine residues indicates that glycine often assumes a conformation that is associated locally with a left-handed helix, or a β-sheet with a twist not found with L-α-amino acids (Fig. 2b). Hence glycine is very important in allowing a particular conformation for a turn or loop in a polypeptide chain.

For the 17 significant conformations identified in Figure 4 (helix, sheet, and 15 turns), the presence of glycine at a particular position is important for 7 of them and the absence of glycine for just one conformation. For the other significantly populated regions one is strongly correlated to the proline residue at Cα(i + 2) and occurs for $\theta_1 = 95°$, $\theta_2 = 110°$, $\tau = -160°$ (peak 8ii). The remaining peaks have no significant correlation to any residue.

The following sections discuss the correlation of the individual peaks with the ϕ and ψ values that the contributing residues occupy on the Ramachandran plot, the assignment of individual peaks to conventional turn types, and the occurrence of particular residues at particular positions in the 4 residue segment. The results are presented in Figure 5a–g. For each peak, we present a composite plot showing the change in Ramachandran angles for residues i + 1 and i + 2 for each of the conformations within the database that make up that particular peak. In addition, we show a representative figure of the conformation taken from the database. It is from these plots that it can be seen that there is a high correlation of the pseudo-Cα angle and torsion with the Ramachandran angles. Some correlation should be expected as the pseudoangles and torsion are dependent to some extent on the four torsion angles that make up the two Ramanchandran points. The high correlation shows that the four residue turn conformation is, in most cases, completely defined by the two Ramachandran angles. However, the spread of values of Ramachandran angles is larger than the corresponding spread for the pseudoangles and torsion calculated from the Cα atom positions.

Peaks 3 and 6 (Fig. 5a)

Both peaks have values of $\theta_1 > \theta_2$ and are overlapping on Figure 4. The two peaks are differentiated by the presence or absence of a glycine at residue i + 2. There are 87 examples of peak 3 in the

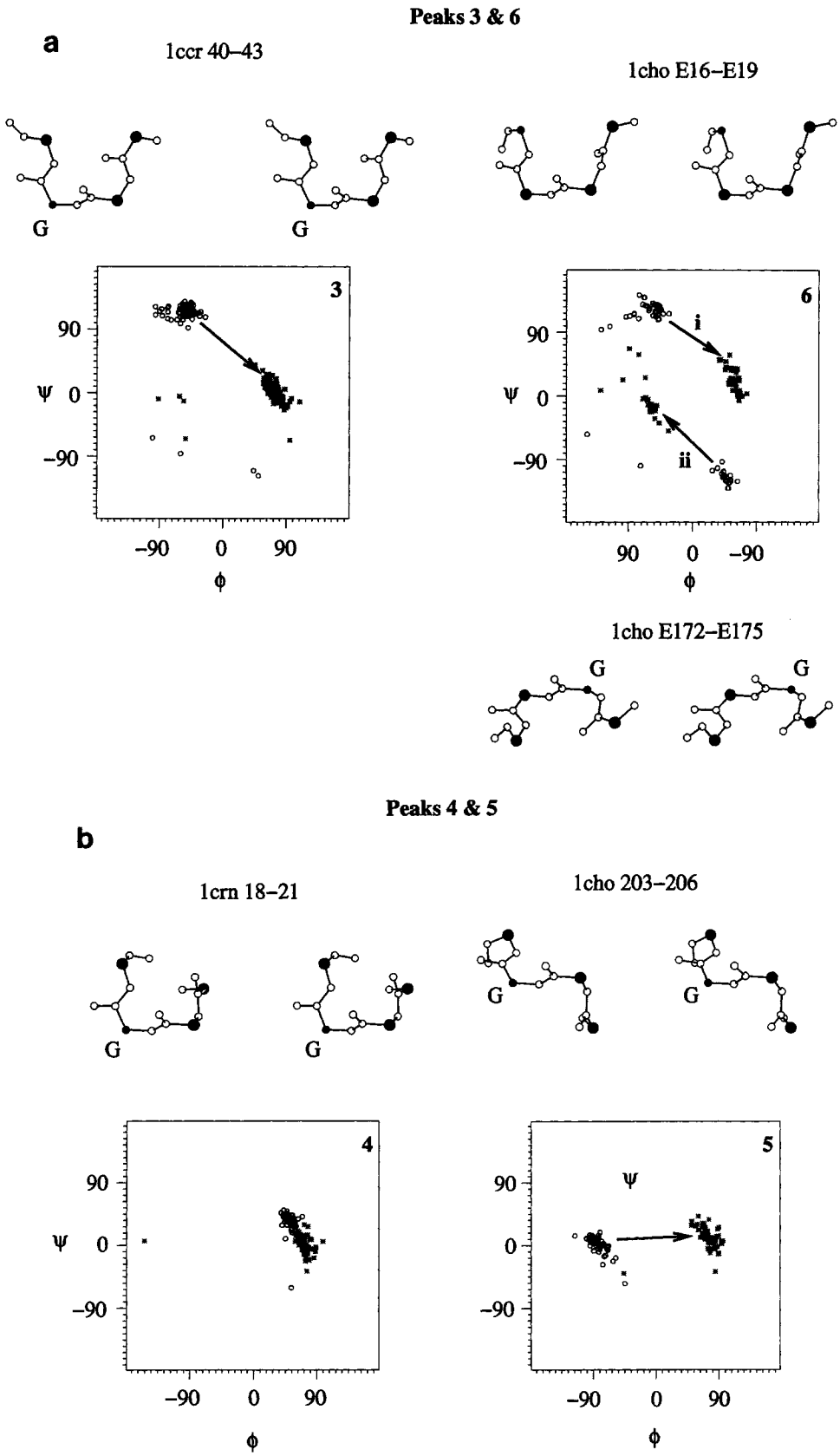


Fig. 5. The 7 figures (a–g) show a summary for all 15 peaks that define different Cα conformations. The figures have been grouped so that similar conformations are shown within the same figure. Ramachandran plots are presented for the $i + 1$ (○) and $i + 2$ residue (*) with arrows showing the connection between clusters of values. A stereo diagram is shown of a representative four residue section of protein.

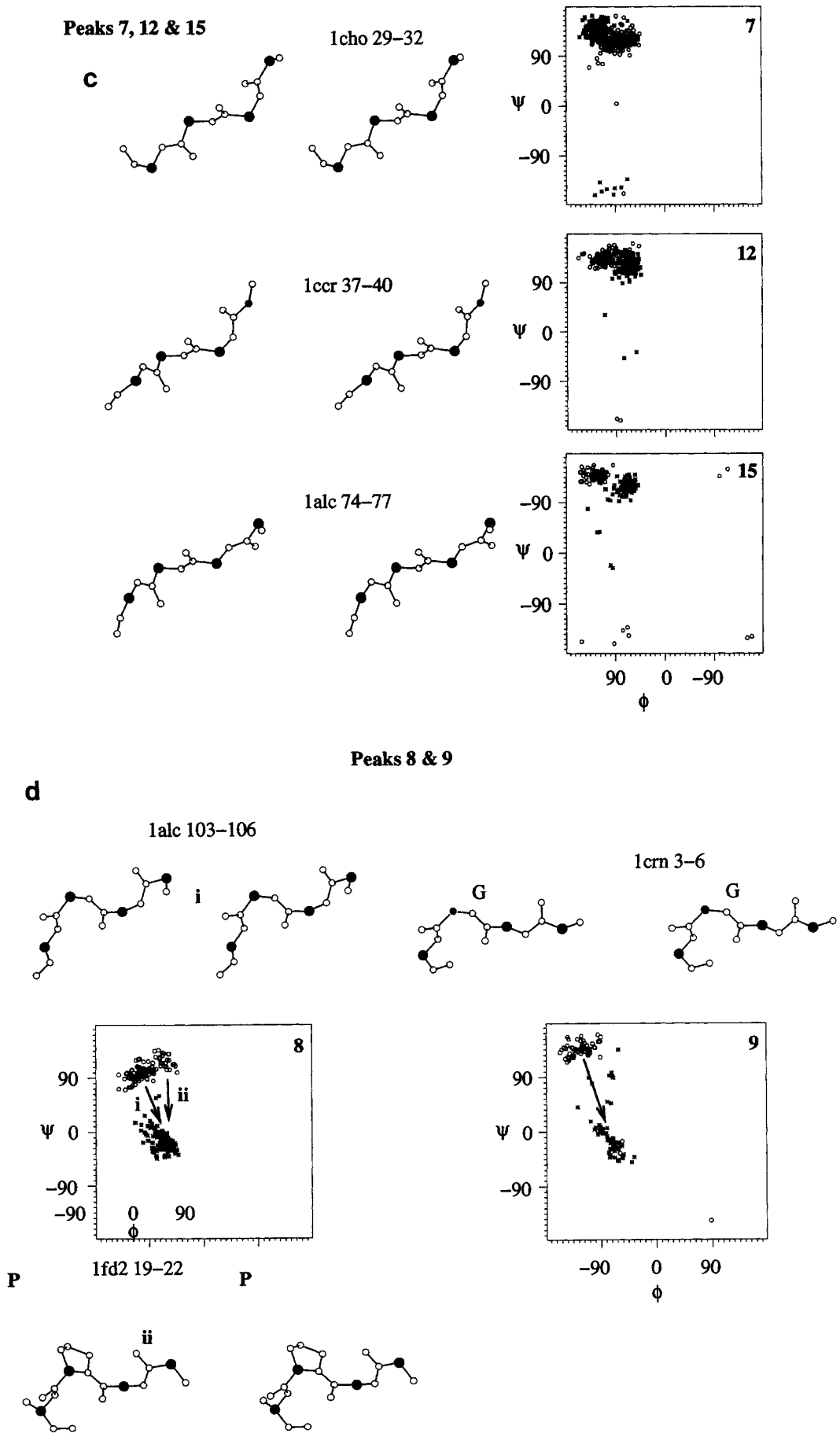
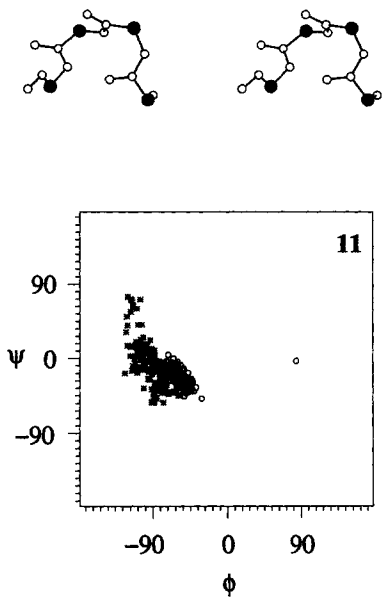

Peaks 7, 12 & 15
c
1cho 29–32
1ccr 37–40
1alc 74–77
7
12
15
90
0
-90
ψ
ϕ
Peaks 8 & 9
d
1alc 103–106
i
1crn 3–6
G
8
9
ii
1fd2 19–22
P
P

Fig. 5c–d.

Peak 11

e

1fx1 26–29



Peaks 13 & 14

f

1cho 60–63

1cho 227–230

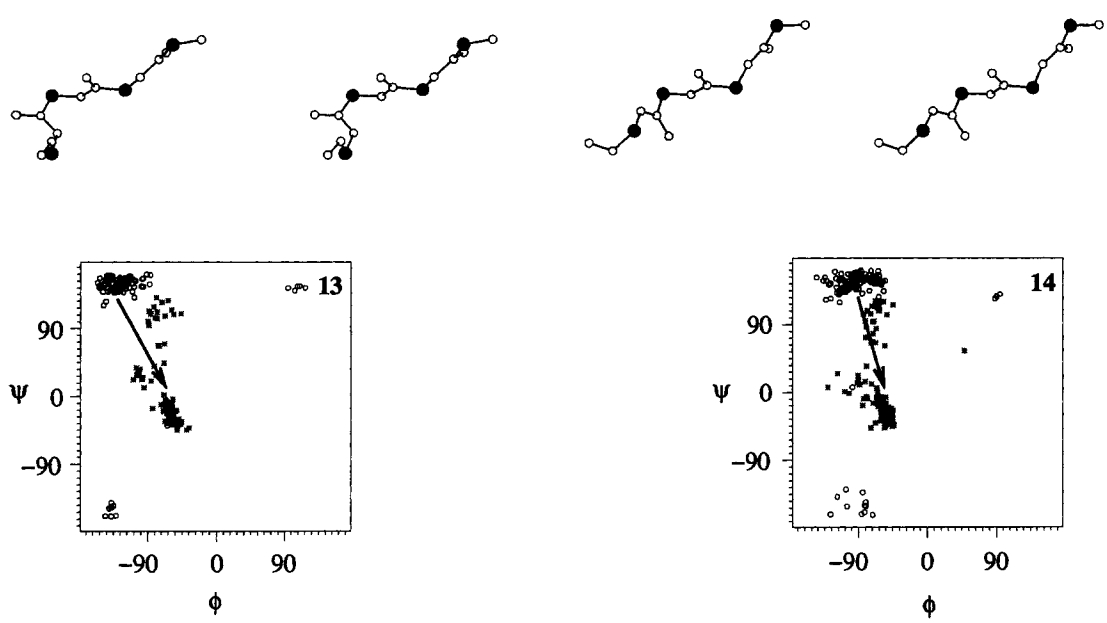


Fig. 5e–f.

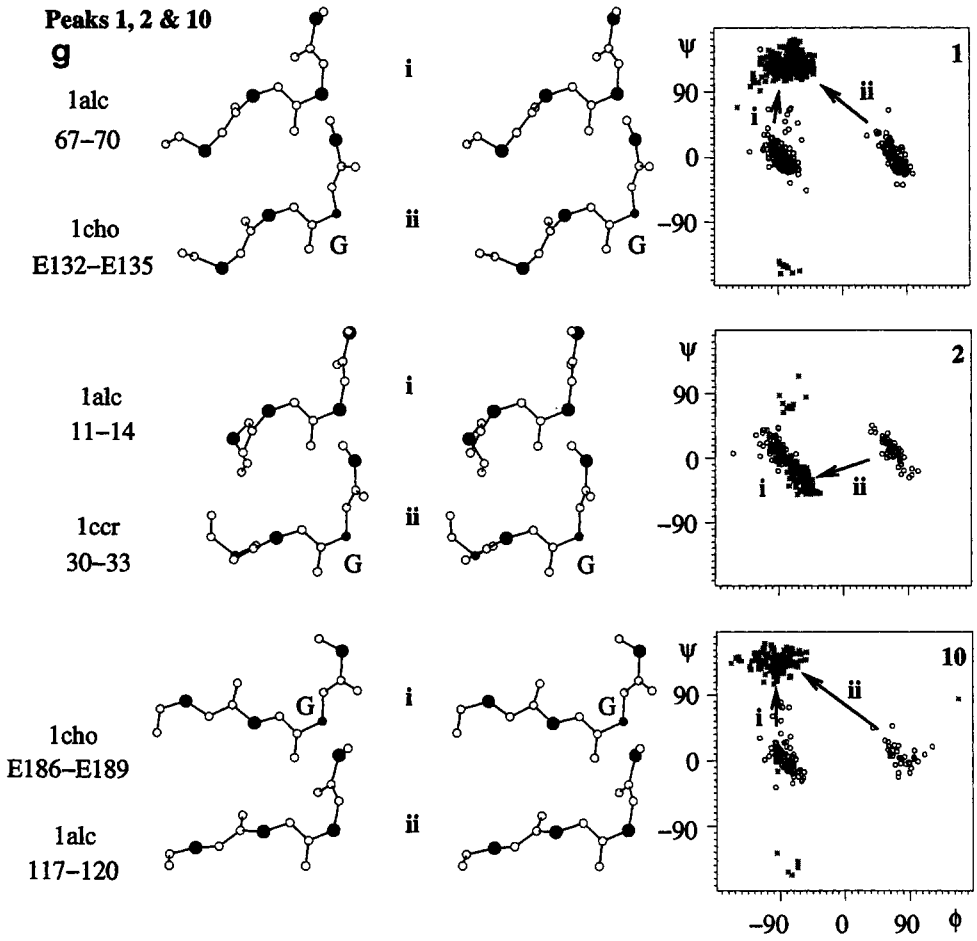


Fig. 5g.

database with a glycine at $i + 2$, and 59 examples of peak 6 without a glycine at that position. Peak 3 corresponds to the Ramachandran angles for the residues $i + 1$ and $i + 2$ of βP → γL, hence the necessity for a glycine in the $i + 2$ position. Peak 6 corresponds to two possible turn types, the first βP → αL and the second ε → αR. The 26 examples of the ε → αR turn conformation require glycine at the $i + 1$ residue position due to steric hindrance attributed to the Cβ atom. It is striking that two turns, which can be thought of as linking protein chains approaching from quite different directions, give rise to the same Cα geometry.

Peak 4 (Fig. 5b)

Peak 4 with 56 examples is equivalent to the type I′ turn consistent with the Ramachandran angles for $i + 1$ and $i + 2$ of αL → γL. This peak is the result of a single turn of a left-handed helix and consequently can occur only if the third residue is a glycine residue. There is a hydrogen bond between the carbonyl O of residue i and the peptide nitrogen of residue $i + 3$. Hence peak 4 is symmetric in conformation with the α-helix previously discussed.

Peak 5 (Fig. 5b)

Peak 5, with 62 examples, appears to have no defined turn type and corresponds to the Ramachandran angle change of αR → γL. Glycine is required at position $i + 2$.

Peak 7, 12, and 15 (Fig. 5c)

Peaks 7, 12, and 15 are variations of the standard β-sheet conformation and are therefore not strictly four residue turn structures as defined by Lewis et al.[5] The conformations result in a separation of the atoms Cα(i) and Cα($i + 3$) of approximately 12 Å. The type of structure normally defined by these conformations are β bulges which result from changes in hydrogen bonding patterns within a β-sheet structure. The two peaks 7 and 12 correspond to a symmetric pair of β bulge structures.

TABLE II. Peaks*

Peak	No. of	Angle θ_1	Angle θ_2	Torsion τ	Symmetry $\alpha = \beta$	Residue association	Ramachandran $\phi\varphi \rightarrow \phi\varphi$†	W&T‡
a	4856	85..105	85..105	35..75	Yes	—	αR	α-helix
b	4356	100..140	100..140	−100..−180	Almost	—	β	β-sheet
1i	184	90..100	110..130	105..135	No	—	αR→β	αβE
1ii	120	90..100	110..130	105..135	No	Gly(i+1)	gL→β	gβ
2i	51	90..100	90..100	100..140	Yes		αR→αR	αα
2ii	63	90..100	90..100	100..140	Yes	Gly(i+1)	γL→αR	γα
3	87	105..120	90..100	−10..20	No	Gly(i+2)	βP→γL	βPγ
4	59	87..98	90..100	−30..−60	Almost	Gly(i+2)	αL→γL	αγ
5	56	90..100	90..100	−95..−115	Yes	Gly(i+2)	αR→γL	αγ
6i	33	105..120	90..100	−10..20	No	≠Gly(i+2)	βP→γL	βPγ
6ii	26	105..120	90..100	−10..20	No	Gly(i+1)	ε→αR	εα
7	176	105..125	130..140	160..180	No	—	β→βb	ββ
8i	28	105..125	85..95	−160..−180	No	—	β→αR	βα
8ii	90	105..125	85..95	−160..−180	No	Pro(i+2)	βP→αR	βα
9	58	130..150	85..100	−125..−150	No	—	β→αR	βα
10i	36	90..100	120..140	80..105	No	Gly(i+1)	γL→β	γβ
10ii	87	90..100	120..140	80..105	No	—	αR→β	αβ
11	319	85..95	98..108	25..50	No	—	αR→αR	αα
12	136	125..135	105..115	−90..−120	No	—	β→β	ββ
13	83	135..145	85..105	−90..−120	No	—	β→αR	bα
14	136	125..135	85..105	−90..−120	No	—	β→αR	bα
15	83	135..145	105..115	−90..−120	No	—	β→β	ββ

*The table includes the number of examples found in the 18,503 values, the range of θ_1 used, the range of θ_2 used, the range of τ used, if $\theta_1 = \theta_2$, residue association and position, and the classification of the Ramachandran angles for the residues (i + 1) and (i + 2). Where a peak is subdivided into (i) and (ii), the conformation of the Cα atoms is identical, but there are 2 different sets of Ramachandran angles for the residues (i + 1) and (i + 2). Hence the chain pathway is the same for the two different backbone conformations.

†The area on the Ramachandran plot where the φ and ψ torsion angles occur for the i + 1 and i + 2 residues.

‡The Wilmot and Thornton[8,24] (W & T) classification of turn conformation for i + 1 and i + 2 residues.

Peak 8 and 9 (Fig. 5d)

Peaks 8 and 9 are highly correlated with the Ramachandran angles i + 1 and i + 2 of β → αR. This would suggest a classical type VI turn which is usually associated with a *cis*-proline at i + 2.[21] However, there are only 49 *cis*-prolines present in the database, and only 28 of them are found in peak 8 and 5 in peak 9. It appears that this conformation is not critically dependent on the presence of a *cis*-proline in the i + 2 position. The difference between peaks 8 and 9 is the value of the Ramachandran angles of the i + 1 residue.

Peak 11 (Fig. 5e)

Peak 11 is the classical type I turn and is highly populated with 132 values. The conformation is not quite that of an α-helix as peak 11 lies a small distance from the true helix peak and $\theta_1 \neq \theta_2$. This peak was distinguished from the α-helix peak by excluding α-helical residues from the search of the database. The Ramachandran angles for this type of turn for the residue i + 1 and i + 2 are αR → αR. There is some difference between the values of the Ramachandran angles i + 1 and i + 2 giving two regions which can just be distinguished from each other. This can be seen in Figure 5e.

Peaks 13 and 14 (Fig. 5f)

Peaks 13 and 14 are different only in the range of θ_1 values used to select examples. The Ramachandran angles for residue i + 1 and residue i + 2 are very similar, but close inspection shows some minor variation in the center of the Ramachandran angles for residue i + 2. This gives rise to two very similar turn types which are quite distinct as seen in Figure 5f.

Peaks 1, 2, and 10 (Fig. 5g)

A total of 304 examples of peak 1, 114 examples of peak 2, and 123 example of peak 10 were observed in the database. For this set of structures, there was a high frequency for glycine at position i + 1 in the four residues. The classification into these three peaks was made by considering the values of θ_1, θ_2, and τ. For peak 1 $\theta_1 \neq \theta_2$ and $105° < \tau < 135°$, for peak 2 $\theta_1 = \theta_2$, and for peak 10 $\theta_1 \neq \theta_2$ and $80° < \tau < 105°$. These three peaks correspond to turn types as defined by the Ramachandran angles of i + 2 and i + 2 as

peak 1:	γL → βP	αR → βP	
peak 2:	γL → αR	αR → αR	
peak 10:	γL → β	αR → β	(VIII)

Initial inspection of the Ramachandran angles shown on Figure 5g suggests that peak 1 and peak 10 describe the same conformation. However, there is a minor, but significant difference in the $\phi(i + 2)$ value of 25° which gives rise to a clearly visible difference in the relative position of residue $i + 3$ as shown in the diagrams. Peak 2 with a Ramachandran angle at $i + 2$ of $\alpha R \rightarrow \alpha R$ is normally found as contributing to the 3_{10}-helix.

DISCUSSION

The results show that the definition of two pseudoangles and a torsion using only the α carbon atoms results in a surprising restriction in the conformation allowed based on these three parameters. The area of the allowed regions is approximately 30% of the total compared to the 22.5% partially allowed region within a Ramachandran plot.[1] The main difference between the overall appearance of the Ramachandran plot and the functions $\tau(\theta_1)$ and $\tau(\theta_2)$ is the distribution within the possible allowed and observed regions. In the case of the plot of $\tau(\theta_1)$ and $\tau(\theta_2)$ the secondary structure (α-helix and β-sheet) is limited to only 0.6 and 7.5% of the total area. The remaining populated regions are subdivided into the many types of turn structure found in proteins. Therefore the presentation of protein conformation in terms of the functions $\tau(\theta_1)$ and $\tau(\theta_2)$ allows a much more comprehensive identification of structure type than is possible from the Ramachandran diagram alone.

A general conclusion of the analysis of the turns is that it is possible to classify most of the major turn types using the three parameters θ_1, θ_2, and τ. Most of the peaks observed using the data presented here provides information on turns that correlates very closely to other methods of determining turn structure. This is not completely surprising as Levitt[9] showed that the torsion angle τ is approximately given by the sum of $\phi(i + 2)$, $\psi(i + 1)$, and 180°.

The number of occurrences of each turn type presented here is high. The regions that define the turns are more localized than the equivalent Ramachandran regions for the residues $i + 1$ and $i + 2$ making the analysis of turns more straightforward. It was necessary to use regions of ±30°, and in some cases up to ±45° for the values of ϕ and ψ to determine turn types from a Ramachandran plot. For only one region (peaks 1, 2, and 10) was it difficult in our analysis to make an assignment between the peaks in the distribution of $\tau(\theta_1)$ and $\tau(\theta_2)$ and particular turn types.

The definition of $\tau(\theta_1)$ and $\tau(\theta_2)$ means that the two functions will be identical only if the four atom sequence $C\alpha(i)$–$C\alpha(i + 1)$–$C\alpha(i + 2)$–$C\alpha(i + 3)$ is equivalent to the sequence $C\alpha(i + 3)$–$C\alpha(a + 2)$–$C\alpha(i + 1)$–$C\alpha(i)$, and hence the polypeptide chain of a protein has no defined direction. The difference plot of $\tau(\theta_1)$–$\tau(\theta_2)$ shows that there is a sense of chain direction relative to the N-terminus and the folding of the protein must be dependent somehow on the order of residues.

The plots also show that conformations are found in proteins that sample most of the intermediate region in $\tau(\theta_1)$ and $\tau(\theta_2)$. So far we have limited our analysis to peaks with more than 50 examples but it is likely that many other types of turns could be identified, particularly as the database of protein structures grows. For example, it is clear that peak 6 corresponds to at least two distinct conformations. One of these occurs 26 times and is for a turn that requires a glycine residue at $i + 2$.

It is interesting to note that there are distinct regions within both the α-helix and β-sheet regions of the Ramachandran map. From the analysis of the turn structures significant populations occur in different parts of the Ramachandran map, and these overlap to form the two main areas known as the β-sheet region and α-helix. Previous work has noted that there are different areas within the β-sheet region that correspond to slightly lower energy and higher populations.[3,4] Our analysis has shown distinct conformations for β-sheet that can be separated on the basis of Cα geometry, but which cannot be identified from the Ramachandran plot. We have identified peaks 7, 12, and 15 corresponding to different β-sheet conformations. It is probable that further analysis on a larger database will reveal additional subtypes.

We have noticed that the preceeding α-carbon atom $(i - 1)$ and subsequent carbon atom $(i + 4)$ appear to have a conformational preference determined by the central 4 residue segment even outside regions of regular secondary structure. This suggests that the description of recurring protein backbone conformations may extend over more than 4 residues to include distinct patterns of conformation for the preceding and subsequent residues around a turn.

This definition of conformation depends on 4 residues rather than the single residue + two single atoms for the Ramachandran plot. Any observed correlation in the plots is the result of the combination of three ϕ, three ψ, and three ω angles and so must be intrinsic to the polypeptide pathway, not just a single residue. In addition, the coordinates for the Cα atoms are often the first to be defined during crystallographic analysis. The effect of errors in Cα positions in our analysis is highly limited in that we see only a slight sharpening of the peaks on studying a higher resolution (< 1.6 Å) set of proteins.

So what is the advantage of this type of plot over a Ramachandran plot? At first inspection, the plot is more complicated as there are more peaks. However, the major peaks for the α-helix and β-sheet are far more restrictive than the equivalent ones in the Ramachandran plot, and the individual turn conformations are found in discrete areas of the plots. The two

plots of the functions $\tau(\theta_1)$ and $\tau(\theta_2)$ are therefore more concise descriptions of protein conformation, and should be of particular use in detailed analysis of residue conformation of a final structure. The two plots and the associated difference plot indicate local structure for 3_{10}-helices, and all the major types of turn. The regions that contain these conformations are discrete and are small in extent allowing a more distinctive classification of protein structure.

There is also the advantage that only Cα atoms are necessary for the calculation. A powerful application of the empirical "rules" for Cα geometry that this analysis suggests is for both building and validating models of protein structure. Three particular examples are currently under investigation. The first is in protein crystallography where the standard technique for building protein structure into an initial electron density map is to produce a skeletonized representation of the map[22] and use this to guide the selection of structural templates from a database. The Cα geometry presented here allows a semiautomatic building of a polypeptide chain into a skeletonized map, growing the peptide chain to coincide with the map and satisfying the Cα geometry rules (manuscript in preparation). For molecular modeling, the probability maps could be used to derive a cost function for assessment of protein structures generated through Monte Carlo simulated annealing techniques. Finally, we have used the probability maps as a guide in building models of proteins from published stereo diagrams (manuscript in preparation).

ACKNOWLEDGMENTS

We thank the SERC and Glaxo Group Research for financial support. In particular, we acknowledge the support of Peter Murray-Rust and Guy Dodson, and thank Martin Karplus for comments on an early manuscript.

REFERENCES

1. Ramachandran, G.N., Sasisekharan, V. Conformation of polypeptides and proteins. Adv. Prot. Chem. 23:283–437, 1968.
2. Ring, C.S., Kneller, D.G., Langridge, R., Cohen, F.E. Taxonomy and conformational analysis of loops in proteins. J. Mol. Biol. 224:685–699, 1992.
3. Salemme, F.R., Weatherford, D.W. Conformational and geometric properties of β-sheets in proteins. I. Parallel β-sheets. J. Mol. Biol. 146:101–117, 1981.
4. Salemme, F.R., Weatherford, D.W. Conformational and geometric properties of β-sheets in proteins. II. Antiparallel and mixed β-sheets. J. Mol. Biol, 146:119–141, 1981.
5. Lewis, P.N., Momany, F.A., Scheraga, H.A. Chain reversals in proteins. Biochim. Biophys. Acta 303:211–229, 1973.
6. Venkatachalam, C.M. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. Biopolymers 6:1425–1436, 1968.
7. Richardson, J.S. The anatomy and taxonomy of protein structure. Adv. Prot. Chem. 34:167–337, 1981.
8. Wilmot, C.M., Thorton, J.M. β-Turns and their distortions: A proposed new nomenclature. Protein Eng. 3:479–493, 1990.
9. Levitt, M. A simplified representation of protein conformations for rapid simulation of protein folding. J. Mol. Biol. 104:59–107, 1976.
10. Gregoret, L.M., Cohen, F.E. Protein folding. Effect of packing density on chain conformation. J. Mol. Biol. 219:109–122, 1991.
11. Rackovsky, S., Scheraga, H.A. Differential geometry and polymer conformation. 1. Comparison of protein conformations. Macromolecules 11(6):1169–1174, 1978.
12. Rackovsky, S., Scheraga, H.A. Differential geometry and polymer conformation. 2. Development of a conformational distance function. Macromolecules 13(6):1440–1453, 1980.
13. Rackovsky, S., Scheraga, H.A. Differential geometry and polymer conformation. 3. Single-site and nearest-neighbor distributions, and nucleation of protein folding. Macromolecules 14:1259–1269, 1981.
14. Rackovsky, S., Scheraga, H.A. Differential geometry and protein folding. Acc. Chem. Res. 17:209–214, 1984.
15. Rackovsky, S., Goldstein, D.A. Differential geometry and protein conformation. V. Medium-range conformational influence of the individual amino acids. Biopolymers 26: 1163–1187, 1987.
16. Rackovsky, S. Quantitative organisation of the known protein X-ray structures. I. Methods and short-length-scale results. Proteins 7:378–402, 1990.
17. Bernstein, F.C., Koetzel, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, K., Tasumi, M.J. The Protein Data Bank: A computer based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
18. Oldfield, T.J. SQUID: A program for the analysis and display of data from crystallography and molecular dynamics. J. Mol. Graphics 10:247–252, 1992.
19. Kabsch, W., Sanders, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12):2577–2593, 1982.
20. Ramachandran, G.N.. Blout, E.R., Bovey, F.A., Goodman, M., Lotan, N. (eds.). In "Pepetides, Polypeptides and Proteins." New York: Wiley Interscience, 1974: 14–34.
21. Aubry, A., Cung, M.T., Marrand, M. βI- and βII turn conformations in model dipeptides with the Pro-Xaa sequences. J. Am. Chem. Soc. 107:7640–7647, 1985.
22. Greer, J. Three-dimensional pattern recognition: An approach to automated interpretation of electron density maps of proteins. J. Mol. Biol. 82:279–301, 1974.
23. Baker, E.N., Hubbard, R.E. Hydrogen bonding in globular proteins. Proc. Biophys. Mol. Biol. 44:97–179, 1984.
24. Wilmot, C.M., Thorton, J.M. Analysis and prediction of the different types of β-turn in proteins. J. Mol. Biol. 203:221–232, 1987.
25. Lewis, P.N., Momany, F.A., Scheraga, H.A. Folding of polypeptide chains in proteins: A proposed mechanism for folding. Proc. Natl. Acad. Sci. U.S.A 68:2293–2297, 1971.