Debnath Pal
Pinak Chakrabarti
Department of Biochemistry,
Bose Institute,
P-1/12 CIT Scheme VIIM,
Calcutta 700 054,
India

Received 13 June 2001;
accepted 12 September 2001

On Residues in the Disallowed Region of the Ramachandran Map

Abstract: *An analysis of the occurrence of nonglycyl residues in conformations disallowed in the Ramachandran plot is presented. Ser, Asn, Thr, and Cys have the highest propensities to exhibit such conformations, and the branched aliphatic residues the lowest. Residues cluster in five regions and there are some trends in the types of residues and their side-chain conformations (χ_1) occupying these. Majority of the residues are found at the edge of helices and strands and in short loops, and are involved in different types of weak, stabilizing interactions. A structural motif has been identified where a residue in disallowed conformation occurs as the first residue of a short 3_{10}-helix. On the basis of the types of neighboring residues, the location in the three-dimensional structure and accessibility, there are similarities with the occurrence of* cis *peptide bonds in protein structures.*
 Biopolymers 63: 195–206, 2002; DOI 10.1002/bip.10051

Keywords: *conformation; residue propensity; loop length; 3_{10}-helix; stabilizing interaction*

INTRODUCTION

The stereochemical quality of a protein model may be judged by the use of ϕ,ψ scatter plots, with incorrect structures generally having a much larger fraction of residues lying in disallowed regions.[1] Excursions into the Ramachandran prohibited regions may induce a strain of up to at least 5 kcal/mol.[2] An amino acid may tolerate small deviations from its ideal conformations in order to optimize stabilizing tertiary interactions in the protein, such as hydrogen bonding or keeping hydrophobic residues buried or interactions with the substrate or ligand at the active site. Regions forbidden for non-Gly residues may be accessible to Gly without any energy penalty, and the strain energy associated with unfavorable ϕ,ψ values has been quantified on the basis of the stability of suitably chosen Gly/Ala mutants in staphylococcal nuclease.[3] Gunasekaran and co-workers[4] identified 66 disallowed residues clustered in distinct regions of the Ramachandran map and in most of the cases the unusual stereochemistry was conserved in related protein structures. As the pool of residues was quite small, we have done a reassessment using a larger repertoire of protein structures available now. Additionally, we have recently shown that occurrence of another "high energy" conformation, viz., the *cis* peptide bonds depend on the neighboring residues and their location in the three-dimensional structure.[5] It would be interesting to investigate the disallowed residues from the same perspective. Some of the interactions shown by such residues and their position

Correspondence to: Pinak Chakrabarti; email: pinak@boseinst.ernet.in
Contract grant sponsor: Council of Scientific and Industrial Research, and the Department of Biotechnology

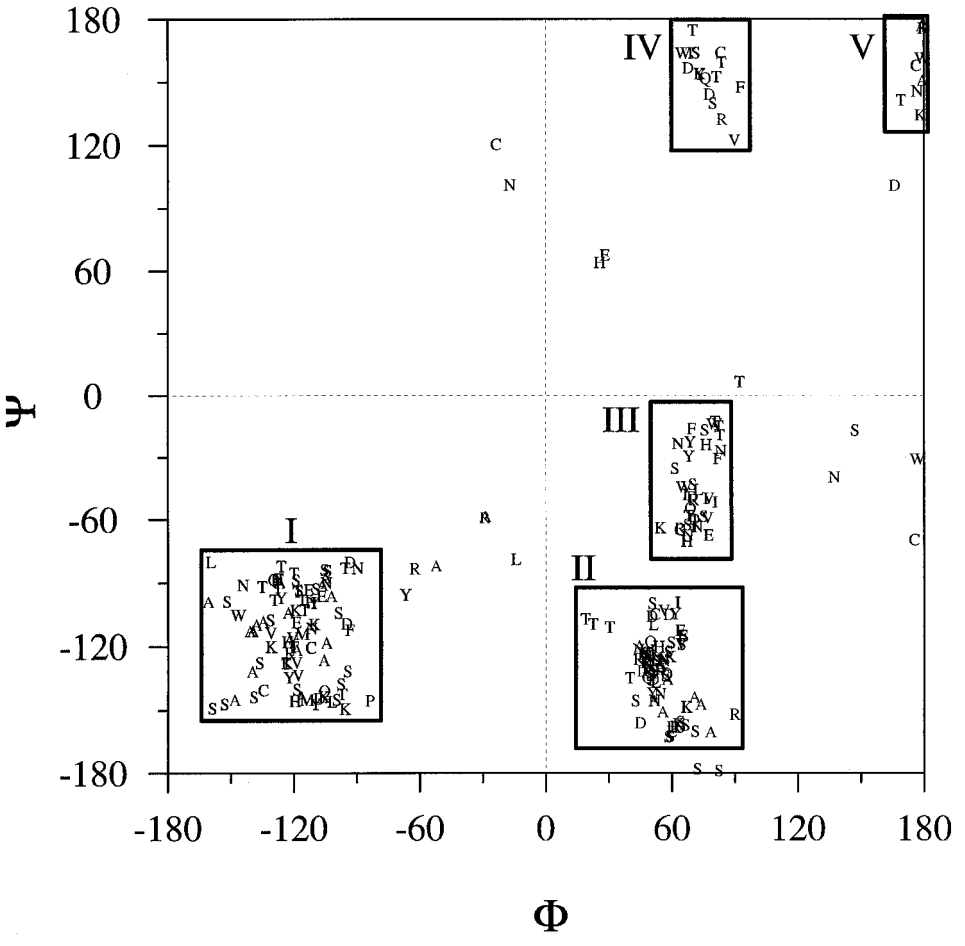

FIGURE 1 Disallowed ϕ,ψ angles (each point is indicated by one-letter amino acid code of the corresponding residue), with delineated clusters identified by Roman numerals.

relative to the nearest secondary structure are also identified so as to facilitate the modeling of normally disallowed conformations in protein structures.

METHODS

The structures were selected from the Protein Data Bank (PDB)[6,7] at the Research Collaboratory for Structural Bioinformatics (RCSB) (http://www.rcsb.org/pdb/), based on the January 2000 release of the representative list found at http://www.sander.embl-heidelberg.de.[8] The list contains structures determined at a resolution 2.0 Å or better, and R factor $\leq$ 20%; the maximum sequence identity between any two of the polypeptide chains is $\leq$25%. To remove proteins having similar folds, the program SARF2[9] was used to structurally compare protein pairs, and those with less than 90% similarity were retained. From the pairs with at least 90% structural similarity, only one representative structure was chosen and the final database compiled. The final list of the PDB codes for 353 structures (with 363 polypeptide chains) along with the chain identifier (following an underscore) used for the analysis is given below. The protein structures containing the conformationally disallowed residues are marked in bold and the residue numbers are given after a hyphen.

1A1I_A; 1A1Y_I; 1A2P_A; **1A2Z_D-73,141**; 1A34_A; **1A3A_B-53,108**; **1A3C-40,109**; **1A48-233**; 1A4I_A; 1A6M; 1A7S; **1A8D-179**; **1A8E-12,294**; 1A9X_F; 1ABA; 1ADS; 1AE9_A; **1AFW_A-124,377,406**; 1AGQ_B; 1AHO; 1AI9_A; **1AJS_A-296**; **1AK1-121,291**; 1ALV_A; 1AMF; 1AMM; **1AOC_B-19**; 1AOH_A; **1APY_A-11**; 1AQ6_B; **1AQB-111**; 1ARU; 1ATL_B; 1AUN; 1AVW_B; 1AWD; 1AXN; 1AY7_B; 1AYF_B; 1AYO_A; 1B0Y; **1B2P_A-20,47**; **1B2V_A-77**; 1B3A_B; **1B4K_A-129**; 1B5E_C; 1B65_E; **1B6G-124,148**; **1B8O_A-221**; 1B93_B; **1BA8_A-7**; 1BBP_A; 1BDO; 1BE9_A; **1BEA-79**; 1BEC; 1BEN_B; **1BF6_B-84**; **1BFD-71**; **1BFG-58**; 1BFT_A; **1BG6-111**; 1BGF; **1BI5_A-338**; 1BJ7; **1BK0-38**; 1BK7_A; **1BKR_A-28**; 1BM8; **1BQC_A-128,198**; **1BRT-9,34,98,236**; **1BS0_A-79,236,240**; 1BS4_A; **1BS9-90,149**; 1BTN; **1BU7_A-436**; **1BW9_B-639**; 1BX7; 1BXA; 1BXO; 1BY2; 1BYI; **1BYQ_A-166**; 1C3D; 1C3W_A; **1C52-16**; 1C53; **1CB8_A-232,233**; 1CBN; 1CCZ_A; **1CEO-317**; **1CEQ_A-57**; 1CEW_I; 1CEX;

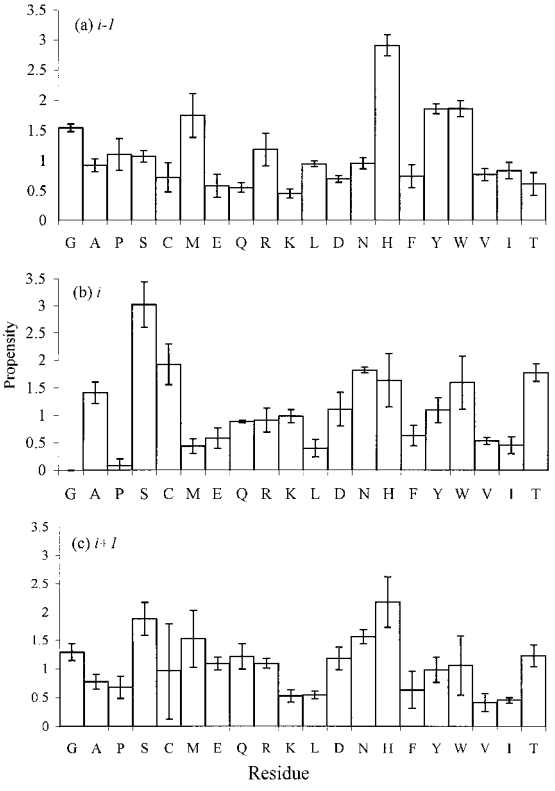

FIGURE 2 Propensities of residues to occur (b) with disallowed Ramachandran angles (Gly excluded) (at position *i*), and (a) and (c), two flanking positions. Standard deviations were obtained as given in Methods.

1CF9_C-274,446,739; 1CFB; 1CJW_A; 1CKA_A; **1CLE_A-18,159,209,243,301,444**; 1CMB_A; **1CNV-93**; **1CPO-102**; 1CPQ; 1CQY_A; **1CS1_A-178,326**; 1CTJ; **1CVL-17,20,87**; **1CXQ_A-178**; 1CY5_A; **1CYD_B-135**; 1DCI_A; **1DCS-155**; 1DHN; **1DIN-123**; **1DLF_H-97,_L-51**; 1DOK_B; 1DPT_B; 1DUN; 1DXG_A; **1ECP_C-207**; **1EDG-30,38,122,169**; 1EDM_B; 1EGP_A; **1EUS-151,369**; 1EXT_A; **1EZM-126,150**; 1FIP_A; 1FLE_I; **1FLT_V-26,_Y**; 1FNA; 1FRP_A; 1FUS; 1FVK_A; 1G3P; **1GAI-411**; 1GCI; 1GCM_C; **1GDO_D-91**; 1GKY; **1GOF-187,188,432,494**; 1GP1_B; **1GPE_A-421**; **1GSA-155**; **1GUQ_A-314**; 1HFC; **1HFE_T-77**; 1HKA; 1HOE; 1HTA; 1HTR_P; 1HUU_B; 1HXN; **1IAB-72**; 1IDA_B; 1IFC; 1IIB_B; 1ISU_A; 1IXH; **1JER-9**; 1JHG_A; **1KNB-404**; 1KOE; 1KPTB; 1KVE_AD; 1LAT_A; **1LCL-73,128**; 1LIS; **1LKF_A-219**; **1LKK_A-221**; 1LOU; 1LTS_AC; 1LUC_A; 1MFM_A; 1MKA_A; **1MLA-92**; **1MML-223**; 1MOF; 1MOL_B; 1MOQ; **1MPG_A-26**; 1MRJ; 1MSI; **1MSK-1193**; 1MUG_A; 1MUN; **1NAR-216**; 1NBC_A; **1NCI_A-1,43**; 1NIF; **1NKD-30**; 1NKR; **1NLR-67,207**; 1NLS; 1NOX; 1NP4; **1NPK-120**; 1NUL_B; **1OAA-42**; 1OPD; **1OPY-120**; 1ORC; 1OTF_D; **1PBE-44**; 1PCF_E; **1PGS-11,86**; **1PHF-143**; 1PHN_A; 1PLC; **1PNE-26,37**; 1POA; 1POC; 1PPN; 1PSR_A; 1PTQ; **1PTY-215**; **1PYM_B-87**; **1QAZ_A-19,240**; **1QB7_A-150**; **1QCX_A-176,187,260**; 1QDD_A; **1QFM_A-311,346,520,554,590**; 1QFO_A; **1QGW_B,_D-75**; **1QH5_B-9,109**; **1QHF_A-180**; **1QKS_A-76,101,109,301,340,345,437,554**; **1QQ4_A-14**; **1QQP_1,_2-36,190,_4**; **1QRE_A-62**; **1QTS_A-732**; 1QTW_A; 1RB9; 1RCF; 1REC; 1REG_X; 1RGE_B; **1RHS-40,274**; 1RIE; 1RZL; 1SCJ_B; 1SFP; 1SGP_I; **1SKF-213,221**; 1SLU_A; **1SMD-102,318,414**; **1SML_A-53**; 1SRA; 1SUR; 1SVF_BC; 1SVP_A; **1SVY-169**; 1TAX_A; **1TCA-51,105**; 1TGX_A; **1TIB-146,199,262**; **1TIF-8**; **1TML-162**; **1TOA_B-39,52**; **1TTB_B-126**; 1TVX_A; **1U9A_A-101**; **1UBP_A,_B-99,_C-23,54,275,411,564**; **1UDC-178**; 1UNK_A; **1VCA_A-76**; 1VFR_A; 1VFY_A; **1VHH-59**; **1VID-68,133,196**; 1VIE; **1VNS-120,124,248,290**; 1WAB; 1WAP_O; 1WDC_A; 1WHI; 1WHO; 1WWC_A; **1XNB-121**; **1YAC_B-174**; **1YCC-27**; **1YGE-294,312,560,687**; 1YTB_B; 256B_B; 2A0B; 2ABK; 2ACY; **2AHJ_A-164**; **2ARC_A-115**; **2AYH-2,84**; **2BC2_A-56,57,204**; 2CCY_B; **2CHS_L-110**; **2CTC-199,273**; 2DTR; 2EBO_B; 2END; 2ERL; 2FDN; 2GAR; 2HDD_B; 2HFT; 2HMZ_C; 2IGD; 2ILK; **2IZH_D-52**; **2KNT-14**; **2MYR-72,142,466**; 2PII; 2PSP_A; **2PTH-66**; 2PVB; **2QWC-404**; 2RN2; 2SAK; 2SIC_I; 2SN3; 2SNS; 2SPC_B; **2TPS_A-93**; 2TRX_A; 2TYS_B; 3B5C; 3CHB_G; **3CHY-62**; 3CLA; **3CYR-51**; 3ENG; 3EZM_A; **3GRS-52,219**; 3LZT; **3PTE-235**; **3PVI_A-82**; 3PYP; 3SDH_B; **3SEB-192**; **3SIL-178,228,230,238,276,278,351**; **3TDT-254**; **3TSS-86**; 3VUB; **4EUG_A-77**; 4MT2; **4PGA_A-177,208,306**; 4TSV_A; **5HPG_A-48**; 5P21; 5PTI; **6CEL-99,210**; 6GSV_A; **7A3H_A-103**; **7RSA-60**; **8ABP-89,232,254**; 8PRK_A; 9WGA_A.

Torsion angles were calculated by means of the DIHDRL program, available from PDB. For the analysis of χ_1 angles, the whole angular range of 360° was divided into three bins centered around the three canonical staggered conformations (180° and ± 60°), and we use the notation: (t) =120°–240°, (g^+) = −120°–0° and (g^-) = 0°–120°. According to IUPAC-IUB Commission recommendations[10] the relative orientation of the two branches on the C^β atom in Val is different from that in Thr and Ile. As a result, at any χ_1 angle the position of the two nonhydrogen atoms at the γ position in Val is different from the other two. To correct for this anomaly the "standard" (t, g^-, g^+) states for Val are listed here as (g^+, t, g^-).

Only those torsion angles which did not involve any atom with a temperature factor, $B > 30$ Å^2 were used; this way only the well-ordered residues were retained. To remove some unrefined structures (with B values for all atoms set uniformly at 1.0), a filter to exclude atoms with $B \leq 1.0$ Å^2 was also used. As demarcated by Gunasekaran et al.,[4] the residues with the allowed conformation were enclosed in the following three regions: (ϕ = −180° to −30°, ψ = −80° to 180°), (ϕ = 30° to 90°, ψ = −10° to 120°), and (ϕ = −180° to −30°, ψ = −180° to −150°). The disallowed region comprised of the rest of the ϕ,ψ space.

The propensity of a non-Gly residue (X) to be in disallowed conformation (j) was calculated using the formula[11]

$$P(X, j) = \frac{f_{x,j}}{\langle f_j \rangle}$$

where

$$f_{x,j} = \frac{n_{x,j}}{n_{x,\text{all}}} = \frac{\text{number of residue } X \text{ in conformation } j}{\text{number of residue } X \text{ in all proteins}}$$

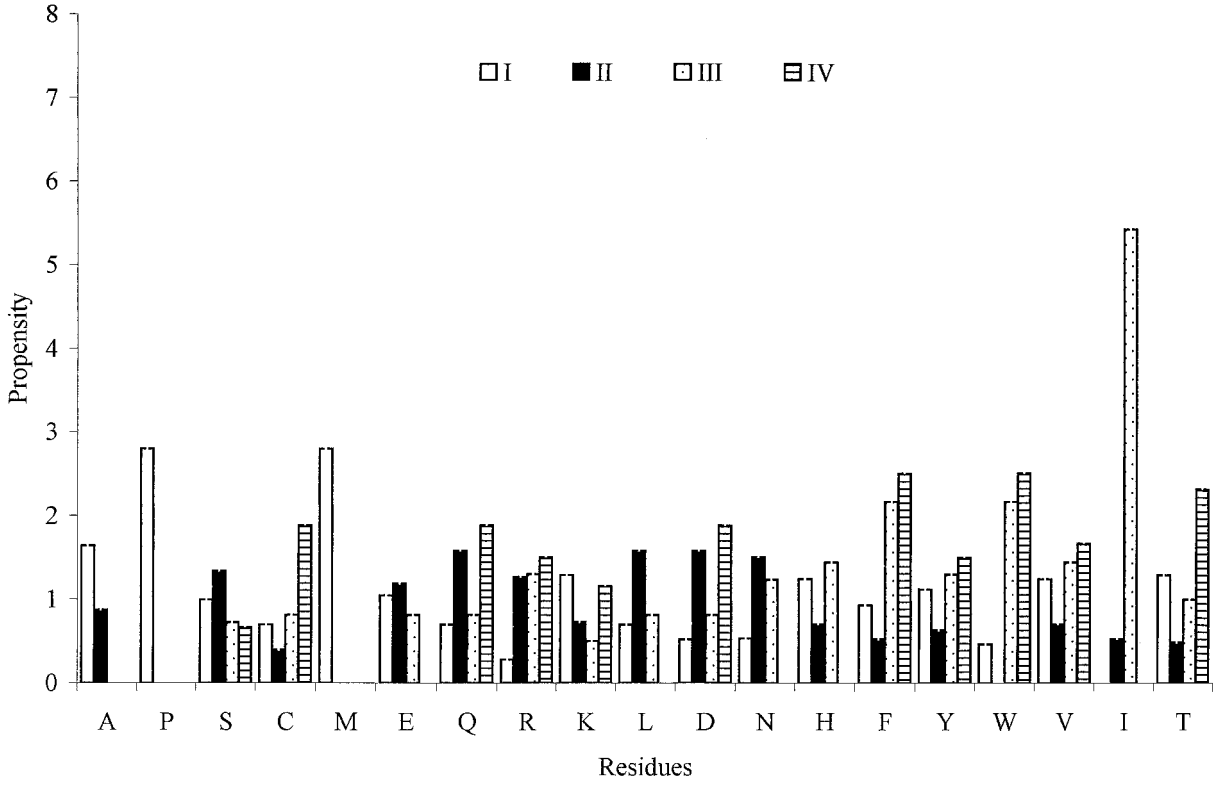

FIGURE 3 Plot showing the distribution of local propensities of residues to be in four clusters as delineated in Figure 1.

$$\langle f_j \rangle = \frac{n_j}{n_{\text{all}}} = \frac{\text{total number of residues in structure } j}{\text{total number of residues in all proteins}}$$

A similar formula was used to derive propensities of residues (Gly included) to be adjacent to a disallowed residue.

The local propensity of a residue (X) to be in one of the (four main) clusters, s, of disallowed residues was calculated in an analogous manner, except that the normalization was based on residues in the whole disallowed region (j) and not in the whole database, i.e.,

$$P_1(X, s) = \frac{n_{X,s}}{n_{X,j}} \Big/ \frac{n_{\text{all},s}}{n_{\text{all},j}}$$

The secondary structural elements were assigned in accordance with the algorithm (DSSP) of Kabsch and Sander,[12] which uses the following notations: B, residue in isolated β-bridge; E, extended strand; G, 3_{10}-helix; I, π-helix; H, α-helix; S, bend; T, hydrogen-bonded turn. Residues without any regular secondary structure were marked C. The solvent-accessible surface area (ASA) was calculated using the program ACCESS,[13] which is an implementation of the Lee and Richards[14] algorithm. We used the default van der Waals radii in the program and the solvent probe size was 1.4Å. The solvent accessibility of a residue was evaluated by the ratio of the summed atomic accessible surface areas of that residue in the protein to that of the same residue (X) in an extended Ala–X–Ala tripeptide. Only one subunit was considered while performing these calculations.

Because results reported here differed from the earlier work,[4] and because the number of data points available for analysis were not very high, it was decided to repeat the calculations using the 1998 and 1999 release of the PDB structures[8] (with 282 and 375 files, respectively, obtained after following the same filtering procedure). Standard deviations were obtained on the basis of these three sets of values. Comparison across three databases indicated that the results were quite stable.

RESULTS AND DISCUSSION

Sterically Disallowed Residues and Their Clusters

A total of 241 residues (0.4%), out of a total of 63949 non-Gly residues in 363 polypeptide chains, were identified to occur in the whole disallowed region. The same percentage was also obtained earlier (when a filter based on B factor was applied).[4] But for a few dispersed points the residues cluster in five regions

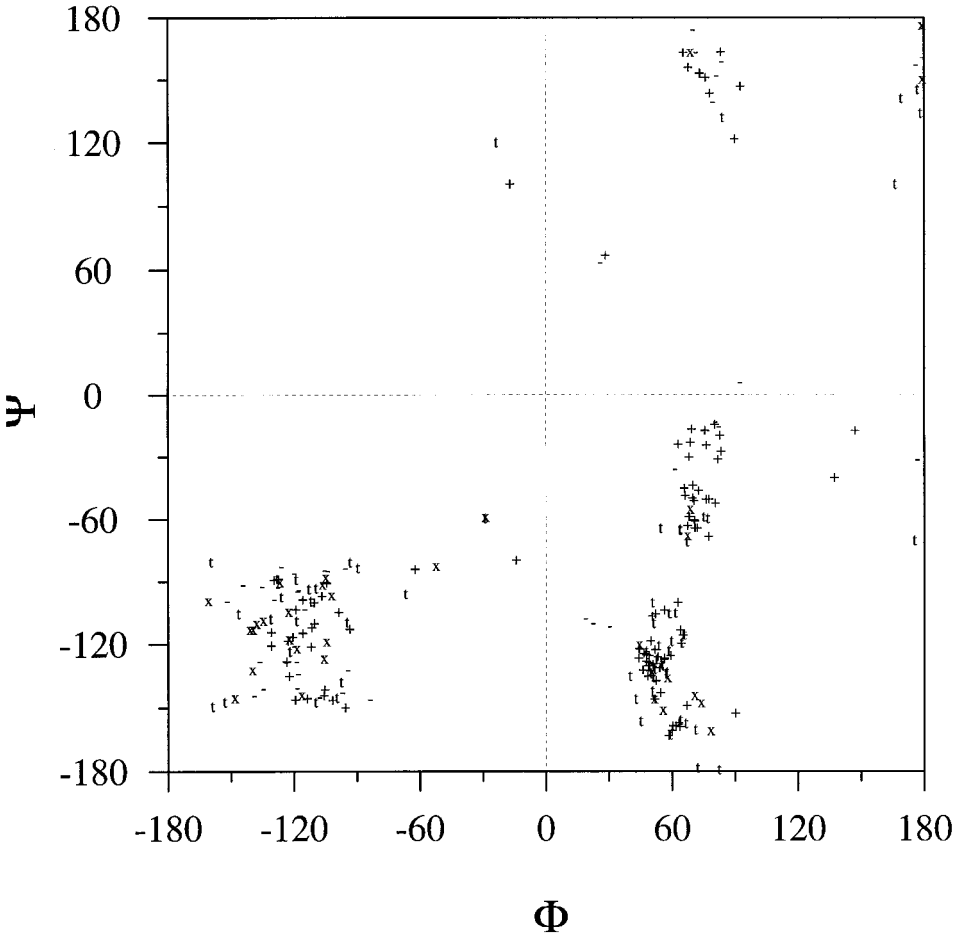

FIGURE 4 ϕ,ψ map of residues (as in Figure 1) with their χ_1 conformational states being indicated by labels + and − (in short for g^+ and $g-$, respectively), *t* and *x* (for Gly and Ala, where χ_1 cannot be defined). The numbers of g^+, *t*, $g-$, and *x* in various clusters are I (25, 19, 25, 17), II (38, 27, 3, 8), III (26, 6, 3, 2), and IV (9, 1, 5, 1).

(Figure 1). With the availability of a larger number of residues the clusters are well populated and to some extent different from the earlier groupings by Gunasekaran et al.[4] Clusters II and IV can also be considered as a continuous streak of points differing in ψ. V may be deleted from the list of disallowed clusters if the first allowed region (see Methods) is expanded along ϕ – these are essentially residues with extended conformations.

Interestingly, consecutive residues can also be present in disallowed conformation. These are (PDB code and residues):1CB8, S232-S233; 1GOF, S187-S188; 2BC2_A, D56-S57. There are a few cases with alternate residues in disallowed conformation: 3SIL, F228, S230, R276, H278.

Amino Acid Propensities to be in and Around the Disallowed Region

The propensities of residues to be in the disallowed region (assumed at position *i*) and the two immediate neighbors (positions i ± 1) are shown in Figure 2. A value of >1 indicates a significant tendency to adopt a disallowed conformation (or be the neighbor of such a residue), whereas values <1 suggest that backbone distortions are unlikely for these residues (or their neighbors). While Gunasekaran et al.[4] found the residue with the highest propensity for disallowed region to be Asn, followed by Asp and His, the highest value has now been assigned to Ser, trailed by Cys, Asn, Thr, His, and Trp. Pro and branched aliphatic residues (Val, Ile, and Leu) disfavor such distortions. Considering the flanking residues, His, Tyr, Trp, and Met have distinctly high tendency to precede a disallowed residue. Ser, His, and Asn not only have high propensities to be in a normally disallowed conformation, but also to follow disallowed residues. Val, Ile, Leu, Phe, and Lys oppose distortions when present as flanking residues.

Interestingly, Val and Ile oppose both the occurrence of a *cis* peptide unit[5] and a disallowed main-

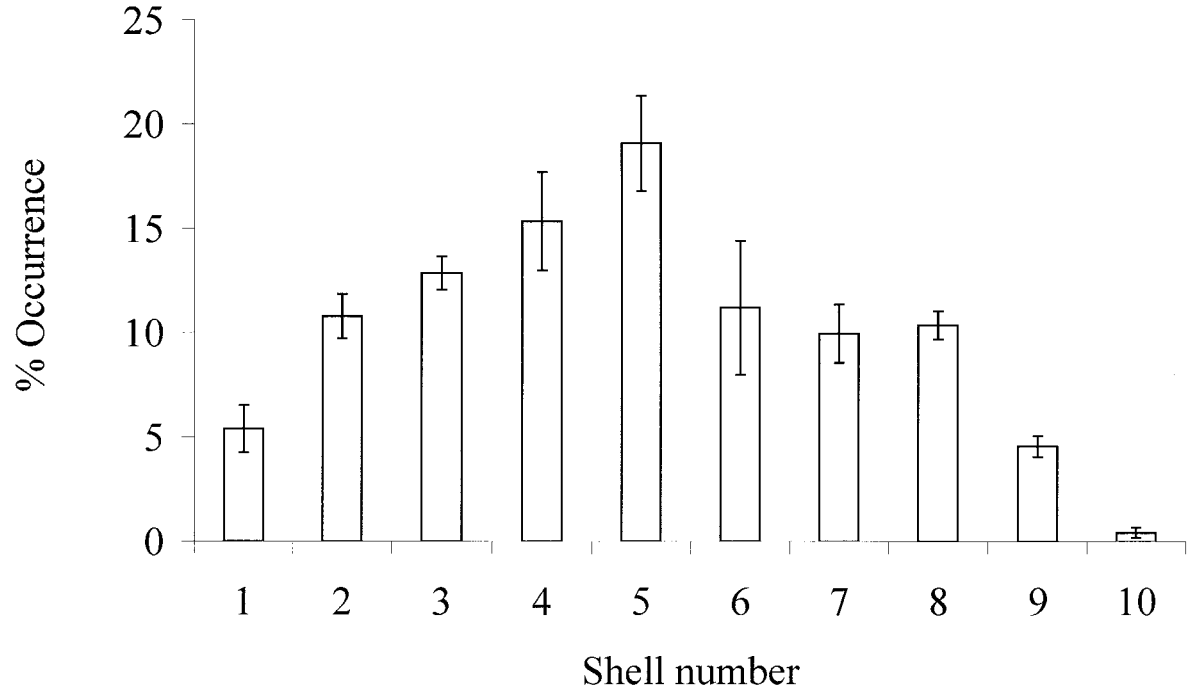

FIGURE 5 Radial distribution of disallowed residues in protein structures.

chain angle. Other residue preferences observed here that are also reflected in *cis* peptides are the use of short polar residues (Ser, Asp, and Asn) in and around X_{np}–X_{np} *cis* bonds (X_{np} = any nonproline residue) and the relative large presence of Trp and Tyr preceding the X–Pro *cis* bond. Gly has a relatively high propen-

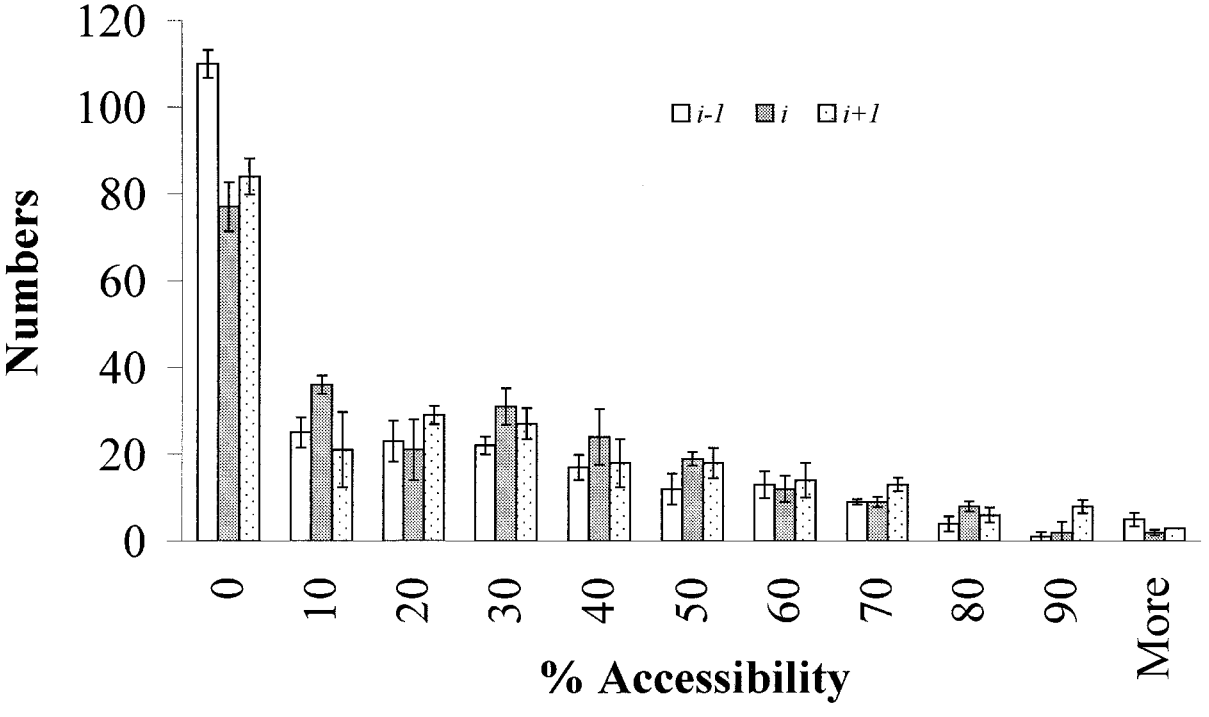

FIGURE 6 Distribution of the accessibilities of residues in the disallowed conformation and their neighbors (at positions i and $i \pm 1$).

Table I Secondary Structural Features of the Disallowed Residue and the Neighbors[a]

Secondary Structure	Position		
	$i-1$	i	$i+1$
H	5	2	10
G	1	6	8
E	27	5	9
B	2	1	1
S	17	29	25
T	20	32	26
C	28	25	21

[a] The number at each position corresponds to the percentage of occurrence of different secondary structural elements. For definitions of secondary structures, see Methods.

sity to be one of the neighbors, which means that, in addition to its well-known characteristic that it can take up a large number of conformations not accessible to others, Gly can also be adjacent to a residue with disallowed conformation.

Distribution of Amino Acid Residues in the Different Disallowed Clusters

The distribution of disallowed residues among the four major clusters (Figure 1) can be expressed in terms of local propensity and is given in Figure 3. Ser, which is present in the largest number, does not seem to prefer any particular cluster. As a class,[15,16] the aromatic residues have a significant presence in clusters III and IV, and Asp and Asn are preferred in cluster II. Of all the clusters, I is in the negative ϕ region and contains many residues with average propensity values (around 1). Arg, Asp, and Asn have low values, and Ala and Met have high. Another feature for Ala is that it is not observed in III and IV. Cluster II is favored by polar residues and Leu. Ile is not a particularly suitable candidate to be observed in disallowed conformation, but whenever there is one it is likely to be in cluster III. Similarly, proline, does not prefer disallowed conformation and is present only once in cluster I, and hence the resultant high local propensity.

χ_1 Distribution

Because of the interrelationship of the side-chain and main-chain conformations,[15,16] it is of interest to study the distribution of χ_1 angles for disallowed residues (Figure 4). The usual order of rotameric population is $g^+ > t > g^-$, and it is maintained in clusters II and III, but is reversed for the t and g^- states in cluster IV. In the negative ϕ range, the most prominent χ_1 rotamer is g^+, however, the other two states are now almost equally favored. Thus residues cannot only be in high-energy backbone conformation, some of them can also have the high-energy side-chain conformation (g^-) as well.

Accessibility and the Location of Disallowed Residues in the Protein Molecule

To have an idea about the location, we calculated the radial distribution of the C^α atoms of the conformationally disallowed residues from the centre of mass of the polypeptide chain. The result is depicted in Figure 5, where the position of the C^α atom in concentric shells is shown, assuming the protein to be a spherical moiety and dividing the distance from the centre of mass to the outer most atom in the structure into ten equal parts. The peak of the distribution occurs at shell number 5, which suggests interior locations. The peak in the radial distribution was found at shell numbers 7 and 3 for *X*–Pro and X_{np}–X_{np} (X = any residue, X_{np} = non-Pro residue) *cis* peptide bonds, respectively.[5] This along with the secondary structural features (next section) suggest that the residues in disallowed conformation have characteristics in between residues in X–Pro and X_{np}–X_{np} *cis* bonds.

About 30% of the residues are completely buried (Figure 6); but a majority have relative accessibilities of 10% or more, indicating solvent exposed locations, as was found for *cis* peptides.[5] However, unlike the latter, where the neighbors are more buried than the residues making up the *cis* peptide bond (it is as if the *cis* peptide moiety is jutted out of the structure), the disallowed residue and both its neighbors are almost equally exposed.

Stabilizing Interactions, Secondary Structural Features and the Length of the Loops Containing the Disallowed Residue

From the structural point of view, one would expect that the disallowed conformation is adequately compensated by some stabilizing interaction in the protein matrix, or such a conformation is forced by the spatial requirements of the folded structure. The secondary structural features around the disallowed conformation provide some clue as to their occurrence. Data presented in Table I and Figure 7 suggest that a majority of the residues in clusters II and III are stabilized by hydrogen-bonded turns; the same is partly true for cluster I, but clusters IV and V are almost devoid of any such structural elements. So for

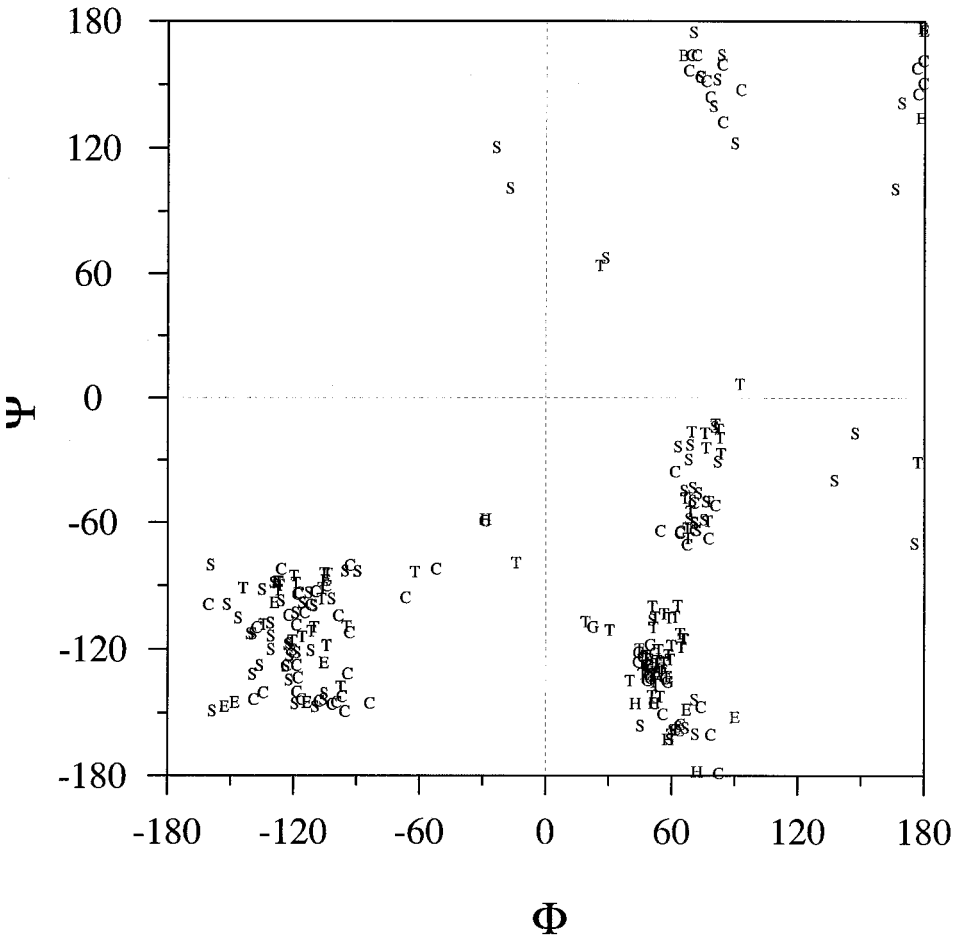

FIGURE 7 ϕ,ψ map of residues, with labels denoting their secondary structural elements (as explained in Methods). The statistics of occurrence of various structures in different clusters is as follows:

Cluster	H	G	E	B	S	T	C
I	0	0	7	0	33	18	28
II	3	14	3	0	8	40	8
III	0	0	0	0	15	13	9
IV	0	0	0	1	6	0	9

a great many residues the penalties for a strained conformation is offset by a hydrogen-bond interaction. Besides, Asp and Asn are found in appreciable numbers in the positive ϕ range,[15,16] and these conformations can be stabilized by the interaction involving the dipoles of the side-chain and main-chain carbonyl groups of these residues;[17] it is thus not unnatural for some of these residues to foray into the disallowed conformation in cluster II, for which they have high local propensities (Figure 3). Figure 8c shows a disallowed Asn residue located close along the central axis of a barrel formed by parallel β-strands.

Residues in cluster I can be fairly extended and some residues in this region can still be part of a β-strand (Figure 7). Figure 8a provides such an example and it is interesting that the residue concerned (Met) participates in an S $\cdots$ O interaction involving its side chain, which has been discussed recently.[18] Other stereospecific interactions can also be seen engaging the disallowed residues. Figure 8(d) shows a tyrosine ring whose side chain is so oriented that not only the hydroxyl group can participate in hydrogen bonds, the aromatic face is also properly positioned for a C—H $\cdots$ π interaction[19] involving the C^{β} protons of another residue.

FIGURE 8 Examples of residues with disallowed conformation in different clusters (diagrams made using MOLSCRIPT[27]). PDB code, cluster number, residue, secondary structures (for residues $i-1$, i, $i+1$), loop length (when in a loop), and details of interactions (pairs of atoms with atom label and residue number) are as follows: (a) 1SMD, I, Met102, EEE, and (N-102:O-166, SD-102: O-101). (b) 1A8E, II, Ser12, EHH, and (O-12:N-16, N-12:O42, OG-12:N-13). (c) 2MYR, III, Asn466, TTT, 13, and (N-466:OH-463, N-466:O62, ND2-466:OD1-201, ND2-466:O-200). (d) 1QFM_A, IV, Tyr311, TCE, 7, and (OH-311:ND2-477, OH-311:ND2-305, π-311:CB-305, N-311: O-308).

Table II Sequence and Secondary Structures Around a Residue in Cluster II with 3_{10}-Helical Conformation

PDB Code	Residue Number	Sequence[a]	Secondary Structure
1AFW_A	124	ALNR**Q**CSSG	EEEC**G**GGHH
1BI5_A	338	YGNM**S**SACV	HCBC**G**GGHH
1BRT	98	LVGF**S**TGTG	EEEE**G**GGHH
1JER	9	HIVG**D**NTGW	EETT**G**GGCS
1NKD	30	LNEL**A**DAAD	HHHT**G**GGHH
1QAZ_A	240	EMTR**H**EQSL	GGGS**G**GGHH
1QFM_A	520	KGGI**L**ANKQ	HTTS**G**GGTH
1TIF	8	DFII**N**EQIR	CCCB**G**GGCC
1TOA_B	52	KNIA**Q**GDVH	HHHH**G**GGSE
1VNS	248	PPGL**R**SNAD	CTTS**G**GGTT
1YGE	560	TFLP**S**KYSV	HSTT**G**GGHH
2CHS_L	110	HVYL**E**KAVV	CEEC**G**GGGG
3PTE	235	VSWA**Q**SAGA	CTTT**G**GGTC
4PGA_A	208	AKRH**T**VNSE	CSCC**G**GGCS

[a] The residue in cluster II is marked in bold.

The observation that 14 residues in cluster II are in 3_{10} conformation (Figure 7) made us look into these cases in greater details (Table II). It was noted that these residues constitute the first position of short 3_{10}-helices (essentially 3-residue turns), which in majority of the cases lead into α-helices, and segments of the polypeptide chains from different structures can be nicely overlaid (Figure 9). While most of the residues in the 3_{10}-stretches, including the one in the disallowed conformation, are hydrophilic, the residue preceding the 3_{10}-helix (at the Ncap position) is mostly hydrophobic. This is in contrast to what is seen in 3_{10}-helices in general, where the Ncap residue is predominantly Asp, Pro, Gly, His, or Asn.[20,21]

Most of the disallowed residues have a regular secondary structure (E or H) or a hydrogen-bonded turn on at least one flank (something that was also observed for X_{np}—X_{np} *cis* bonds,[5] suggesting a role of the secondary structure in their occurrence. When the length of the loop, connecting the disallowed residue to the nearest secondary structures (only the two main types, H and E, are considered) on either side, is analyzed (Figure 10), one finds that for a considerable number of cases the loop is very short and the residue may also be at the edge of a secondary structure (Figure 8b). It should be noted that residues in clusters II–IV have a positive ϕ value around $+60°$, which can cause a sharp turn in the chain direction; consequently when the loop linking two secondary structures is too short, one of the residues can be squeezed into an unfavorable conformation and thereby achieving a turn. Thus in some cases, the disallowed conformation may be a consequence of the packing re-

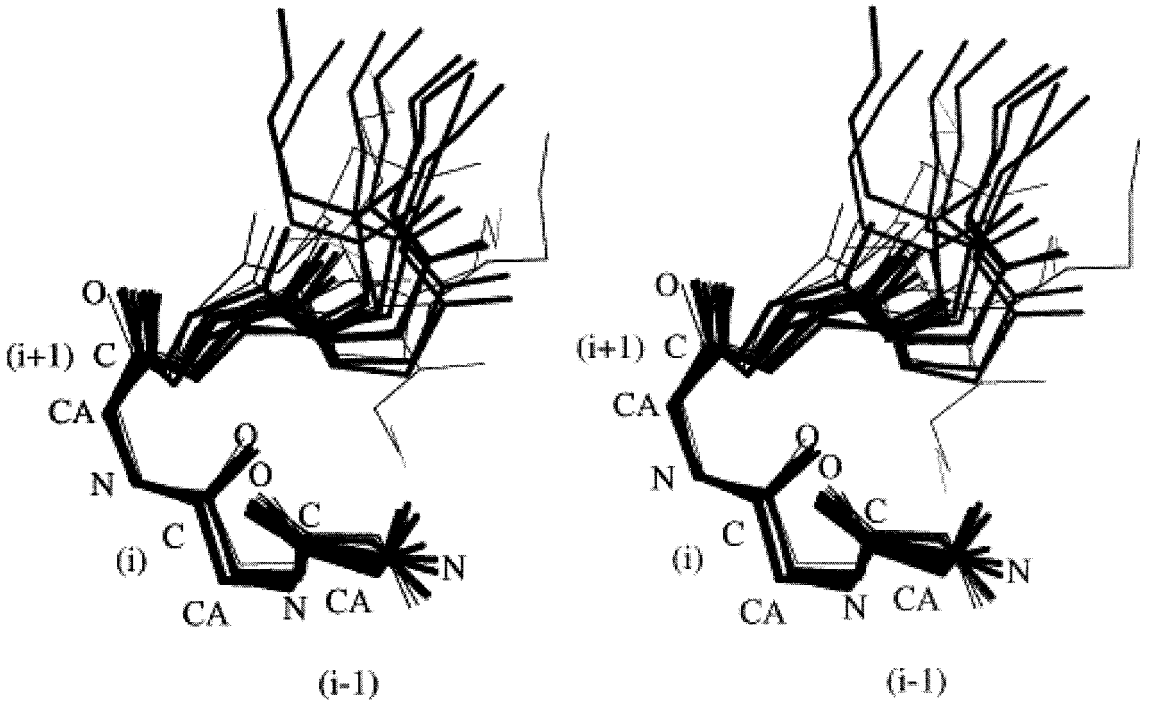

FIGURE 9 Best fit least-squares superimposition using the backbone atoms of the six residues (one before the disallowed residue and four after) of the polypeptide chains (given in Table II), shown in stereo. The chain traces, which generally follow the same direction, are given in bold.

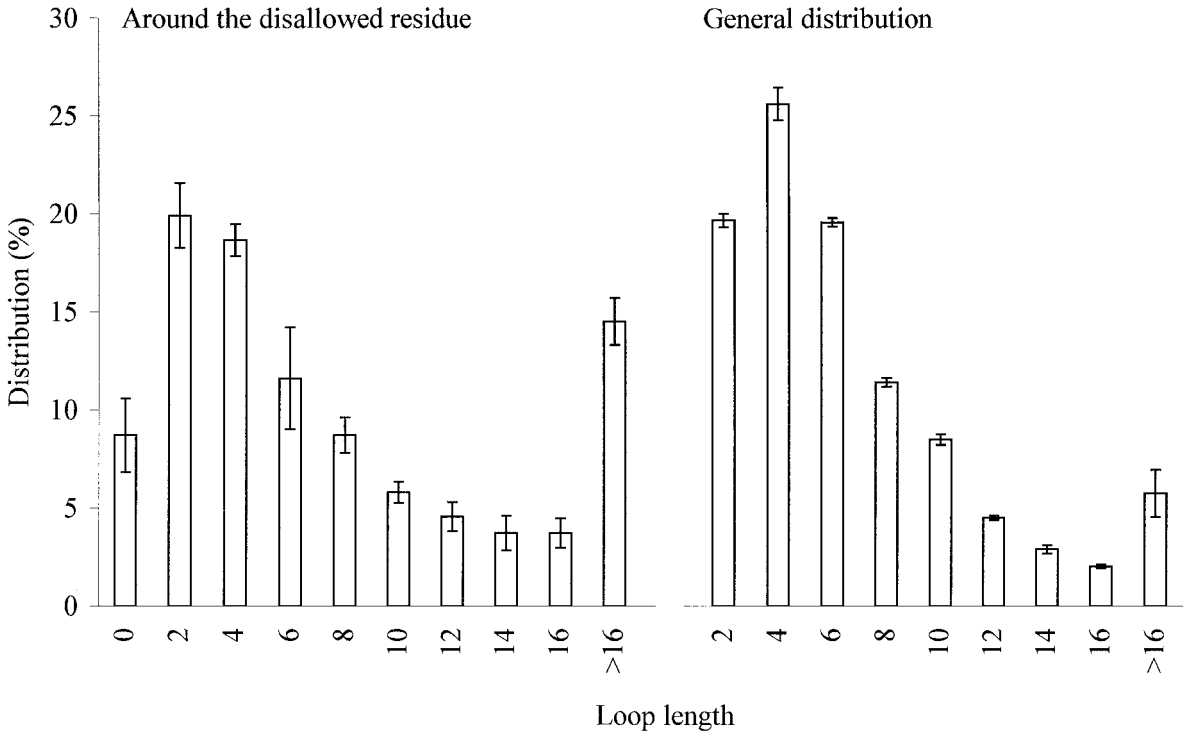

FIGURE 10 The distribution of the length (in terms of the number of residues contained on the flanking sides) of the loop containing the disallowed residue and connecting the two nearest secondary structural elements (E or H). The bins used are 0 (no intervening residue on either flanking side connecting the nearest secondary structure), 2 (loop length of 1–2), etc. The general distribution of loop lengths is shown on the right.

quirement of two secondary structures linked by a short loop. In this connection it can be pointed out that one of the factors leading to the enhancement of protein thermostability, the molecular basis of which is still not clearly understood,[22–24] has been the truncation of loop length, which lowers the entropy of unfolding.[25] From a comparative genomic analysis, Thomson and Eisenberg[26] found a trend toward shortened thermophilic proteins relative to their mesophilic homologs. However, our reasoning indicates that making the loop too short may introduce a high-energy disallowed conformation and there has to be a compromise between the two factors.

CONCLUSION

0.4% of all non-Gly residues deviate from the allowed region of the Ramachandran plot. Propensities of residues to be in the disallowed region decrease in the order Ser, Cys, Asn, Thr, and His. The disallowed residues form clusters in the ϕ,ψ space. As in *cis* peptide bonds, aliphatic branched residues have the minimum inclination for being in the disallowed conformation and aromatic residues have a high preference to precede a disallowed residue. Identification of such residue characteristics in and around disallowed conformation would facilitate in the modeling of protein structures. Disallowed residues are usually located close to the surface of the molecule and in short loops, being in many cases at the junction of two secondary structural elements. Though one would not normally expect a disallowed residue to be part of any secondary structure, a structural motif has been identified where a residue with disallowed conformation constitutes the first position of a short stretch of a 3_{10}-helix.

We thank the Council of Scientific and Industrial Research (for a fellowship to DP and a grant to PC) and the Department of Biotechnology (for the National Facility for Biomolecular Modeling).

REFERENCES

1. Kleywegt, G. J. Acta Crystallogr 2000, D56, 249–265.
2. Herzberg, O.; Moult, J. Proteins 1991, 11, 223–229.
3. Stites, W. E.; Meeker, A. K.; Shortle, D. J Mol Biol 1994, 235, 27–32.

4. Gunasekaran, K.; Ramakrishnan, C.; Balaram, P. J Mol Biol 1996, 264, 191–198.
5. Pal, D.; Chakrabarti, P. J Mol Biol 1999, 294, 271–288.
6. Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. J Mol Biol 1977, 112**,** 535–542.
7. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucleic Acids Res 2000, 28, 235–242.
8. Hobohm, U.; Sander C. Protein Sci 1994, 3, 522–524.
9. Alexandrov, N. N. Protein Engng 1996, 9, 727–732.
10. IUPAC-IUB Commission on Biochemical Nomenclature. Biochemistry 1970, 9**,** 3471–3479.
11. Chou, P. Y.; Fasman, G. D. Adv Enzymol 1978, 47, 45–148.
12. Kabsch, W.; Sander, C. Biopolymers 1983, 22, 2577–2637.
13. Hubbard, S. ACCESS: A Program for Calculating Accessibilities. Department of Biochemistry and Molecular Biology, University College of London, 1992.
14. Lee, B.; Richards, F. M. J Mol Biol 1971, 55, 379–400.
15. Chakrabarti, P.; Pal, D. Protein Eng 1998, 11, 631–647.
16. Chakrabarti, P.; Pal, D. Prog Biophys Mol Biol 2001, 76, 1–102.
17. Deane, C. M.; Allen, F. H.; Taylor, R.; Blundell, T. L. Protein Eng 1999, 12, 1025–1028.
18. Pal, D.; Chakrabarti, P. J Biomol Struct Dynam 2001, 19, 115–128.
19. Samanta, U.; Pal, D.; Chakrabarti, P. Proteins 2000, 38, 288–300.
20. Karpen, M. E.; de Haseth, P. L.; Neet, K. E. Protein Sci 1992, 1, 1333–1342.
21. Doig, A. J.; MacArthur, M. W.; Stapley, B. J.; Thornton, J. M. Protein Sci 1997, 6, 147–155.
22. Jaenicke, R.; Böhm, G. Curr Opin Struct Biol 1998, 8, 738–748.
23. Ladenstein, R.; Antranikian, G. Adv Biochem Eng Biotech 1998, 61, 37–85.
24. Szilagyi, A.; Zavodszky, P. Structure 2000, 8, 493–504.
25. Usher, K. C.; de la Cruz, A. F. A.; Dahlquist, F. W.; Swanson, R. V.; Simon, M. .; Remington, S. J.; Protein Sci 1998, 7, 403–412.
26. Thompson, M. J.; Eisenberg, D. J Mol Biol 1999, 290, 595–604.
27. Kraulis, P. J. J Appl Crystallogr 1991, 24, 946–950.