PERGAMON

Progress in Biophysics & Molecular Biology 76 (2001) 1–102

Progress in Biophysics & Molecular Biology

www.elsevier.com/locate/pbiomolbio

Review

The interrelationships of side-chain and main-chain conformations in proteins

Pinak Chakrabarti*, Debnath Pal

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

Abstract

The accurate determination of a large number of protein structures by X-ray crystallography makes it possible to conduct a reliable statistical analysis of the distribution of the main-chain and side-chain conformational angles, how these are dependent on residue type, adjacent residue in the sequence, secondary structure, residue–residue interactions and location at the polypeptide chain termini. The interrelationship between the main-chain (ϕ, ψ) and side-chain (χ_1) torsion angles leads to a classification of amino acid residues that simplify the folding alphabet considerably and can be a guide to the design of new proteins or mutational studies. Analyses of residues occurring with disallowed main-chain conformation or with multiple conformations shed some light on why some residues are less favoured in thermophiles.

Keywords: Protein folding; Conformation; Secondary structure propensity; Protein engineering; Residue flexibility; Polypeptide chain termini; Thermostability

Contents

*Corresponding author. Fax: +91-33-334-3886.
E-mail address: pinak@boseinst.ernet.in (P. Chakrabarti).

0079-6107/01/$ - see front matter
PII: S0079-6107(01)00005-0

1. Introduction

Protein folding can be viewed as being driven by the burial of apolar side chains without compromising the potentials of the hydrogen bond donors and acceptors. In addition to hydrogen bonding (Baker and Hubbard, 1984; Jeffrey and Saenger, 1991; McDonald and Thornton, 1994), the native structure of the protein, which is remarkable in the compactness of its core (Richards, 1977; Hubbard et al., 1994), exhibits various other noncovalent interactions involving the side chains (Burley and Petsko, 1988; Samanta et al., 2000). As the secondary structures are characterized by conformational features or the hydrogen-bonding pattern of the main-chain atoms, to the first approximation it appears that that the role of the side-chains is mainly to form a stable tertiary structure through the proper packing of the secondary structural elements. In fact, the key properties of the native state—namely, compactness (Hue and Dill, 1991), uniqueness (Shakhonovich and Gutin, 1990) and characteristic folding motifs (Finkelstein and Ptitsyn, 1987; Chothia and Finkelstein, 1990)—may result more from the general physical properties of the polypeptide chain than from specific sequence features (Wodak and Rooman, 1993). Justification for this comes from the landmark work by Ramachandran and coworkers (Ramachandran et al., 1963; Ramakrishnan and Ramachandran, 1965; Ramachandran and Sasisekharan, 1968), who showed how considerations based on simple spatial exclusion place major limitations on the conformation accessible to polypeptides. The general agreement of the allowed regions in the plot of the main-chain torsion angles, ϕ and ψ, with the observed conformations in proteins has provided strong evidence that local interactions within a single dipeptide unit are sufficiently strong to provide powerful restraints on torsional freedom, irrespective of the nature of the side chain. Indeed, as compared to the plot for Ala (where the side chain extends only up to C^{β} atom), the addition of a C^{γ} atom (from a longer side chain) was found to have the effect of removing the regions which are not highly populated (Ramachandran et al., 1965); 90% of ϕ, ψ angles of all non-Gly residues lie in only 14% of ϕ, ψ space (Morris et al., 1992). However, in spite of the near equivalence of ϕ, ψ distributions of different residues, there are sequence-specific features that are important, as has been revealed by comparing designed proteins to the targets (Fedorov et al., 1992; Betz et al., 1993; Quinn et al., 1994; Smith et al., 1995). Although hydrophobic interactions appear to be sufficient to drive folding and achieve a compact structure, it does not necessarily lead to a unique structure for these molecules; uniqueness requires a number of specific interations to be present in the protein core.

Ramachandran and Sasisekharan (1968) also outlined the use of van der Waals potentials to convert the simple allowed/disallowed distinction to a continuous function of conformational energy. Such ϕ, ψ energy contour plots for the backbone of each of the 20 amino acid residues have been computed (Ponnuswamy and Sasisekharan, 1971; Sasisekharan and Ponnuswamy,

1971; Zimmerman et al., 1977; Finkelstein and Ptitsyn, 1977). Calculations have been carried out with various empirical force fields and quantum mechanics, with and without solvation terms, keeping the covalent peptide geometry fixed or allowing it to relax at each point of the ϕ, ψ map (Brooks and Case, 1993; Hermans, 1993; Vásquez et al., 1994; Lazardis et al., 1995). While the plots capture the basic features of protein conformations, the calculated energetics may not conform to the observed ϕ, ψ distribution in entirety (Karplus, 1996), suggesting that the current force fields are still not free of defects (Roterman et al., 1989). As long as energy-minimized structures do not match crystal structures accurately (Whitlow and Teeter, 1986) there is always scope for improving molecular mechanics calculations and data base potentials (Wodak and Rooman, 1993; Kuszewski et al., 1996; Moult, 1997) by analysing the observed distribution of conformational parameters.

There are evidences to indicate that small peptides, some of which correspond to early folding regions, can have well-defined conformations in solution (Brown and Klee, 1971; Chakrabartty and Baldwin, 1995; Dyson et al., 1988; Serrano, 2000), leading to the suggestion that local effects may dominate in some regions of the protein and thereby play an important role in determining the folding pathway (Wright et al., 1988). Moreover, statistical analyses have revealed that the distribution of amino acid residues is not random along the polypeptide chain, but different residues have different propensities to occupy the secondary structural elements (Chou and Fasman, 1978; Levitt, 1978). This could be because of the chemical nature of the side chain (and the resulting difference in how it interacts with the rest of the molecule and the solvent), and/or the effect of the side chain (more specifically, its conformation) on the conformation of the main chain.

From the early protein structures it was evident that side-chain torsion angles tend to cluster around the three staggered positions of the γ-atom ($\chi_1 \approx 60^\circ$, 180° and -60°) (Chandrasekaran and Ramachandran, 1970; Sasisekharan and Ponnuswamy, 1970), and the distributions of all individual χ angles that define the side-chain conformation have been studied (Janin et al., 1978; Bhat et al., 1979; Benedetti et al., 1983; James and Sielecki, 1983). There are clusters in the n-dimensional χ-angles space corresponding to the local minima of potential energy, and Ponder and Richards (1987) identified these to derive a library of rotamers. Several groups have since compiled updated rotamer libraries (Schrauber et al., 1993; Tuffery et al., 1991, 1997; De Maeyer et al., 1997; Dunbrack and Cohen, 1997). Although rotamers generally correspond to the relaxed state of the side chain where no atomic contacts are made to the backbone atoms (Ponnuswamy and Sasisekharan, 1971; Janin et al., 1978; Gelin and Karplus, 1979), the libraries usually contain some rotamers that exhibit impossible atomic overlaps. Recently, Lovell et al. (2000) have developed a new library after removing uncertain residues (with temperature factor $\geqslant 40$ Å^2 or van der Waals overlaps $\geqslant 0.4$ Å) and flipping of the planar side chains (when required by atomic overlaps or hydrogen bonding), a procedure that greatly improved the clustering of rotamer populations. The rotameric preferences are also known to get altered depending on the secondary structure in which a residue is located (McGregor et al., 1987).

In this article, we study the effect of the side chain on the main-chain conformation by analysing the interrelationship between the side-chain and main-chain conformations in three dimensions. As all the three torsion angles, ϕ, ψ and χ_1 involve a common bond ($N-C^\alpha$, Fig. 1), they are not independent of each other over their whole range (Dunbrack and Karplus, 1993, 1994; Chakrabarti and Pal, 1998). As will be shown, such a distribution is dominated by the local short-

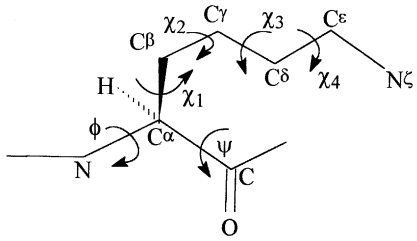

Fig. 1. Using Lys as an example, the main-chain (ϕ, ψ) and side-chain (χ_s) torsion angles are defined. The atom C of the preceding residue and N of the following residue are used in specifying ϕ and ψ, respectively, while χ_1 is defined starting from N.

range interactions between the side-chain atoms (usually up to C^{γ}/O^{γ}) with the main-chain atoms. Consequently, each residue has its own unique joint distribution of the three torsion angles ϕ, ψ and χ_1 and the parameters derived from these distributions reflect on the secondary structural propensities of the residues. For some residues, notably Pro, the flanking residues can also influence its main-chain conformation, and can contribute to the occurrence of the peptide bond preceding the Pro residue in the *cis* conformation. Beyond the two immediate neighbours the secondary structure in which a residue resides not only determines its main-chain conformation but also restricts the χ_1 torsion angle. The side-chain rotamer of a pair of residues can be further restricted if there is any specific interaction between them brought about by their close proximity in the secondary structure. In short, the dependence of the conformational angles on the residue type, its neighbours and the environment in the protein structure will be analysed in a progressive manner so as to bring out the similarity and dissimilarity between different residues from the structural perspective. The relevance of the results in our understanding of protein folding, stability and mutational studies will be discussed. Rather than being an exhaustive review of protein conformations, the thrust is to use conformational features to understand residue-specific properties in protein structure and folding.

2. Conventions, methodology and representation of data

Analyses based on databases require continuous update as ever more high quality structural data become available. Conclusions drawn from the weak statistics derived from small dataset are of doubtful validity. So while compiling results from different papers we have tried to reproduce some of them (and also calculate new parameters) with the data available now, so that much more robust statistics are ensured. The structures were selected from the Protein Data Bank (PDB) (Bernstein et al., 1977; Abola et al., 1997) at the Research Collaboratory for Structural Bioinformatics (RCSB) (http://www.rcsb.org/pdb/) (Berman et al., 2000), based on the January-2000 release of the representative list found at http://www.sander.embl-heidelberg.de (Hobohm and Sander, 1994). The list contains structures determined at a resolution of 2.0 Å or better, and *R*-factor $\leqslant 20\%$; the maximum sequence identity between any two of the polypeptide chains is $\leqslant 25\%$. Table 1 lists the PDB codes for 393 structures (with 408 polypeptide chains) used in the

Table 1
PDB codes (with the polypeptide identifier, if any, separated by a hyphen) for the structures used

153L	16PK	1A1I-A	1A1Y-I	1A28-B	1A2P-A	1A2Z-D	1A34-A
1A3A-B	1A3C	1A48	1A4I-A	1A6M	1A7S	1A8D	1A8E
1A9X-F	1ABA	1ADO-D	1ADS	1AE9-A	1AFW-A	1AGQ-B	1AHO
1AI9-A	1AIE	1AJS-A	1AK1	1ALV-A	1AMF	1AMM	1AOC-B
1AOH-A	1APY-A	1AQ6-B	1AQB	1ARU	1ATL-B	1AUN	1AVW-B
1AWD	1AXN	1AY7-B	1AYF-B	1AYL	1AYO-A	1AZO	1B0N-AB
1B0Y	1B2P-A	1B2V-A	1B3A-B	1B4K-A	1B5E-C	1B65-E	1B6G
1B8O-A	1B93-B	1BA8-A	1BAB-B	1BBH-B	1BBP-A	1BDO	1BE9-A
1BEA	1BEC	1BEN-B	1BF6-B	1BFD	1BFG	1BFT-A	1BG6
1BGF	1BI5-A	1BJ7	1BK0	1BK7-A	1BKR-A	1BM8	1BQC-A
1BRT	1BS0-A	1BS4-A	1BS9	1BSM-B	1BTN	1BU7-A	1BW9-B
1BX7	1BXA	1BXO	1BY2	1BYI	1BYQ-A	1C3D	1C3W-A
1C52	1C53	1CB8-A	1CBN	1CCZ-A	1CEO	1CEQ-A	1CEW-I
1CEX	1CF9-C	1CFB	1CJW-A	1CKA-A	1CLE-A	1CMB-A	1CNV
1CPO	1CPQ	1CQY-A	1CS1-A	1CSH	1CTJ	1CV8	1CVL
1CXQ-A	1CXY-A	1CY5-A	1CYD-B	1DCI-A	1DCS	1DFN-B	1DHN
1DIN	1DLF-HL	1DOK-B	1DOS-A	1DPS-I	1DPT-B	1DUN	1DXG-A
1ECD	1ECP-C	1EDG	1EDM-B	1EGP-A	1EUS	1EXT-A	1EZM
1FIP-A	1FIT	1FLE-I	1FLT-VY	1FNA	1FRP-A	1FUS	1FVK-A
1G3P	1GAI	1GCI	1GCM-C	1GDO-D	1GKY	1GOF	1GP1-B
1GPE-A	1GSA	1GUQ-A	1HFC	1HFE-T	1HKA	1HOE	1HTA
1HTR-P	1HUU-B	1HXN	1IAB	1IDA-B	1IFC	1IIB-B	1ISU-A
1IXH	1JDW	1JER	1JHGA	1KNB	1KOE	1KP6A	1KPTB
1KVE-AD	1LAM	1LAT-A	1LCL	1LIS	1LKF-A	1LKI	1LKK-A
1LOU	1LTS-AC	1LUC-A	1MDC	1MFM-A	1MKA-A	1MLA	1MML
1MOF	1MOL-B	1MOQ	1MPG-A	1MRJ	1MRO-CDE	1MSI	1MSK
1MTY-BG	1MUG-A	1MUN	1NAR	1NBC-A	1NCI-A	1NIF	1NKD
1NKR	1NLR	1NLS	1NOX	1NP4	1NPK	1NUL-B	1OAA
1OBW-B	1OPD	1OPY	1ORC	1OTF-D	1PBE	1PCF-E	1PGS
1PHF	1PHN-A	1PLC	1PNE	1POA	1POC	1PPN	1PSR-A
1PTQ	1PTY	1PYM-B	1QAZ-A	1QB7-A	1QBZ-A	1QCX-A	1QDD-A
1QFM-A	1QFO-A	1QGW-BD	1QH5-B	1QHF-A	1QKS-A	1QQ4-A	1QQP-124
1QRE-A	1QTS-A	1QTW-A	1RB9	1RCF	1REC	1REG-X	1RGE-B
1RHS	1RIE	1RZL	1SCJ-B	1SFP	1SGP-I	1SKF	1SLU-A
1SMD	1SML-A	1SRA	1SUR	1SVF-BC	1SVP-A	1SVY	1TAF-AB
1TAX-A	1TCA	1TEN	1TGX-A	1TIB	1TIF	1TML	1TOA-B
1TTB-B	1TVX-A	1U9A-A	1UBP-ABC	1UDC	1UNK-A	1UOX	1VCA-A
1VFR-A	1VFY-A	1VHH	1VID	1VIE	1VLS	1VNS	1WAB
1WAP-O	1WDC-A	1WHI	1WHO	1WWC-A	1XNB	1YAC-B	1YCC
1YGE	1YTB-B	256B-B	2A0B	2ABK	2ACY	2AHJ-A	2ARC-A
2AYH	2BC2-A	2BOP-A	2CBP	2CCY-B	2CHS-L	2CTC	2DRI
2DTR	2EBN	2EBO-B	2END	2ERL	2FDN	2GAR	2GDM
2HBG	2HDD-B	2HFT	2HMZ-C	2IGD	2ILK	2IZH-D	2KNT
2MSB-A	2MYR	2PII	2PSP-A	2PTH	2PVB	2QWC	2RN2
2SAK	2SIC-I	2SN3	2SNS	2SPC-B	2TPS-A	2TRX-A	2TYS-B
3B5C	3CHB-G	3CHY	3CLA	3CYR	3ENG	3EZM-A	3GRS
3LZT	3PTE	3PVI-A	3PYP	3SDH-B	3SEB	3SIL	3TDT
3TSS	3VUB	451C	4EUG-A	4MT2	4PGA-A	4TSV-A	5HPG-A
5P21	5PTI	6CEL	6GSV-A	7A3H-A	7RSA	8ABP	8PRK-A
9WGA-A							

analysis. The number of occurrences and the percentage composition of residues in the dataset are given in Table 2.

Torsion angles were calculated by means of the DIHDRL program, available from PDB. These are normally given in the range -180° to $+180^\circ$; however, to have a continuous distribution of points the ranges used in some of the plots and tables are from -120° to 240° for ψ and from -240° to 120° for χ_1. For the analysis of χ_1 angles, we use the formalism: $t = 180^\circ$, $g^+ = -60^\circ$ and $g^- = +60^\circ$ (Fig. 2). The whole angular range of 360° is divided into three bins centred around these three canonical values to define the three χ_1 rotamers: $t = 120^\circ$–240°, $g^+ = -120^\circ$ to 0° and $g^- = 0$–120°. According to IUPAC–IUB Commission recommendations (IUPAC–IUB Commission on Biochemical Nomenclature, 1970) the relative orientation of the two branches on the C^β atom in Val is different from that in Thr and Ile (Fig. 3). As a result, at any χ_1 angle the position of the two non-hydrogen 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^-).

Table 2
Number of occurrences and percentage of total population of residues in the dataset[a]

Residue	Number	%
Ala	7050	8.59
Gly	6511	7.93
Pro	3966	4.83
Class I	31470	38.36
Ser	5197	6.33
Cys	1338	1.63
Met	1690	2.06
Glu	4874	5.94
Gln	3161	3.85
Arg	3743	4.56
Lys	4698	5.72
Leu	6769	8.25
Class II	8773	10.69
Asp	4867	5.93
Asn	3906	4.76
Class III	9311	11.35
His	1885	2.30
Phe	3166	3.86
Tyr	3019	3.68
Trp	1241	1.51
Class IV	14950	18.22
Val	5680	6.92
Ile	4341	5.29
Thr	4929	6.01

[a] Residues are also grouped in classes as defined in Table 5. The numbers given here are based on the sequence record of the PDB files. However, some of the residues might be disordered or not located in the electron density maps, and would thus be excluded from Table 6.

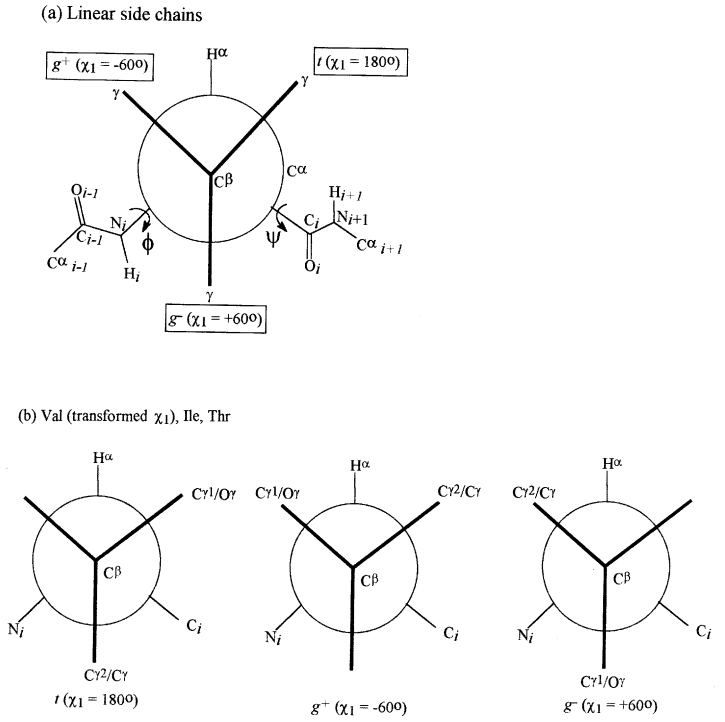

Fig. 2. Newman projections, down the C^{β}–C^{α} bond, for residues other than Ala, Gly and Pro. (a) The three staggered positions of the γ atom, their χ_1 torsion angles (N–C^{α}–C^{β}–C^{γ}/O^{γ}) and designations (t, g^-, g^+), for residues with no branching up to the β position. (b) Three conformational states for the side chain in Val, Ile and Thr with branching at C^{β}. The atom labels correspond to Ile (and Val)/Thr. χ_1 values of Val have been calculated after interchanging the atom labels of $C^{\gamma 1}$ and $C^{\gamma 2}$ (this is equivalent to increasing the "standard" χ_1 value by 120°).

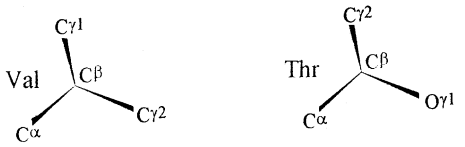

Fig. 3. Numbering and disposition of atoms in the branched side chain of Val and Thr, in accordance with the IUPAC–IUB Commission recommendations (1970).

2.1. Regions in the Ramachandran plot

For the discussion of conformation and modelling purposes the Ramachandran plot has been divided into regions (usually 6–7, but can be as large as 16) on the basis of observed maxima in the ϕ, ψ angles (Zimmerman et al., 1977; Richardson and Richardson, 1989; Wilmot and Thornton, 1990; Rooman et al., 1992; Topham et al., 1993; Efimov, 1993; Karplus, 1996). The one used by Karplus (1996) is shown in Fig. 4. We use a simplified, four rectangular domains to designate the gross conformational features (Fig. 5a). Additionally, for discussion in Section 14.2.2, three

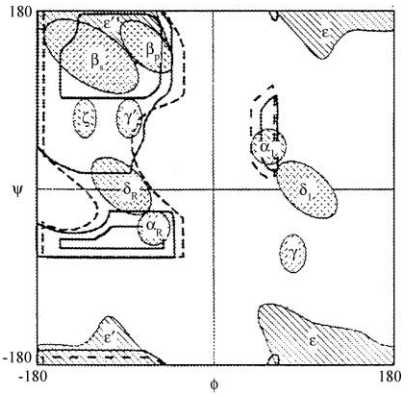

Fig. 4. Fully allowed and partly allowed regions for an Ala dipeptide with N–C^{α}–C angle = 110° (thick solid lines) and the partly allowed regions for an N–C^{α}–C angle of 115° (thick dashed lines) based on the hard sphere model (Ramachandran and Sasisekharan, 1968) are superimposed with a nomenclature for various regions (shaded areas with central Greek letters) of the plot, as used by Karplus (1996). The regions are as follows: α_R, right-handed α-helix region; α_L, mirror image of α_R; β_S, region largely involved in β-sheet formation; β_P, region associated with extended polyproline-like helices, but also observed in β-sheets; γ and γ', regions forming tight turns known as γ and inverse-γ turns; δ_R, right-handed region commonly referred to as the bridge region; δ_L, mirror image of δ_R region; ε, extensive region with $\phi > 0, \psi = \pm 180$ that is predominantly observed for Gly; ε' and ε'', mirror images of the two parts of the ε region, given distinct designations because ε'' overlaps heavily with the β_S and β_P regions that are observed commonly for other residues; ξ, a region that is largely associated with residues preceding Pro. The 3_{10}-helix conformation is located near the approach of the δ_R and α_R regions. (Reproduced from Karplus, 1996).

regions, α, β (broadly encompassing regions populated by α-helical and β-sheet residues, respectively) and bridging (the intervening region), have been demarcated in Fig. 5b.

2.2. CONFPLOT

A schematic representation of the ϕ, ψ and χ_1 angles (Fig. 1) in two dimensions is desirable to visualize the interrelationship between them and how they change along the sequence. Such a diagram can be produced using the program CONFPLOT (Pal and Chakrabarti, 1999a) (Fig. 6). The y-axis of the plot is divided into four major bands, corresponding to the four regions of the Ramachandran plot containing non-overlapping clusters of ϕ, ψ points (Fig. 5a), each of which is further subdivided into four groups based on χ_1 (Table 3). There are a number of panels below the x-axis. The first indicates the sequence number of the residues, the second indicates the amino acid type, while the third shows the secondary structural notifier based on the program PROCHECK (Laskowski et al., 1993).

Structures determined under different experimental conditions may show localized changes in the backbone and/or side-chain conformations, which can be visualized using such a plot, for example, lysozyme determined at different temperatures (Kurinov and Harrison, 1995) (Fig. 6b). The observation that the crystal structures of identical proteins can contain residues in different rotamer positions (Faber and Matthews, 1990; Kossiakoff et al., 1992; Kishan et al., 1994) can be highlighted using CONFPLOT.

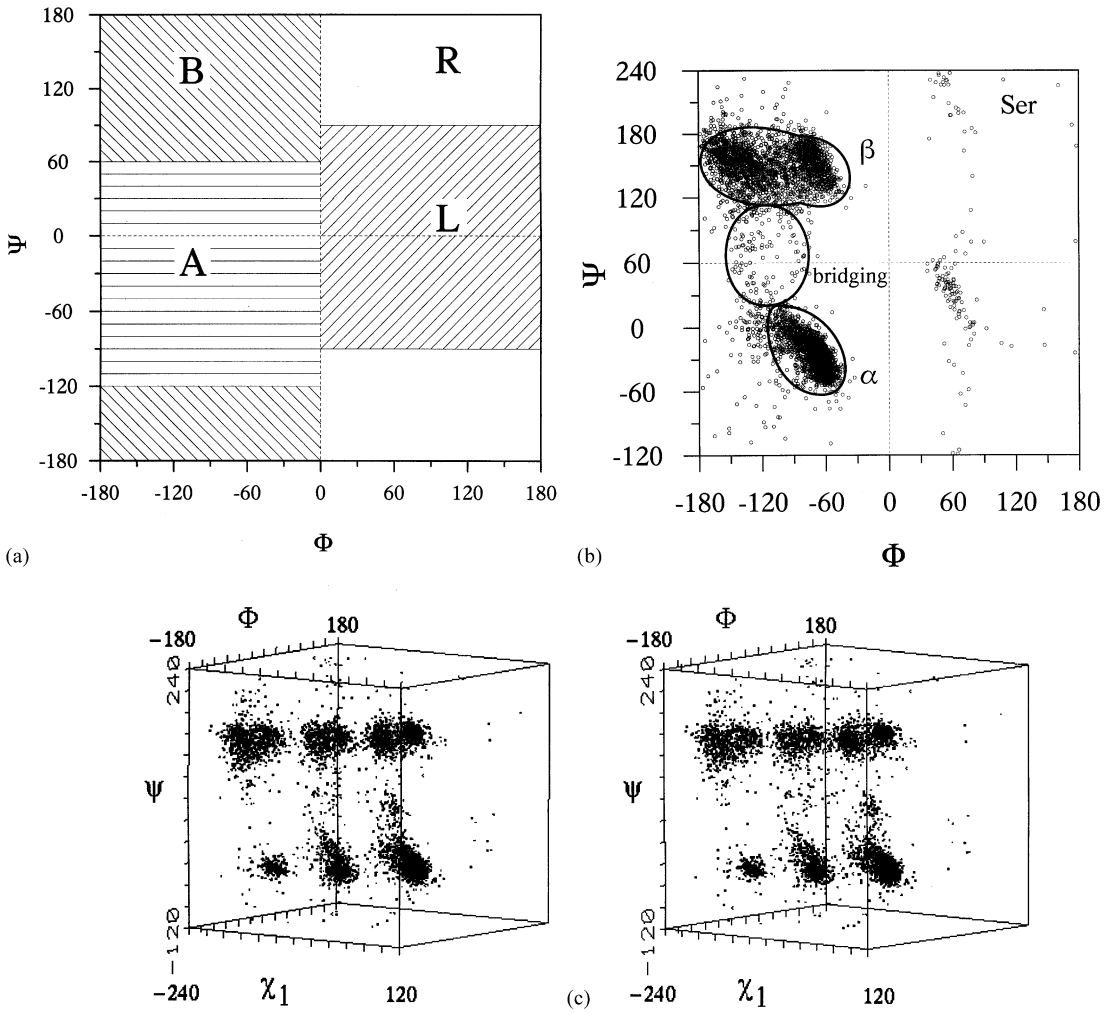

Fig. 5. (a) Four regions in the ϕ, ψ space (ranges are also given in Table 3). (b) Using Ser as an example, α, β and bridging regions are indicated. α and β regions are delineated by residues in α-helices and β-sheets and connected by the bridging region. These will be of different sizes depending on the type of residue and its side-chain conformation, as can be seen in the corresponding three-dimensional distribution (c), shown in stereo.

2.3. Percentage of the ϕ, ψ space occupied

Two methods were employed to determine what fraction of the total area was occupied with ϕ, ψ points (of Gly, Ala and Pro, and those at a given χ_1 for other residues). The number of residues in each $10^\circ \times 10^\circ$ pixel of the available space was counted. In one, the number of grids containing two or more points gave the area occupied (Chakrabarti and Pal, 1998). In another, if a grid contained more than 0.25% of the data points it was assumed to be occupied (for Gly, however, as the points are distributed in all the four quadrants of the map, a threshold value of 0.1% was used) and the total number of such grids gave the occupied area (Pal and Chakrabarti, 1999c); the

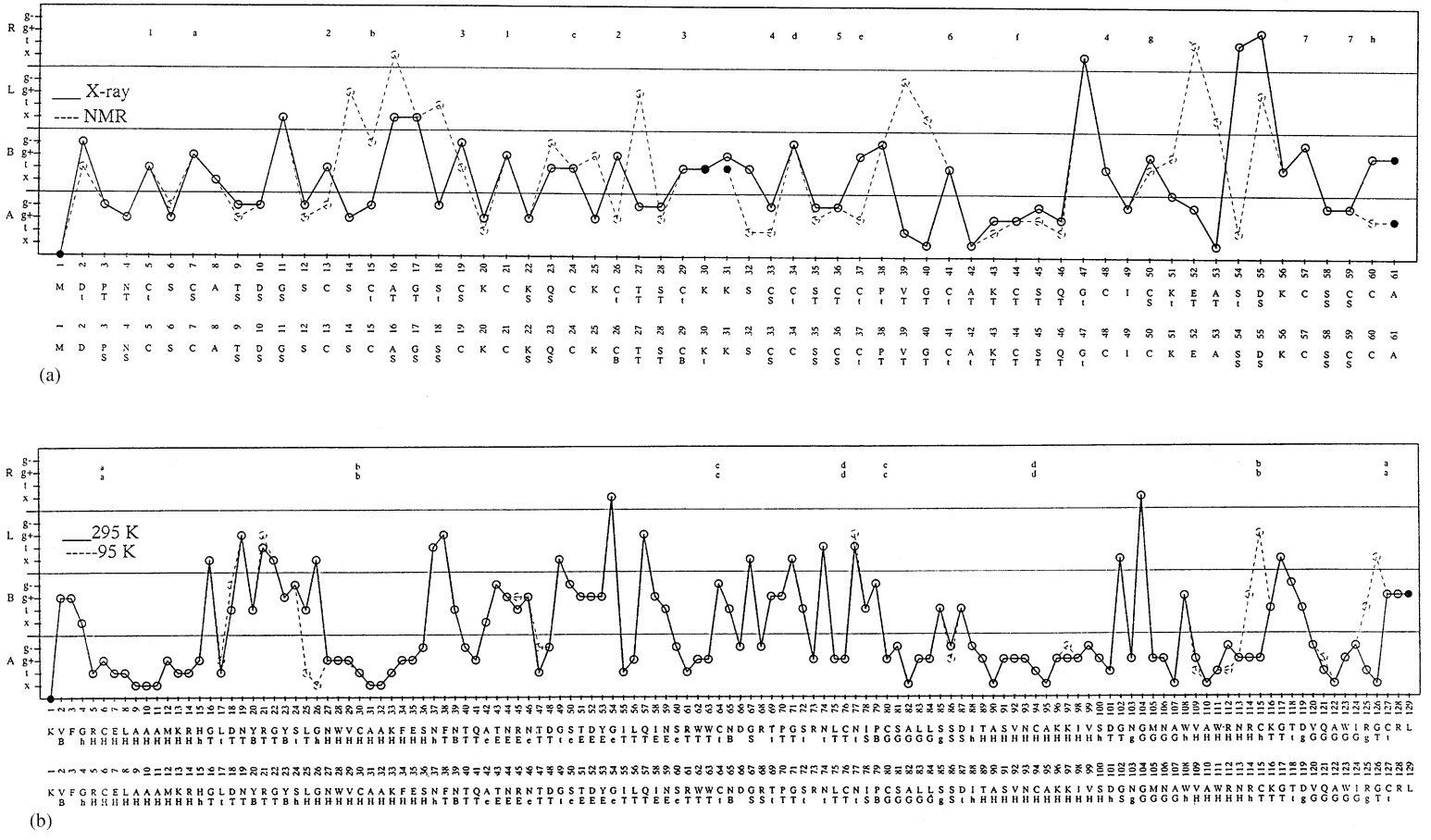

Fig. 6. CONFPLOT to show the differences between (a) the X-ray and the NMR structures of Cd,Zn-metallothionein (PDB files: 4MT2, 1MRT and 2MRT), and (b) hen egg white lysozyme structures determined at two temperatures (PDB files: 1LSE and 1LSF). In (a), metal binding ligands are identified as $1, 2, \ldots, 7$ depending on the serial number of the cations [the first three constitute the $CdZn_2(S^7)_9$ cluster, and the remaining four, the $Cd_4(S^7)_{11}$ cluster] they are coordinated to, or as a (coordinating simultaneously to 1 and 2), b (2,3), c (1,3), d (4,5), e (5,6), f (4,6), g (5,7) and h (6,7). The NMR structure has a break in the peptide bond between residues 30 and 31. In (b) the labels (a–d) in the top panel are the designations of the four disulphide linkages.

Table 3
Designations for different regions (Fig. 5a) in the ϕ, ψ space and the side-chain conformations[a]

	ϕ (°)	ψ (°)	χ_1 (°)
A	−180 to 0	−120 to 60	
B	−180 to 0	60–240	
L	0–180	−90 to 90	
R	0–180	90–270	
x			For Gly and Ala with no χ_1
t			120–240
g^+			−120 to 0
g^-			0–120

[a] These labels are used in Fig. 6.

percentage of the population enclosed in this area was also evaluated. The first procedure generally gave a higher value than the other.

2.4. Conformational similarity indices, $CS_{XX'}$

$CS_{XX'}$ values (Pal and Chakrabarti, 2000c) were computed by finding out the correlation coefficients between the three-dimensional ϕ, ψ, χ_1 distributions of the two residues (X and X'):

$$CS_{XX'} = \frac{\sum_i (N_{Xi} - \langle N_X \rangle)(N_{X'i} - \langle N_{X'} \rangle)}{\sqrt{\sum_i (N_{Xi} - \langle N_X \rangle)^2 \sum_i (N_{X'i} - \langle N_{X'} \rangle)^2}}$$

where N_{Xi} is the number of a residue X at grid i (of size $10^\circ \times 10^\circ \times 10^\circ$) and $N_{X'i}$ is the number at the equivalent position for residue X' and $\langle N_X \rangle$, $\langle N_{X'} \rangle$ are the averages of the numbers of the two residues. The choice of a 10° grid size has been found to be suitable in other studies (Niefind and Schomburg, 1991; Stites and Pranata, 1995).

For comparing residues (Gly, Ala and Pro) with no (or restricted) χ_1, only the two-dimensional ϕ, ψ distribution was used. When comparing the three-dimensional ϕ, ψ, χ_1 distribution of a residue with the two-dimensional ϕ, ψ distribution of Gly/Ala/Pro, the former was divided into three ϕ, ψ distributions corresponding to the three rotameric states of χ_1, and each of them was independently compared with the latter. The weighted average (on the basis of the relative population of X in the three χ_1 states) provided the similarity index. CS_{AX} values relating Ala to all other residues were thus calculated.

2.5. Secondary structural features

The secondary structural elements were assigned in accordance with the algorithm (DSSP) of Kabsch and Sander (1983), 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. Statistics on α-helix are based on all residues marked H, although some data dealing with helices in general also include G and I. For statistics on β-sheet, residues considered had designation E and these were further grouped into those belonging to parallel and antiparallel β-sheets following the

convention of DSSP. Among the turns, only the most prominent category, viz. type I β-turn was considered and the four residues were selected by identifying the two central residues with designations TT and with ϕ, ψ values not deviating by more than 30° from the standard angles (Hutchinson and Thornton, 1994).

2.6. Propensity values

The propensity of a residue (X) to be in a secondary structural element (j) was calculated using the formula (Chou and Fasman, 1974)

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

where

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

and

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

The α-helix propensity (P_α), β-sheet propensity (P_β) and propensity to be in the four individual locations making up a type I β-turn (P_{TI}) were thus calculated. Residues occurring more frequently than the average have propensity values greater than unity. The associated standard deviations were derived (Williams et al., 1987) as

$$\sigma_{X,j} = (1.0/\langle f_j \rangle)\sqrt{[f_{X,j}(1 - f_{X,j})]/n_{X,\mathrm{all}}}.$$

In α-helices the preference of a residue to be located in the N-terminus (or N-end, i.e., the first three residues, marked H according to DSSP notation), C-terminus (C-end—the last three positions) and in between (Interior) were determined by finding the local propensity within α-helices which is 7 residues or more long. For the calculation of these local propensities, $P_{\alpha/\mathrm{l}}$ (as distinct from the normal or global propensity, P_α, calculated above) the terms $n_{X,\mathrm{all}}$ and n_{all} for all proteins are replaced by $n_{X,\mathrm{helix}}$ and n_{helix} (data restricted to α-helices only), respectively, and instead of connoting a secondary structure, j stands for one of the three regions in the helix.

The propensity of a residue (X) in a secondary structure (j) to be in one of the three χ_1 rotameric states, s, was given by

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

where all numbers correspond to the given amino acid; $n_{X/j}$ is the number in the secondary structure j, of which $n_{X/j,\ s}$ are in the conformational state; $s\, n_{X,\ \mathrm{all}}$ and $n_{X,\ s}$ are the corresponding numbers in the whole database. Like $P_{\alpha/\mathrm{l}}$, the calculation of the local propensity, $P_{s/\mathrm{l}}$, in a given region of the helix was implemented in a similar manner—the first two terms mentioned above are corresponding to a given helical region and the last two, to the whole helix.

2.7. Analysis of peptide structures

The Cambridge Structural Database (CSD) (Allen and Kennard, 1993) is a repertoire of structural data on small molecules, including peptides. These structures are much more accurately determined than the proteins and are devoid of any effect due to the secondary structure. These have been used to confirm some trends in conformational features observed in protein structures. Only structures with R-factor $\leqslant 10\%$ and with 'no error' (a flag used in CSD to indicate if the entry contains residual unresolved numerical errors) in coordinates were extracted.

3. ϕ, ψ Distributions for glycine and alanine

The Ramachandran map as applicable to all non-Gly L-amino acid residues was derived using Ala in the model dipeptide unit (Ramachandran et al., 1963). The ϕ, ψ points for Ala residues in protein structure are, however, not spread evenly across the allowed region (Fig. 7a), and occupy 17% or 6% of the total space (depending on the method used), short of the 20% allowed (7.5% fully, and the rest partly, Fig. 4) based on steric considerations alone (Ramakrishnan and Ramachandran, 1965).

The map appropriate for the glycyl residue was worked out by Ramakrishnan and Ramachandran (1965). Due to the lack of a β-carbon, the glycyl map, spanning both the right and left halves of the ϕ, ψ plane, is centrosymmetric with respect to the origin (ϕ, $\psi = 0,0$) and the ϕ, ψ region allowed by extreme limit (57%) is more than double of that for Ala (20%) (Ramakrishnan and Srinivasan, 1990). The minima from energy calculations occur at

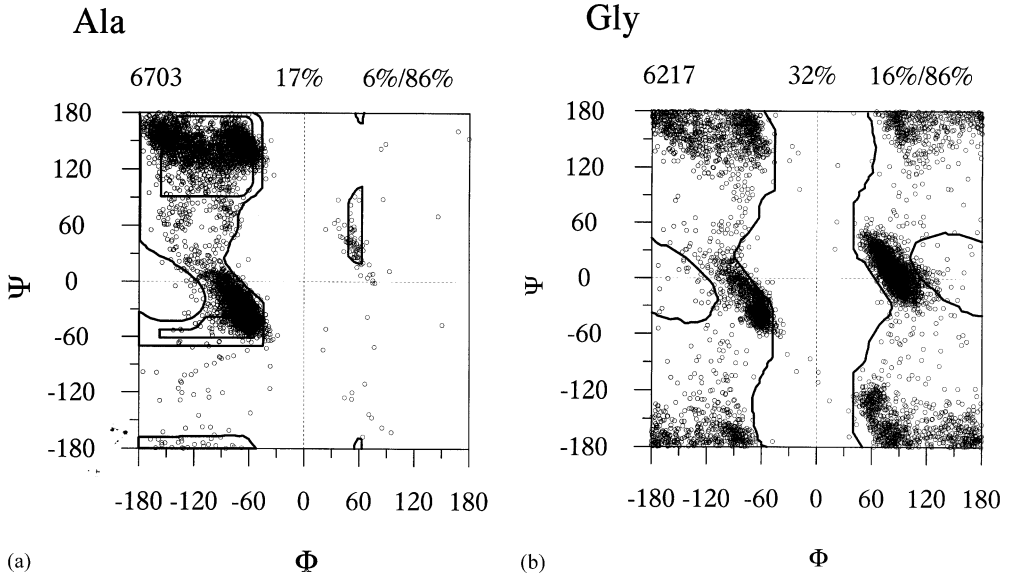

Fig. 7. The ϕ, ψ distributions of (a) Ala and (b) Gly residues superimposed on the respective Ramachandran plot. Values on the top are explained in Fig. 12.

$(-90°, +90°)$ and $(+90°, -90°)$ (Hermans et al., 1992); but these regions are poorly populated (Fig. 7b) in the distribution of glycyl conformations occurring in proteins (Richardson and Richardson, 1989; Nicholson et al., 1989). Karplus (1996) observed that 90% of Gly residues fall within only 18% of the total conformational space (the values are 86% and 16%, respectively, according to Fig. 7b), less than double the 10% observed for non-Gly residues. Another feature of the Gly distribution is that it is asymmetric around the origin; there is a higher concentration of points around the $\psi = 0$ axis with positive ϕ than with negative ϕ. Rather than reflecting energetics, this asymmetry may point to the evolutionary pressure to select Gly in situations where a residue has to be in the left-handed α-helical conformation, like in helix termination (Gunasekaran et al., 1998; Schellman, 1980; Richardson and Richardson, 1988) or in specific locations of different types of β-turn (Hutchinson and Thornton, 1994; Ramakrishnan and Srinivasan, 1990).

4. φ, ψ Distributions for proline and α-aminoisobutyric acid

Proline is an imino (rather than an amino) acid, with a five-membered pyrrolidine ring containing the $N-C^{\alpha}$ bond, the rotation about which (the angle ϕ) is thus constrained to be near $-60°$. As a result the conformational energy of a Pro residue depends largely on ψ, whose values corresponding to the two minima are $-55°$ and $+145°$ (Schimmel and Flory, 1968; Summers and Karplus, 1990). The distribution of ϕ, ψ angles of *trans* Pro residues in protein structures has been reported by MacArthur and Thornton (1991), Nicholson et al. (1992) and Karplus (1996). Ninety-four per cent of Pro residues are restricted to 4% of ϕ, ψ space (Fig. 8). As noted by Nicholson et al. (1992) there is a discrepancy of 10–30° between the values of ϕ that correspond to the potential energy minima and the values that are most frequently observed. In the case of ψ, the potential energy surface has a minima extending from $\psi \sim -60°$ to $\sim 180°$, whereas prolines in proteins segregate into two distinct groups, one with $\psi \sim -30°$ and the other with $\psi \sim 150°$. The topic to be taken up in Section 8.1 is how ϕ and ψ angles can be tuned by the magnitude and sign of χ_1.

Though not a component of proteins, α-aminoisobutyric acid (Aib, α-methylalanine, α,α-dimethylglycine) is another stereochemically constrained amino acid, which is found in diverse fungal polypeptides, and it is instructive to study its conformational features (Prasad and Balaram, 1984; Kaul and Balaram, 1999). Substitution of the α-hydrogen in L-Ala by a methyl group leads to this achiral residue. Consequently, the ϕ, ψ space allowed for both L-Ala and D-Ala residues would define the allowed region for Aib; the superposition of the Ramachandran maps for L-Ala and D-Ala residues, which are related by inversion about the origin indicates that the allowed conformations would lie in a small region around $\phi = \pm 60°$, $\psi = \pm 30°$ (Fig. 9). This line of reasoning first led Ramachandran and Chandrasekaran (1972) to suggest that Aib would be a conformationally restricted residue favouring helical conformations. Similar conclusions were drawn by Marshall and Bosshard (1972) using conformational energy calculations, and more recently by Hermans et al. (1992) by molecular dynamics simulations. The crystal structures of Aib-containing peptides demonstrate the overwhelming tendency of these residues to promote helical conformations (Fig. 9). The ability of Aib residues to nucleate and stabilize helical

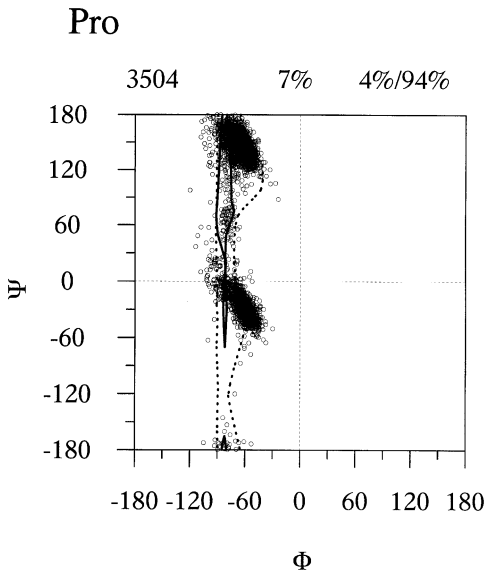

Fig. 8. ϕ, ψ angles of proline with its energy surface (Summers and Karplus, 1990) superimposed. Continuous and dotted contours enclose regions that are within 5 and 10 kcal/mol of the global minimum, respectively. Values on the top are explained in Fig. 12.

conformations in oligopeptides has been used to construct helical modules which can then be assembled into supersecondary structures (Kaul and Balaram, 1999).

5. χ_1-Dependent ϕ, ψ distributions for residues other than glycine, alanine and proline

5.1. Correlation between ψ and χ_1

χ_1, ψ plots for all residues are provided in Chakrabarti and Pal (1998) and the updated diagrams for a selected few are shown in Fig. 10. At each χ_1 angle the points are clustered in A and B regions, but the distributions of points among the side-chain conformational states are not uniform, being in general more diffused along the ψ direction in the g^+ state as compared with the g^- and t states. The Newman projection of a side chain with a single γ atom in a dipeptide unit (Fig. 2a), showing the positions of the three χ_1 rotamers relative to the protein backbone, is helpful in understanding the distribution on steric grounds. The lesser dispersion in the g^- and t states is due to the steric interaction brought about by the proximity of the γ atom and the main-chain carbonyl group. The concerned main- and side-chain atoms are further apart in the g^+ state and this gives a greater freedom for placing the main-chain atoms, making the distribution more diffused along the ψ axis. Between the g^- and t states, the points are scarcer in the former (more so in the A region) as it has two gauche interactions involving both N_i and C_i. For the β-branched residues (Fig. 2b), one of the branches is placed in between N_i and C_i in the g^- and t states, making the remaining (g^+) state the most populated one (Fig. 10). However, going against this trend, Thr (and also Ser) have the maximum population in the g^- state indicating that for these

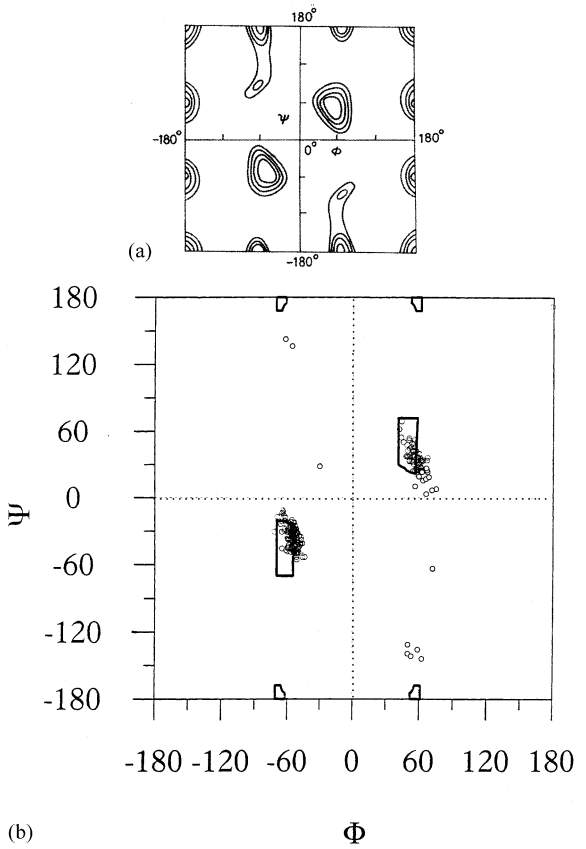

Fig. 9. (a) Potential energy map for Ac–Aib–NHMe; the contours are drawn at 1 kca/mol intervals with respect to the innermost contour enclosing the minimum (from Prasad and Balaram, 1984). (b) Crystallograhically observed ϕ, ψ values of Aib residues (a total of 267 from 114 independent crystal structures of Aib-containing peptides, with no reported error and R factor $<10\%$). In the case of achiral peptides crystallizing in a centrosymmetric space group, the sign of the torsion angles has been chosen arbitrarily. The contours surround regions common to the Ramachandran plots for L-Ala and D-Ala residues, as given in Kaul and Balaram (1999).

residues electrostatic interactions involving the γ hydroxyl group may have a more decisive role than steric interaction.

5.2. Correlation between ϕ and χ_1

Some typical χ_1, ϕ plots are shown in Fig. 11. The g^+ and g^- states bring the γ position and the main-chain N_i atom close to each other (Fig. 2) and the remaining t state shows a wider and/or more evenly spread distribution. In addition to the 1,4-interaction between the γ atom and N_i, there can also be 1,5-interaction of the γ atom with C_{i-1}. The terminal carbons of the organic molecule, pentane, are very close to each other if the two torsion angles in the moiety are $\{+60^\circ, -60^\circ\}$ or $\{-60^\circ, +60^\circ\}$, so that the *syn*-pentane conformations extending to a range of about $\pm 30^\circ$ of the above values are of higher energy than the minimum energy conformation $\{t,t\}$

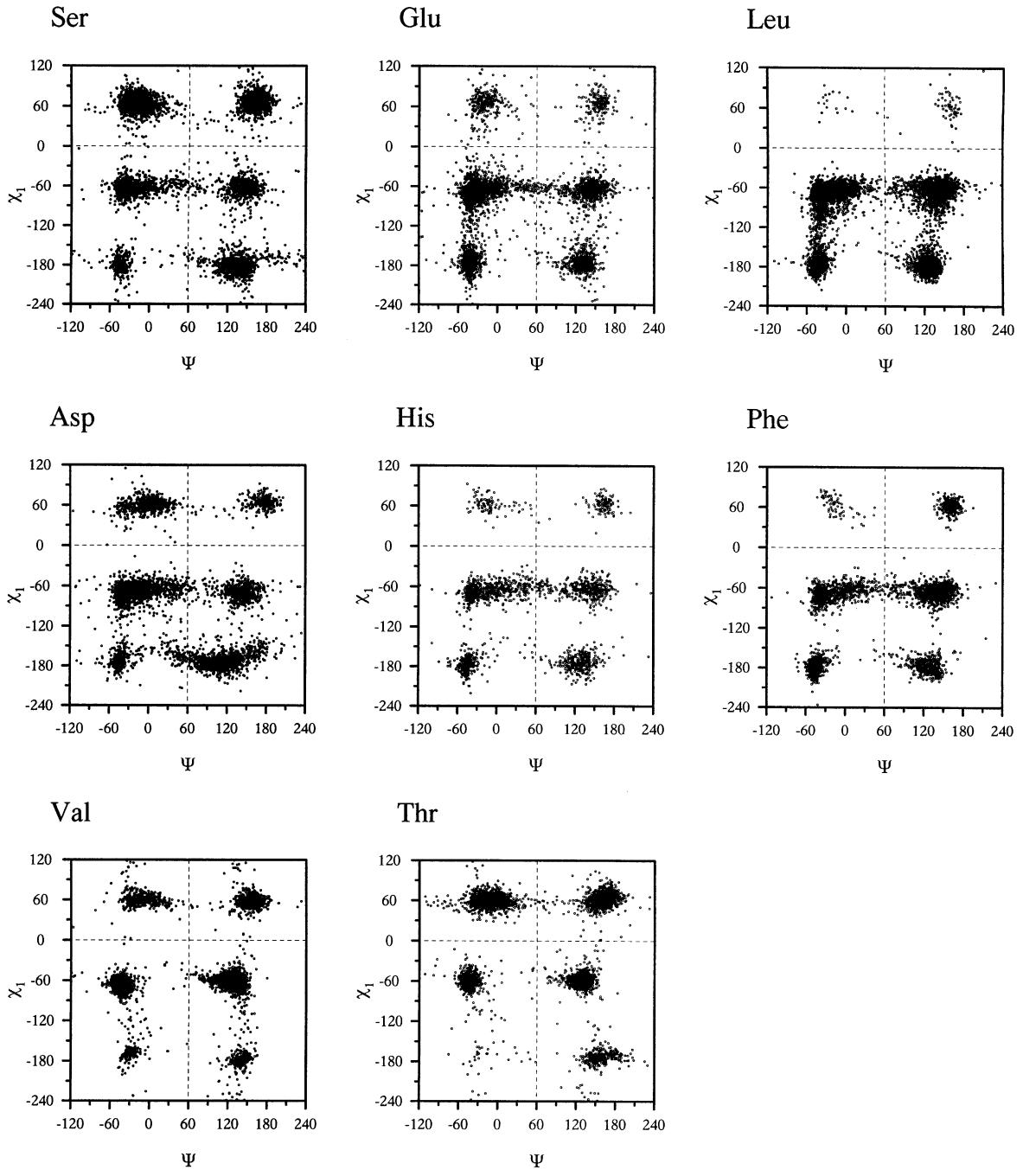

Fig. 10. Joint distribution of χ_1 and ψ for some representative residues.

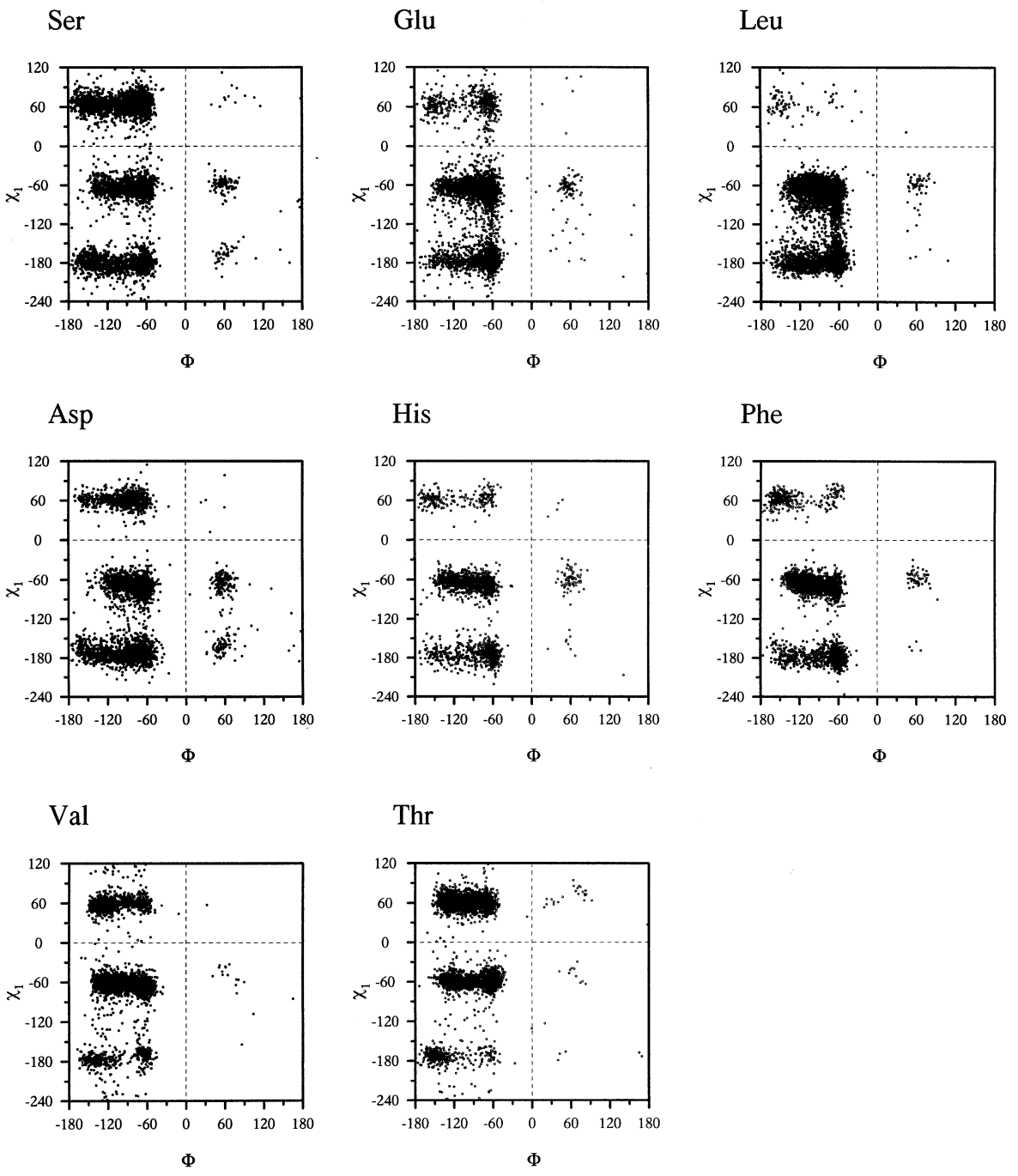

Fig. 11. Joint distribution of χ_1 and ϕ for some representative residues.

(Wiberg and Murcko, 1988). Due to this *syn*-pentane effect, although the g^+ state of the side chain is the most populous, ϕ is not observed within the range -180° to -150° (Dunbrack and Karplus, 1994).

5.3. ϕ, ψ maps at different χ_1 angles

Discussions in Sections 5.1 and 5.2 suggest that the ranges of ϕ, ψ will be different for each χ_1 rotamer (or from a converse point of view not all χ_1 states are possible for all values of ϕ and ψ). The percentage occurrences of all the residues in each χ_1 state is given in Table 6, and at each of these rotameric states the distributions of ϕ, ψ angles are shown in Fig. 12. In general, in the t state there are hardly any points with ψ greater than -30° in the region A, and 150° in the region B. As in the t state the band encompassing the distribution of points in the region B is quite narrow in the g^- state, but it has moved up to lie within the ψ range of 150–180°. Only in the g^+ state the points are rather widely scattered to take up the whole space of what is normally assumed to be the allowed region of the Ramachandran plot. Concerning the spread of points along the ϕ direction, -150° seems to be the extreme lower limit in the g^+ state; even this is brought up to -130° for Asp and Asn. As regards to the upper limit it is around -60° in the g^- state (and $\sim -40^\circ$ in the other two states). In the g^- state there are indications of the points in the region B to bunch either in two clusters or, as for the aromatic residues, to lie in a compact space with ϕ less than -120°; Thr is an exception to this.

In the g^+ state the maximum fraction of the total area is filled up (for Thr it is observed in the g^- state), whereas the occupancy is the least for g^- state with Leu representing the extreme situation. The percentage of the area covered is equally high in the t and g^+ states for Asp and Asn, and equally low in the t and g^- states for Ile. Generally, the bridging area (Fig. 5b) is populated only for the g^+ state. Positive values of ϕ in the L region can be taken up only in the g^+ state. Asp and Asn which have a higher tendency to occupy this region (Richardson, 1981; Srinivasan et al., 1994; Deane et al., 1999) can do so with the side chain in both t and g^+ states.

5.4. Comparison of ϕ, ψ map of Ala with χ_1-dependent ϕ, ψ maps of other residues

It is interesting to know how each of the maps at the three χ_1 angles are different from the ϕ, ψ map of Ala. In Fig. 13 the three difference maps for the class I residues (with the maximum number of cases, Table 6) are shown. The g^+ state (with fewer boxes with large values) bears the closest resemblance. Overall, relative to Ala, the introduction of the γ-atom has the effect of moving the points towards regions with higher ψ and lower ϕ values (i.e., along the lower-right to the upper-left direction, the former region containing progressively more negative values, and the latter, more positive) as the side-chain conformation changes from t to g^+, and then to g^-. Section 6 deals with a more detailed comparison of Ala map with the three-dimensional ϕ, ψ, χ_1 maps of other residues.

5.5. Average helical ϕ, ψ values at three χ_1 rotamers

The main-chain conformational parameters for the different classes of residues in α-helices are presented in Table 7. The overall ϕ, ψ values (neglecting χ_1) in various classes are nearly identical to those for Ala. But when the residues are separated into groups of three χ_1 angles, each group

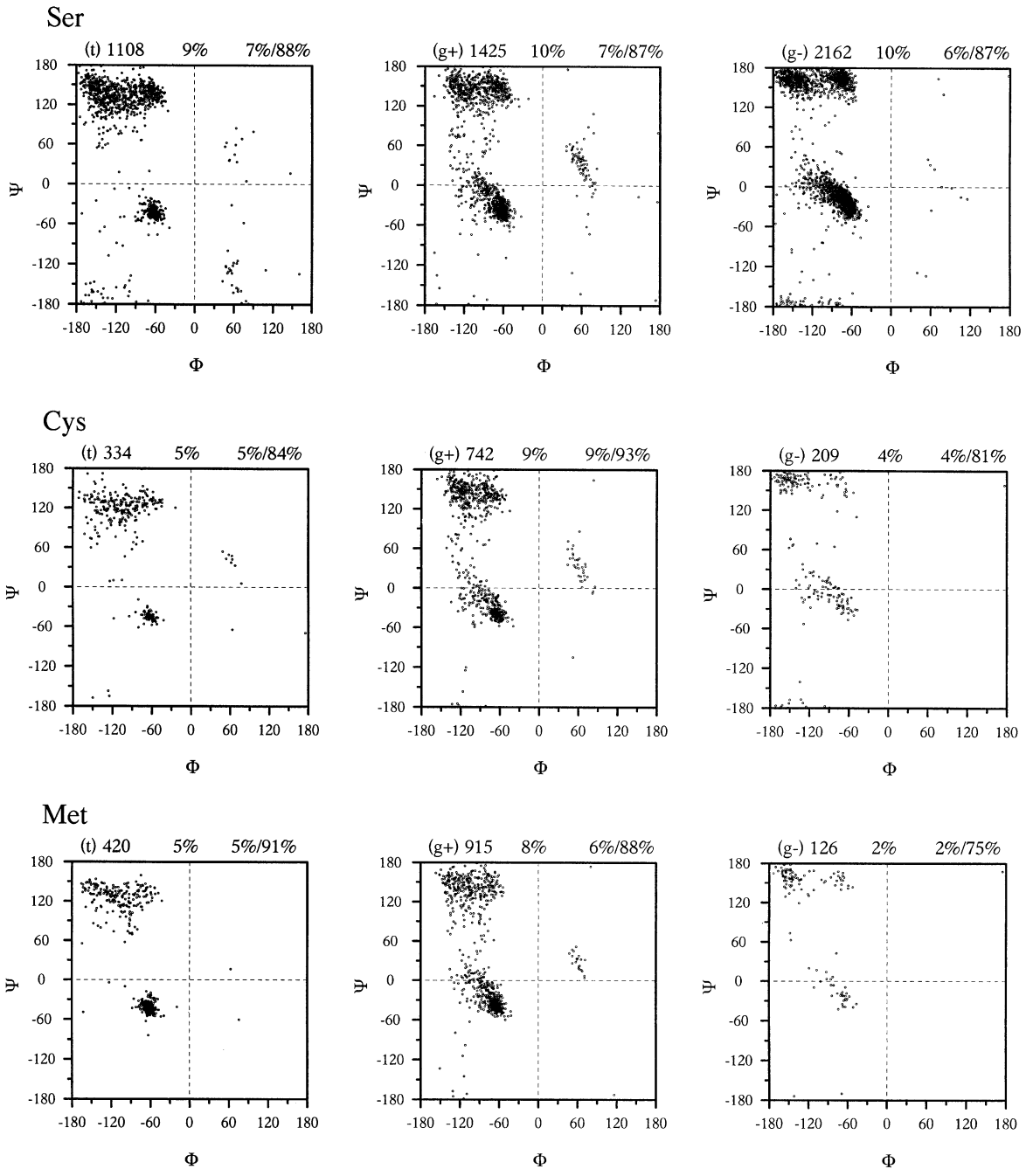

Fig. 12. ϕ, ψ maps of different residues corresponding to the three χ_1 rotameric states. Against each diagram are marked the conformational state, the number of data points, percentages of the plot areas occupied (counting the number of grids with points above threshold values specified in two different ways, as elaborated in Section 2.3) and the percentage of points contained in one of the areas calculated.

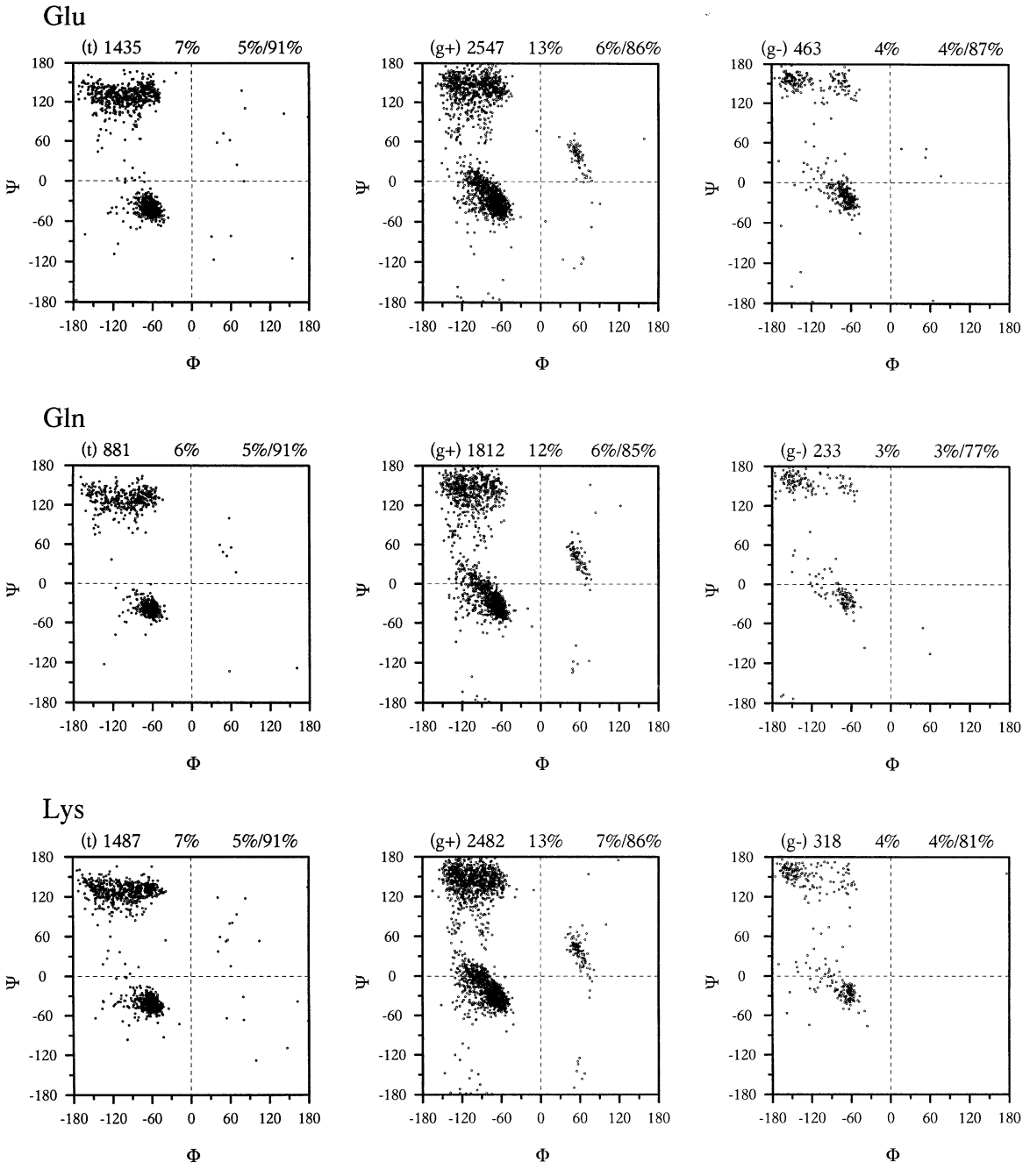

Fig. 12. (*Continued*).

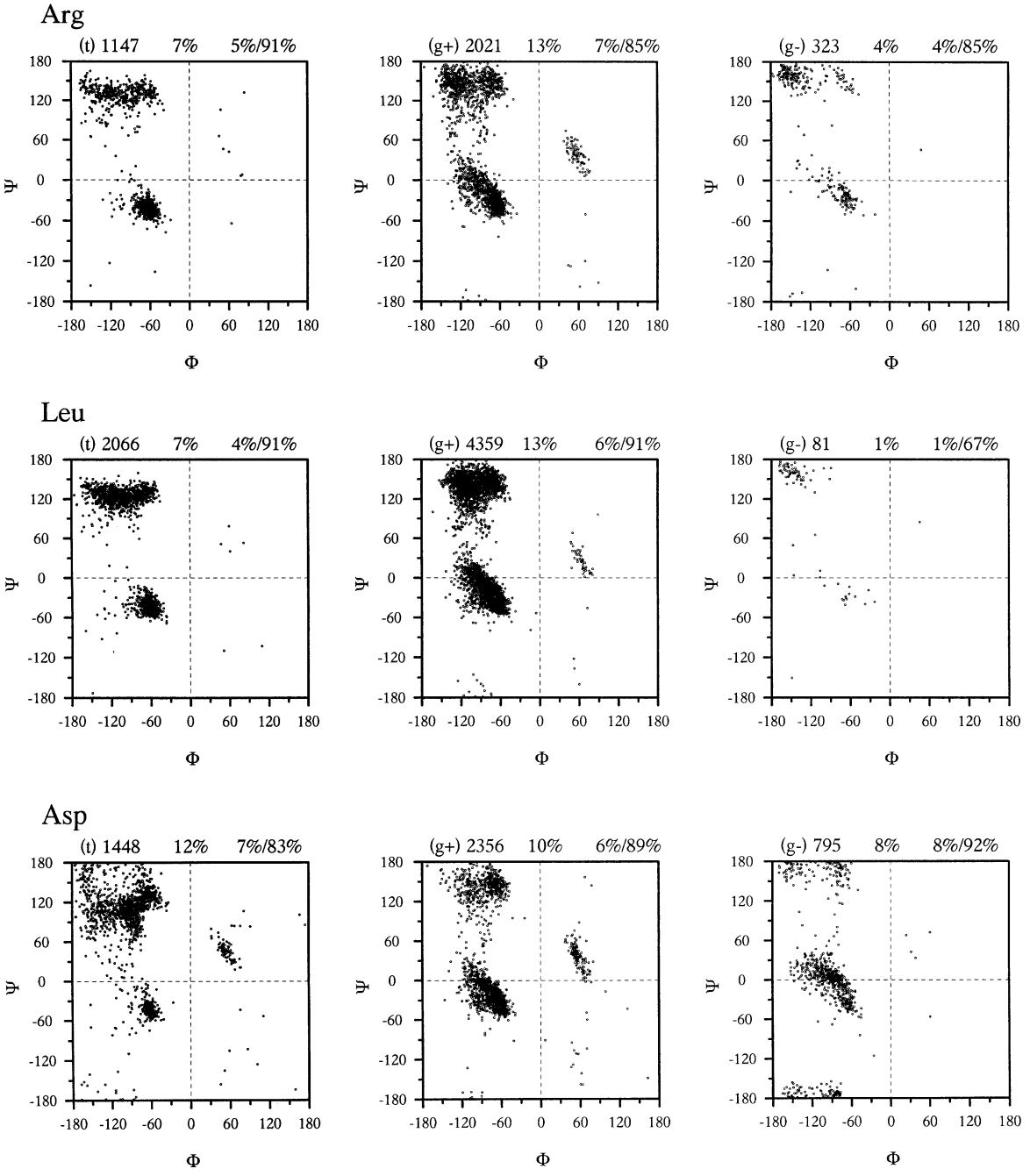

Fig. 12. (*Continued*).

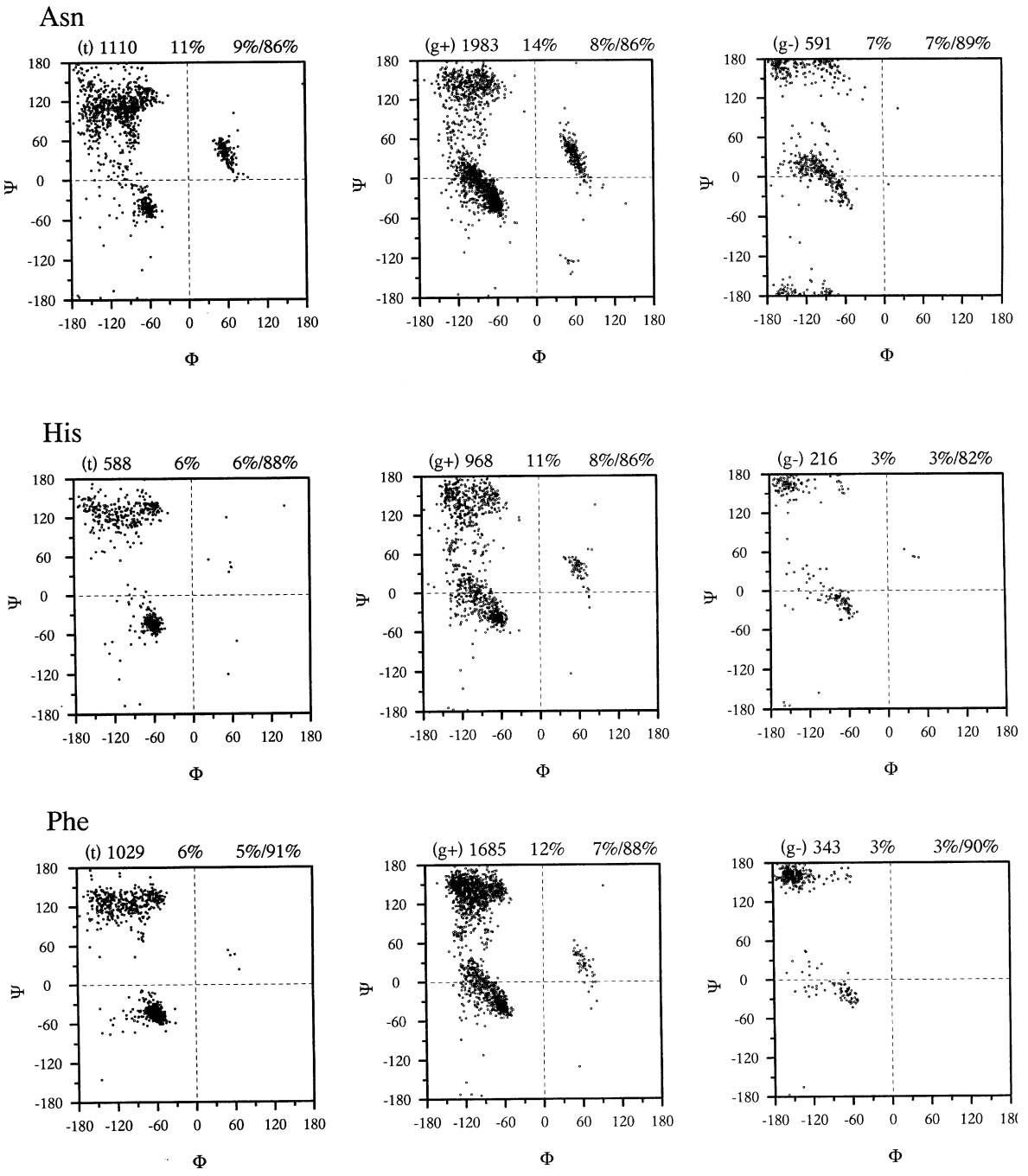

Fig. 12. (*Continued*).

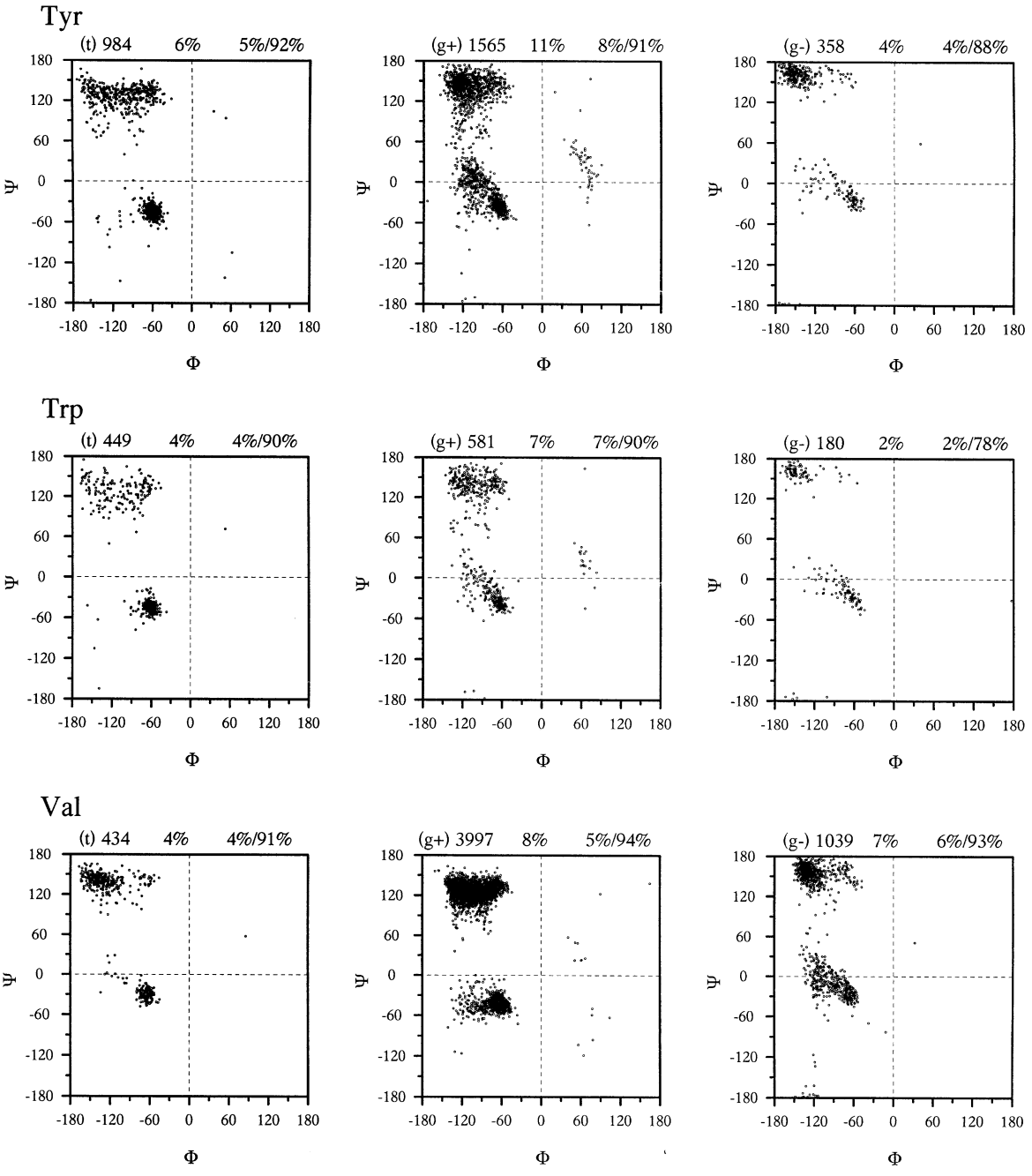

Fig. 12. (*Continued*).

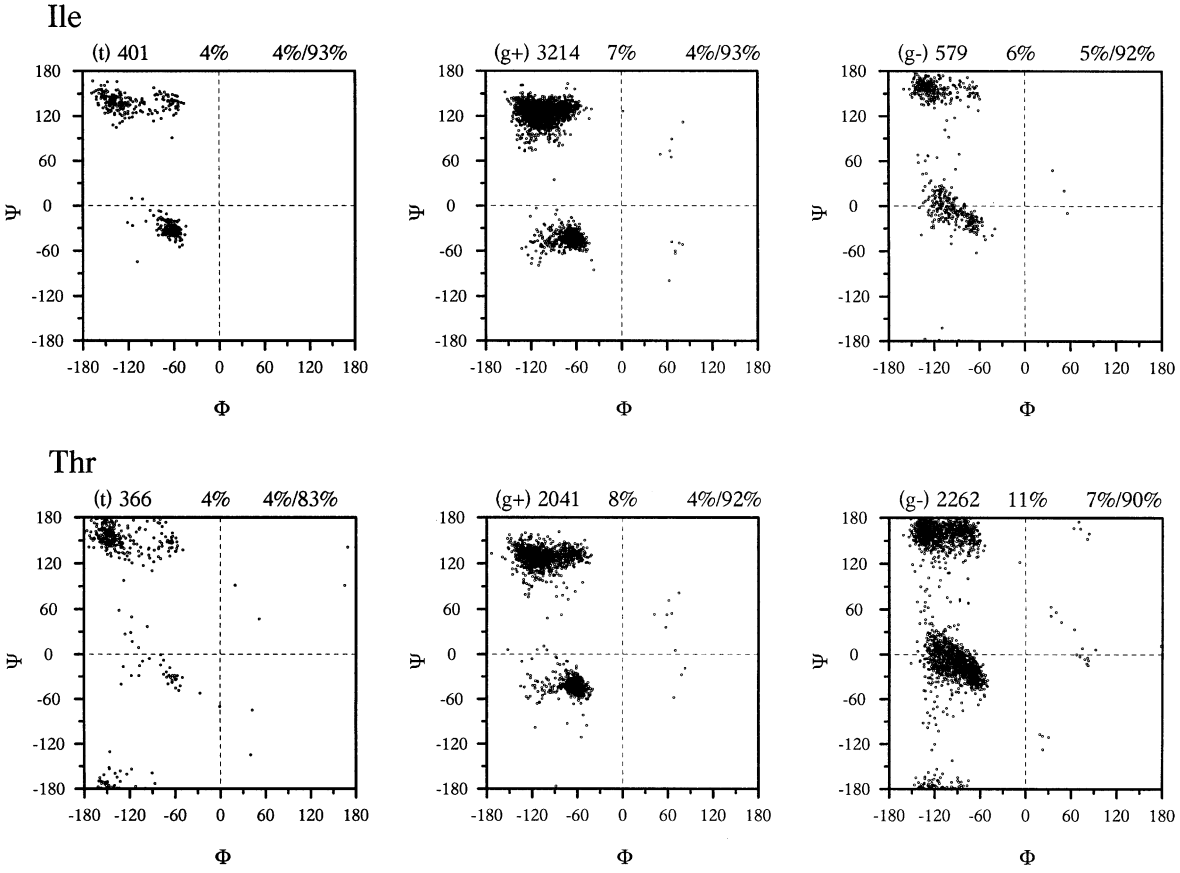

Fig. 12. (*Continued*).

takes up rather distinct ϕ, ψ angles. In particular, the changes between the three states of aromatic (class III) and the β-branched (class IV) residues are quite striking. Considering classes I–III residues, a change from t to g^+ to g^- states results in the average ϕ, ψ values becoming less negative in ψ (and to some extent, more negative in ϕ). The shift away from the average helical conformation is the maximum in the g^- state. A steric clash involving the C^γ atom and the i-3 carbonyl group (Section 13.1.2) is an obvious explanation, but it is likely that there is also a contribution from the interaction between the side-chain and main-chain atoms of the residue, as the same trend was observed—not only for α-helical, but also for the β-sheet region (where the possibility of the aforementioned steric clash does not exist)—in Section 5.4 (Fig. 13), when the ϕ, ψ distributions in different χ_1 states were compared to Ala.

For classes I–III residues in α-helices, g^- is the least occupied state, but it is t in 3_{10}-helices (Table 7). However, as these exist in very short stretches (Smith et al., 1996; Pal and Basu, 1999) there is considerable variation about the average ϕ, ψ values. Smith et al. (1996) noted that 9% of residues in 3_{10}-helices and also a considerable number in β-strands (mostly the terminal residues in a given secondary structure) can have positive ϕ angles; these points have not been considered while calculating the average values.

5.6. Average β-sheet φ, ψ values at three χ_1 rotamers

Table 8 shows that there is no significant difference in the average ϕ, ψ values between residues in parallel and antiparallel β-sheets. However, within classes I–III of residues a change in χ_1 from *t*

Table 4
Shift (°) in the mean ψ and ϕ values as the side-chain conformation is changed from *t* to g^+ to g^- states[a]

Residue	Class	Region A		Region B	
		$\Delta\psi$	$\Delta\phi$	$\Delta\psi$	$\Delta\phi$
Ser	I	−14,−11	4,5	−14,−17	−15,22
Cys	I	−13,−16	8,7	−24,−15	−1,33
Met	I	−9,−16	6,1	−17,−17	2,40
Glu	I	−8,−12	6,1	−13,−13	−3,17
Gln	I	−9,−8	9,−1	−14,−17	14,30
Lys	I	−14,−7	9,−1	−17,−9	2,20
Arg	I	−16,−12	12,1	−16,−13	−1,28
Leu	I	−12, –	10, –	−17,−17	−3,47
Asp	II	−12,−37	4,29	−33,−30	−15,25
Asn	II	−21,−26	15,23	−27,−30	−4,29
His	III	−28,−3	22,−7	−15,−21	7,35
Phe	III	−26,0	19,1	−15,−21	0,38
Tyr	III	−26,1	21,−13	−14,−18	10, 31
Trp	III	−14,−14	9,4	−11,−23	3,47
Val	IV	13,−26	−2,24	13,−30	−18,14
Ile	IV	13,−39	−2,26	13,−30	−18,13
Thr	IV	–,−31	– ,27	31,−31	−33,7

[a] From Chakrabarti and Pal (1998), except that Leu, which was given in a separate class, has been merged with class I

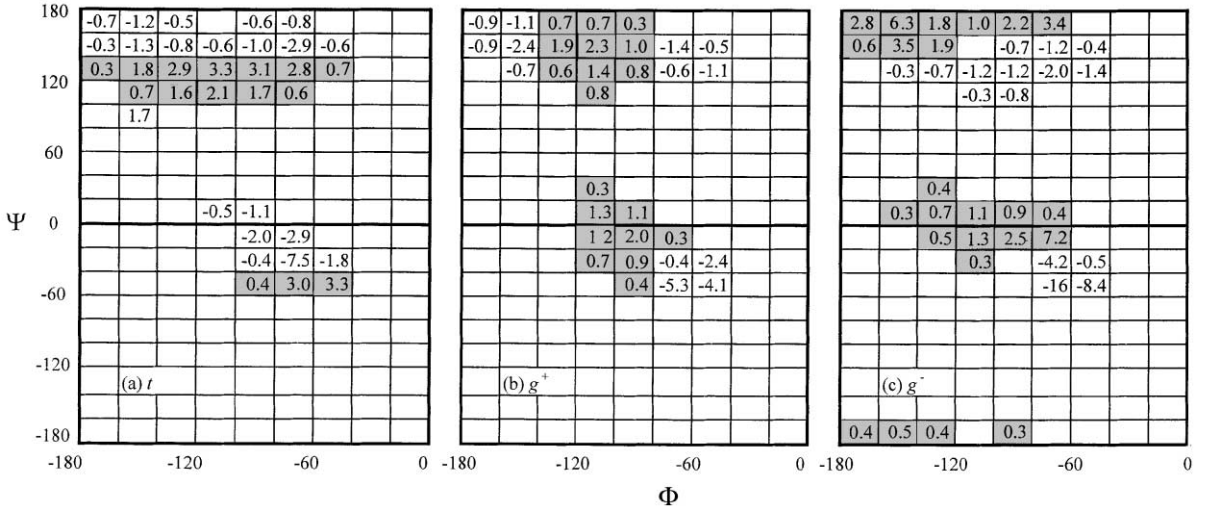

Fig. 13. The difference in the percentage distribution of points in the ϕ, ψ maps (only the negative ϕ region is shown) between the class I residues (Table 5) and Ala (Fig. 7a) (the former minus the latter). Values (< -0.2 and > 0.2) in the individual 20×20 blocks are indicated and those with positive values are shaded.

Table 5
Amino acid classes[a]

Class	Residues
I	Ser, Cys, Met, Glu, Gln, Lys, Arg, Leu
II	Asp, Asn
III	His, Phe, Tyr, Trp
IV	Val, Ile, Thr
V	Ala
VI	Gly
VII	Pro

[a] The classification is the same as in Chakrabarti and Pal (1998), except that Leu has been merged with class I (details are in Section 6.1; also see Table 4). Results from Section 6.3 indicate that Ser can also be made into a separate class.

to g^+ to g^-states results in ϕ becoming more negative and ψ, more positive—the value of ψ can change by more than 30°; the trend in the variation of torsion angles is similar to that found in α-helices (Section 5.5). It is to be noted that for Gly the ψ angle is closer to 180°.

5.7. Average χ_1 values for the three side-chain rotamers in helices and sheets

Chakrabarti and Pal (1998) observed some systematic changes in the average χ_1 values for some residues depending on the location in the A and B regions of the Ramachandran plot. However, when the residues are grouped into classes (Table 9), the trends are less clear, except that in the two *gauche* conformations the class IV residues take up values closer to $\pm 60°$. For others, the mean χ_1 in the g^+ state is close to $-70°$. Also, in the t state the distribution is not symmetric about 180°, the average χ_1 magnitude lying between $-170°$ and $-180°$.

6. Classification of amino acid residues based on conformation

6.1. Classification based on the dependence of ϕ and ψ on χ_1

From Figs. 10 and 11 it can be seen that various residues have different mean values of ϕ and ψ in the two regions (A and B) at a given χ_1. The effect of the side-chain conformation on the main-chain geometry can be quantified by noting the change in these ψ and ϕ values as χ_1 is changed from t to g^+ to g^- states (Table 4) (Chakrabarti and Pal, 1998). These values can be used to group the amino acid residues (excluding Ala, Gly and Pro) into four classes (Table 5): (I) The major class consists of most of the aliphatic residues with no branching in the side-chain before the γ position. Ser, Cys, Met, Glu, Gln, Lys, Arg and Leu belong to this category. Although on the basis of the results in Table 4 and those discussed in Section 6.3, Leu is a member of class I, it was earlier (Chakrabarti and Pal, 1998) placed as a separate class as it has, unlike other class I members, only a few points in the g^- state. (II) Short polar/acidic residues Asp and Asn. (III) Aromatic residues His, Phe, Tyr and Trp (however, the last one has values in the region A which

Table 6
Number of different amino acid residues, their occurrences in the three χ_1 states and their distribution among secondary structural elements

	Total				Helix				Sheet				Turn				Rest			
	Number	%t	%g^+	%g^-	Number (%)[a]	%t	%g^+	%g^-	Number (%)[a]	%t	%g^+	%g^-	Number (%)[a]	%t	%g^+	%g^-	Number (%)[a]	%t	%g^+	%g^-
Ala	6703				3424 (51.1)				1183 (17.6)				1025 (15.3)				1071 (16.0)			
Gly	6217				1079 (17.4)				961 (15.5)				2937 (47.2)				1240 (19.9)			
Pro	3502		50.4	49.6	680 (19.4)		68.5	31.5	362 (10.3)		45.3	54.7	969 (27.7)		47.6	52.4	1491 (42.6)		45.3	54.7
Class I	29096	30.5	56.0	13.5	12109(41.6)	34.0	58.2	7.8	6410 (22.0)	38.9	47.9	13.2	5377 (18.5)	17.6	60.5	21.9	5200 (17.9)	25.4	56.5	18.1
Ser	4695	23.6	30.4	46.0	1349 (28.7)	18.9	39.9	41.2	978 (20.8)	34.8	27.2	38.0	1173 (25.0)	11.8	29.2	59.1	1195 (25.5)	31.4	23.3	45.3
Cys	1285	26.0	57.7	16.3	356 (27.7)	22.8	72.2	5.1	385 (30.0)	29.6	55.3	15.1	204 (15.9)	12.7	53.9	33.3	340 (26.5)	33.2	47.6	19.1
Met	1461	28.7	62.6	8.6	669 (45.8)	29.3	67.4	3.3	382 (26.1)	36.1	50.8	13.1	189 (12.9)	19.6	66.7	13.8	221 (15.1)	22.2	65.2	12.7
Glu	4445	32.3	57.3	10.4	2178 (49.0)	34.5	58.6	6.9	735 (16.5)	42.6	44.9	12.5	882 (19.8)	21.3	62.6	16.1	650 (14.6)	28.0	59.8	12.2
Gln	2926	30.1	61.9	8.0	1351 (46.2)	32.6	64.1	3.3	552 (18.9)	38.4	48.7	12.9	535 (18.3)	18.7	69.0	12.3	488 (16.7)	26.4	63.1	10.5
Arg	3491	32.9	57.9	9.3	1456 (41.7)	42.7	52.3	4.9	788 (22.6)	35.0	52.7	12.3	655 (18.8)	20.8	66.9	12.4	592 (17.0)	19.1	68.6	12.3
Lys	4287	34.7	57.9	7.4	1771 (41.3)	43.3	52.5	4.2	837 (19.5)	42.1	50.1	7.9	931 (21.7)	20.6	69.0	10.4	748 (17.4)	23.5	65.8	10.7
Leu	6506	31.8	67.0	1.2	2979 (45.8)	33.6	66.0	0.3	1753 (26.9)	42.8	54.8	2.4	808 (12.4)	15.8	83.4	0.7	966 (14.8)	19.3	78.4	2.4
Class II	8283	30.9	52.4	16.7	2524 (30.5)	16.6	76.8	6.7	1149 (13.9)	47.5	41.3	11.1	2437 (29.4)	19.6	53.8	26.6	2173 (26.2)	51.4	28.3	20.3
Asp	4599	31.5	51.2	17.3	1511 (32.9)	15.2	77.2	7.6	567 (12.3)	52.0	39.3	8.6	1263 (27.5)	17.4	50.7	31.9	1258 (27.4)	56.0	25.9	18.1
Asn	3684	30.1	53.8	16.0	1013 (27.5)	18.7	76.1	5.2	582 (15.8)	43.1	43.3	13.6	1174 (31.9)	22.0	57.2	20.9	915 (24.8)	45.0	31.6	23.4
Class III	8946	34.1	53.6	12.3	3099 (34.6)	50.2	44.4	5.4	2849 (31.8)	25.9	55.6	18.5	1448 (16.2)	19.3	66.0	14.8	1550 (17.3)	30.8	57.0	12.2
His	1772	33.2	54.6	12.2	586 (33.1)	41.0	52.4	6.7	416 (23.5)	38.5	46.2	15.4	404 (22.8)	18.6	67.1	14.4	366 (20.7)	30.9	54.1	15.0
Phe	3057	33.7	55.1	11.2	1064 (34.8)	53.8	42.9	3.4	1070 (35.0)	22.7	57.4	19.9	426 (13.9)	19.0	70.2	10.8	497 (16.3)	26.8	63.6	9.7
Tyr	2907	33.8	53.8	12.3	954 (32.8)	50.6	44.8	4.6	1017 (35.0)	24.1	57.7	18.2	454 (15.6)	20.7	63.7	15.6	482 (16.6)	33.6	54.4	12.0
Trp	1210	37.1	48.0	14.9	495 (40.9)	52.7	37.6	9.7	346 (28.6)	25.7	55.5	18.8	164 (13.6)	17.7	58.5	23.8	205 (16.9)	34.1	52.2	13.7
Class IV	14332	8.4	64.6	27.1	4446 (31.0)	5.6	79.2	15.3	5409 (37.7)	10.1	69.4	20.4	1851 (12.9)	8.6	37.9	53.5	2626 (18.3)	9.3	48.6	42.1
Val	5469	7.9	73.1	19.0	1707 (31.2)	6.1	82.8	11.1	2390 (43.7)	9.1	72.8	18.0	529 (9.7)	9.6	55.0	35.3	843 (15.4)	7.2	65.4	27.4
Ile	4194	9.6	76.6	13.8	1489 (35.5)	8.3	84.2	7.6	1694 (40.4)	9.5	78.6	11.9	392 (9.3)	15.3	54.1	30.6	619 (14.8)	9.2	67.4	23.4
Thr	4669	7.8	43.7	48.4	1250 (26.8)	1.6	68.2	30.2	1325 (28.4)	12.8	51.5	35.7	930 (19.9)	5.3	21.3	73.4	1164 (24.9)	10.9	26.5	62.6
Overall	77079	24.5	56.8	18.7	27361(35.5)	27.7	62.8	9.5	18323(23.8)	26.7	55.9	17.3	16044(20.8)	15.4	55.3	29.3	15351(19.9)	24.2	49.0	26.8

[a] The value within parentheses corresponds to the percentage occurrence of a residue in a given secondary structure. Under Helix all types of helices are considered; all residues with tag E and B (in the DSSP output) are denominated as Sheet; and S and T constitute Turn.

Table 7
Mean ϕ, ψ values (and the associated standard deviations) for different residues in α- and 3_{10}-helices at three χ_1 states[a]

Residue	Number $t/g^+/g^-$	t		g^+		g^-	
		ϕ	ψ	ϕ	ψ	ϕ	ψ
α-helix							
Ala	3096	−64(7)	−39(8)				
Gly	851	−64(9)	−40(12)				
Pro	−/338/141	—	—	−56(5)	−38(7)	−61(5)	−32(7)
Class I	3914/6173/616	−63(7)	−43(7)	−67(10)	−37(10)	−68(12)	−30(11)
Class II	383/1606/90	−64(10)	−42(15)	−67(12)	−37(10)	−66(11)	−35(12)
Class III	1499/1107/83	−62(7)	−46(7)	−74(18)	−32(15)	−70(22)	−26(15)
Class IV	191/3430/497	−66(10)	−32(15)	−64(7)	−44(6)	−80(19)	−25(13)
3_{10}-helix							
Ala	322	−66(15)	−20(27)				
Gly	114	−70(18)	−16(17)				
Pro	−/128/73			−57(7)	−28(19)	−63(8)	−16(37)
Class I	193/844/327	−59(12)	−29(31)	−77(19)	−13(20)	−69(15)	−16(15)
Class II	26/309/78	−79(17)	−11(16)	−93(57)	−9(20)	−77(16)	−5(15)
Class III	53/253/83	−61(13)	−19(45)	−88(20)	−2(20)	−68(14)	−15(25)
Class IV	55/85/178	−61(8)	−26(7)	−59(12)	−30(32)	−85(22)	−3(26)

[a] Residues other than Gly, Ala and Pro have been grouped into classes as specified in Table 5. Values for Gly and Ala are given under t. Statistics restricted to points in the negative ϕ region only. (The numbers of Gly would have been increased by 37 and 74 in two types of helices, if positive ϕ values were included. The other residues, especially in α-helix, are hardly affected by this condition).

Table 8
Mean ϕ, ψ values (and the associated standard deviations) for different residues in a β-sheet at three χ_1 states[a]

Residue	Type	Number $t/g^+/g^-$	t		g^+		g^-	
			ϕ	ψ	ϕ	ψ	ϕ	ψ
Ala	P	285	−120(28)	140(21)				
	A	733	−126(29)	142(18)				
Gly	P	166	−121(33)	154(26)				
	A	359	−136(31)	162(24)				
Pro	P	−/22/30			−67(7)	136(10)	−76(10)	138(16)
	A	−/98/123			−64(8)	141(9)	−73(8)	140(14)
Class I	P	482/536/103	−111(23)	126(16)	−108(19)	134(19)	−133(25)	157(17)
	A	1691/1959/630	−115(27)	129(14)	−113(20)	140(17)	−140(22)	157(13)
Class II	P	103/67/12	−105(25)	121(24)	−101(21)	131(19)	−129(30)	155(17)
	A	345/302/80	−115(29)	121(24)	−105(22)	136(19)	−133(29)	166(24)
Class III	P	189/280/65	−118(23)	126(17)	−110(19)	131(18)	−143(16)	159(9)
	A	452/1057/398	−121(27)	127(15)	−114(18)	140(17)	−148(14)	159(13)
Class IV	P	114/1071/173	−135(18)	144(16)	−112(15)	125(11)	−122(16)	153(14)
	A	364/2194/710	−135(20)	145(15)	−112(18)	127(11)	−125(16)	156(13)

[a] β-sheet type: P = parallel, A = antiparallel. Also, see Table 7 footnotes. (The residue with a considerable number of positive ϕ angles—40 in P and 224 in A—is Gly. Such points are excluded.)

Table 9
Average values of χ_1 at the three conformational states of the side chain with the main chain having a defined conformation

Residue	Residue secondary structure						Main chain conformation[a]					
	α-Helix			β-Sheet			A region			B region		
	t	g^+	g^-	t	g^+	g^-	t	g^+	g^-	t	g^+	g^-
Pro		−27(8)	20(8)		−20(9)	26(8)		−25(9)	26(10)		−23(9)	27(9)
Class I	−176(15)	−70(12)	67(14)	−178(13)	−65(13)	63(12)	−175(15)	−68(13)	64(14)	−177(14)	−65(13)	64(13)
Class II	−172(13)	−72(9)	57(17)	−174(12)	−69(12)	64(9)	−170(14)	−71(11)	61(11)	−173(11)	−69(12)	63(9)
Class III	−180(10)	−71(11)	67(12)	−178(11)	−66(10)	63(9)	−179(11)	−68(12)	63(12)	−177(11)	−66(10)	62(10)
Class IV	−169(16)	−66(7)	63(12)	−176(14)	−61(8)	60(10)	−168(18)	−65(9)	60(10)	−176(15)	−61(8)	60(11)

[a] As defined in Fig. 5a.

are quite similar to the ones in class I). (IV) Residues branched at the C^{β} position: Val, Ile and Thr.

6.2. Comparison of conformations of residues based on ϕ, ψ, χ_1 distribution

Of the three variables ϕ, ψ and χ_1 one has been held fixed to find the dependence between the other two in Figs. 10–12, and the visually identifiable features of these variations have been used to group amino acid residues in Table 4. Due to the interdependence of the three variables a more rigorous method of comparing the conformations of two residues should involve all the three dimensions. Conformational similarity indices, $CS_{XX'}$ (where X and X' are any two residues), were calculated by grid-wise comparison of ϕ, ψ, χ_1 maps (Section 2.4) (Pal and Chakrabarti, 2000c). Essentially, the index is a measure of the common volume in the two maps. When X' is Ala, Gly or Pro (with no or restricted χ_1), the map of any other residue, X, was divided into three ϕ, ψ distributions corresponding to three rotameric states of χ_1 which were individually compared to the X' map; the weighted average (on the basis of the relative population of the residue in the three states) provided $CS_{XX'}$. This method of calculation, using three χ_1-dependent ϕ, ψ maps, takes into account the effect of the side-chain on the ϕ, ψ distribution of X. $CS_{XX'}$ values are presented in Fig. 14 (lower-left triangle).

6.2.1. Comparison of conformations of residues based on ϕ, ψ distribution

Statistical analysis of the protein database has shown that the 20 amino acids are found at the allowed ϕ,ψ regions of the Ramachandran plot with different probabilities (Muñoz and Serrano, 1994; Swindells et al., 1995; Stites and Pranata, 1995). Searching the whole ϕ,ψ space for modelling polypeptide chains can be computationally very expensive. To reduce the problem Kang et al. (1993) have estimated the probabilities of ϕ, ψ angles for each residue from a database of high-resolution structures and have shown that these probabilities can be used to efficiently sample the conformational space of short polypeptides. In a similar exercise, Abagyan and Totrov (1994) and Evans et al. (1995) have calculated dihedral probability zones for different residues. While these procedures make the conformational sampling very efficient, there was no attempt to compare the distributions of individual residues.

SER	CYS	MET	GLU	GLN	LYS	ARG	LEU	ASP	ASN	HIS	PHE	TYR	TRP	VAL	ILE	THR	PRO	ALA	GLY	
1.00	0.89	0.88	0.89	0.89	0.90	0.90	0.87	0.90	0.86	0.89	0.83	0.85	0.85	0.72	0.72	0.86	0.62	0.88	0.60	SER
0.59	1.00	0.90	0.89	0.90	0.92	0.92	0.92	0.89	0.86	0.93	0.93	0.93	0.91	0.87	0.86	0.93	0.43	0.87	0.57	CYS
0.62	0.78	1.00	0.98	0.99	0.98	0.98	0.98	0.94	0.86	0.91	0.90	0.90	0.93	0.85	0.86	0.88	0.37	0.97	0.63	MET
0.65	0.77	0.93	1.00	0.99	0.99	0.98	0.98	0.95	0.86	0.91	0.89	0.88	0.92	0.83	0.85	0.87	0.41	0.99	0.64	GLU
0.62	0.77	0.93	0.94	1.00	0.99	0.99	0.99	0.95	0.87	0.91	0.90	0.90	0.93	0.84	0.87	0.88	0.38	0.98	0.64	GLN
0.62	0.75	0.87	0.93	0.90	1.00	0.99	0.98	0.95	0.89	0.93	0.92	0.92	0.95	0.85	0.87	0.90	0.43	0.98	0.63	LYS
0.62	0.75	0.87	0.92	0.89	0.94	1.00	0.98	0.94	0.87	0.95	0.94	0.93	0.96	0.87	0.89	0.92	0.40	0.98	0.63	ARG
0.59	0.80	0.92	0.94	0.93	0.93	0.91	1.00	0.94	0.86	0.92	0.92	0.91	0.94	0.88	0.91	0.90	0.37	0.97	0.62	LEU
0.56	0.71	0.82	0.81	0.82	0.75	0.75	0.81	1.00	0.94	0.91	0.86	0.85	0.89	0.77	0.80	0.84	0.47	0.94	0.62	ASP
0.51	0.67	0.73	0.72	0.75	0.73	0.73	0.74	0.86	1.00	0.88	0.80	0.81	0.82	0.71	0.73	0.80	0.42	0.84	0.60	ASN
0.56	0.64	0.67	0.73	0.69	0.79	0.81	0.71	0.60	0.64	1.00	0.95	0.94	0.94	0.87	0.88	0.93	0.43	0.90	0.59	HIS
0.51	0.63	0.63	0.68	0.65	0.79	0.80	0.73	0.48	0.53	0.79	1.00	0.98	0.97	0.91	0.91	0.94	0.35	0.89	0.57	PHE
0.51	0.61	0.62	0.67	0.63	0.79	0.79	0.71	0.47	0.53	0.76	0.89	1.00	0.96	0.89	0.88	0.94	0.36	0.88	0.57	TYR
0.54	0.59	0.62	0.69	0.63	0.79	0.78	0.71	0.47	0.49	0.72	0.86	0.83	1.00	0.87	0.89	0.91	0.39	0.93	0.60	TRP
0.47	0.58	0.66	0.59	0.67	0.50	0.52	0.61	0.62	0.50	0.40	0.35	0.32	0.29	1.00	0.99	0.91	0.23	0.81	0.50	VAL
0.47	0.57	0.67	0.59	0.69	0.50	0.51	0.62	0.61	0.48	0.37	0.33	0.29	0.27	0.96	1.00	0.90	0.24	0.85	0.52	ILE
0.59	0.45	0.50	0.45	0.50	0.38	0.38	0.44	0.46	0.35	0.30	0.27	0.26	0.25	0.74	0.78	1.00	0.35	0.85	0.56	THR
0.53	0.36	0.34	0.39	0.36	0.40	0.39	0.36	0.38	0.35	0.36	0.30	0.29	0.35	0.20	0.21	0.28	1.00	0.44	0.29	PRO
0.75	0.72	0.88	0.91	0.92	0.89	0.89	0.91	0.74	0.70	0.75	0.70	0.67	0.76	0.67	0.70	0.56	0.44	1.00	0.64	ALA
0.51	0.48	0.57	0.59	0.59	0.58	0.58	0.59	0.49	0.50	0.50	0.45	0.44	0.50	0.41	0.44	0.37	0.29	0.64	1.00	GLY

Fig. 14. Matrix of conformational similarity indices relating different residues. The lower triangle is based on ϕ, ψ, χ_1 distribution ($CS_{XX'}$ values), whereas the upper triangle is based on ϕ, ψ distribution ($CS_{XX'/2D}$).

By way of developing scoring matrices for protein sequence alignment there have been attempts to characterize the conformational features of residues using ϕ, ψ distributions. Niefind and Schomburg (1991) developed the structure-derived correlation coefficients (SCCs) using three steps. (i) The ϕ, ψ graph of each residue was normalized by dividing the population of $12^\circ \times 12^\circ$ grids by the overall frequency. (ii) The average ϕ, ψ distribution was calculated by pointwise addition (with weights of 1/20) of the 20 normalized amino acid distributions. For each amino acid the mean value was subtracted from the amino acid specific value in the corresponding grid. (iii) Finally, a correlation coefficient for each couple of amino acid residues was calculated from the two different plots. It has been observed that the use of the difference rather than the direct ϕ, ψ distribution enhances the individuality of a given residue in the SCC value. In their approach, Kolaskar and Kulkarni-Kale (1992) calculated ϕ, ψ probability maps by finding the percentage-occurrence (P_{ij}) of each residue in each (i, j) of the $20^\circ \times 20^\circ$ grids in the ϕ, ψ plane; the sum of absolute differences over all the grids between any two residues (A and B) provided their similarity index, ΔP_{AB}.

6.2.2. *Usefulness of the χ_1 dimension in discriminating residue conformations*

As SCC and ΔP_{AB} are based on ϕ, ψ distribution only and $CS_{XX'}$ values are calculated by incorporating χ_1 also, it is relevant to ask to what extent it is necessary to advocate the effect of the side chain torsion angles on the main chain conformations in the calculation of the similarity indices. To answer this $CS_{XX'/2D}$ parameters were calculated by comparing only the two-dimensional ϕ, ψ maps for each pair of residues and the values are given in the upper right triangle of Fig. 14. The correlation coefficient between $CS_{XX'}$ and $CS_{XX'/2D}$ is 0.70 (Table 10), showing that

Table 10
Comparison (using correlation coefficients) between different conformational similarity indices[a]

	$CS_{XX'/2D}$	SCC	ΔP_{AB}
$CS_{XX'}$	0.70	0.27	−0.56
$CS_{XX'/2D}$		0.25	−0.93
SCC			−0.32

[a] SCC values are from Niefind and Schomburg (1991) and ΔP_{AB} from Kolaskar and Kulkarni-Kale (1992).

χ_1 has its signature on $CS_{XX'}$ making them considerably different from the other set of values. Additionally, CS_{AX} (X' being Ala) correlate better with α-helix propensities (Section 14.1.1) than $CS_{AX/2D}$. Another parameter, D_V (based on ϕ, ψ, χ_1 distribution) correlates better with β-sheet propensities than the corresponding parameter in two-dimension, $D_{A/2D}$ (Section 14.2.1). All these vindicate the inclusion of χ_1 in the derivation of parameters characterizing the conformation of individual residues. Of the three parameters calculated on the basis of ϕ, ψ distributions, $CS_{XX''/2D}$ and ΔP_{AB}, which are highly similar, show very little correlation with SCC.

6.3. $CS_{XX'}$ and residue classification

$CS_{XX'}$ values can range between 0 and 1, corresponding to no and complete similarity, respectively. Subtracting these from 1 we get conformational distances between amino acid residues (0 closest, 1 farthest). A complete-linkage cluster analysis using these distances provide a pictorial representation of the residue clusters based on ϕ, ψ, χ_1 distribution (Fig. 15) (Pal and Chakrabarti, 2000c). The clusters essentially reproduce the classes of residues that were delineated based on two-dimensional ψ, χ_1 and ϕ, χ_1 plots (Section 6.1), and additionally indicate how individual residues differ within a class. Ser which was found to be a constituent of class I is now shown to have a distribution of torsion angles fairly distinct from the other members, which must have been caused by its ability to form short-range hydrogen-bonded contact due to the presence of a hydroxyl group rather close to the main-chain atoms. Chakrabarti and Pal (1998) originally put Leu as a separate class, in spite of it being very similar to class I members, as it did not have significant presence in the g^- state of the side-chain. Based on $CS_{XX'}$, Leu can be merged with class I. Additionally, residues (with no or restricted χ_1, Ala, Gly and Pro), which were earlier left out, are also placed relative to other residues. It is interesting to note that Ala can indeed be placed along with other class I members; the relevance of this is discussed in Section 14.1.1.

6.4. Similarity indices and sequence comparison

Specifying an appropriate amino acid substitution matrix is central to protein comparison methods and much effort has been devoted to defining, analysing and refining such matrices (see Altschul, 1991). Tomii and Kanehisa (1996) have collected 42 published similarity (or mutational) matrices derived using different physicochemical and biochemical properties of amino acids and which have been used for protein sequence alignments and similarity searches. On pairwise comparison (based on correlation coefficient), the matrix of $CS_{XX'}$ elements is found to be quite

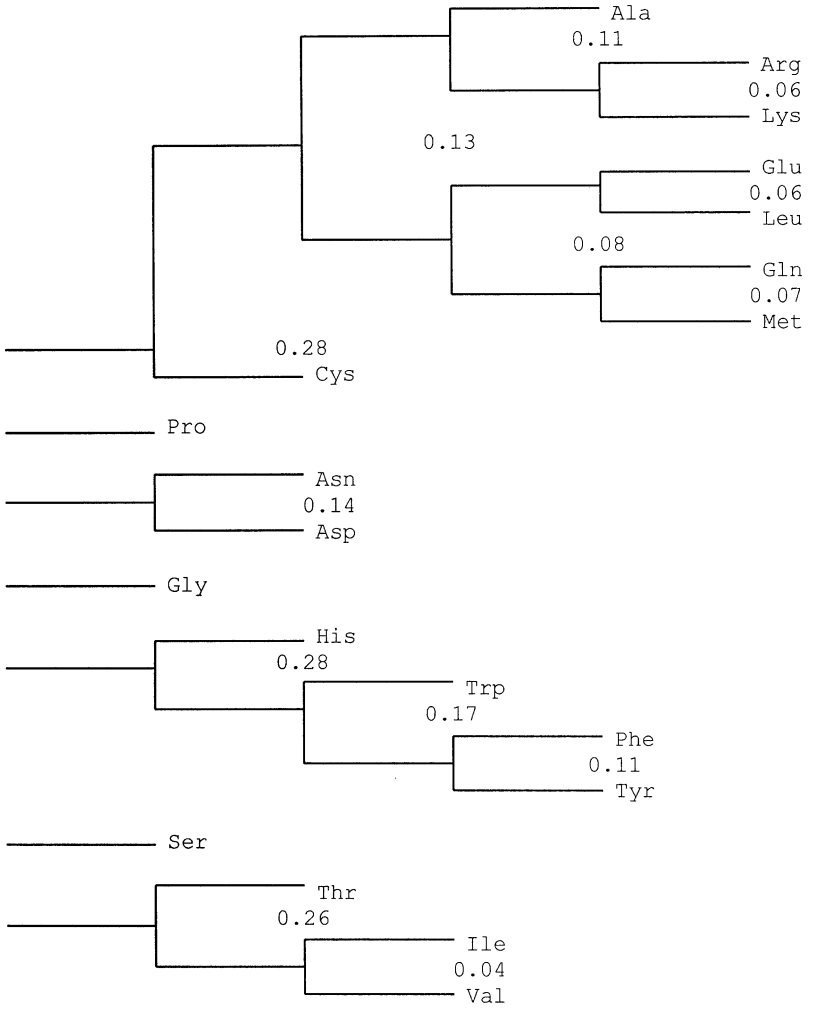

Fig. 15. Minimum spanning tree obtained for $(1-CS_{XX'})$ values using complete-linkage cluster analysis with a threshold distance of 0.30. The distance between two residues or the maximum of all the distances between two clusters is indicated when they are below the threshold.

different from others, the closest resemblance being with the conformational similarity weight matrix (Kolaskar and Kulkarni-Kale, 1992) derived from ΔP_{AB} values discussed in Section 6.2.1. But even with this the correlation coefficient is quite low (0.42), suggesting that the similarity indices based on ϕ, ψ, χ_1 distribution have features not incorporated in other commonly used matrices. Interestingly, however, residues in many of the conformationally similar clusters are found to be highly exchangeable within evolutionarily related proteins. For example, by analysing the replacement pattern between amino acids in structurally similar proteins Risler et al. (1988) delineated four strong clusters: (i) Ile and Val, (ii) Leu and Met, (iii) Lys, Arg and Gln, and (iv) Tyr and Phe. These residues are also shown to be conformationally similar (Fig. 15), thus

suggesting that during evolution the substitution of residues is strongly dictated by conformational consideration, such that amino acids are preferentially replaced by others having similar local folding requirements. It has been shown by Risler et al. (1988) that the exchangeability of many residues cannot be described simply in terms of their chemical properties. Although their charges are opposite, Arg and Glu were found to have a short distance between them and also belong to the same class (Table 5). However, the chemical properties are important when considering the functional aspects. For example, Cys in disulphide bonds or His in active sites cannot be replaced by other residues, and this can be clearly seen in the Euclidean representation of the distance matrix of Risler et al. (1988). All conformational similarity based matrices would be impervious to such effects. However, as they contain information on the similarity of local folding requirements, their use in the alignment of two sequences can identify conformationally similar protein fragments (Kolaskar and Kulkarni-Kale, 1992). It has been argued that although the most widely used scoring matrix, Dayhoff's PAM matrix (Dayhoff et al., 1978, 1983), can reveal phylogenetic relationships, similarity matrices that reflect structural aspects of the amino acids might provide more information for protein structure prediction by homology, protein design (by introducing substitutions that need not necessarily be frequent in nature) and modelling (Niefind and Schomburg, 1991).

6.5. Minimum number of residues required for protein folding

The aim of Sections 6.1–6.4 has been to simplify protein sequences so as to treat some residues as equivalent and group 20 amino acids in a small number of classes based on the general similarity of local folding at specific amino acid residues given by the ϕ, ψ, χ_1 distributions. Besides Ala, Gly and Pro there are four classes of residues, making a total of 7 (Table 5). In a more rigorous classification Ser can be taken out of class I to constitute a separate class, because of the short-range hydrogen bonding capability of its side chain. Also, in Fig. 15 Ala is shown to be quite similar to class I, but because of its small size (no χ_1) its separate identity can be retained.

The prospect of achieving protein-like properties using an alphabet with lesser than 20 amino acids has been an attractive proposition for structural biochemists (Wolynes, 1997; Wang and Wang, 1999). Motivated by the experimental finding of Riddle et al. (1997) that a small β-sheet protein, the SH3 domain, can be largely encoded by a five letter amino acid alphabet, Ile, Ala, Gly, Glu and Lys (IAGEK), Wang and Wang (1999) proposed a theoretical procedure for grouping residues based on a 'minimal mismatch' principle which ensures that all interactions between amino acids belonging to any two given groups are as similar to one another as possible. It was found that sequences with 5 types of residues can indeed have protein-like properties, folding into unique native structures in a reasonable amount of time. One of the optimally reduced sets was the same (IAGEK) as determined by Riddle et al. (1997), two others being IAGDK and IAPDK. Although these groupings are such that they cover the entire hydrophobicity spectrum (Chan, 1999) it is interesting to note that except the first set, which has Glu and Lys from the same class I (Table 5), each letter belongs to a separate class. Consequently, it can be suggested that picking a representative from each class in Table 5 would give a reduced alphabet that can meet all the local folding requirements of protein structures (although the size of the alphabet may need to be expanded to meet the diversity required for tertiary interaction and especially, function).

7. *Cis* peptide bonds

Due to the partial double bond character and the consequent restriction of rotation about it, the peptide bond linking two residues [(1) and (1′) in Fig. 16a] can exist in two conformations—*trans* in which the dihedral angle ω, [$C^{\alpha}(1)-C(1)-N(1')-C^{\alpha}(1')$], is close to 180° and *cis* in which ω is around 0°. The *cis* and *trans* isomers of *X*–Pro peptide bonds (where *X* is any residue) differ in free energy by only 0.5 kcal/mol (Maigret et al., 1970), because of very similar

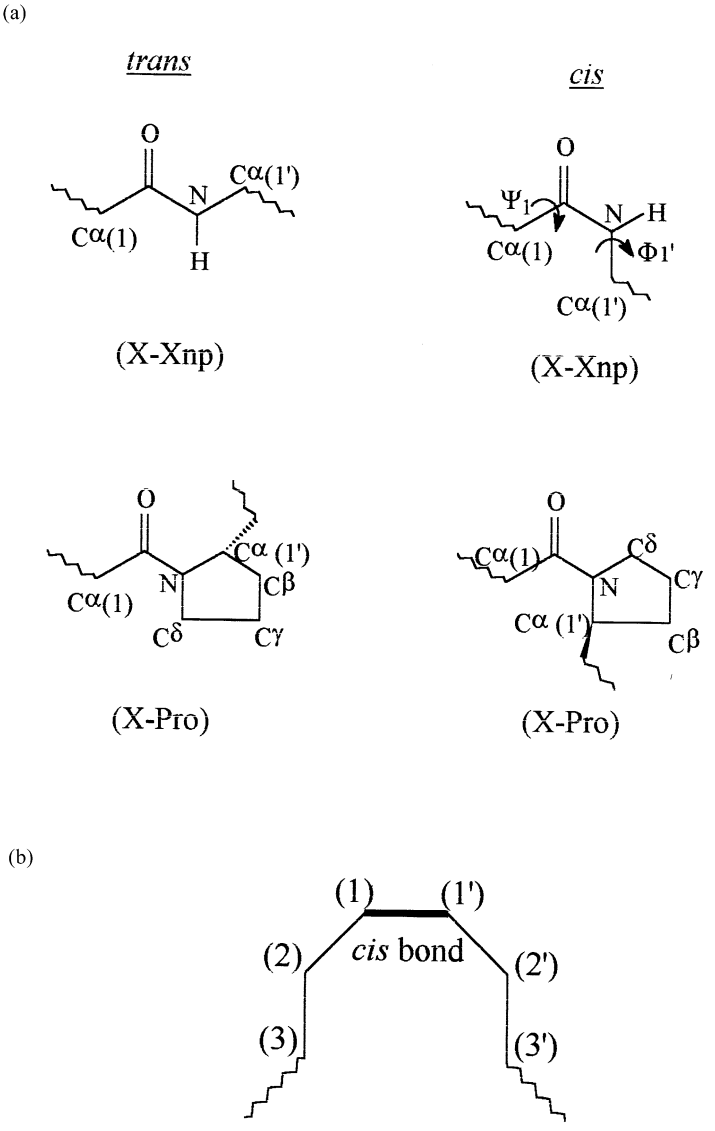

Fig. 16. Schematic representation of *cis* and *trans* conformations around $X-X_{np}$ and $X-$Pro peptide bonds (where X = any residue, X_{np} = any non-Pro residue). (b) Convention for numbering residues flanking a *cis* peptide bond.

steric clashes between $C^{\alpha}(1)$ and $C^{\alpha}(1')$ in the *cis* conformation and between $C^{\alpha}(1)$ and $C^{\delta}(1')$ in the *trans* conformation. Accordingly, 5–6.5% of *X*–Pro peptide bonds are found to be *cis* in different statistical analyses (Stewart et al., 1990; MacArthur and Thornton, 1991; Reimer et al., 1998; Pal and Chakrabarti, 1999b). In contrast, the *cis* and *trans* isomers of nonprolyl peptide bonds differ in free energy by approximately 2.5 kcal/mol (Radzicka et al., 1988). The greater difference arises from the absence of a steric clash in the *trans* conformation. Only 0.03–0.05% nonprolyl bonds are *cis* in known protein structures (Stewart et al., 1990; Jabs et al., 1999), although depending on the residues involved a much higher percentage has been observed (Pal and Chakrabarti;, 1999b), and many *cis* peptide bonds might have gone unrecognized due to the limited resolution of the data and the refinement protocol used (Weiss et al., 1998). Ramachandran and Mitra (1976) used conformational energy calculations of tripeptide units to derive expected frequencies of 0.1% and 30% for an Ala–Ala and Ala–Pro peptide bond, respectively. Using dimensions of the peptide units (Ramachandran and Venkatachalam, 1968; Engh and Huber, 1991), it is possible to identify some of the incorrectly assigned *trans* peptide bonds (Weiss and Hilgenfeld, 1999).

The occurrence of non-Pro *cis* peptide bonds has been associated with steric strain in proteins (Herzberg and Moult, 1991) similar to the occurrence of residues with unfavourable ϕ, ψ angles, and it has been speculated that the location of these *cis* peptide bonds is often a peculiar one with respect to the function of the molecule (Weiss et al., 1998; Stoddard and Pietrokovski, 1998). *Cis–trans* isomerizations in peptides and proteins are characterized by relaxation times from ten to hundreds of seconds at 25°C (Grathwohl and Wütrich, 1981; Schmid et al., 1993). Fifty per cent of all well-defined protein structures contain at least one *cis* peptide bond (Pal and Chakrabarti, 1999b). A *trans* → *cis* isomerization of the concerned bonds is needed to achieve the native state of the protein and has often been found as the rate limiting step in in vitro protein folding (Brandts and Lin, 1986; Kim and Baldwin, 1990). Numerous enzymes termed peptidyl–prolyl *cis–trans* isomerases have been found to accelerate this interconversion powerfully (Schmid et al., 1993).

7.1. Residues involved

The intrinsic probability of a residue (*X*) to cause a *cis* conformation of the *X*–Pro linkage, given by the fraction of occurrence of the bond in the *cis* form, is provided in Fig. 17 (Pal and Chakrabarti, 1999b). Residues with high frequency in the *cis* form are (i) Pro, (ii) aromatic residues, (iii) small residues, Gly and Ala, and (iv) polar residues Ser, Gln and Arg. Branched aliphatic residues Val, Ile, Thr and Leu are less frequent. That the prolyl bond conformation is mainly determined by local effects is indicated by the rough correlation between the *cis* content in the pentapeptide series, acetyl–Ala–X–Pro–Ala–Lys–carboxamide and the propensity of *X*–Pro *cis* prolyl bonds in proteins (Reimer et al., 1998). The favourable interaction taking place between the aromatic ring and the proline residue in the *cis* conformer has been shown in a set of tetrapeptides of general sequence acetyl–Gly–*X*–Pro–Gly–carboxamide (Wu and Raleigh, 1998).

7.2. Neighbouring residues

For the prediction of prolyl residues in *cis* conformation it is necessary to analyse the local amino acid sequence (Frömmel and Preissner, 1990). However, as will be discussed in Section 7.4, in spite of being constrained, a *cis* peptide can mediate in a variety of reverse turns, so that the

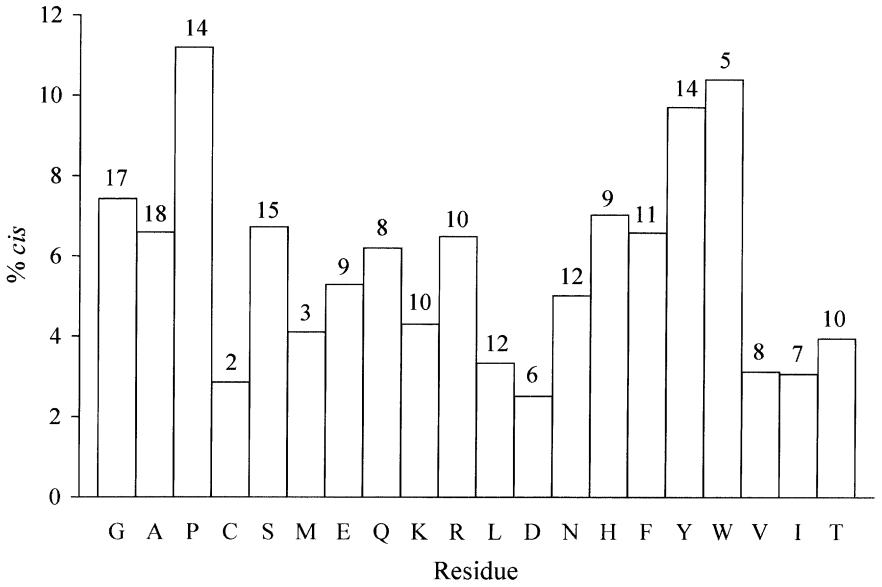

Fig. 17. Histogram showing the percentage of occurrence of various residues in the *cis* conformation of the X–Pro peptide bond; the numbers of *cis* cases are given on top of each bar (Pal and Chakrabarti, 1999b).

sequence-based modelling is not just a simple question of choosing between the *cis* and *trans* forms, but also what local conformation a given sequence can adopt. Of all the structure-based sequence preferences (Pal and Chakrabarti, 1999b) the strongest one is the relative presence of aromatic and β-branched (classes III and IV) residues (Fig. 18). Aromatic residues have high occurrences at positions (1) and (2′) which decrease sharply on moving outward. On the contrary, the β-branched residues are less at (1) and (2′) (especially in the former position, which is also indicated in Fig. 17), and increase along the outward locations (especially upstream). Although Fig. 18 is based on X–Pro residues in VIa turn type, VIb turns as well as X_{np}–X_{np} (X_{np} means any residue other than proline) cases also show similar position-specific variations. The importance of aromatic residues on either side of Pro which can provide C–H groups to have C–H$\cdots\pi$ interactions with the flanking π systems (Pal and Chakrabarti, 1999b) is exemplified by the occurrence of a high population of the *cis* isomeric form in solution of the polypeptide, Ser–Tyr–Pro–Tyr–Asp–Val (Yao et al., 1994). Small residues have relatively higher occurrences in all the positions around X_{np}–X_{np} *cis* bonds, and it has been suggested that the presence of these bonds may be dictated to a greater extent by the secondary structure around them, whereas X–Pro *cis* peptides are controlled more by surrounding residues (Pal and Chakrabarti, 1999b).

7.3. Variation of ϕ and ψ, with χ_1 of residues involved in cis peptide bonds

Comparison of Fig. 19a with the general distribution (Fig. 10) shows that the residue preceding *cis* peptide unit can occupy only the B region with a tight clustering (even in the g^+ state). However, the trend in the shift of ψ towards a more extended value, as χ is changed from t to g^+

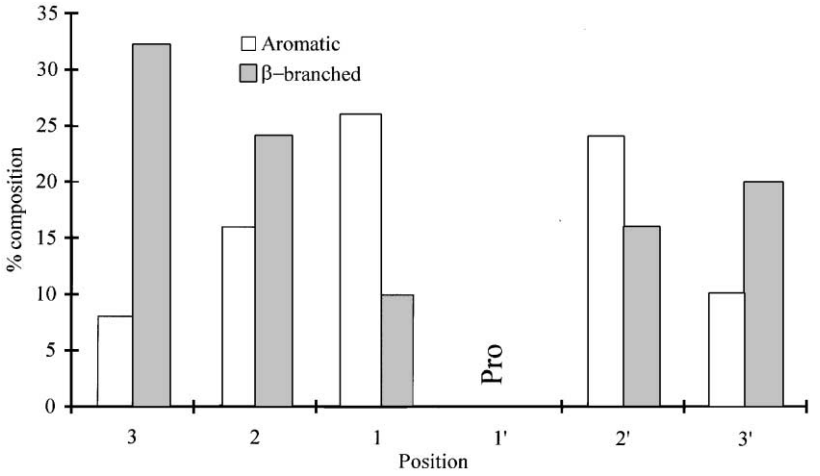

Fig. 18. Histogram showing the variation of percentage composition of aromatic and β-branched residues at each position in the neighbourhood (Fig. 16b) of $X-$Pro *cis* units in type VIa β-turns (based on data in Pal and Chakrabarti (1999b)).

to g^-, is retained. The ϕ, χ_1 plots (Figs. 19b and 11) are similar, except that points move towards a more negative value of ϕ in the *cis* form.

Pro in *cis* X–Pro has a noteworthy dependence of χ_1 on ϕ and ψ (Figs. 19c and d). Residues predominantly have a positive χ_1 (positive : negative $\approx 6:1$). Notably, however, when ψ is less than 60°, a positive value of χ_1 is the norm, and only when ψ is $\sim 120°$ or more a few points are also observed in the negative range of χ_1. Starting at $-60°$ the ϕ values go up to $-80°$ when χ_1 is negative, whereas for positive χ_1 it can extend up to $-110°$.

Although the residue X in both X–Pro and X–X_{np} peptide units has similar conformational features, those for Pro and X_{np} are considerably different. Apart from the obvious difference in the χ_1 angles, which are restricted in the range -30 to $+40°$ for Pro, whereas the non-Pro residues have three conformational states, the ϕ values of the latter, as compared to the former, are shifted towards more negative region (Fig. 19d). Without the constraint on ϕ imposed by the pyrrolidine ring, non-Pro residues, by taking a more extended value of ϕ reduce the steric clash between C^α of (1) and the carbonyl group of (1′) (Fig. 20). Interestingly, irrespective of whether it is a Pro or a non-Pro residue, there is a near linear relationship between $\chi_{1'}$ and $\phi_{1'}$ in the g^- state (Pal and Chakrabarti, 1999b).

7.4. Dependence of the turn geometry on the residues involved

A *cis* peptide bond can cause reversal of chain direction (Lewis et al., 1973) leading to two types of turns, VIa and VIb, with the two residues (1) and (1′) occurring in regions B and A, respectively (Fig. 5a), in the former, and both occupying the region B in the latter (Richardson, 1981; Rose et al., 1985). These two types have been further subdivided depending on the presence or absence of hydrogen bond and residues like Gly (Fig. 21) (Pal and Chakrabarti, 1999b). For most of the residues with the BA conformation when ϕ_1 is greater than $\sim -90°$ there is a hydrogen bond between residues (2) and (2′) (sometimes between (2) and (3′)), which is absent when ϕ is decreased

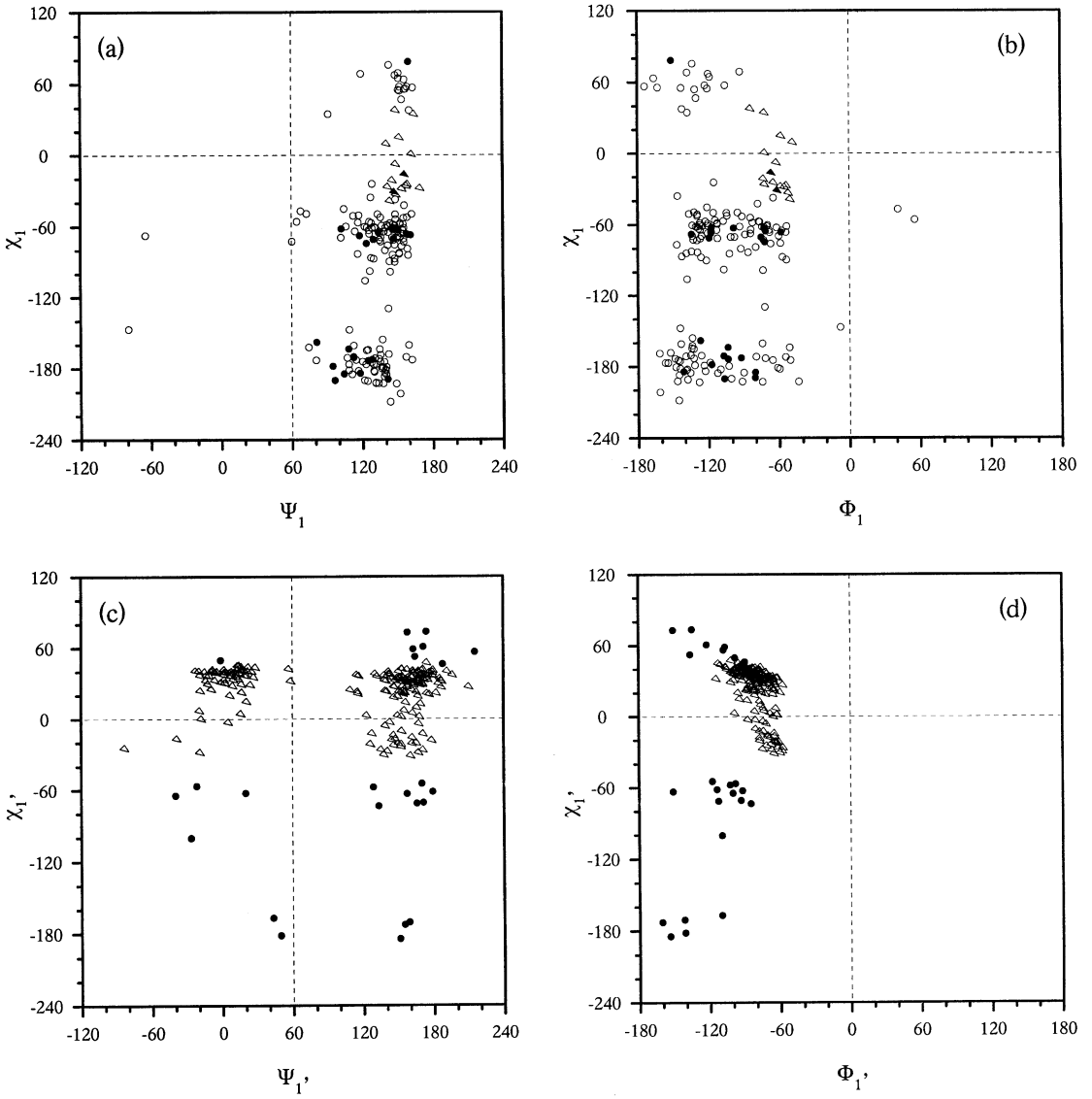

Fig. 19. Joint distributions of χ_1 with ϕ and ψ for residues at positions (1) and (1′). Symbols used: (Δ) Pro, (O) non-Pro, and these are open for $X-$Pro and filled for $X-X_{np}$ cases (Pal and Chakrabarti, 1999b).

below -90°. Consequently, type VIa turn type can be broken into two groups, VIa-1 and VIa-2, the former with hydrogen bonding and the latter without (Fig. 22), and their average ϕ, ψ values are listed in Table 11. Similarly, the VIb turns (with the central residues in the extended conformation (B)) have been subdivided into the predominant VIb-1 type and a minor VIb-2 type, the former with no hydrogen bond and the latter with ψ_1 below 100° and hydrogen bonding between residues (3) and (3′) or (2) and (3′). As Gly residues at position (1) stand out from the rest in having ψ close to 180° (Fig. 21b), these were grouped into a separate class of turn, type VIb-3.

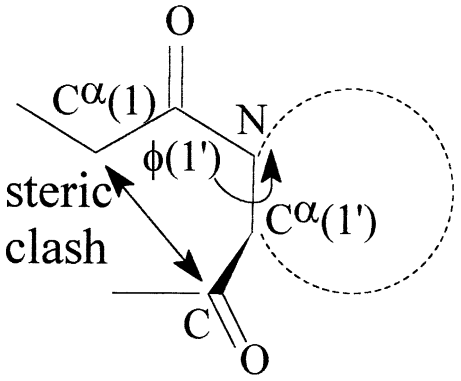

Fig. 20. The steric clash between residues linked by *cis* unit is released by $\phi(1')$ taking up a more extended conformation if the constraint of a pyrrolidine ring (shown schematically with dashed line) is removed by changing the residue at (1′) by a non-Pro residue.

Beyond these categories, residue (1) in the remaining cases is mostly Gly in region R (Table 3), while (1′) occurs in the region A or B (Fig. 21c), and these constitute two other turn types, VIc and VId.

As discussed in Section 7.3 and also seen in Fig. 21b, the non-Pro residues at (1′) have a more negative ϕ value than prolines. Hence, the average values (Table 11) for X_{np}–X_{np} cases, as also for Pro–Pro cases (which have a more restricted conformational parameters), are different from the X–Pro cases. Due to the more extended nature of ϕ, the turn opens up in X_{np}–X_{np} cases (in Fig. 22, compare (c) and (h), both having the same turn type, but different sequences), which thus have a longer (2)−(2′) distance (between C^{α} atoms) than what is observed in the corresponding X−Pro motif.

8. Pyrrolidine ring puckering

Unsaturated five-membered rings can, in principle, adopt a continuum of possible puckered conformations (Altona and Sundaralingam, 1972; Cremer and Pople, 1975; DeTar and Luthra, 1977) that may be described using the concept of pseudorotation. The pyrrolidine ring of Pro residues has two preferred conformations that are energetically equivalent (Ramachandran et al., 1970; Balasubramanian et al., 1971; Ashida and Kakudo, 1974; Momany et al., 1975). The atoms N, C^{α}, C^{β} and C^{δ} are essentially planar, and the two conformations are distinguished by the direction in which the C^{γ} atom protrudes from this plane. In the UP (or A or C^{γ}-exo) conformation C^{γ} and the carbonyl group are on the opposite sides relative to the plane, whereas they are on the same side in the DOWN (or B or C^{γ}-endo) conformation (Fig. 23). The UP conformation is characterized by negative χ_1 and χ_3, and positive χ_2 and χ_4, the signs are reversed for the DOWN conformation.

Though some rotamer libraries consider it possible for Pro ring to occur in the planar conformation also (Ponder and Richards, 1987; Tuffery et al., 1997) Lovell et al. (2000) treat Pro as having only two acceptable puckers. This is based on results from small-molecule structures

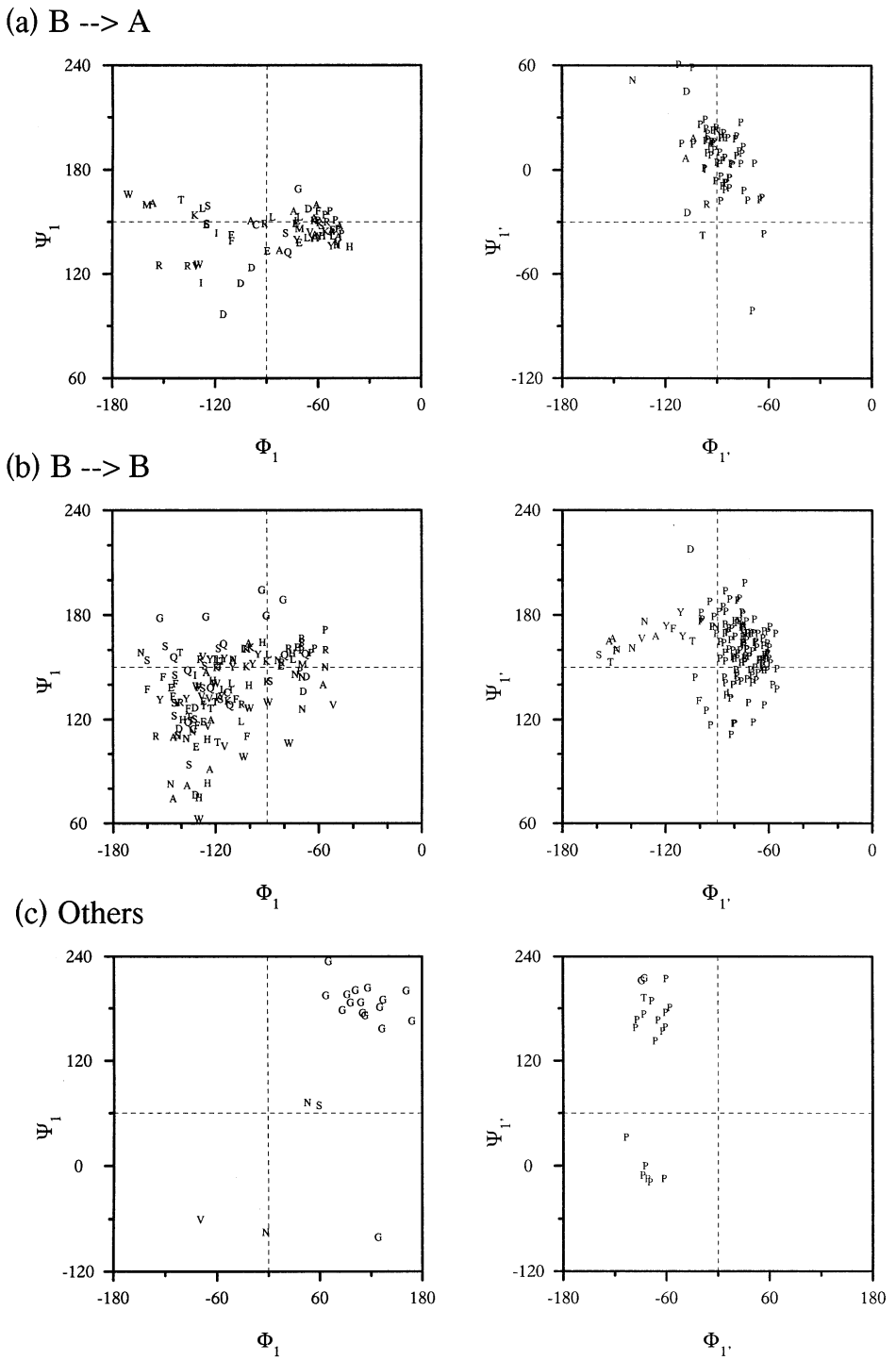

Fig. 21. ϕ, ψ plots for residue pairs at positions (1) and (1′) (each point is indicated by the one-letter amino acid code of the corresponding residue). (a) The first residue is in the region B and the second in A; (b) both are in the region B; and (c) the rest (except 5 cases, the first residue is in R, whereas the second is either in A or B region). Only the specified regions (Fig. 5a) are shown in (a) and (b) (Pal and Chakrabarti, 1999b).

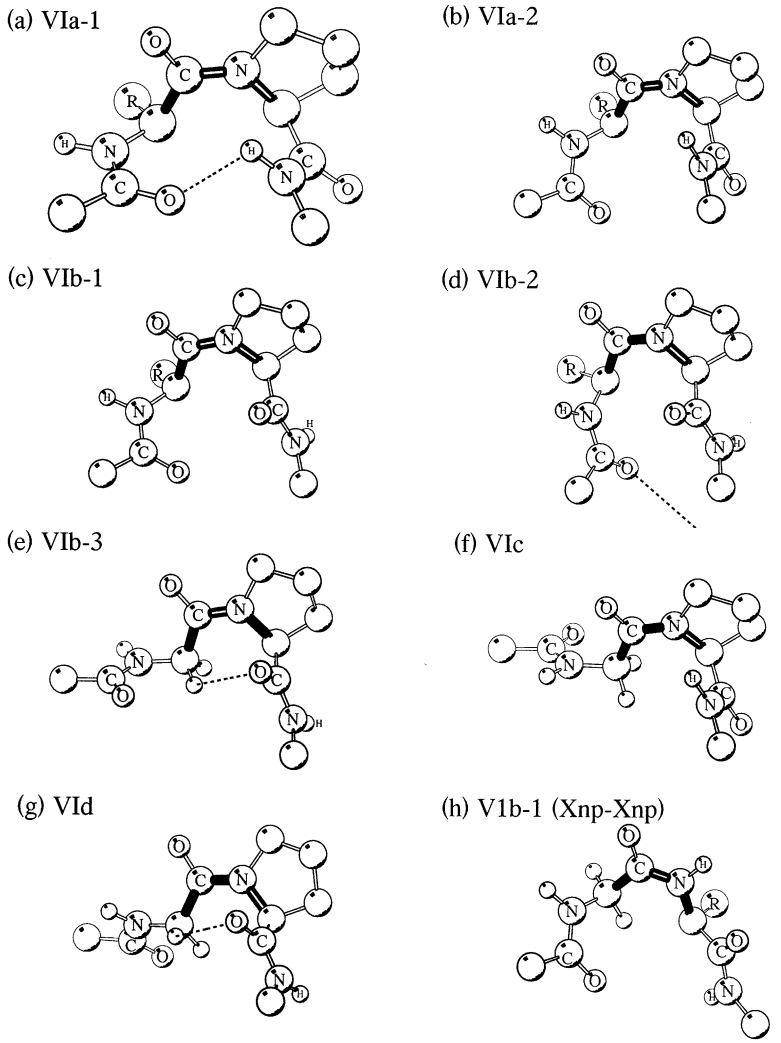

Fig. 22. Molecular representations of the different classes of β-turns around *cis* $X_{np}-$Pro bond (a)–(g), and (h) one case of $X_{np}-X_{np}$ bond. The *cis* peptide is shown in thick lines; hydrogen bond (including C–H$\cdots$O hydrogen bond), if present, is presented as broken lines (in (d) the bond involves the NH group at position (3′), which is not shown) (Pal and Chakrabarti, 1999b). For ease of comparison, two overlaid diagrams (in stereo) are provided, where the superimposition has been done using the five atoms of the *cis* peptide fragment: (i) combining (a) and (b) (the former drawn with thicker sticks), and (j) combining (c) and (h).

(Némethy et al., 1992) and the observation that long-range clashes are substantially decreased by substituting either the UP or DOWN pucker for the planar state (Word et al., 1999). They further suggest that the planar electron density for Pro rings is probably caused by averaging between the two pucker states and is better modelled as two alternate conformations. While atomic resolution protein structures do indicate that about 18% of residues can exist in two distinct conformations, very flat or 'not very puckered' rings are also seen (Wilson et al., 1998).

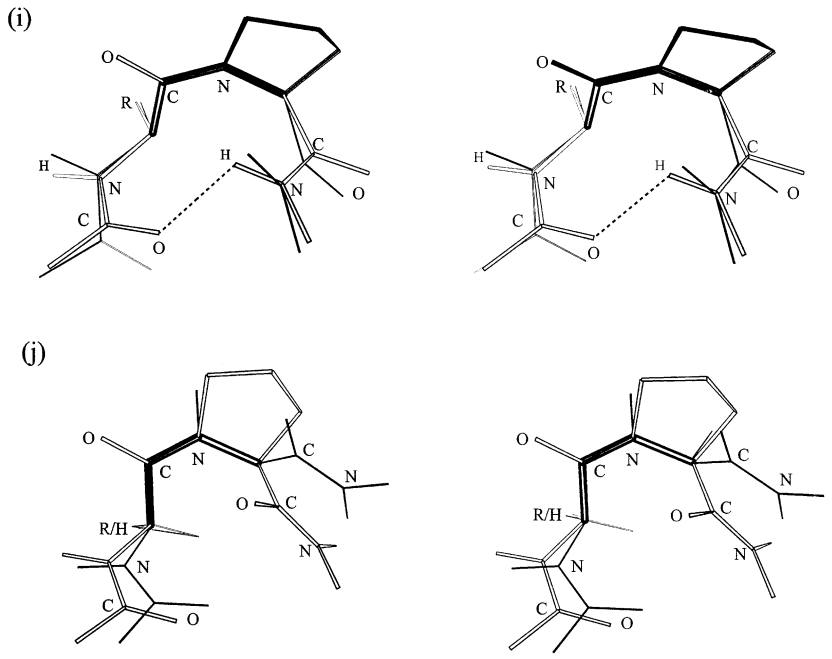

Fig. 22. (*Continued*).

Based on a systematic analysis of 68 proteins, Milner-White et al. (1992) found that over a half of all *cis* and *trans* proline residues could be unambiguously assigned to the UP or DOWN ring puckering. Considering the cases where the average of the absolute value of a side-chain torsion angle is greater than 10°, the average values of χ_1, χ_2, χ_3 and χ_4 are as follows: for *trans* Pro residues, UP: −21, 29, −31, 16 and DOWN: 22, −30, 25, −11 [the two sets of values as obtained from small-molecule structures (Némethy et al., 1992) are −28, 39, −35, 18 and 27, −36, 29, −12, respectively, and individual structures vary by less than $\pm 10^\circ$ from these values]; for *cis* Pro residues, DOWN: 30, −36, 24, −8. 89% of the *cis* proline residues exhibit the DOWN pucker, while the *trans* proline residues, on average, are about evenly distributed between the two forms. However, when located in α-helices, 79% of *trans* Pro residues (69% according to Table 6) are found to have the UP ring pucker. Though there have been theoretical studies (for example, Kang et al., 1999) on the puckering of prolyl ring, simple conformational analysis, discussed below, can be quite illuminating on this issue.

8.1. Differences in the variation of ϕ and ψ with χ_1 for trans and cis proline residues

As the form of the puckering is based on the sign of χ_1 angle, it is important to compare the variation of ϕ and ψ with χ_1 in *cis* Pro residues (Fig. 19) to what is found in *trans* residues (Fig. 24). *Cis* Pro residues with ψ in the A region are almost non-existent when χ_1 is negative and only a few are found in the B region. In contrast, there is not much difference in the density of cluster pairs (with opposite signs of χ_1) in the A or B region of ψ for *trans* Pro residues. However,

Table 11
Types of turns mediated by *cis* peptide bonds and their geometries (representative diagrams are given in Fig. 22)

Turn type[a]	Conf.[b]	No.	ϕ_1	ψ_1	$\phi_{1'}$	$\psi_{1'}$	Distance[c] (Å) (2)–(2′)
$X_{np}-P$							
VIa-1	BA	39	−74(24)	141(9)	−93(9)	12(16)	5.9(6)
VIa-2	BA	13	−131(24)	145(16)	−79(9)	−16(24)	6(1)
VIb-1	BB	100	−117(26)	138(16)	−77(10)	158(17)	6.3(8)
VIb-2	BB	12	−134(12)	98(23)	−78(12)	165(9)	4.5(7)
VIb-3	BB	4	−100(20)	183(8)	−72(10)	154(2)	7.7(2)
VIc	RA	5	104(38)	188(8)	−83(9)	−16(7)	8.4(4)
VId	RB	7	102(20)	186(25)	−69(8)	171(23)	8.3(3)
$P-P$							
VIa-1	BA	7	−54(5)	147(5)	−81(5)	9(10)	5.6(3)
VIb-1	BB	6	−69(6)	160(8)	−77(11)	149(14)	7.4(7)
VIb-2	BB	1	−84	149	−96	115	6.3
$X_{np}-X_{np}$							
VIa-1	BA	5	−89(21)	134(30)	−111(17)	14(36)	6.4(9)
VIa-2	BA	3	−113(41)	149(9)	−106(7)	−15(17)	7(1)
VIb-1	BB	15	−108(29)	121(23)	−134(21)	168(15)	8(1)
VIb-2	BB	2	−123(6)	121(57)	−102(23)	152(26)	6(1)
VIb-3	BB	1	−155	176	−102	129	8.6
VId	RB	3	131(30)	174(11)	−91(2)	202(13)	9.1(6)

[a] VIb-3, VIc and VId turns have Gly at position (1).
[b] Conformation based on the location of the two residues in the regions defined in Table 3.
[c] The distance between C^{α} atoms at indicated locations corresponding to Fig. 16b.

all the points in the narrow strip ($\psi \approx 10$–120°) linking A and B regions in Fig. 8 have positive χ_1 angles (Fig. 24) (and about a quarter of such prolines have a preceding Leu or Val—a percentage significantly larger than the expected value from Table 2). In the context of puckering the most crucial difference is the displacement to a more negative ϕ value in the *cis* proline residues compared to *trans* (Table 12)—this is to reduce the steric clash between the C^{α} group of the preceding residue and the carbonyl group of Pro (Fig. 20) (MacArthur and Thornton, 1991; Pal and Chakrabarti, 1999b). It can additionally be noted that for both *cis* and *trans* prolines a value of χ_1 close to 0° (when the ring is almost planar and consequently strained) is less likely (Figs. 19d and 24), as already discussed.

8.1.1. Variation of ϕ and ψ with χ_1 for proline residues in small peptides

As ϕ and χ_1 torsion angles are about two adjacent bonds in the pyrrolidine ring (Fig. 25) they are correlated. To have a clear perspective, the angle, ϕ_R (which is defined using the atoms within the ring, as is the case with all χs for Pro), is calculated for all accurately determined small molecule structures, plotted against the conventional ϕ (Fig. 26a), and the two parameters are correlated. As in general, the ring torsion angles are alternatively positive and negative, when χ_1 is negative, ϕ_R is in the range -10° to 20° (mostly positive) and when χ_1 is positive, ϕ_R is within -30° to 5° (mostly negative) and two lines can be fitted through the points (data not shown). This

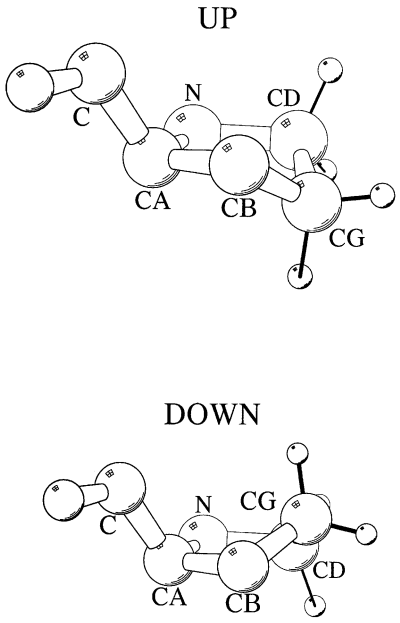

Fig. 23. UP (or A or C^{γ}-exo) and DOWN (or B or C^{γ}-endo) conformations representing the two limiting cases of ring puckering of Pro residues (only the protons at C^{γ} and C^{δ} are shown).

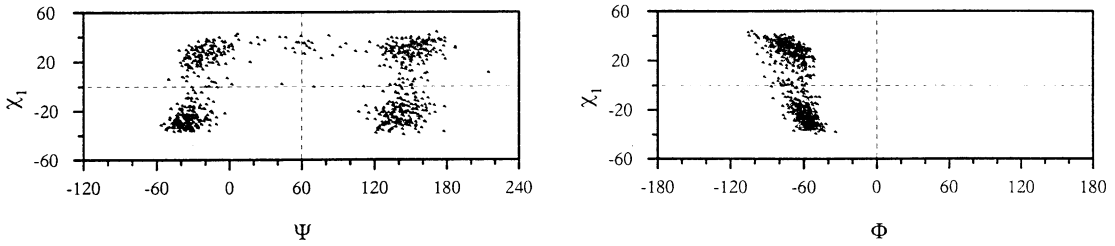

Fig. 24. χ_1, ψ and χ_1, ϕ plots for *trans* proline residues (Pal and Chakrabarti, 1999b).

Table 12
Range of ϕ angles for proline residues involved in both *trans* and *cis* peptide bonds and with positive and negative χ_1 angles[a]

χ_1 value	Conformation	ϕ range (°)
Negative	*trans*	−75 to −40
	cis	−81 to −58
Positive	*trans*	−100 to −50
	cis	−110 to −60

[a] Based on data in Figs. 19d and 24b. Also, see Fig. 26. The ranges given contain >95% of total data points in each case.

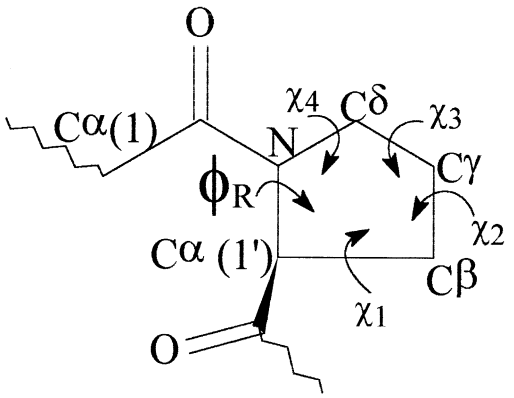

Fig. 25. Schematic representation of the ring torsion angles of a Pro residue. ϕ_R (torsion C^{β}–C^{α}–N–C^{δ} is related to the standard ϕ (C–N–C^{α}–C), as shown in Fig. 26.

shows that χ_1 is related to ϕ_R and, in turn, to ϕ. Additionally, the ranges of ϕ angles observed in small molecule structures, vis-à-vis the sign of χ_1 angles (Figs. 26b and c), are identical to those found in proteins (Table 12), suggesting that there is no long-range factor controlling the magnitude of these angles in proteins.

8.2. DOWN puckering in cis (X–Pro) proline residues

It can be seen that for *trans* Pro residues the range of ϕ is somewhat smaller when χ_1 is negative than when it is positive (Table 12). As already mentioned, a change of the *trans* to *cis* form pushes ϕ to a more negative value, such that the ϕ range is only about 20° when χ_1 is negative (as compared to 50° for a positive χ_1). A longer available range of ϕ means that a greater number of *cis* residues will have positive χ_1, i.e., the DOWN puckering. Thus the local steric interaction about the *cis* peptide bond causes ϕ to be more negative, which in turn shifts the puckering to the DOWN conformation.

8.3. UP puckering in helical proline residues

In spite of lacking an >N–H group, Pro residues are found in the interior of about 10% α-helices having a minimum of 9 residues (Piela et al., 1987; Barlow and Thornton, 1988; Chakrabarti and Chakrabarti, 1998). A helix containing a Pro (at position *i*) lacks the $N_i \cdots O_{i-4}$ hydrogen bond, and in 73% of the cases the hydrogen bond involving the atoms N_{i+1} and O_{i-3} is also disrupted. However, both these carbonyl groups (at $i-4$ and $i-3$), with no matching NH donors, are usually engaged in two C–H$\cdots$O hydrogen bonds (Fig. 27) involving the protons at the C^{δ} position of Pro, which being adjacent to the electron-withdrawing N atom carry a higher positive charge facilitating the interaction (Chakrabarti and Chakrabarti, 1998). Besides partially making up for the loss in conventional hydrogen bond, the C–H$\cdots$O interaction also fixes the puckering of the ring in the UP conformation. The optimum geometry for a C–H$\cdots$O interaction is linear. An interconversion between the UP and DOWN forms not only changes the position of the C^{γ} atom, but also displaces the protons attached to the neighbouring C^{β} and C^{δ} atoms

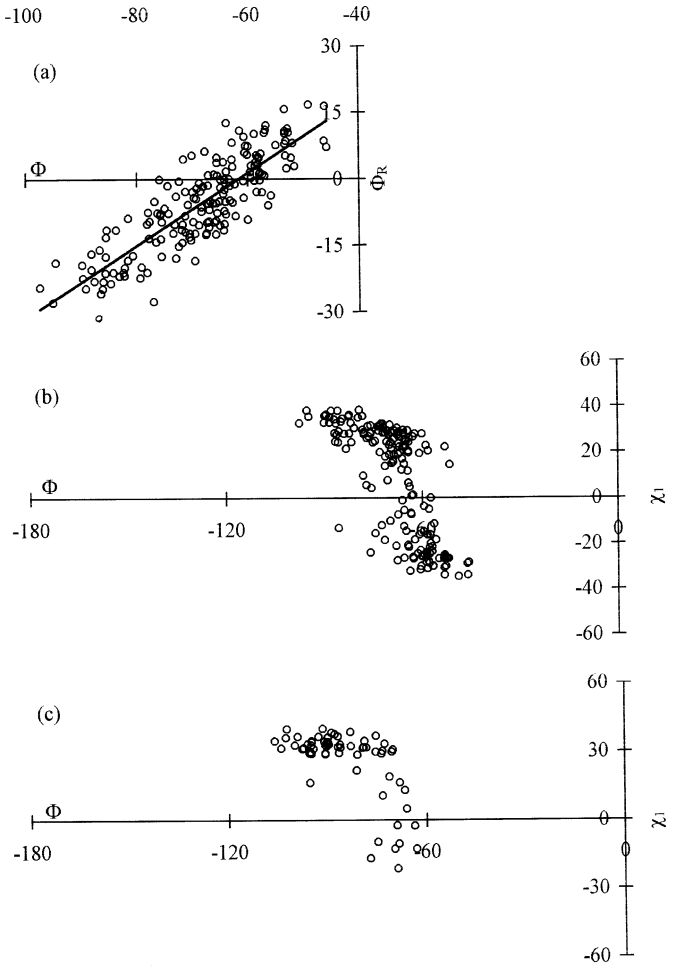

Fig. 26. Interrelationship between torsion angles, as observed in peptide structures containing the fragment shown in Fig. 25, retrieved from CSD (Section 2.7). (a) and (b) are for *trans* peptide units, and (c) for *cis*. (The *cis* isomer is found mostly in cyclic peptides, and to exclude cases with severe ring constraints those larger than tripeptides were considered.) The correlation coefficient between the parameters for the *trans* proline residues in (a) is 0.85, and the equation of the best-fit line is $\phi_R = 0.82\phi + 50.47$ (the corresponding data for the *cis* proline residues are 0.65 and $\phi_R = 0.42\phi + 17.53$).

(Fig. 23). Depending on the puckering, the geometry (H···O distance and C–H···O angle) of the C–H···O interaction would be different and the UP conformation usually has a better geometry and is consequently favoured (Chakrabarti and Chakrabarti, 1998).

8.4. Puckering in cis Pro–X residues

From Fig. 19b it can be seen that most of the Pro residues at position (1) of *cis* peptides have a negative value of χ_1, i.e., UP puckering (UP : DOWN = 2.2 : 1). Except two, in other cases the *cis* bond is between two Pro residues. The UP puckering is the most prominent (negative χ_1 in 5 out

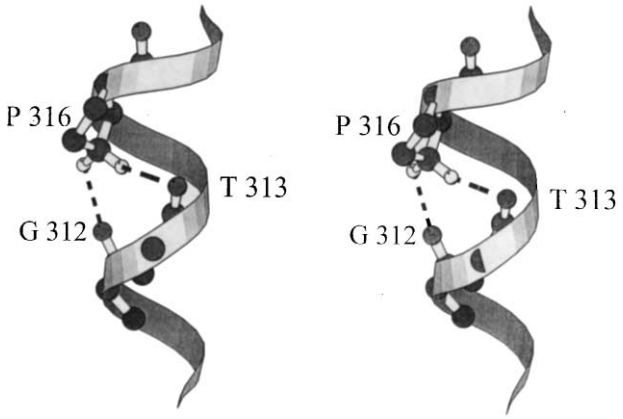

Fig. 27. Stereoplot showing C–H···O interactions (broken lines) taken from the structure, 1HPM; the participating residues are labelled, and a ribbon representation of a part (residues 310–318) of the helix is shown. The pyrrolidine ring has the UP puckering, such that relative to the average plane of the ring the C^{γ} atom and the helix axis are on the opposite side (Chakrabarti and Chakrabarti, 1998).

of 7 cases) when the Pro–Pro bond is in a turn of type VIa-1 for which the average value of ϕ is -54° (Table 11). It is to be noted that the first Pro residue has a *trans* peptide bond linking it to the preceding residue, and for *trans* Pro residues, this value of ϕ is very close to the upper limit when χ_1 is positive, but is within the range of allowed ϕ when χ_1 is negative (Table 12). So when ϕ is -54°, a negative χ_1 is preferred.

9. Specific short-range interactions affecting the conformation

Some conformations of the side chain or the main chain are observed more frequently than expected because they bring two chemical groups in the right orientation for a favourable interaction which, besides structural significance, may have functional implications also.

9.1. Cysteine residues

For Cys residues bound to metal ions (when the side-chain sulfhydryl group exists as the thiolate anion), the population of the g^+ state is drastically reduced (only 17% in Fig. 28, as opposed to 57.7% in Table 6 for all cysteines) in favour of the g^- and t states (Chakrabarti, 1989; Chakrabarti and Pal, 1997). In the latter two states the S atom is positioned above or below the peptide group (linking the Cys carbonyl group) such that the S···C length is 3.2 ($\pm$ 2) Å and the S···C=O angle is 109° ($\pm$ 15°). In the g^+ conformation, the two atoms concerned are farthest from each other (see the curve in Fig. 28) and there cannot be any intra-residue interaction, but a few of these S atoms are in close proximity to a carbonyl C atom belonging to a different residue (Chakrabarti and Pal, 1997). The close to perpendicular orientation of the S atom relative to the carbonyl group facilitates the overlap between the highest occupied molecular orbital (HOMO) of the former (usually a lone pair of electrons) and the lowest unoccupied molecular orbital (LUMO)

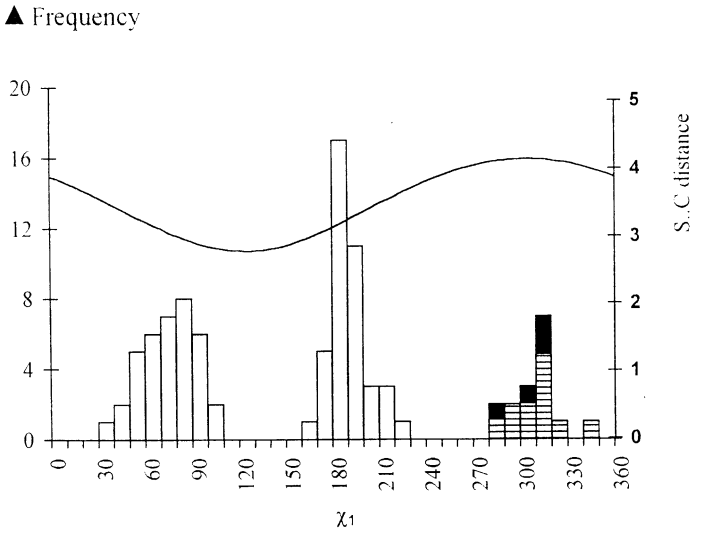

Fig. 28. Distribution of χ_1 torsion angle (°) of metal-bound Cys residues. Open and closed bars are for the cases with intra- and inter-residue S···C=O interactions; hatched bar represents the examples showing no interaction. The curve shows the variation of the intra-residue S···C distance (Å) (right side ordinate) as the torsion is changed through 360° (Chakrabarti and Pal, 1997).

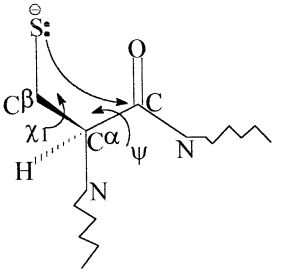

Fig. 29. The positioning of the sulphur atom on top of the carbonyl carbon atom restricts the available range of χ_1 and ψ angles and leads to the delocalization of the negative charge from S to the carbonyl oxygen atom.

of the latter (Fujimoto, 1987) and is a prerequisite for the nucleophilic addition of an amino, hydroxyl or sulfhydryl group to an electrophilic carbonyl group. The interaction between an electrophile and nucleophile is stablizing (Bürgi et al., 1974) and by analysing their geometry in small molecule structures it has been possible to derive information on reaction pathways (Bürgi and Dunitz, 1983). The placement of the S atom over a peptide plane results in a delocalization of the electronic charge from the S atom (and by extension, from the bound cation) all the way to the carbonyl oxygen atom (Fig. 29), thereby modulating the redox potential of the metal centre (Chakrabarti and Pal, 1997).

Semi-empirical calculations at the MNDO/3 level indicate that the intra-residue S···C=O interaction is stabilizing even for a free Cys residue. Moreover, the restriction imposed on the

values of χ_1 and ψ (Fig. 29) for the S atom to be on top of the carbonyl C atom is, by and large, adhered to in the distribution of these two angles (Pal and Chakrabarti, 1998). The identification of cases with the interacting S atom and the carbonyl group coming from different residues also points to the gain in energy that such a relative orientation leads to. As regard to the functionality, it has been suggested that the S···C=O interaction, by providing a way to delocalize the excess negative charge carried by the ionized sulphydryl group (Fig. 29), can make the deprotonation of the –SH group facile and thus lower its pK_a (Pal and Chakrabarti, 1998).

9.1.1. Other nucleophile-electrophile interactions

In a few instances the proper juxtaposition of nucleophile and electrophile in a protein structure may lead to a chemical reaction and alteration of the structure. For example, the green fluorescent protein has a chromophore formed through a rarely observed posttranslational cyclization of a peptide from its own backbone structure (Branchini et al., 1998). A tight turn conformation in the immature protein with a distance of less than 2.9 Å between a carbonyl carbon (at i) and amide nitrogen ($i+2$) leads to a nucleophilic attack of the amino group of Gly67 on the carbonyl group of Ser65 leading to the cyclization necessary for chromophore formation. Moreover, a variety of proteins have been found to activate functions by self-catalysed peptide bond rearrangements from single-chain precursors. These include autocleavage of Hedgehog proteins (Lee et al., 1994a), protein splicing (Paulus, 1998; Perler, 1998), maturation of pyruvoyl-dependent enzymes (Recsei et al., 1983), etc. The structure of the prototype protein, glycosylasparaginase (Xu et al., 1999) shows how the side chain of Cys, Ser or Thr is utilized as a nucleophile in different autoprocessing proteins to attack the carbonyl group of the immediate upstream peptide bond which is located in a highly strained tight turn conformation. Thus nucleophile–electrophile interactions are of structural and functional utility in proteins.

9.2. Asparagine and aspartic acid residues

A large number of Asn and Asp residues have ϕ, ψ values (Fig. 12) beyond the fully allowed region of the Ramachandran plot, notably in the bridging and L regions (Fig. 5) (Srinivasan et al., 1994). Deane et al. (1999) sought to explain this feature in terms of attractive interactions between pairs of $>C(\delta+)=O(\delta-)$ (carbonyl) dipoles, based on a systematic study of the alignment of ketonic groups in the Cambridge Structural Database (Allen et al., 1998). There are three main types of motifs: (a) a sheared antiparallel motif with two short carbon–oxygen interactions; (b) a perpendicular motif with only one short carbon–oxygen interaction; and (c) a highly sheared parallel motif with only one short carbon–oxygen interaction (Fig. 30). About 70% of Asp and Asn side chains could be considered to be in the sheared parallel motif such that they were stacked against their own backbone carbonyl or the backbone carbonyl of the previous residue (at a separation of less than 4 Å), and the former was usually associated with the t state of χ_1 and the latter, g^+. The sheared parallel motif has an attractive energy of about -7.6 kJ/mol at a separation of 3.07 Å (Allen et al., 1998), comparable to the secondary structure stabilization due to the Coulombic interactions between backbone carbonyls proposed by Maccallum et al. (1995).

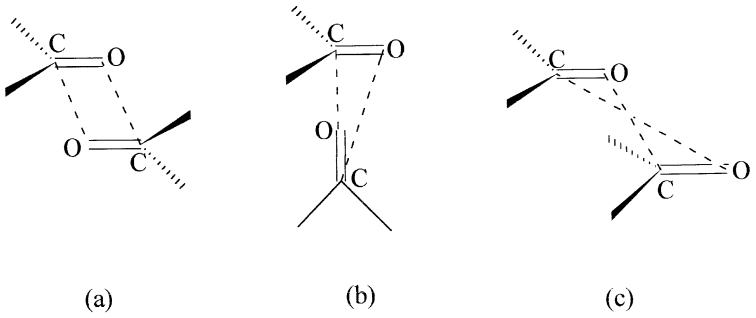

Fig. 30. (a) Antiparallel, (b) perpendicular and (c) sheared parallel motifs, commonly observed in the interactions between two carbonyl groups (adapted from Allen et al., 1998).

10. Effect of the neighbouring residue in the sequence

The ab initio prediction of protein tertiary structure from sequence information remains a distant and elusive goal. This is because the local sequence (short-range interactions) contributes on average up to 65% of the conformation of the residues in protein. The rest of the contributions come from long-range interactions along the sequence, i.e., between residues brought close by the folding of the polypeptide chain (Gibrat et al., 1991). As a result, hexamers that have identical sequences may have completely different conformations (Kabsch and Sander, 1984; Cohen et al., 1993), though their numbers are relatively small (8 pairs out of 59 pairs of identical hexapeptides culled from 366 polypeptide chains were found to form α-helical structure in one and β-strand in the other). There are methodologies to identify protein segments, of length 5–15 residues, that adopt well-defined conformations in the absence of tertiary interactions (Rooman et al., 1992). Sudarsanam and Srinivasan (1997) have devised a procedure for predicting the backbone conformations of hexamers in a sequence-dependent manner by starting with the distributions of ϕ_{i+1} and ψ_i angles for 400 dimers of naturally occurring amino acids, which were further subgrouped based on the homology of the two amino acids on either side of the dimer. At the most local level if one studies the influence of a residue on the conformation of the neighbouring residue, the effect of Pro is the only one that is clearly identifiable.

10.1. Effect of proline

The conformation of the residue preceding Pro (Fig. 31) is severely curtailed in the α region due to the steric conflicts between its $>$NH and $-C^{\beta}H_2$ groups and the $C^{\delta}H_2$ group of Pro (MacArthur and Thornton, 1991). Calculations based on ideal covalent bond lengths and angles indicate that the α conformation of the residue preceding a proline is about 7 kcal/mol less favourable than the β conformation (Schimmel and Flory, 1968; Summers and Karplus, 1990). Although this should preclude the occurrence of prolines in α-helices the observation is contrary to the expectation, as a substantial number are found in the middle of helices both in globular (Piela et al., 1987; Richardson and Richardson, 1988; MacArthur and Thornton, 1991; Kumar and Bansal, 1996) and membrane (von Heijne, 1991; Williams and Deber, 1991) proteins, although this leads to a kink in the helix (Barlow and Thornton, 1988; Sankararamakrishnan and

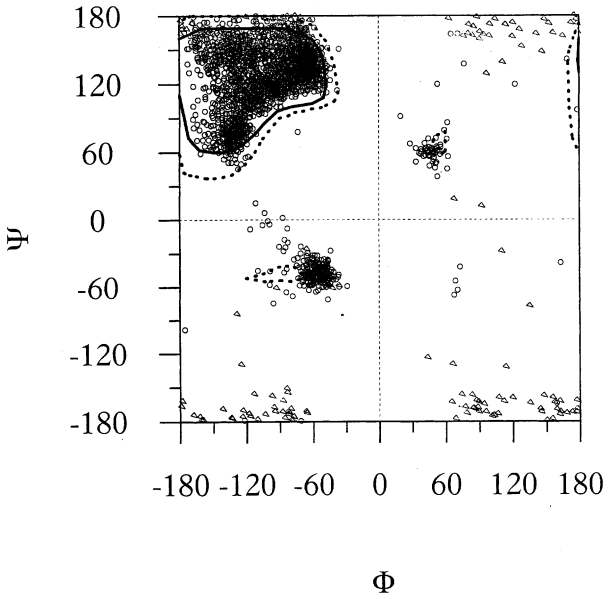

Fig. 31. ϕ, ψ angles for residues (Δ indicates Gly and o, non-Gly) preceding proline. Superimposed is the potential energy surface for Ala preceding a Pro (Summers and Karplus, 1990) with the conformation of Pro defined by $\omega = -180^\circ$, $\phi = -60^\circ$ and $\psi = -145^\circ$. Continuous and dotted contours enclose regions that are within 5 and 10 kcal/mol of the global minimum, respectively.

Vishveshwara, 1992). Chakrabarti and Chakrabarti (1998) have found that about 10% of all helices (nine residues or more long) contain a Pro (with at least one helical turn on its either side). A free-energy perturbation calculation in an explicit solvent model showed that the free energy in the α conformation of a proline dipeptide is only 1.6 kcal/mol higher than the β conformation (Yun et al., 1991), so that the former conformation can also be expected to be populated. Conformational energy calculations yielded result in agreement with the experiment if the bond lengths and angles, and torsion angles within the pyrrolidine ring were allowed to relax (Hurley et al., 1992) or when improved geometrical parameters were used (Némethy et al., 1992).

Curiously enough, Gly residues preceding Pro are rarely found in the A region, and even in the B or R regions (Fig. 5a) they are very close to $\psi = 180^\circ$ (Fig. 31) (also see Section 10.2). Another feature of Fig. 31 is the large number of points near -120, 80 designated as ξ (Fig. 4) (Karplus, 1996).

10.2. Uniqueness of Gly and the influence of its neighbours on its conformation

Gly is the only residue for which ϕ is equally allowed both in the negative and positive regions, and consequently 57% of the available ϕ, ψ space is allowed for the residue (Section 3). Consequently, Gly can assume conformations normally forbidden to other residues. However, within the allowed space the distribution (Fig. 7b) is quite nonuniform. Most of the Gly residues in B and R regions (Fig. 5a) are within 30° of the fully extended conformation. This also shows up in its average ψ value in β-sheet (Table 8), which is generally more extended than any other

residue. Gly residues preceding Pro are also extended, both when the peptide bond linking them is *trans* (Section 10.1) and *cis* (Section 7.4). A survey of crystal structures (Nicholson et al., 1989) showed that the energy minimum for left-handed helical (α_L in Fig. 4) Gly residues is near $\phi = 90°$, $\psi = 0°$, whereas that for non-Gly residues is close to $\phi = 60°$, $\psi = 30°$. Consequently, the replacement of one such residue, Asn55 to Gly in phage T4 lysozyme, gave a mutant protein which, though marginally less stable (by 0.5 kcal/mol), has the ϕ, ψ angles of residue 55 change by about 20°. Thus Gly usually has ϕ, ψ values not exactly conforming to other residues.

10.2.1. Conformation of Gly in X–Gly–Y triplets

Ramakrishnan et al. (1987) addressed the question of whether Gly has a greater tendency to occur either in the positive (P) or negative (N) ϕ region depending on the flanking residues. They could designate 25 triplets (for example, Asp–Gly–Lys, Asn–Gly–Ser, etc.) as P-predominant and 19 (like Leu–Gly–Phe and Pro–Gly–Val) as N-predominant. Though no explanation has been offered for the observation, one can use the results in modelling.

11. Terminal residues in polypeptide chains and their conformation

One of the backbone torsion angles cannot be defined for residues occupying the terminal positions in a polypeptide chain (Fig. 32). Moreover, at the pH values normally used for crystallographic experiments, the carboxy-(C-) terminal carries a negative charge and the amino-(N-) terminal, with the usual pK_a in the range 6.8–8.0 (Creighton, 1993), may also be positively charged. Consequently, both the steric and electrostatic factors prevailing on the terminal residues are different from the rest of the polypeptide chain. Potential energy calculations have been carried out to determine the possible conformations of N- and C-terminal Gly and Ala residues (Ponnuswamy and Sasisekharan, 1970). Recently, a detailed analysis of the preference of residues to occupy the two terminal positions, their solvent accessibility and hydrogen bonding features, the distribution of their main- and side-chain conformations and comparison of these to the general pattern have been carried out (Pal and Chakrabarti, 2000a).

11.1. Residue preference for the terminal positions

The propensities of residues to occupy the terminal positions, given in Fig. 33, show that Met is overwhelmingly the first residue of the chain. In eukaryotes all proteins are initiated with a

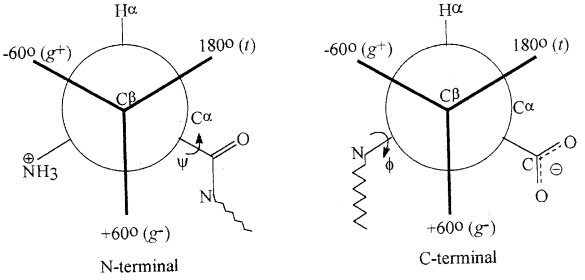

Fig. 32. Newman projections down the C_β–C_α bond for the two terminal residues showing the ϕ and ψ torsion angles and the three positions of the γ-atom corresponding to the three χ_1 angles.

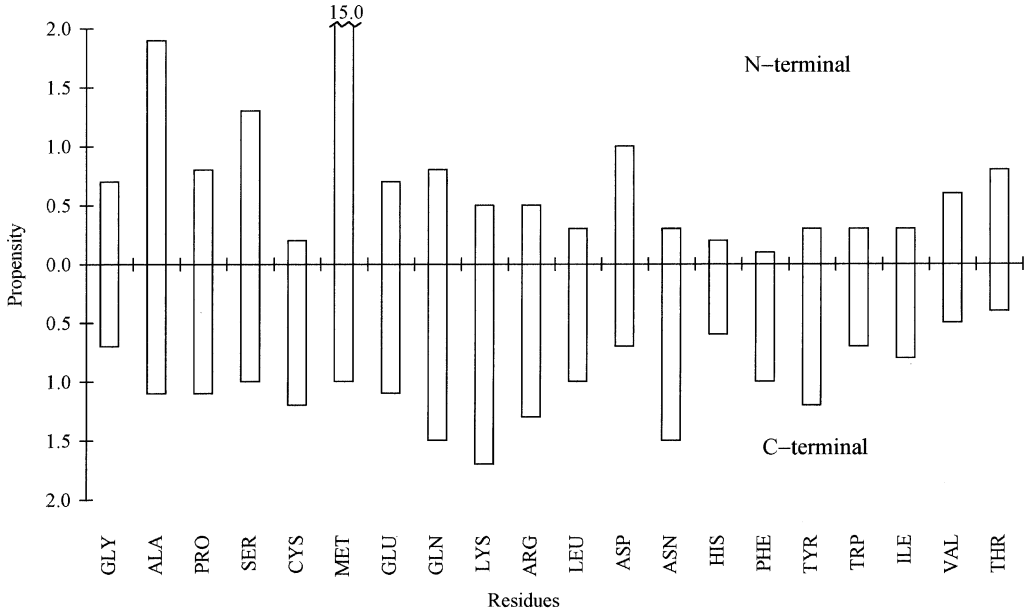

Fig. 33. Histogram showing the propensity of residues to occur at chain termini. Based on the sequence information on 393 polypeptide chains from 385 PDB files (Pal and Chakrabarti, 2000a).

methionine (*N*-formyl methionine in prokaryotes). In about half the proteins of both prokaryotic and eukaryotic cells, the initiating Met residue is removed from the nascent chain by a ribosome-associated Met-aminopeptidase (Creighton, 1993). Whether it is removed or not depends primarily on the second amino acid residue. Small residues (Gly, Ala, Ser, Cys and Thr) favour removal of the Met residue in prokaryotes; large, hydrophobic and charged residues seem to prevent removal (Hirel et al., 1989). Consistent with this, Ala and Ser have the highest propensities (next to Met). No particular physiological role is associated with the C-terminal residue. But still long-chain basic residues (Lys and Arg) and those with amide side-chains (Gln and Asn) have marked inclinations to occur at this end.

11.2. Conformation

For non-Gly/Ala residues the interdependence between the main-chain torsion angle (ψ or ϕ and the side-chain torsion χ_1 is shown in Figs. 34a and b. Considering only the ordered residues (the average temperature factor $\leqslant 40\,\text{Å}^2$), the most striking feature of Fig. 34a is that there are only a few points below ψ of 60°; these are mostly in the range 110–200° corresponding to an extended conformation. The preferred range for Ala is 140–170°, and for Gly, 170–210° (Fig. 35a). The origin for the inclination of the N-terminal residue to assume a value of ψ close to 180° is likely to be electrostatics, as this conformation is stabilized by the *syn* orientation of the $-NH_3^+$ group and the carbonyl oxygen (carrying a partial negative charge) (Fig. 36). The pK_a of the α-amino group is 6.8–8.0 (Creighton, 1993), depending on its environment and the identity of the terminal residue and, depending on the pH of the crystallization medium, even if it exists as

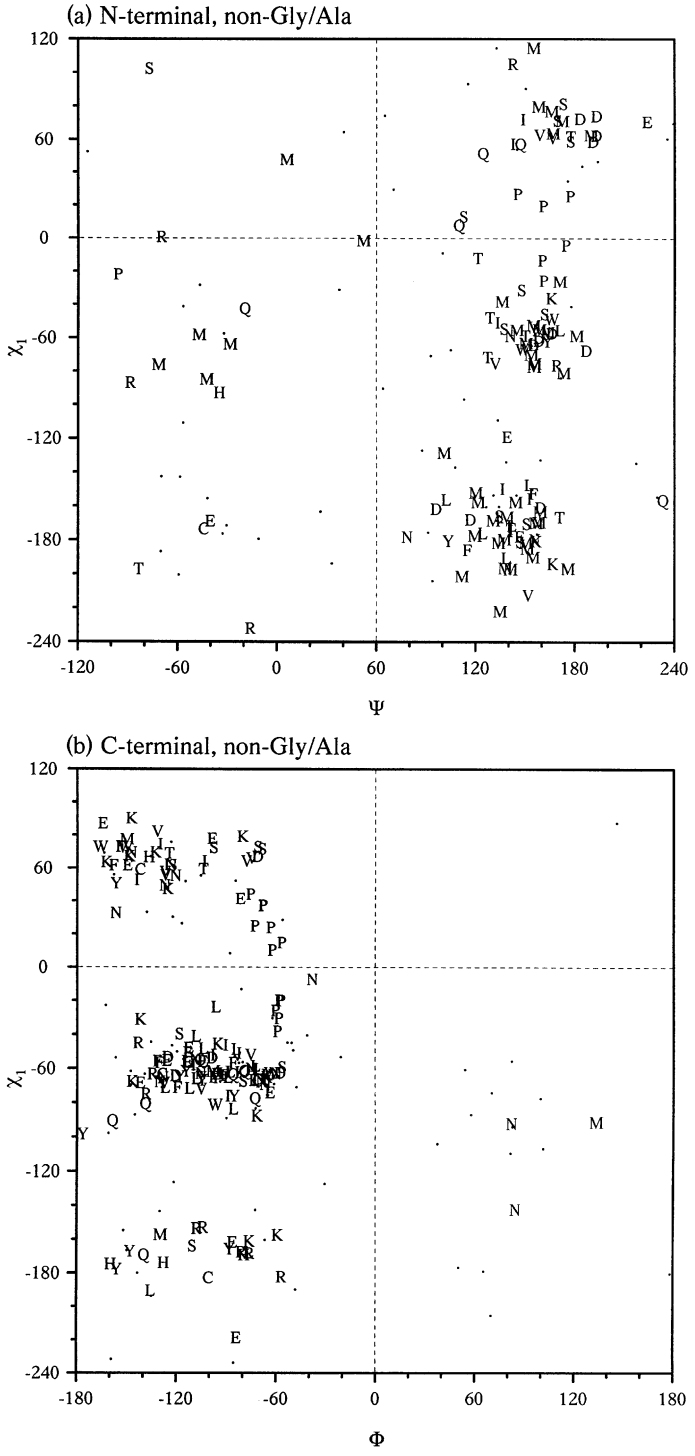

Fig. 34. Joint distribution of (a) χ_1 and ψ, (b) χ_1 and ϕ (°) for residues occupying the first and the last positions of the polypeptide chains. Positions are indicated by the one-letter amino acid code of the corresponding residue if the average B-factor is $\leqslant 40\,\text{Å}^2$ (127 and 158 cases, respectively); otherwise a dot is used (64 and 81, respectively) (Pal and Chakrabarti, 2000a).

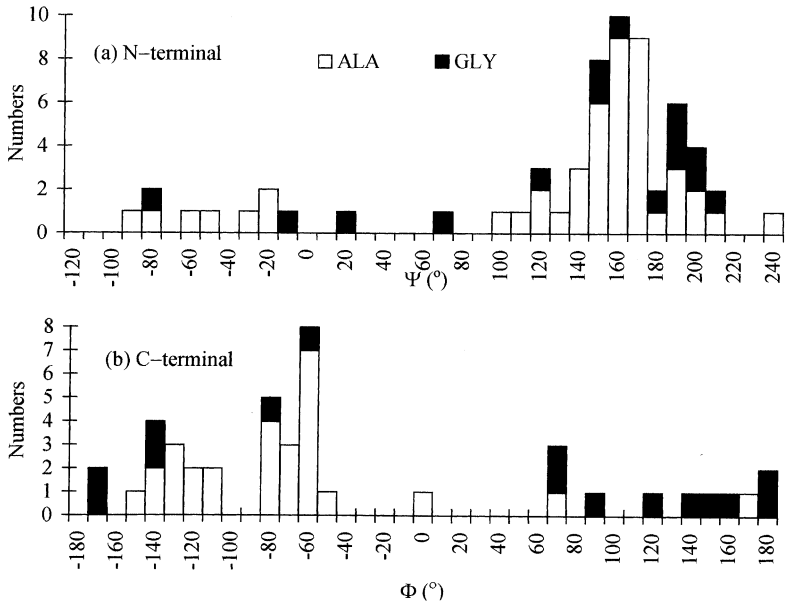

Fig. 35. Histograms showing the distribution of (a) ψ and (b) ϕ angles for Gly and Ala at the two termini (Pal and Chakrabarti, 2000a).

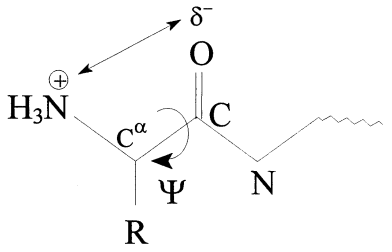

Fig. 36. Hydrogen bonding and electrostatic interaction (double-headed arrow) at the N-terminal residue that causes ψ to be near 180°.

$-NH_2$, rather than $-NH_3^+$, an amino proton may be suitably placed to interact favourably with the carbonyl oxygen atom when $\psi \approx 180°$. This conformation was also shown to be the most favourable for the N-terminal glycyl or alanyl residue when electrostatic term was included in potential energy calculations (Ponnuswamy and Sasisekharan, 1970).

As is found for most of the non-terminal residues (Fig. 10), the mean of the ψ distribution shifts towards a more extended value as the side-chain conformation is changed from t to g^+ to g^- states. The total numbers of points occurring in the t, g^+ and g^- states are 51, 47 and 29, respectively. Unlike the general distribution, where the population decreases in the order $g^+ > t > g^-$ (Table 6), for the N-terminal residue the maximally occupied state is t, which places the side-chain atoms opposite to the $-NH_3^+$ group (Fig. 32). As most of the residues are Met, it is unlikely that electrostatics is the primary reason for this observation. On the contrary, it is plausible that the relative preference for the t state increases because it allows the long side-chains to have van der Waals contacts with the rest of the molecule, whereas in the other two states

(especially, in g^+) these would point away from the main body of the molecule. Indeed, a simple calculation involving the number of contacts (within 3.8 Å) made by the side-chain atoms (C^γ onwards) with the rest of the protein molecule shows that 35%, 37% and 13% of residues in g^+, g^- and t states, respectively, have 0 contact, whereas 20%, 30% and 35%, respectively, have more than 5 contacts—thereby, indicating that compared to g^+, the t state has a smaller number of residues with no contact and a higher number with more contacts.

For the C-terminal residues, although ψ cannot be defined, the steric interaction may not be much different from a non-terminal residue as there is a carboxylate oxygen in place of the N atom (Fig. 32). The χ_1, ϕ plot (Fig. 34b) is essentially identical to similar plots for non-terminal residues (Fig. 11). However, the t state is the least occupied. The reason offered for the higher occurrence of the t state at the N-terminal may also be applied to explain its lower occurrence here. This conformation of the side-chain, as compared to the other two conformations, would provide it with the least opportunity to come in contact with the rest of the molecule. This need for the optimum surface to pack against explains why polypeptide chains can crystallize only when they have some threshold length. For Ala, the ϕ values are distributed in two ranges, -160° to -110° and -90° to -50°, whereas for Gly the points are widely spread, including the positive region of the ϕ (Fig. 35b).

11.3. Conformation of terminal residues in small peptides

Due to the paucity of small peptide structures with non-Gly/Ala residues occupying the terminal positions, the joint distribution of χ_1 with the backbone angles could not be considered. Interestingly, however, even in the small peptides, devoid of the influence of any secondary structure, the N-terminal residue has an extended conformation (as in proteins), a preference which is also retained when the end is acetylated, and the distribution of the ψ angle in these cases (Figs. 37a and b) is different from the general distribution (Fig. 38b). The C-terminal residue, on the other hand, shows a ϕ distribution (Fig. 37c) which is very similar to the general distribution (Fig. 38a).

11.4. Secondary structural features

The preference for the secondary structural elements in the terminal regions (consisting of 10 residues) and the location of the chain termini in the three-dimensional structure has been studied (Thornton and Chakauya, 1982; Thornton and Sibanda, 1983) and it was found that the N-terminal preferentially adopts a β-sheet conformation and the C-terminal is usually helical, and this led to the suggestion that $\beta\alpha$ is the basic unit using which all α/β proteins are constructed. In a recent study Pal and Chakrabarti (2000a) found that the only the N-terminal region has a secondary structural preference different from the rest of the protein. They also found out at what position along the sequence is the first secondary structure encountered (Fig. 39). Interestingly, the residue occurring next to the terminal has a very high propensity to be in the β conformation ($\beta:\alpha = 3.8$ and 2.1 for the two termini). The greater proclivity towards taking up the β conformation continues in the N-terminal region till the relative position of 6, beyond which for about 3 positions there is no particular preference, and then $\beta:\alpha$ ratio nears the average value (0.70). In contrast, in the C-terminus the preference for the β over the α structure shown at the

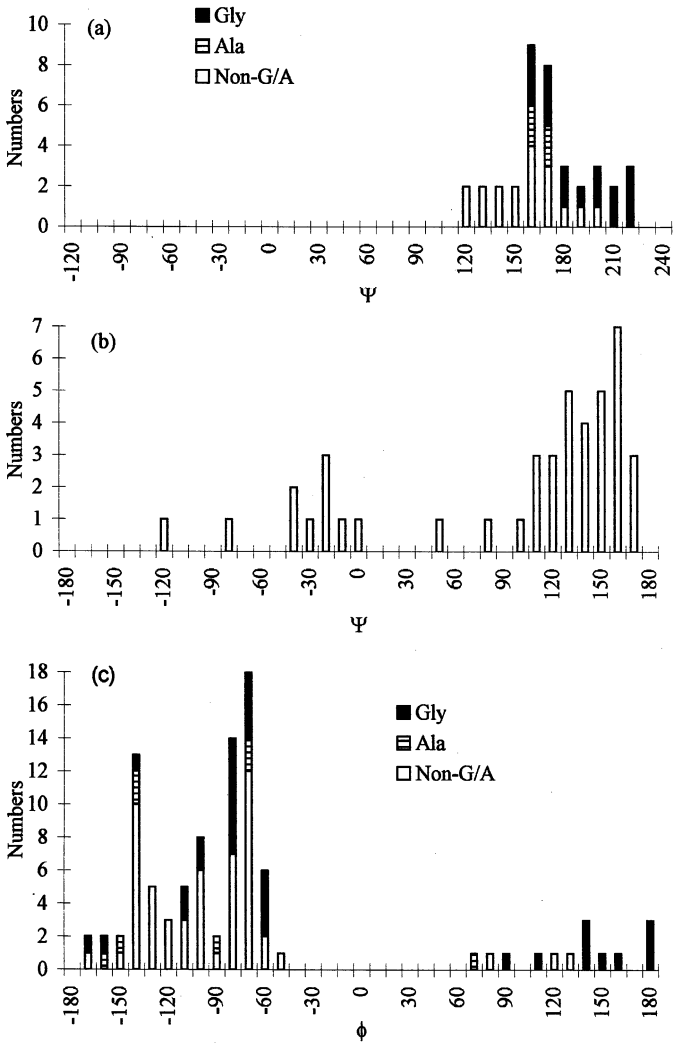

Fig. 37. Histograms showing the distributions of (a) ψ for free N-terminal residues, (b) ψ for acetylated N-terminal residues and (c) ϕ for the C-terminal residues in small peptide structures. The linear fragments searched in CSD (Section 2.7) were: (a) NH_2–CH(R)–CO–N–$\cdots$, (b) CH_3–CO–NH–CH(R)–CO–N–$\cdots$ and (c) $\cdots$–CO–NH–CH(R)–CO_2H.

relative position 1 is reversed in the next residue, and beyond 2 the $\beta:\alpha$ ratio approximates the average value.

It has been proposed (Pal and Chakrabarti, 2000a) that the electrostatic interaction and hydrogen bonding that makes the N-terminal residue assume an extended conformation (Fig. 36) is propagated along the chain so that the higher occurrence of β structure is exhibited by the first few residues in the N-terminal region. On the other hand, the ϕ value of the C-terminal residue is not restricted to the extended conformation only (Figs. 34b and 35b) and, therefore, can lead to both helix and sheet. As to why the preference for helical structure becomes conspicuous from position 3 onwards, it is plausible that 2 to 3 residues are needed at the free end to satisfy the

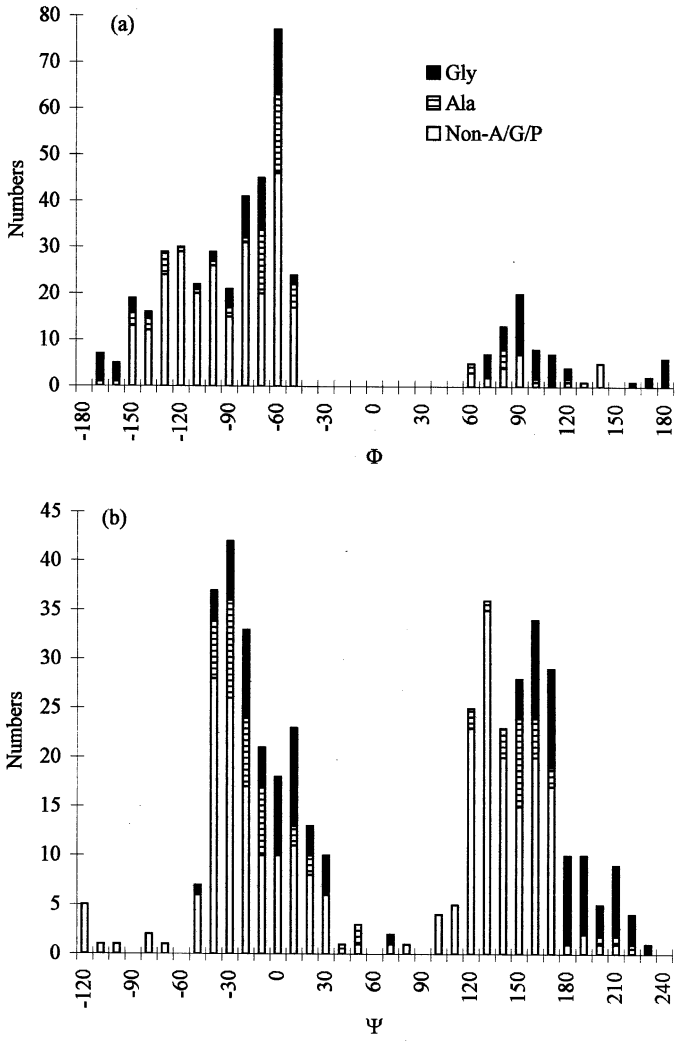

Fig. 38. Histograms depicting the distributions of (a) ϕ and (b) ψ angles for non-terminal residues in small peptides. The fragment searched in CSD was ···–CH(R)–CO–NH–CH(R)–CO–NH–CH(R)–···, such that it was not a part of a cyclic structure.

capping requirement (Aurora and Rose, 1998) of the helical C-terminus. Due to its role the free amino end of a polypeptide chain can be described as a β-strand initiator, like the influence Pro exerts in initiating an α-helix (Section 13.1.1).

It has been shown (Muñoz and Serrano, 1995) that the helical content of the polyalanine-based peptides, $(AAQAA)_n$ (where n is the number of repeats of the unit), is more when the N-terminus is acetylated than when it is unprotected. The disappearance of the electrostatic repulsion with the helix macrodipole (Hol, 1985) is assumed to contribute to the increase in the helical content of the N-terminal blocked peptide. But it is conceivable that the decrease in helical population is also brought about by a shift towards the extended conformation which is stabilized by the free

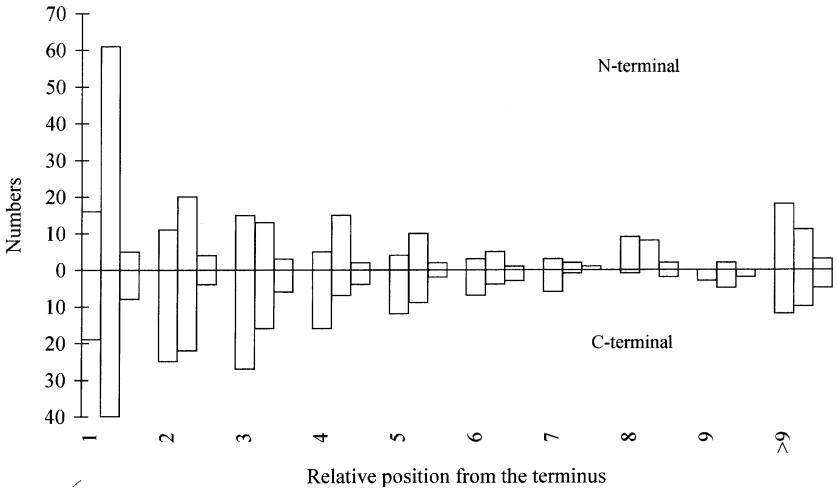

Fig. 39. Histogram showing the variation of the sequence gap between the terminal position and the starting position of the nearest regular secondary structural element (α-helix, β-strand and 3_{10}-helix, shown in this order against each position) along the polypeptide chain.

N-terminus. In this connection it is interesting to see that while ψ is invariably in the extended conformation for the N-terminal residue in small peptides, values of $\psi \leqslant 0^{\circ}$ start appearing when the end is acetylated (Figs. 37a and b).

11.5. Conformation at the cleavage sites

Proteolytic cleavage of polypeptide chains after synthesis is a common occurrence with certain classes of proteins, primarily those destined for cellular organelles or for secretion, in addition to removal of the signal peptide (Creighton, 1993). All the precursor *pro* proteins have specific cleavage sites. The proteolysis creates a new N-terminal end, which as the above discussion suggests tends to have an extended ψ value. Consequently, to have the minimum change in conformation at this end the terminal residue should have an extended ψ in the precursor protein. This premise has been found to hold good in an analysis of structures of a few zymogens and the corresponding active enzymes (Pal and Chakrabarti, 2000a). Thus in addition to easy accessibility to the protease molecule, the cleavage site generally has a conformation which is predisposed towards a value it would assume on breaking of the peptide bond.

12. Residues in the disallowed region

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 (Section 19). Excursions into the Ramachandran prohibited regions may induce a strain of up to at least 5 kcal/mol (Herzberg and Moult, 1991). 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. A residue (Ala16) having normally disallowed Ramachandran angles in the active site has been observed in the structure of histidine-containing phosphocarrier protein (Jia et al., 1993), where His15 is the target of phosphorylation. However, as the crystals were obtained at pH 5.0, below the pKa of histidine, the physiological relevance of the observation is not clear, though it shows that the protonation of a His residue (at lower pH) may sometimes induce strained main-chain conformation in the neighbouring residue.

Regions forbidden for non-Gly residues may be accessible to Gly without any energy penalty, and the strain energy associated with unfavourable ϕ, ψ values has been quantified on the basis of the stability of suitably chosen Gly/Ala mutants in staphylococcal nuclease (Stites et al., 1994). Gunasekaran et al. (1996) 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, the analysis was repeated with the presently available larger dataset, and also considering the neighbours of such residues (Pal and Chakrabarti, 2001).

12.1. Sterically disallowed clusters

A total of 285 residues (0.4%), with no atom involved in the definition of ϕ and ψ having a thermal parameter $> 30\,\text{Å}^2$, were identified to occur in the whole disallowed region as demarcated by Gunasekaran et al. (1996). With the availability of a larger number of residues the clusters (Fig. 40) are well-populated and to some extent different from the earlier groupings. There are five clusters, besides a few dispersed points. 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 allowed region B (Fig. 5a) is expanded along ϕ.

12.2. Amino acid propensities to be in the disallowed region

The propensities of residues to be in the disallowed region (assumed at position i) and the two immediate neighbours (positions $i \pm 1$) are shown in Fig. 41. A value of > 1 indicates a significant tendency to adopt a disallowed conformation (or be the neighbour of such a residue), whereas values < 1 suggest that such deviations are unlikely for these residues. While Gunasekaran et al. (1996) 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 His, Asn, Asp, Thr, Tyr and Trp. Pro, Phe and branched aliphatic residues (Val, Ile and Leu) disfavour such deviations. Considering the flanking residues, His, Tyr and Trp have distinctly high tendency to precede a disallowed residue. The propensity of a residue to follow a disallowed residue roughly parallels its own propensity to occupy one such position. Val, Ile, Leu and Ala oppose distortions when present as flanking residues.

Interestingly, both the occurrence of a *cis* peptide unit (Section 7.2) and a disallowed main-chain angle are opposed by Val and Ile. 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 $X_{\text{np}} - X_{\text{np}}$ *cis* bonds and the relative large presence of Trp and Tyr preceding the

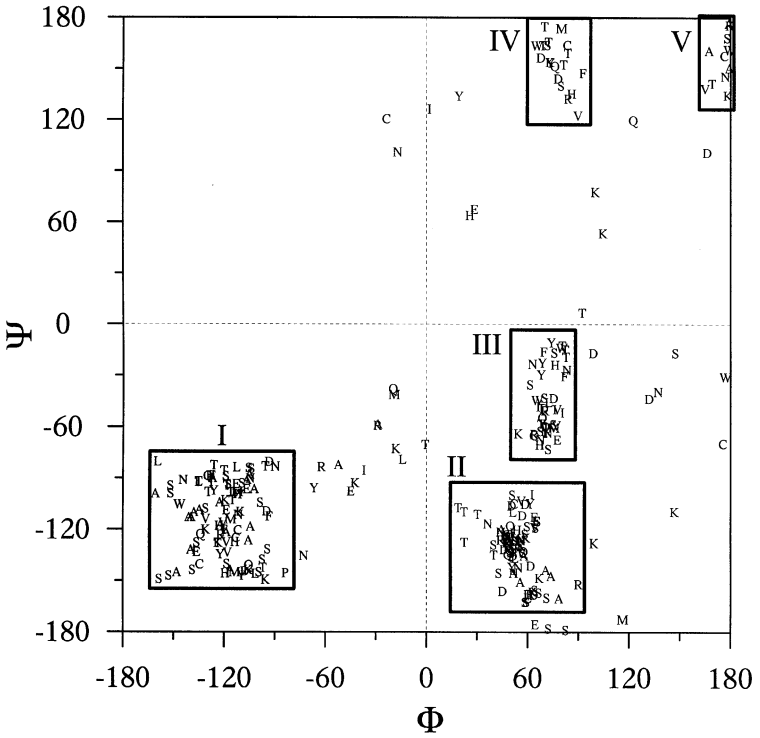

Fig. 40. Disallowed ϕ, ψ angles (each point is indicated by the one-letter amino acid code of the corresponding residue), with delineated clusters identified with Roman numerals.

X–Pro *cis* bond. Gly has a relatively high propensity to be one of the neighbours, 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.

13. Residue secondary structure and its effect on the distribution of ϕ, ψ, χ_1 angles

The relationship between side-chain conformation of individual residues and secondary structure has been analysed (McGregor et al., 1987; Summers et al., 1987; Schrauber et al., 1993). It would be of interest to study the differences in residue preferences in different positions/regions of a particular secondary structure along with the concomitant differences in χ_1 preferences (measured in terms of percentage distribution (Table 6) and χ_1 propensities, P_s (Tables 13–15), as defined in Section 2.6), and to analyse if the systematics are uniform across all the members of a particular class.

13.1. Different regions in α-helix

Helices and their flanking residues are labelled as follows (Richardson and Richardson, 1988; Presta and Rose, 1988):

$$N'' - N' - N_{cap} - N1 - N2 - N3 - \cdots - C3 - C2 - C1 - C_{cap} - C' - C'',$$

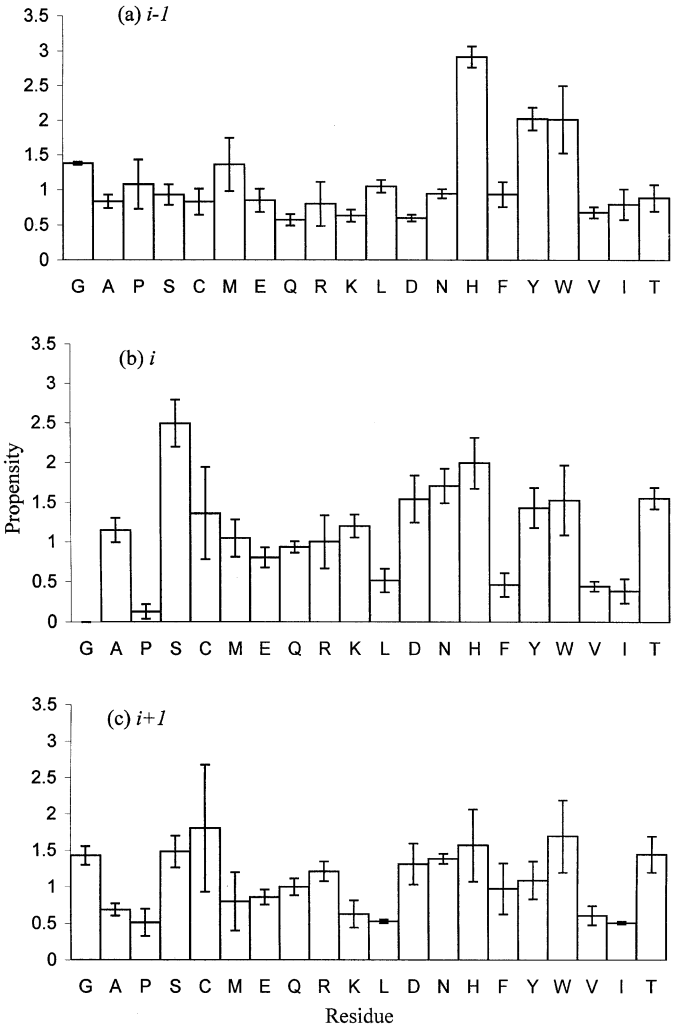

Fig. 41. (b) Propensities of residues (Gly excluded) to occur with disallowed Ramachandran angles (at position *i*), and (a) and (c), two flanking positions. Standard deviations were obtained by repeating the calculations with the 1998 and 1999 PDB structures (with 294 and 385 files, respectively).

where N_{cap} (and C_{cap}) are the residues with non-helical ϕ, ψ angles immediately preceding (following) the N (C) termini of an α-helix. (A slightly different definition not based on angles was used by Richardson and Richardson (1988), according to which the N_{cap} (or C_{cap}) is the first (or last) residue whose α carbon lies approximately in the cylinder formed by the helix backbone.) The characteristic i, $i+4$ hydrogen bonds between each carbonyl oxygen and an amide hydrogen are missing involving the $>$N–H groups of N1, N2 and N3, and $>$C=O of C1, C2 and C3 in the initial and final turns of the helix. The term helix "capping" has been used to describe the alternative hydrogen bond patterns of these groups. Since the hydrogen bonding environments are different, it is convenient to distinguish three regions in a helix, N-end (consisting of the first

Table 13
Propensities (P_α) of residues to occur in α-helices and different locations within helices, along with their χ_1 propensities (P_S)[a]

Residue	Overall				N-end				Interior				C-end			
	P_α	P_S			$P_{\alpha/l}$	$P_{S/l}$			$P_{\alpha/l}$	$P_{S/l}$			$P_{\alpha/l}$	$P_{S/l}$		
		t	g^+	g^-		t	g^+	g^-		t	g^+	g^-		t	g^+	g^-
Ala	**1.48**(2)				0.85(3)				1.04(2)				1.06(4)			
Gly	0.46(1)				**1.62**(9)				0.90(4)				0.60(6)			
Pro	0.44(2)		**1.40**(4)	0.6(2)	**3.7**(2)		1.02(4)	0.95(8)	0.24(3)		0.9(1)	1.3(2)	0.02(2)		0.7(5)	**2**(1)
Class I	1.18(1)	1.20(2)	1.03(1)	0.4(2)	0.91(2)	1.12(3)	0.84(2)	**2.0**(1)	0.98(1)	1.08(2)	0.99(1)	0.57(5)	1.13(2)	0.75(2)	1.15(2)	1.07(9)
Ser	0.75(2)	0.93(5)	**1.43**(5)	0.8(2)	1.24(7)	0.9(1)	0.89(7)	1.25(9)	0.87(4)	1.16(9)	1.14(5)	0.67(6)	1.06(7)	0.8(1)	0.86(7)	**1.3**(1)
Cys	0.74(4)	1.0(1)	1.20(5)	0.27(5)	0.9(1)	0.8(2)	1.0(1)	**2**(1)	1.07(9)	1.2(1)	0.93(6)	0.5(3)	1.0(1)	0.6(2)	1.15(8)	1.2(7)
Met	**1.32**(4)	1.06(7)	1.08(3)	0.21(2)	0.70(7)	**1.4**(2)	0.77(8)	**3**(1)	1.13(6)	1.03(8)	1.00(4)	0.6(3)	1.00(9)	0.6(1)	1.17(5)	1.0(7)
Glu	**1.38**(2)	1.15(3)	1.02(2)	0.41(6)	**1.44**(6)	0.98(5)	0.96(4)	**1.9**(3)	0.80(3)	1.13(5)	0.95(3)	0.5(1)	1.02(5)	0.79(6)	1.15(4)	0.8(2)
Gln	**1.33**(3)	1.14(5)	1.02(2)	0.29(4)	1.09(6)	1.07(8)	0.94(5)	**1.9**(6)	0.97(4)	1.02(6)	1.00(3)	0.8(3)	0.98(6)	0.88(8)	1.08(5)	0.6(3)
Lys	1.18(2)	**1.34**(4)	0.88(2)	0.38(6)	0.80(5)	1.12(7)	0.81(6)	**2.7**(6)	0.94(3)	1.02(4)	1.01(4)	0.4(2)	**1.33**(6)	0.89(5)	1.11(5)	0.9(3)
Arg	1.20(3)	**1.39**(4)	0.88(2)	0.40(6)	0.66(5)	1.10(8)	0.79(7)	**3.0**(7)	1.06(4)	1.03(4)	1.00(4)	0.6(2)	1.21(6)	0.89(6)	1.12(5)	0.8(3)
Leu	1.33(2)	1.12(3)	0.96(1)	0.27(2)	0.57(3)	**1.65**(8)	0.63(4)	**2**(1)	1.10(3)	1.11(4)	0.94(2)	0.9(4)	1.20(4)	0.47(4)	**1.30**(2)	0.9(6)
Class II	0.80(2)	0.60(3)	**1.47**(2)	0.3(1)	**1.32**(5)	0.7(7)	0.98(2)	**2.9**(4)	0.88(3)	1.20(7)	0.99(2)	0.06(4)	0.96(5)	1.0(1)	1.04(3)	0.3(1)
Asp	0.86(2)	0.54(3)	**1.51**(2)	0.32(9)	**1.61**(8)	0.55(8)	1.02(3)	**2.4**(3)	0.82(4)	**1.4**(1)	0.98(3)	0.09(6)	0.80(6)	1.1(2)	1.02(4)	0.4(2)
Asn	0.73(2)	0.69(5)	**1.43**(3)	0.16(6)	0.89(7)	1.1(2)	0.88(5)	**5**(1)	0.97(5)	1.0(1)	1.02(3)	0(0)	1.18(8)	0.9(1)	1.06(4)	0.2(2)
Class III	0.96(2)	**1.63**(3)	0.77(2)	0.25(9)	0.87(4)	1.00(4)	0.85(5)	**3.4**(5)	1.06(3)	1.12(2)	0.88(3)	0.31(9)	0.99(4)	0.71(4)	**1.43**(5)	0.6(2)
His	0.89(3)	**1.40**(7)	0.92(4)	0.27(4)	1.0(1)	0.9(1)	1.0(1)	**3**(1)	0.88(6)	1.21(7)	0.83(7)	0.6(3)	**1.3**(1)	0.76(9)	1.27(8)	0(0)
Phe	0.98(3)	**1.74**(5)	0.71(3)	0.17(4)	0.82(6)	1.15(6)	0.67(9)	**4**(1)	1.08(5)	1.09(4)	0.90(6)	0.1(1)	0.99(7)	0.67(6)	**1.53**(9)	0.7(5)
Tyr	0.91(3)	**1.67**(5)	0.76(3)	0.21(5)	0.78(7)	0.97(7)	0.9(1)	**3**(1)	1.13(5)	1.12(4)	0.88(6)	0.2(1)	0.93(7)	0.69(7)	**1.4**(1)	**1.3**(6)
Trp	1.16(4)	**1.56**(6)	0.74(5)	0.43(4)	1.0(1)	0.90(9)	0.8(1)	**3.2**(8)	1.10(7)	1.09(5)	0.95(9)	0.4(2)	0.8(1)	0.9(1)	**1.4**(2)	0.3(3)
Class IV	0.92(1)	0.56(4)	**1.29**(1)	0.5(2)	0.86(3)	**2.3**(3)	0.92(2)	1.1(1)	1.14(2)	0.68(8)	1.08(1)	0.51(4)	0.82(3)	0.6(1)	0.83(2)	**2.5**(2)
Val	0.93(2)	0.67(7)	1.18(1)	0.4(1)	0.84(5)	**2.4**(4)	0.92(3)	1.0(2)	1.18(4)	0.7(1)	1.06(1)	0.53(9)	0.74(5)	0.7(2)	0.87(3)	**2.7**(3)
Ile	1.06(2)	0.70(7)	1.15(1)	0.37(8)	0.62(5)	**3.0**(5)	0.88(3)	0.9(3)	1.23(4)	0.7(1)	1.05(1)	0.4(1)	0.84(5)	0.7(2)	0.92(3)	**3.0**(5)
Thr	0.78(2)	0.17(4)	**1.67**(3)	0.5(2)	1.17(7)	**2.4**(8)	1.00(3)	0.9(1)	0.97(4)	0.6(3)	1.13(2)	0.59(6)	0.90(6)	0.3(3)	0.67(5)	**2.1**(2)

[a] The three residues at the two ends define N-end and C-end, and Interior is the region in between. The first set of values are global propensities (P_α) and (P_S). The last three sets are local propensities ($P_{\alpha/l}$) and ($P_{S/l}$) based on data from 1646 α-helices at least 7 residues long. Figure in parentheses gives the standard deviation in the final digit. Values $\geqslant 1.30$ are in bold, and $\leqslant 0.70$ are underlined.

Table 14
Propensities (P_β) of residues to occur in β-sheets, along with their χ_1 propensities (P_S)[a]

Residue	Overall[a]				Antiparallel				Parallel			
	P_β	P_S			P_β	P_S			P_β	P_S		
		t	g^+	g^-		t	g^+	g^-		t	g^+	g^-
Ala	0.76(2)				0.72(3)				0.88(5)			
Gly	0.64(2)				0.62(2)				0.68(5)			
Pro	0.40(2)		0.88(6)	1.13(6)	0.41(3)		0.88(7)	1.12(7)	0.31(4)		0.8(1)	1.2(1)
Class I	0.93(1)	1.29(2)	0.85(1)	0.97(3)	0.97(1)	1.29(2)	0.82(1)	1.09(4)	0.79(2)	**1.41**(5)	0.85(3)	0.68(6)
Ser	0.87(3)	**1.48**(7)	0.91(5)	0.81(3)	0.98(4)	**1.45**(8)	0.91(6)	0.83(4)	0.65(5)	**1.9**(2)	0.8(1)	0.70(8)
Cys	1.23(6)	1.11(9)	0.96(5)	1.0(1)	**1.42**(8)	1.1(1)	0.99(5)	0.8(1)	0.8(1)	**1.4**(3)	0.8(1)	1.2(3)
Met	1.10(5)	1.25(9)	0.81(4)	**1.6**(2)	1.06(7)	**1.3**(1)	0.76(5)	**1.9**(3)	1.1(1)	**1.4**(2)	0.77(9)	**1.4**(4)
Glu	0.71(2)	**1.31**(6)	0.78(3)	1.2(1)	0.76(3)	1.29(7)	0.79(4)	**1.3**(1)	0.53(5)	**1.6**(1)	0.65(8)	1.2(3)
Gln	0.79(3)	**1.30**(7)	0.78(4)	**1.5**(2)	0.88(4)	1.28(8)	0.77(4)	**1.8**(2)	0.49(6)	**1.3**(2)	0.9(1)	0.7(4)
Lys	0.82(3)	1.22(5)	0.86(3)	1.1(1)	0.92(4)	1.27(6)	0.81(4)	1.2(2)	0.54(5)	1.1(1)	1.02(8)	0.2(2)
Arg	0.93(3)	1.09(5)	0.90(3)	**1.3**(1)	1.05(4)	1.12(6)	0.87(4)	**1.4**(2)	0.63(6)	1.1(1)	0.96(8)	1.0(3)
Leu	1.15(2)	**1.38**(4)	0.80(2)	**2.0**(3)	1.02(3)	**1.42**(5)	0.77(2)	**2.9**(5)	**1.38**(6)	**1.43**(8)	0.80(4)	0.9(4)
Class II	0.56(2)	**1.50**(5)	0.82(3)	0.63(6)	0.58(2)	**1.54**(6)	0.79(3)	0.66(7)	0.46(3)	**1.8**(1)	0.72(7)	0.4(1)
Asp	0.50(2)	**1.59**(7)	0.80(4)	0.51(7)	0.51(3)	**1.62**(8)	0.79(5)	0.49(9)	0.38(4)	**2.2**(2)	0.6(1)	0.2(1)
Asn	0.64(3)	**1.42**(7)	0.84(4)	0.77(9)	0.67(3)	**1.47**(9)	0.79(5)	0.8(1)	0.55(5)	1.5(2)	0.84(9)	0.6(2)
Class III	**1.34**(2)	0.76(2)	1.03(2)	**1.54**(6)	**1.40**(3)	0.70(3)	1.03(2)	**1.70**(8)	1.23(5)	1.04(6)	0.98(4)	1.0(1)
His	0.99(4)	1.18(7)	0.84(5)	1.2(2)	0.95(6)	1.02(9)	0.87(6)	**1.5**(2)	1.1(1)	**1.5**(2)	0.75(9)	0.7(2)
Phe	**1.47**(4)	0.68(4)	1.03(3)	**1.8**(1)	**1.49**(5)	0.60(5)	1.04(3)	**2.0**(1)	**1.4**(1)	0.95(9)	1.00(6)	1.2(2)
Tyr	**1.47**(4)	0.71(4)	1.07(3)	**1.5**(1)	**1.60**(6)	0.71(5)	1.04(3)	**1.6**(1)	1.26(9)	0.8(1)	1.09(7)	1.1(2)
Trp	1.21(6)	0.69(6)	1.14(6)	**1.3**(2)	**1.37**(9)	0.61(7)	1.17(7)	**1.4**(2)	0.8(1)	**1.3**(2)	0.9(2)	0.7(3)
Class IV	**1.61**(2)	1.21(5)	1.09(1)	0.73(2)	**1.50**(2)	**1.33**(7)	1.04(1)	0.80(3)	**1.95**(4)	1.00(9)	1.22(2)	0.47(3)
Val	**1.88**(3)	1.17(8)	1.00(1)	0.94(4)	**1.64**(4)	**1.4**(1)	0.95(2)	1.04(6)	**2.55**(9)	0.9(1)	1.10(2)	0.67(7)
Ile	**1.73**(3)	0.97(8)	1.03(1)	0.86(6)	**1.53**(5)	1.1(1)	0.99(2)	1.01(8)	**2.3**(1)	0.8(1)	1.12(2)	0.47(8)
Thr	1.19(3)	**1.6**(1)	1.23(3)	0.70(3)	**1.30**(4)	**1.6**(1)	1.25(4)	0.68(3)	0.90(6)	**2.0**(3)	**1.31**(8)	0.57(6)

[a] Including all residues in parallel, anitiparallel and mixed β-sheets (with designation E in the DSSP output). Also, see Table 13 footnote.

three residues, N1, N2 and N3), C-end (containing the last three helical residues, C1, C2 and C3) and Interior (between N3 and C3).

13.1.1. Amino acid preferences for different regions of α-helices

Amino acid residues show strong, specific preferences for the N/C_{cap}, N/C1, N/C2, N/C3 and interior positions of the α-helix (Argos and Palau, 1982; Richardson and Richardson, 1988; Petukhov et al., 1998; Aurora and Rose, 1998; Kumar and Bansal, 1998; Penel et al., 1999). Attempts have also been made to estimate residue preferences for the N-end, C-end and Interior regions (Gunasekaran et al., 1998) using propensity values. Helix propensity, as "normalized

Table 15
Propensities (P_{TI}) of residues to occur at different positions (i to $i+3$) of type I β turn and their χ_1 propensities (P_S)[a]

Residue	i				$i+1$				$i+2$				$i+3$			
	P_{TI}	P_S			P_{TI}	P_S			P_{TI}	P_S			P_{TI}	P_S		
		t	g^+	g^-		t	g^+	g^-		t	g^+	g^-		t	g^+	g^-
Ala	0.70(2)				1.16(2)				0.69(2)				0.89(2)			
Gly	0.97(2)				0.44(2)				0.75(2)				**2.54**(3)			
Pro	1.07(2)		1.05(8)	0.95(8)	**4.55**(4)		0.90(4)	1.10(4)	0.26(1)		1.1(2)	0.9(2)	0(0)		0(0)	0(0)
Class I	0.86(2)	0.73(3)	0.99(3)	**1.63**(5)	1.03(2)	0.68(3)	0.85(2)	**2.35**(5)	1.01(2)	0.15(1)	1.17(2)	**2.21**(5)	0.87(2)	0.93(3)	1.08(3)	0.81(4)
Ser	**1.64**(3)	**1.47**(8)	0.35(4)	1.19(6)	**1.52**(3)	0.25(4)	0.68(6)	**1.60**(5)	**1.60**(3)	0.07(2)	0.69(6)	**1.68**(5)	0.93(2)	**1.6**(1)	0.87(8)	0.77(7)
Cys	**1.55**(3)	**1.5**(1)	0.8(1)	0.8(1)	0.52(2)	0(0)	1.2(2)	**1.9**(3)	0.81(2)	0(0)	0.6(1)	**4.1**(2)	**1.71**(3)	**1.5**(1)	0.73(9)	1.1(1)
Met	0.59(2)	0.4(1)	**1.3**(1)	1.2(2)	0.57(2)	0.7(2)	1.0(1)	**2.3**(3)	0.39(1)	0(0)	**1.47**(9)	0.9(3)	0.81(2)	1.2(1)	1.0(1)	0.9(2)
Glu	0.66(2)	0.67(9)	1.15(8)	1.2(1)	**1.46**(3)	0.84(6)	0.89(5)	**2.1**(1)	1.07(2)	0.14(3)	**1.35**(5)	**1.8**(1)	0.68(2)	0.81(9)	1.29(7)	0(0)
Gln	0.60(2)	0.7(1)	1.16(9)	1.0(2)	0.93(2)	0.9(1)	0.84(8)	**2.5**(2)	1.05(2)	0.33(6)	1.18(7)	**2.1**(2)	0.78(2)	0.6(1)	1.15(8)	**1.5**(2)
Lys	0.68(2)	0.36(7)	**1.38**(7)	1.1(1)	1.22(2)	0.95(7)	0.76(6)	**3.1**(1)	1.09(2)	0.16(4)	**1.47**(4)	**1.3**(1)	0.86(2)	0.55(7)	**1.30**(6)	0.80(9)
Arg	0.70(2)	0.17(5)	**1.54**(6)	0.6(1)	0.90(2)	0.57(8)	1.06(7)	**2.2**(2)	0.87(2)	0.13(4)	**1.46**(6)	1.2(1)	0.91(2)	0.85(9)	1.10(7)	0.9(1)
Leu	0.67(2)	0.39(6)	1.29(4)	0.84(9)	0.57(2)	0.71(8)	1.14(5)	0.9(1)	0.71(2)	0.18(4)	**1.41**(3)	0(0)	0.83(2)	0.87(7)	1.08(5)	0(0)
Class II	**2.40**(3)	**1.88**(4)	0.31(2)	**1.52**(5)	0.79(2)	0.33(4)	**1.31**(5)	1.26(8)	**2.64**(3)	0.15(2)	0.97(3)	**2.65**(5)	1.15(2)	**1.57**(6)	0.78(5)	0.64(5)
Asp	**2.72**(4)	**2.05**(5)	0.23(3)	**1.38**(6)	0.89(2)	0.29(5)	**1.31**(7)	**1.4**(1)	**2.91**(4)	0.13(2)	0.83(4)	**3.07**(7)	1.12(2)	**1.57**(8)	0.83(6)	0.44(6)
Asn	**2.00**(3)	**1.57**(7)	0.46(5)	**1.76**(9)	0.67(2)	0.39(8)	**1.32**(8)	1.1(1)	**2.31**(3)	0.19(3)	1.19(5)	**1.91**(8)	1.18(3)	**1.58**(9)	0.71(7)	0.88(9)
Class III	0.95(2)	0.69(5)	1.28(5)	0.65(6)	0.62(2)	0.71(6)	0.97(6)	**1.9**(1)	0.97(2)	0.09(2)	**1.34**(4)	**2.04**(9)	1.11(2)	1.21(6)	0.96(5)	0.58(5)
His	**1.56**(3)	1.2(1)	0.93(9)	0.7(1)	0.61(2)	0.9(2)	0.7(1)	**2.5**(3)	**1.30**(3)	0.06(3)	1.17(9)	**2.8**(2)	1.29(3)	1.1(1)	1.02(9)	0.6(1)
Phe	0.84(2)	0.31(7)	**1.56**(6)	0.31(7)	0.51(2)	0.7(1)	1.2(1)	1.0(2)	0.72(2)	0.17(6)	**1.39**(8)	**1.6**(2)	0.98(2)	**1.3**(1)	0.96(8)	0.26(6)
Tyr	0.76(2)	0.5(1)	**1.33**(9)	0.8(1)	0.76(2)	0.7(1)	0.88(9)	**2.3**(2)	0.92(2)	0.10(4)	**1.53**(7)	1.2(1)	1.07(2)	**1.3**(1)	0.89(8)	0.69(9)
Trp	0.75(2)	0.5(2)	**1.4**(1)	1.0(2)	0.55(2)	0.3(1)	1.3(2)	**1.7**(3)	1.22(3)	0(0)	1.2(1)	**2.8**(2)	1.23(3)	1.0(1)	1.0(1)	0.8(1)
Class IV	0.65(2)	0.52(5)	0.43(4)	**2.52**(6)	0.60(2)	**1.99**(9)	0.49(4)	**1.92**(7)	0.49(2)	0.37(5)	0.05(2)	**3.47**(4)	0.74(2)	1.04(6)	1.04(4)	0.90(5)
Val	0.36(1)	0.29(8)	0.83(9)	**2.0**(2)	0.51(2)	**2.5**(2)	0.55(7)	**2.1**(1)	0.17(1)	0.6(2)	0.20(9)	**4.3**(2)	0.65(2)	1.1(1)	1.08(5)	0.65(8)
Ile	0.39(1)	0.9(2)	0.7(1)	**2.8**(2)	0.46(2)	**3.9**(2)	0.37(8)	**2.5**(2)	0.12(1)	1.7(4)	0.11(9)	**5.4**(3)	0.79(2)	**1.7**(1)	0.92(6)	1.0(1)
Thr	1.24(3)	0.50(6)	0.21(4)	**1.79**(4)	0.82(2)	0.56(8)	0.60(7)	**1.43**(7)	1.20(2)	0.20(4)	0.02(1)	**2.02**(2)	0.81(2)	0.30(6)	1.20(8)	0.94(8)

[a] The two central residues together form a subset of data given under 'Turn' in Table 6. Also, see Table 13 footnote.

frequency'' is usually calculated as the ratio of the fractional occurrence of an amino acid residue in helix and the fractional occurrence of all residues (in the database) in helix. This global propensity reflects the preference of a residue to be in helix compared to the rest of the molecule. However, it does not tell, once in the helix, which of the three regions it is likely to occupy. To better understand the preference to be in a given region inside the helix compared to other regions, the local propensity has been calculated (where the normalization is based on data restricted to helices only—Section 2.6).

By considering region-specific local propensities (Table 13) it is possible to discern patterns not seen in studies dealing with global propensities at individual positions. Of the two residues, Ala and Leu, with high P_α and no non-polar atom in the side chain, Ala is more uniform in its occurrence along the whole length of the helix, whereas based on $P_{\alpha/l}$ it can be seen that Leu is less likely to be found at N-end and has a higher preference for C-end. This is possibly due to the hydrophobic interaction that a Leu side chain can get involved in, in one of the capping motifs found at helix C-termini (Aurora and Rose, 1998). Although Gly is known to be helix terminator occupying the C_{cap} position (Richardson and Richardson, 1988; Preissner and Bork, 1991; Aurora et al., 1994; Gunasekaran et al., 1998) and has a low propensity to be in helix, but those present are mostly found at N-end, which may be due to the location of the binding sites of anion or anionic substrates at the helical N-termini (Hol, 1985; Chakrabarti, 1994), for which Gly is an ideal ligand because of the availability of its free $>$N–H group for hydrogen bonding and the absence of any steric resistance involving the side chain (Chakrabarti, 1993). Pro has a high local propensity to be at N-end because its constrained ϕ value matches with the helical ϕ and thus it can initiate the helix formation (Richardson and Richardson, 1988). Two other residues with low helix propensity but high local propensity for N-end are Ser and Thr. These side chains, as well as those of Asp, Asn, Glu and Gln can be used for capping interactions (Presta and Rose, 1988; Kumar and Bansal, 1998; Penel et al., 1999). Of the residue pairs Glu and Gln, and Asp and Asn, the charged residue has a higher local propensity for N-end. Of the three regions, Asn has the highest preference for C-end. This and a similar behaviour by His can be explained by the ability of the $>$NH groups of their side chains (at i) forming hydrogen bond with the main-chain $>$C=O (at $i-4$) (Chakrabarti, 1991; Kumar and Bansal, 1998). Excluding Thr, the two other class IV members have the highest local propensity for the interior region.

13.1.2. χ_1 Distribution. Consistent with the study by McGregor et al. (1987), the χ_1 distribution for most side chains in an α-helix show the avoidance of the g^- conformation, due to the steric clash between the C^γ atom with the carbonyl group of the i-3 residue (Table 6). The notable exceptions are Ser and Thr, whose hydroxyl group in the g^- state can form hydrogen bonds with the carbonyl group in the preceding turn of the helix (Gray and Matthews, 1984). The change in the relative proportions of the three χ_1 states in helices as compared to the overall values (Table 6) is conveniently expressed as χ_1 propensities, P_s, and given in Table 13. P_s for all residues are considerably less than 1 in the g^- state. This is true even for Ser and Thr, which have a higher population relative to other residues in the g^- state, but with respect to their own overall distributions the population in this state is reduced. The lowering of P_s below 1 in the g^- state results in a shift in the population to the t state ($P_s > 1$) for classes I and III residues in general, and to the g^+ state for classes II and IV residues. Of the class I residues, Ser and Cys, whose side-chain hydroxyl or sulfhydryl group is capable of hydrogen bonding with the CO group in the

preceding turn, show a shift towards the g^+ state. Though less severe than in the g^- state, in the g^+ conformation too there is some steric clash between C^γ and O ($i-4$) (McGregor et al., 1987) making it less favourable than the t state, especially in class III. For Pro the g^+ state gains at the expense of the g^- state, and is discussed in the context of the puckering of the pyrrolidine ring (Section 8). Some of the shifts mentioned above are determined by the helical region the residues are preferentially located in, and discussed next.

The local (in a given region of helix) χ_1 propensity, P_s, is the ratio of the percentage occurrence of a particular χ_1 state of a residue in the given region, and the percentage occurrence of this χ_1 state adopted by the same residue in all regions of the helix. It reflects the χ_1 preference of an amino acid in a given helical region compared to the average value over the whole helix. Values of P_s, given in Table 13, show a clear trend involving the distribution of the χ_1 state in different helical regions. Though in general, rare in the helix, the g^- state, when observed in this secondary structure, is more likely to be found in the N-end, where the possibility of a steric clash involving the side chain and the atoms in the preceding turn does not exist. The t state is preferred in the interior, and g^+ in the C-end. The preferred side-chain conformations for class IV residues are t, g^- and g^+ in N-end, C-end and Interior, respectively. There have been discussions on the side-chain rotameric states that are implicated in specific hydrogen-bond interactions in the first helical turn (Penel et al., 1999) and contacts involving pairs of side chains in helices (Klingler and Brutlag, 1994; Walther and Argos, 1996).

13.2. Propensities of residues to occur in β-sheet and their χ_1 preferences

The propensities of residues to be in parallel, antiparallel and mixed β-sheets taken together, and in each of the first two categories are given in Table 14. In the first exercise of this type, Lifson and Sander (1979) observed that the parallel structure was more selective in the sense that there were more extreme large and small values (the range being 2.63–0.28), and it favoured the hydrophobic side chains more than the polar and charged groups. The trend is generally maintained in the present data (the most notable exception being His, with a previous value of 0.38 is now found to have a value of 1.1 in the parallel structure). The best makers of antiparallel β-sheets are Val > Tyr > Ile > Phe > Cys > Trp, and of parallel, Val > Ile > Phe > Leu > Tyr. The preference of parallel β-sheets for hydrophobic residues is exemplified by lower propensity values of Glu, Gln, Lys, Arg and especially Thr, and a higher value of Leu to occur in the parallel than the antiparallel structure. By considering disulphide-bonded cystines and free Cys residues separately, Wouters and Curmi (1995) found that P_β value for the former to occur in antiparallel β-sheets, 1.92, is much higher than that (1.06) for the latter. Multiple-stranded parallel β-sheets are typically situated in the protein interior where the parallel strands are interconnected by α-helices which pack on the sheet surface(s); in contrast, antiparallel sheets are generally found on the surface (Richardson, 1981; Salemme, 1983). This explains why more hydrophobic residues are preferred in the parallel sheet.

13.2.1. χ_1 propensities

As was done in Section 13.1.2 for α-helices, the P_s propensities of different residues in parallel, antiparallel and all types of β-sheets taken together were calculated and given in Table 14. For class II members, the t state gains at the expense of the other two states. In class III, the g^- state

has a higher P_s in antiparallel β-sheet, but in the parallel structure the largest value is observed in the t state for His and Trp. Though g^+ is the state with the highest percentage of occurrence in β-sheet for class I residues, the value (47.9%, Table 6) is lower than the overall value (56%); consequently, P_s for this state is smaller than 1. The most prominent change among class IV members is the reduction of the g^- state for Thr. The rotameric preferences for specific inter-strand residue pairs in antiparallel sheets have been discussed by Hutchinson et al. (1998).

13.3. Propensities of residues to occur in type I β-turns and their χ_1 preferences

β-Turns are the most common type of non-repetitive structure recognized in proteins and comprise, on average, 21% of the residues (Table 6). First recognized by Venkatachalam (1968) these have been categorized into different classes based on the ϕ, ψ angles of the two central residues (Lewis et al., 1973; Richardson, 1981; Rose et al., 1985; Hutchinson and Thornton, 1994). Unlike the helix or sheet, the four positions in a turn are not equivalent and are usually characterized by different sequence preferences (Hutchinson and Thornton, 1994). In the following discussion (Table 15), only the most populous, type I β-turn is considered.

The trend among the propensity values to occur at different positions of the turn is quite similar to what was observed by Hutchinson and Thornton (1994), though there have been some changes in the magnitude. The alternating pattern of hydrophobic and hydrophilic residues in β-strands is well known (Lim, 1974; Sun et al., 1995; West and Hecht, 1995; Hutchinson et al., 1998). In an analogous manner, positions i and $i+2$ in turn show similar behaviour in the type of residues which are favoured (Asp, Asn, Ser, His, etc.) or disfavoured (Val, Ile). Within a class at a given position, the members tend to have similar propensity values, the exceptions are generally the hydrophilic residues with higher values, like Ser, His and Thr at positions i and $i+2$, and Ser, Glu, Lys at $i+1$.

13.3.1. χ_1 propensities

Of the four positions defining the β-turn, the two central ones have the most distinct preference with regard to the χ_1 state that is avoided (Table 15). At these positions class IV residues have the smallest P_s value in the g^+ state, whereas for the other classes it occurs in the t state. The less-than-one value of P_s in these states is compensated by a greater-than-one value in the g^- state. Although class II members have high residue propensities at i and $i+2$ positions, their χ_1 propensities are different; as mentioned, P_s is quite small in the t state at the $i+2$ position, while the smallest P_s is found in the g^+ state at position i.

14. Signature of secondary structural propensities in the overall ϕ, ψ, χ_1 distribution

Elements of secondary structure—α-helix, β-sheet and tight turns—are ubiquitous in proteins. It is observed that the three-dimensional structure of a protein is hierarchical, with a local organization of the amino acids into secondary structure elements, which are themselves organized in space to form the tertiary structure. Hence, to unravel the protein folding problem it is important to understand the physical basis for the correlation between sequence and the presence of an α-helix or a β-sheet in the structure. Yet, a simple physicochemical theory of secondary structure in peptide and proteins has proved elusive (Srinivasan and Rose, 1999).

14.1. α-helix propensities

Davies (1964) first noted that some amino acid residues occur more often than others in the helix. This was later quantified in terms of propensity values based on statistical analysis of known structures (Chou and Fasman, 1974; Levitt, 1978; Williams et al., 1987). These are the ratio of the percentage occurrence of a given amino acid in a particular secondary structure and its percentage occurrence in the whole structure (Section 2.6). A propensity greater than one means that the proportion of the amino acid in the specific secondary structure is bigger than in the rest of the structure.

Swindells et al. (1995) calculated propensities ($P_{a/coil}$ and $P_{B/coil}$) of residues not belonging to helices and strands (i.e., considering those not having the interactions associated with these elements of secondary structure) to occur in broad α-helix and β-sheet regions, and found these to correlate reasonably with classic Chou and Fasman type propensities. Likewise, Muñoz and Serrano (1994) calculated tendencies of residues (irrespective of whether they have any secondary structure or not) to populate dihedral angles typical for α-helix and β-sheet structures and converted them into pseudoenergy empirical scales, which agreed very well with the experimental ones in relative and absolute terms.

Amino acid preferences in the α-helix has the added complexity in that they are also dependent upon the specific location in the helix, as discussed in Section 13.1.1. The helices can be further stabilized by capping interaction (for review, see Aurora and Rose, 1998; Serrano, 2000), like the double hydrogen-bonded pattern between >N–H (at C″ and C′) and >C=O (at C3 and C2, respectively) (see Section 13.1 for the atom labels), in the Schellman motif (Schellman, 1980; Milner-White, 1988; Dasgupta and Bell, 1993; Gunasekaran et al., 1998). Besides various local interactions (Aurora et al., 1997), there can be several specific interactions between pairs of side chains that stabilize helix formation (Chakrabartty and Baldwin, 1995), which is thus influenced by the context of the sequence and tertiary interactions (Horovitz et al., 1992).

Helix propensities have been measured in different monomeric peptide systems (Lyu et al., 1990; Padmanabhan et al., 1990; Park et al., 1993; Muñoz and Serrano, 1995; Yang et al., 1997) and small, single-domain proteins (Horovitz et al., 1992; Blaber et al., 1993; Myers et al., 1997), as well as a coiled-coil leucine zipper peptide of de novo design (O'Neil and DeGrado, 1990). Although the different sets of values obtained do not agree numerically, they are significantly correlated between themselves and with the statistical propensity values derived from the structure database (Chakrabartty and Baldwin, 1995; Blaber et al., 1994; Pace and Scholtz, 1998; Serrano, 2000). Indeed, one can have a consensus rank order of helix propensities (Chakrabartty and Baldwin, 1995). Ala has the highest value, followed by amino acids with long side chains (Arg, Leu, Lys, Gln, Glu, Met). The other amino acids, except Gly and Pro, have intermediate to low propensities, and Gly and Pro have the lowest.

14.1.1. α-Helix propensities and correlation with conformational similarity with Ala

In Section 6.2 the conformational similarity indices relating the ϕ, ψ, χ_1 distributions of different residues have been evaluated. As Ala has the highest helix propensity, we determined if the CS_{AX} values (Fig. 14) have any bearing on propensities (Pal and Chakrabarti, 2000c). Indeed, CS_{AX} and Chou–Fasman type propensities, P_α (Table 13) are strongly correlated (Fig. 42). As these are descriptors of residue conformation, doubts may be cast that they are two different ways

of looking at the same thing and the correlation between them is trivial. However, there are counter arguments (Pal and Chakrabarti, 2000c) the most important of which is that the parameter, CS_{IX} (Fig. 14) the conformational similarity defined relative to Ile has a poor correlation (0.43) with the β-sheet propensity. This means that the propensity of a residue to be in β-sheet does not depend on how similar its conformational map is with that of Ile, the residue with one of the highest β-sheet propensities. These results indicate that the structural requirements for the formation of different secondary structures are different. A contiguous stretch in the polypeptide chain, in tandem, forms the helix and should contain residues with high CS_{AX}, which is not true for β-sheet formation where residues involved are from non-contiguous regions of the chain. Indeed, as will be shown in Section 14.2.1, there are other residue characteristics which correlate with β-sheet propensities.

Other helix propensity scales based on both experimental data and theoretical consideration were also compared. Pace and Scholtz (1998) have derived a scale using the available experimental data on 11 systems, including both proteins and peptides. A scale based only on data from peptides was developed by Muñoz and Serrano (1995). Other scales considered were the

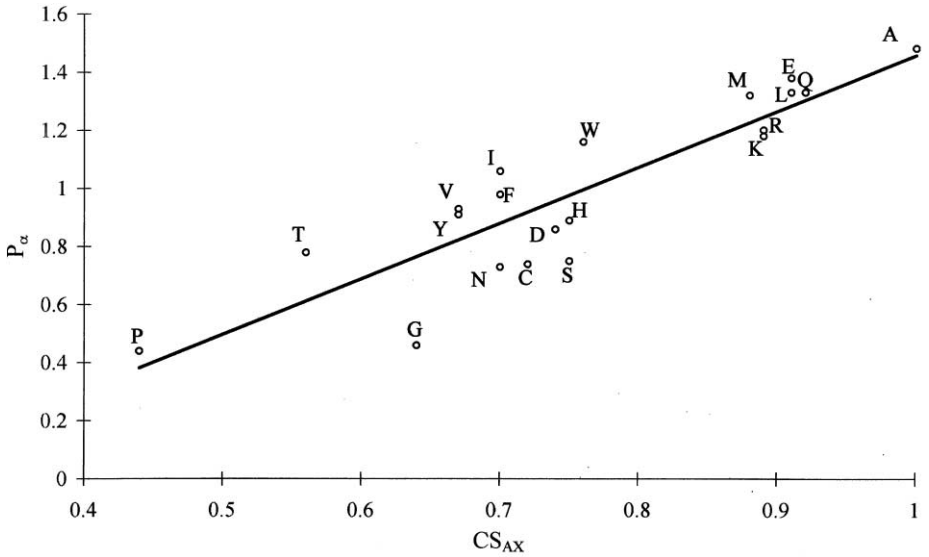

Fig. 42. Plot of P_α against CS_{AX} (values taken from Table 13 and Fig. 14). The fitted line has an equation $P_\alpha = 1.92CS_{AX} - 0.46$.

Table 16
Comparison (using correlation coefficients) between CS_{AX} (and $CS_{AX/2D}$) and some α-helix propensity scales[a]

	P_α	Pace	Agadir	Luque	Design
CS_{AX}	0.89	−0.83	−0.78	−0.82	−0.58
$CS_{AX/2D}$	0.82	−0.83	−0.86	−0.80	−0.76

[a] P_α values are taken from Table 13. Other references are: Pace and Scholtz (1998), Agadir (Muñoz and Serrano, 1995), Luque et al. (1996) and Design (Koehl and Levitt, 1999). Only P_α, CS_{AX} and $CS_{AX/2D}$ contain values for Pro.

thermodynamic scale derived by Luque et al. (1996) using a structure-based optimization scheme and another one by Koehl and Levitt (1999) generated using computer-designed sequences. Results given in Table 16 show that CS_{AX} values are in excellent agreement with all but one (marked 'Design') of the scales.

The ranking of residues in terms of CS_{AX} (Fig. 42) is chemically intuitive. Aliphatic side-chains (not branched before the γ position) have high CS_{AX} values; only Ser and Cys, with oxygen/sulphur atom at the γ position, have lower values. Other classes (Table 5) of residues (like aromatic, β-branched, etc.) have values in distinct ranges of CS_{AX}. This suggests that the topology of the side chain (linear, β-branched, γ-branched-aliphatic, γ-branched-aromatic, etc.) has the strongest bearing on the CS_{AX} values and in turn, on the helix propensities, as they are highly correlated. Asp and Glu, which are similarly charged and should have similar electrostatic effects when placed in a helix, have very different helix propensities. Interestingly, these two residues have quite different conformational features and belong to different classes. CS_{AX} is a distinctive characteristic of the steric features (rather than the charge or hydrophobicity) of the side-chain (as the chain is extended beyond the C^{β} position of Ala) and residues with high CS_{AX}, occurring in sequence, can cooperatively fold into a conformation where $i \cdots i+4$ hydrogen bond can form, and thus leading into a helix. As to why Ala has the highest helix propensity it may be noted that for residues with the side chain beyond C^{β}, a change in χ_1 causes a change in the helical ϕ, ψ angles (Section 5.5), and such structural fluctuations that may impede helix formation or propagation is absent in the case of Ala.

To see if χ_1 could be dispensed with in the derivation of similarity indices, $CS_{AX/2D}$ was calculated using only the two-dimensional ϕ, ψ distributions (Fig. 14). Compared to CS_{AX}, $CS_{AX/2D}$ has a poorer correlation with P_{α} (Table 16). However, both compare favourably with other scales, and for the propensity scale marked Design, the correlation coefficient is actually better with $CS_{AX/2D}$.

14.2. β-Sheet propensities

β-Sheets have elements of both secondary (a stretch of residues form a β-strand) and tertiary (two or more strands have to come together to constitute a β-sheet) structures. As a result, although different amino acid residues have been found to have measurably different propensities for forming β-sheets (Kim and Berg, 1993; Minor and Kim, 1994a; Smith et al., 1994), the value is also found to depend on whether the residue is located in the central or the edge strand of the β-sheet (Garratt et al., 1991; Minor and Kim, 1994b). Moreover, statistical surveys revealed a nonrandom pairwise distribution of amino acids in cross-strand positions in antiparallel β-sheets (von Heijne and Blomberg, 1977; Lifson and Sander, 1980). The specific pairing of amino acids depends on their positions in hydrogen-bonded or non-hydrogen-bonded sites (Wouters et al., 1995; Hutchinson et al., 1998), and the energetic contributions of side-chain interactions to β-sheet stability have been measured experimentally (Smith and Regan, 1995).

Experimental studies of β-sheet preferences have been addressed mainly in two protein model systems, a zinc-finger peptide (Kim and Berg, 1993) and the B1 domain protein G (Minor and Kim, 1994a, b; Smith et al., 1994). Other model systems based on homooligopeptide have proven to be inappropriate for the study of β-sheet propensities. In spite of the context dependence, the experimental results show overall correlations with statistical and theoretical analyses (Chou and

Fasman, 1974; Muñoz and Serrano, 1994; Finkelstein, 1995; Avbelj and Moult, 1995; Smith and Regan, 1997) and indicate that the amino acids have different intrinsic propensities to adopt the β-sheet conformation.

14.2.1. β-Sheet propensities and the inverse correlation with the volumes of dispersion of ϕ, ψ, χ_1 points

As a measure of entropy of the distribution, Pal and Chakrabarti (2000b) calculated the volume of dispersion (D_V)—the volume encompassing the ϕ, ψ, χ_1 points. The whole space was divided into four regions of ϕ, ψ (Fig. 5a), each with three possibilities of χ_1. The volume of clusters (of the 12 regions, only those with at least 1% of the total data points) was determined as the product of the standard deviations associated with the means of ϕ, ψ and χ_1 angles contained therein, and the sum provided D_V (Table 17), which shows a strong inverse correlation (Fig. 43a) with P_β, the updated Chou–Fasman propensities (Table 14). The values are also compared with four experimental and one theoretical scales (Table 18). While all experimental scales correspond to one given environment, Chou–Fasman propensities are average over all the positions in a secondary structure, and thus eliminate system-specific effects while retaining the trends. It is thus expected that the best correlation is shown against P_β. However, the correlation with the scale of Kim and Berg (1993) is equally good, but other scales do not compare very well. Like D_V, another

Table 17
Volume (D_V) and area ($D_{A/2D}$) of dispersion of ϕ, ψ (and χ_1) points, absolute (S) and relative (ΔS) conformational entropies of different residues[a]

Residue	D_V	$D_{A/2D}$	Entropy (S/R)	$-T\,\Delta S$ (kcal/mol)
Ser	3.19	1.87	6.74	−0.66
Cys	2.49	2.88	6.21	−1.02
Met	2.05	1.53	5.87	−1.17
Glu	2.71	1.44	6.31	−0.91
Gln	2.23	1.83	6.19	−1.00
Lys	2.60	2.12	6.46	−0.79
Arg	2.39	2.09	6.39	−0.86
Leu	1.48	1.19	6.14	−0.88
Asp	2.89	2.70	6.51	−0.63
Asn	3.36	2.88	6.78	−0.52
His	2.29	2.60	6.46	−0.57
Phe	1.82	1.74	6.30	−0.63
Tyr	1.96	1.83	6.40	−0.57
Trp	1.61	1.57	5.96	−0.79
Val	1.36	1.09	5.65	−0.98
Ile	1.0	1.0	5.47	−1.11
Thr	2.66	1.68	6.19	−0.68
Pro	—	0.64	—	—
Ala	—	1.57	—	−1.45
Gly	—	4.65	—	−1.41

[a] D_V and S values, as given in Pal and Chakrabarti (2000b), have been updated. ΔS values are from Pal and Chakrabarti (1999c) (Gly and Ala are using Method 3 and the rest using Method 2).

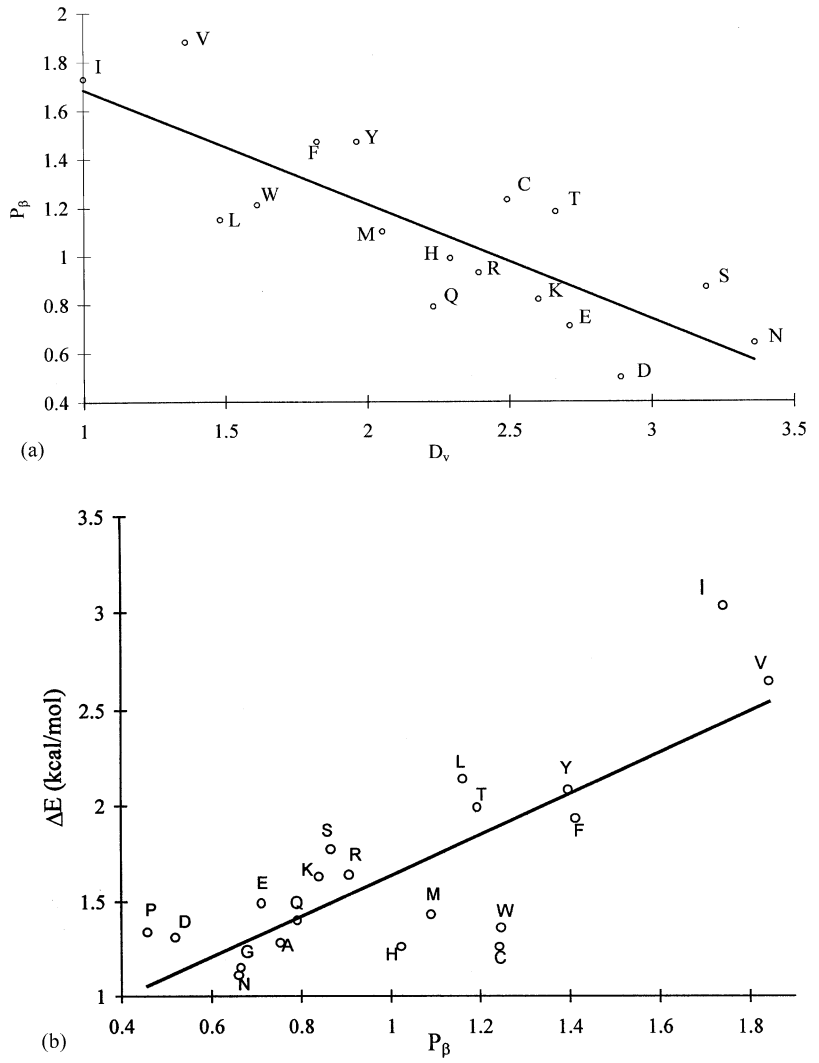

Fig. 43. (a) Plot of P_β against D_V (values taken from Tables 14 and 17). The fitted line has the equation: $P_\beta = -4.73D_V + 2.16$. (b) ΔE (kcal/mol) plotted against P_β (the correlation coefficient between the two variables is 0.80). The equation from the regression analysis is $\Delta E = 1.07P_\beta + 0.56$ (Pal and Chakrabarti, 2000b).

parameter, S, the absolute entropy of the distribution of points (Table 17) also have good correlation with P_β (Pal and Chakrabarti, 2000b).

To appreciate the importance of including χ_1 in the calculation of any residue-specific conformational quality, an equivalent parameter in two-dimension (not including χ_1), the area of dispersion, $D_{A/2D}$, was calculated (Table 17), and compared to different propensity scales, both by excluding Gly, Ala and Pro (when the results are directly comparable to those obtained with D_V), and on including Gly and Ala (Table 18). The former set gives inferior correlation with P_β and Kim (not much change is seen with others), as compared to D_v. (However, the inclusion of Gly and Ala markedly improves the correlation that $D_{A/2D}$ has with scales marked Minor and Smith).

Table 18
Comparison (using correlation coefficients) between D_V (and $D_{A/2D}$) and some β-sheet propensity scales[a]

	P_β	Minor	Minor2	Smith	Kim	Design
D_V	−0.81	−0.52	0.17	0.43	0.84	0.60
$D_{A/2D}$	−0.59	−0.60	−0.23	0.45	0.54	0.43
	−0.35	−0.75	−0.53	0.66	0.57	0.40

[a] P_β values are taken from Table 14. Other references are: Minor and Kim (1994a, b), Smith et al. (1994), Kim and Berg (1993) and Design (Koehl and Levitt, 1999). Two sets of values are given against $D_{A/2D}$; for the first, Gly, Ala and Pro are excluded, whereas the second includes all residues (except Pro).

Thus, as was observed with CS_{AX} (Section 14.1.1) the correlation with Chou–Fasman propensities improves on inclusion of χ_1. The relevance of the use of χ_1 in defining parameters that characterize the protein backbone can be appreciated by comparing Figs. 5b and 5c, showing the distribution of points in two- and three-dimensions for the same residue, Ser. The clustering of points, based on which D_v is calculated, is much more distinct when χ_1 is used as the third dimension.

14.2.2. Energy barrier for the conversion of the β conformation to α and its correlation with P_β

Residues which are poor β-sheet makers have a more diffused distribution of points (larger D_V and S), especially in the region bridging α and β regions (Fig. 5b). It is known that the low-energy regions of molecular potential energy surfaces can be recognized and mapped from the observed distributions (Bürgi and Dunitz, 1983). Inasmuch as the bridging region straddles the β and α regions it can be expected to lie on the reaction path for the transformation of the β conformation to α and vice versa. If it is assumed that the number (N) of points in different regions follows the Boltzmann distribution, the energy ($\Delta E = E_{\text{bridging}} - E$) of the bridging region relative to the β region is given by

$$N_{\text{bridging}}/N_\beta = \exp^{-\Delta E/RT}.$$

ΔE is then an estimate of the energy barrier that the points lying in the β region have to overcome to reach the α conformation. Pal and Chakrabarti (2000b) found that this energy barrier (ΔE) for the conversion of the β conformation to α could be as high as 3 kcal/mol (a large enough value to preclude a thermal equilibrium between α and β conformations) for Ile and Val—residues with the highest β-sheet propensities, whereas the ones with lower propensity have lower barrier. Furthermore, ΔE and P_β values are strongly correlated (Fig. 43b) with the three residues Gly, Ala and Pro (not considered in Fig. 43a) also falling into the general pattern. This suggests that the propensity of a residue to be in the β structure is directly related to the energy barrier that separates its α conformation from β. A portion of the polypeptide chain containing residues having high barriers is likely to have an extended conformation and constitute a strand of the β-sheet in the folded structure. The two correlations given in Fig. 43 can be reconciled on the basis of the chemical structures of the side-chains. The groups of residues with high P_β are branched (at C^β) aliphatic or aromatic (branched at C^γ) residues. Branching in the side-chain close to the main-chain means that there will be greater steric clash in these cases (as compared to long-chain residues) resulting in the reduction in the available conformational space and the concomitant lowering of entropy (first correlation). On the same steric ground, there will be higher energy

barrier linking one backbone conformation (corresponding to a secondary structure) to another (second correlation).

14.3. Implications for protein folding

Analysis in Sections 14.1 and 14.2 reveals that α-helix and β-sheet propensities are more a function of the structural framework of the side chains than their chemical nature. Two parameters reflecting features of the framework—the similarity of the overall ϕ, ψ (and χ_1) distribution of different residues with Ala and the volume enclosing this distribution—are brought into play during the formation of α-helix and β-sheet. The chemical nature has a role in the context dependence—for example, the location of Arg and Lys in helix C-terminus and short, polar residues at the N-terminus for capping interactions, and the determination of which pair of residues in adjacent strands would register in the hydrogen-bonded sites, as opposed to non-hydrogen-bonded sites in β-sheets. The energy barrier for the interconversion between the α and β conformations (Section 14.2.2) can act as an intrinsic constraint to limit the effective size of the conformational space that has to be searched during the folding process (Levinthal, 1968).

15. Loss of the main-chain conformational entropy on folding

When a protein folds into a compact globule the residues lose degrees of freedom as lesser number of conformations can be accessed by the main- and the side-chain—this reduction in conformational entropy opposes the folding process (Brady and Sharp, 1997). As most globular proteins are only marginally stable (the free energy for a folding-unfolding reaction is around 5–20 kcal/mol) (Dill, 1990), it appears that the conformational entropy is the prime deterrent to folding (Privalov and Gill, 1988). The change in the conformational entropy, ΔS, in going from the unfolded state (U) to the folded (F) state is given by

$$\Delta S = S(F) - S(U)$$

and is often subdivided into backbone and side-chain contributions. As side chains adopt restricted number of rotameric states, with torsion angles tightly clustered, it has been relatively easier to determine scales for the conformational entropy change of side chains during protein folding (Creamer and Rose, 1992; Pickett and Sternberg, 1993; Lee et al., 1994b; Blaber et al., 1994; Koehl and Delarue, 1994; Doig and Sternberg, 1995). The contribution of the backbone is difficult to determine accurately and Yang and Honig (1995) estimated that about 2 kcal/mol (at room temperature) are lost upon helix, sheet or turn formation. Backbone conformational entropy for a residue relative to Gly has been evaluated using the sterically allowed area on a ϕ, ψ map for the two residues (Némethy et al., 1966), as well as the probability distribution over the ϕ, ψ space (Stites and Pranata, 1995).

Due to the interdependence of the side-chain and backbone torsion angles (Section 5), and especially since the conformational parameters calculated including χ_1 correlate better with secondary structural propensities (Section 14), χ_1 cannot be neglected in the estimation of the main-chain conformational entropy. Another important consideration is the representation of the native and the unfolded states and there is no unanimity in this regard. Backbone entropies in the unfolded state have been approximated from the ϕ, ψ distribution in crystallographic structures

(Stites and Pranata, 1995). D'Aquino et al. (1996) found the conformational entropies to be extremely similar (6.7 cal/K mol) for Gly and Ala in the folded state (α-helix), whereas Creamer and Rose (1992, 1994), while investigating the difference in conformational entropy between the unfolded and the α-helical states for a series of apolar side chains, simulated χ_1 distributions in the two states using Monte Carlo calculations and found distinct values for both the states. Likewise, the conformational entropy, $S(F)$, has been set to 0 for the side chain in the folded form (Pickett and Sternberg, 1993), which has again been disputed by Koehl and Delarue (1994) for exposed residues. Along the line taken by Creamer and Rose (1992, 1994) to consider two conformational-angle distributions, Pal and Chakrabarti (1999c) devised a procedure to include χ_1 in the estimation of the average loss in main-chain conformational entropy by considering ϕ, ψ, χ_1 distributions in the folded and unfolded states—the former obtained from known structures, and for the latter they took recourse to the conformation-based residue classification (Table 5). As the classification is a reflection of the topological arrangements at least up to the γ position of the side chain, which would mainly contribute to the short-range steric interactions with the backbone, and thus define the ϕ, ψ, χ_1 distribution in the unfolded state, they argued that all the residues in a class would have similar ϕ, ψ, χ_1 distribution, which was obtained by combining the observed distributions of the individual members. This translates to the assumption that the limit of ϕ, ψ, χ_1 angles in the denatured state is set by the extremes of these angles as observed in the native state by any member in the class.

Entropy in the native and unfolded states was obtained using the classical definition proposed by Boltzmann,

$$S = -R\sum_{i=1}^{N} p_i \ln p_i,$$

where p_i is the probability of occurrence in the grid i, with the sum being performed over all the N grid points (of size $10^\circ \times 10^\circ \times 10^\circ$) into which the whole ϕ, ψ, χ_1 space is divided; R is the gas constant. As Gly and Ala, the two residues without χ_1, cannot be included in the above procedure (besides Pro), an alternate methodology in which the fraction of the ϕ, ψ map occupied in the folded and unfolded states—the latter obtained from an energy calculation—was utilized (Pal and Chakrabarti, 1999c). Values presented in Table 17 show that Gly and Ala have nearly identical ΔS, which is in contrast to the results of D'Aquino et al. (1996), which indicate that the presence of the methyl group in Ala reduces the conformational entropy of the peptide backbone by 2.46 cal/K mol with respect to that of Gly. The near equivalence of the two ΔS values is due to the fact that there is almost 33% reduction in going from the accessible area as delineated by the Ramachandran plot and the area actually occupied (Fig. 7)—the respective numbers are 58 and 18% for Gly, and 19 and 6% for Ala (Pal and Chakrabarti, 1999c). As has been discussed in Section 10.2, though a large ϕ, ψ range is available to Gly, in a way it is more restrained than other residues in the values that it takes up in various structures. A similar view has been put forward by Yang and Honig (1995), who attribute the differences in helical propensities between Ala and Gly to hydrophobic and packing interactions involving the C^β with a smaller contribution arising from the increased conformational freedom for Gly in the coil state.

16. Implications for protein engineering

A wealth of information on the dependence of protein stability on the structural context of a residue has been obtained from protein engineering studies on T4 lysozyme (Matthews, 1995). Important lessons can also be learnt from the conformational features of different residues. From Section 6 it is clear that residues with values of $CS_{XX'}$ close to 1 (or generally belonging to the same class) would occur in similarly folded parts of proteins. Thus in protein design, if for some functional reason an amino acid has to be exchanged by another (in a region in which the conservation of torsion angles is important), the two should have high $CS_{XX'}$. If, for example, an extra negative charge should be introduced in a protein by protein engineering, so that the folding is changed as little as possible, Glu would be a better replacement for Ala than Asp (assuming both the side chains can be fitted equally well). [The greater similarity between Ala and Glu than between Ala and Asp is also revealed in their secondary structural propensities (Tables 13 and 14)]. Likewise, residue pairs with small $CS_{XX'}$ values are likely to be located in differently folded regions and are not mutually replaceable without compromising the stability. On the basis of branching at the C^{γ} position, an interchange between Leu and the aromatic residues is deemed to be conservative (Karpusas et al., 1989). However, they belong to different classes (Table 5) with ϕ, ψ angles showing distinctly different behaviour as χ_1 is changed (Figs. 10–12). Leu has very few points in the g^- state, and consequently, if an aromatic residue in this state is mutated to Leu it may be rather destabilizing. Similarly, though they both carry a negative charge, Asp and Glu are not alike in their local interactions with the backbone. Isostructural residues, Val and Thr, are also members of the same class (Table 5), but there are distinct differences in the ϕ, ψ distributions in the three χ_1 states (Fig. 12). While replacing a residue one should check that its main-chain conformation is equally attainable by the substitute residue (at the given χ_1 value). For Gly to non-Gly exchanges (for some representative references, see Pal and Chakrabarti, 1999c), it is to be remembered that besides the aspect of conformational entropy (Section 15), the average ϕ, ψ values of Gly in a particular structural context could be different from non-Gly residues (Section 10.2).

16.1. Cis peptide

Experimental data have been derived on the thermodynamics and kinetics of *cis-trans* isomerization by substituting a Pro at (1′) (Fig. 16a) by a non-Pro residue (see Pal and Chakrabarti, 1999b). However, in addition to the ϕ, ψ values, the (2)–(2′) distances involving $X_{np} - X_{np}$ *cis* residues are consistently longer than in the $X_{np}-$Pro cases in all the turn classes (Table 11), which means that if the Pro in the latter type is mutated to make it $X_{np}-X_{np}$, there may be a need for considerable movement of the main-chain atoms not only of the concerned residue but its neighbours also. Pal and Chakrabarti (1999b) discussed the structural implications of some protein engineering experiments involving *cis* peptide bonds in the wild type protein.

17. Flexibility and residues with multiple conformations

Flexibility of proteins, often referred to as mobility or dynamics, plays a crucial role in the function, for example, in allowing a substrate to enter buried cavities or in the motion of loops

and hinge bending resulting in the formation of an active site (Frauenfelder et al., 1991; Lesk and Chothia, 1984, Feher et al., 1996). With the advent of sophisticated techniques for data collection, meaningful information on protein dynamics can now be extracted from X-ray diffraction analyses (Ringe and Petsko, 1986; Rejto and Freer, 1996). Crystallographic methods provide, in addition to the positional coordinates for protein atoms, atomic displacement parameters (also known more commonly, but less precisely as temperature factors or *B*-values). While Cartesian coordinates define the position at which the probability to find a given atom is maximal, temperature factors describe how diffuse the vibration of the atom is around its equilibrium location (Glusker and Trueblood, 1985). By treating proteins as elastic ellipsoids, a differential equation model has been derived for the increase of the *B*-values from the core to the surface of the protein (Bhaskaran and Ponnuswamy, 1988). Parthsarathy and Murthy (1999) have expressed the frequency distribution of *B*-values as the summation of two Gaussian functions which are characteristic of protein structures. These values have been utilized to calculate flexibility indices of amino acids useful for the prediction of antigenicity and flexibility of polypeptide segments (Karplus and Schulz, 1985; Ragone et al., 1989; Vihinen et al., 1994).

B-factors for an individual side chain may be high for various reasons, including thermal motion, static disorder and side-chain misfitting. For some residues there may be error in the interpretation of electron-density maps, especially at low resolution. For example, Leu can have nearly coincident side-chain atom positions and yet exhibit very different χ angles. Lee and Subbiah (1991) have noted that if χ_1 is altered by 30–40° and χ_2 is changed by 140–150°, the C^{δ} atoms are nearly superimposable on the initial structure, though C^{γ} gets shifted slightly. This, however, results in a poor fit of C^{γ} into the electron density and a higher *B*-factor (than those for the C^{δ}s) for the misfit rotamer relative to the genuine one (Lovell et al., 2000). Carugo and Argos (1997) have shown that the temperature factors show a tendency to be larger in side chains with unfavourable local conformations rather than in those displaying conformational energy minima. However, when the nonrotameric residues are packed in clusters in protein core, they exhibit lower average temperature factors compared to isolated nonrotameric residues (Heringa and Argos, 1999). MacArthur and Thornton (1999) also analysed the *B*-factors in the disfavoured high-energy barrier region between the rotameric wells and found these to have larger than average values. However this, according to them, reflects local conformational flexibility and static disorder, which at low resolution is interpreted as a single distorted conformer, but can be resolved at high resolution into the individual components.

Crystal structure is the average of the contents of all the unit cells in the crystals (typically of the order of 10^{15}) over the data collection period. If appreciable populations of a side chain or a loop region exist over different but distinct states it is possible to identify the alternate conformations in the electron density map (Rejto and Freer, 1996). Accessibility of alternative conformational states is important for protein function, including assembly, regulation of biological activity and enzymatic catalysis (Gerstein et al., 1994). Roughly 30% of all side chains in the X-ray structure of the small protein, crambin, have multiple conformational substates (Stec et al., 1995). Freezing the crystals and the use of synchroton radiation may provide diffraction data at atomic resolution, and this usually results in a significant reduction of the average temperature factors of protein atoms (Longhi et al., 1998). For example, 0.94 Å data from the crystals of 2[4Fe–4S] ferredoxin from *C. acidurici* helped to identify the electron density of the exposed loop encompassing residues 25–29, for which no clear electron density was visible in the 1.84 Å structure—the

disordered loop was modelled as two alternative main-chain conformations with equal occupancy (Dauter et al., 1997). Packing forces can stabilize alternative conformational substates in protein crystals, and comparison of protein structures from different crystal forms has been used to infer the existence of distinct low energy substates in solution (Rejto and Freer, 1996). Kossiakoff et al. (1992) performed a molecular dynamics study of bovine pancreatic trypsin inhibitor (BPTI) and compared with five high-resolution X-ray structures of BPTI, each in a different crystal packing environment. There was significant correlation between the flexible regions observed in the simulation and the differences in the five BPTI crystal forms.

Chakrabarti and Pal (1998) analysed the residues which are observed in different conformational states of the side chain and features of their ϕ, ψ, χ_1 distribution. The same analysis is repeated here with a larger dataset of structures (Table 1) and the results, along with those from Section 12, are used to comment upon the etiology of protein thermostability.

17.1. Residues exhibiting two different conformations of the side chain, their ϕ, ψ values and secondary structures

Seven hundred and sixty nine residues were identified in two distinct conformations of the side chain, of which 245 had $\Delta\chi_1 \leqslant 30^\circ$. (A further 18 were found in more than two orientations and are not considered here.) A greater percentage of long chain residues have χ_1 varying by less than 30° (i.e., the side chain stays in the same conformational well) (Fig. 44a). When $\Delta\chi_1 > 30^\circ$, the torsion angle usually spans two distinct conformational states of χ_1, and Ser is overwhelmingly amenable to such side-chain rotations. The facile movement of its side chain must be made possible by the short length, coupled with the ease with which the hydroxyl group can be involved in the hydrogen bond interaction with a neighbouring group. Compared to the observation of Chakrabarti and Pal (1998) the relative number of Thr has come down considerably. Next to Ser, Glu and Val have the highest numbers of cases spanning two χ_1 states. From a comparison of χ_1 distribution of residues in high-resolution structures (resolution $\leqslant 2.0$ Å) with that from lower resolution (> 2.0 Å) structures MacArthur and Thornton (1999) found that two prominent residues which are likely to be modelled as occupying two distinct side-chain conformations as the resolution improves are Ser and Leu. Though only high-resolution structures have been used here, Leu is not among the top 5 residues with alternate conformations. It can also be mentioned that although Leu is known to suffer from the misfitting of the side chain (discussed earlier), it mainly affects the conformational state of χ_2 angle, and not χ_1 (Lovell et al., 2000), which is considered here.

When two different conformational states are taken up in the two orientations of the side chain, the most favourable combination is tg^+ for classes I–III residues (Fig. 44b). This is because, for these residues t and g^+ are the most populated states (Table 6), and consequently, a change in conformation between these two states has a higher probability. For class IV residues and Ser, t is the least populated state and g^+g^- is the most favoured combination. The percentage occurrences in different secondary structural elements of residues with side chains existing in two conformational states are Helix : Sheet : Turn : Rest = 44 : 22 : 19 : 15 (Fig. 44c), whereas these are found in the ratio 36 : 24 : 21 : 20 in the whole database (Table 6). These show that residues in regular secondary structures and not in irregular structures have a higher number of side chains in multiple conformations.

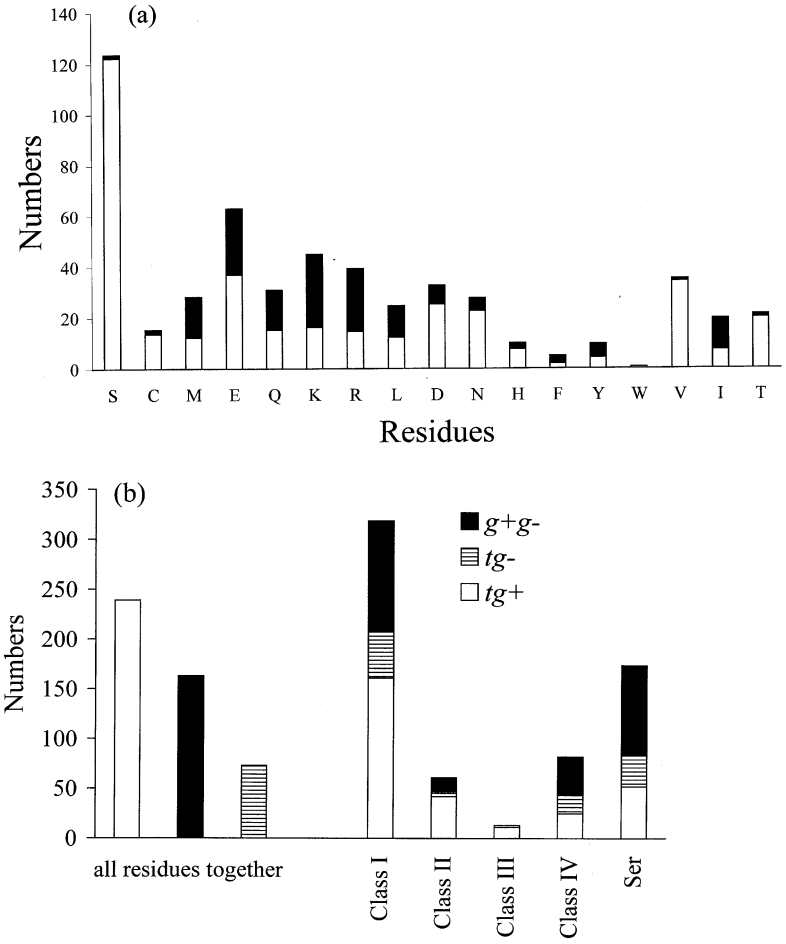

Fig. 44. (a) Histogram of residues with side chains modelled in two different orientations (with distinct χ_1 angles). The shaded part corresponds to the cases where the two χ_1 angles differ by less than 30°. (b) Histogram showing the various combinations of distinct χ_1 states that different residues can occupy (overall and broken into classes, as well as of Ser). (c) The secondary structures of residues existing in two χ_1 states; Helix includes all types of helices; Sheet corresponds to residues with tag E and B in the DSSP output; Turn, S and T, and Rest denotes the remaining residues with no regular secondary structure. The overall distribution of the secondary structural elements among all such residues is given in the inset. (d) ϕ, ψ points for class I residues with the side chain occupying two conformational states in the crystal structure. Contiguous boundaries for the core regions in the three Ramachandran plots (enclosing the $5 \times 5\,\phi$, ψ blocks containing more than 2 points) at different χ_1 angles, as derived by Chakrabarti and Pal (1998), are indicated.

When the side-chain flips between two conformational states the ϕ, ψ values of the residue must lie in a region that will allow the side-chain to adopt both the χ_1 angles, i.e., a region in the map that is common to the Ramachandran plots of the residue at the two χ_1 angles. Taking class I residues as an example, the Ramachandran plots in the t and g^+ states have a larger fraction of the total map area occupied by points (as compared to the g^- state) (Fig. 12), and consequently a change in conformation between these two states should have the highest probability, as is indeed

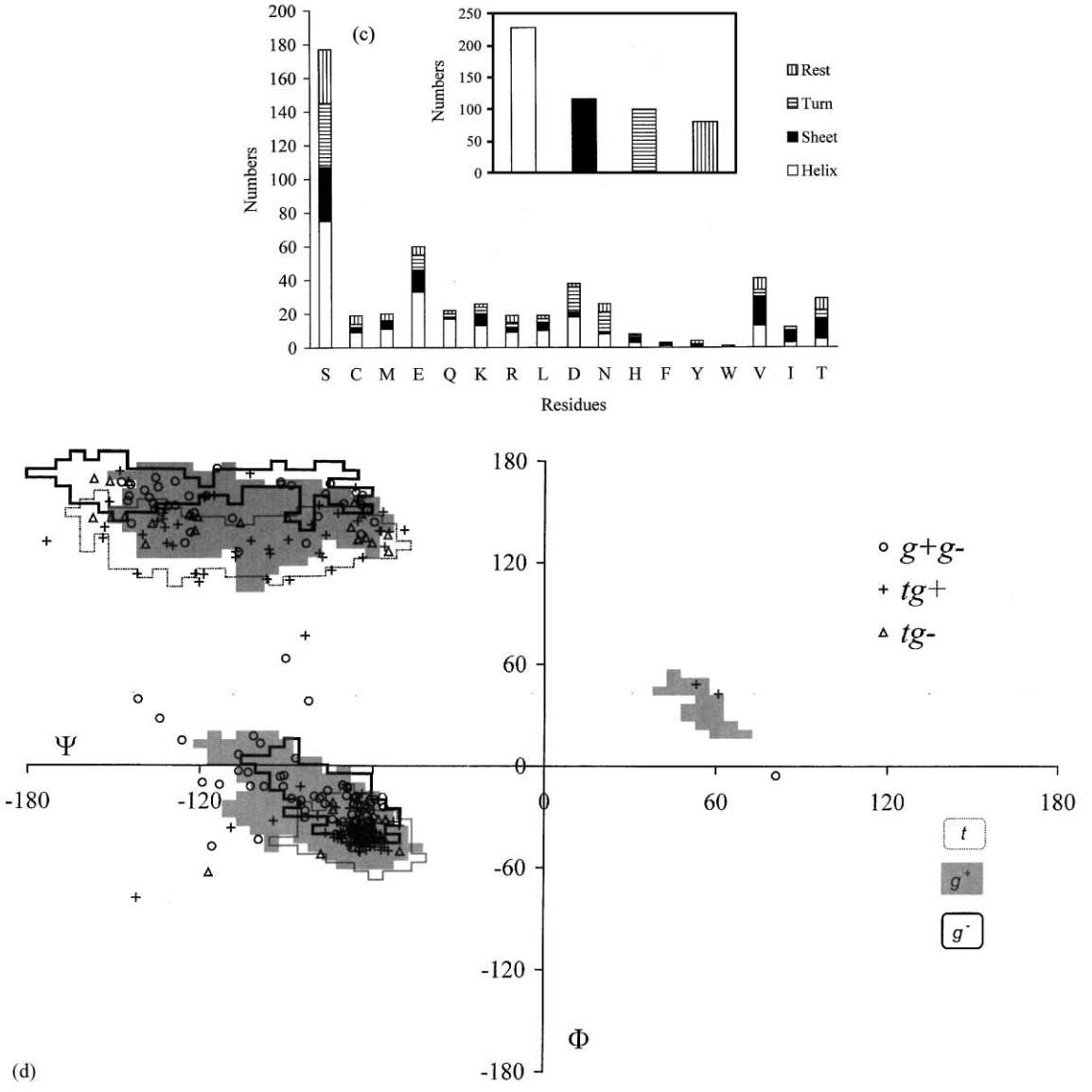

Fig. 44. (*Continued*).

the case (Fig. 44d). Besides a lower percentage of the plot area being occupied, there is very little overlap in the B region of the Ramachandran plots in the t and g^- states, and consequently the number of residues that can simultaneously reside in these two states is low. This shows that there is some restriction in the backbone torsion angles if the side chain has to exist in two conformational states without a change in the main-chain conformation. Interestingly, in an analogous situation involving the flexibility of residues in different crystal forms, it is observed that the four residues (Phe4, Glu64, Asp72 and Phe104), that show the maximum side-chain adaptability in the 25 crystal forms of T4 lysozyme (Zhang et al., 1995), span the states g^+

and t with the main-chain geometry ($\varphi \approx -65^\circ$, $\psi \approx -45^\circ$) quite close to the core helical region.

17.2. Residues with multiple backbone conformations

The simultaneous occurrence of one or more residues (in a stretch) in more than one location (such that either one or both the ϕ, ψ angles of the residues are altered) was investigated. The inherent tendency of a residue to have a flexible backbone is indicated if it can have more than one position for the backbone atoms, independent of its neighbours in the chain. As with the side chain (Section 17.1), Ser is also quite prominent in exhibiting a flexible backbone even when its neighbours are not disordered (unshaded part of the bars in Fig. 45a). It is interesting to see that a constrained residue like Pro can also have alternate locations for its backbone atoms. In a way it may be easier to identify such positions for Pro, whereas other residues, especially in loops, may be so flexible that the individual atomic locations cannot be seen in electron density maps. Gly and

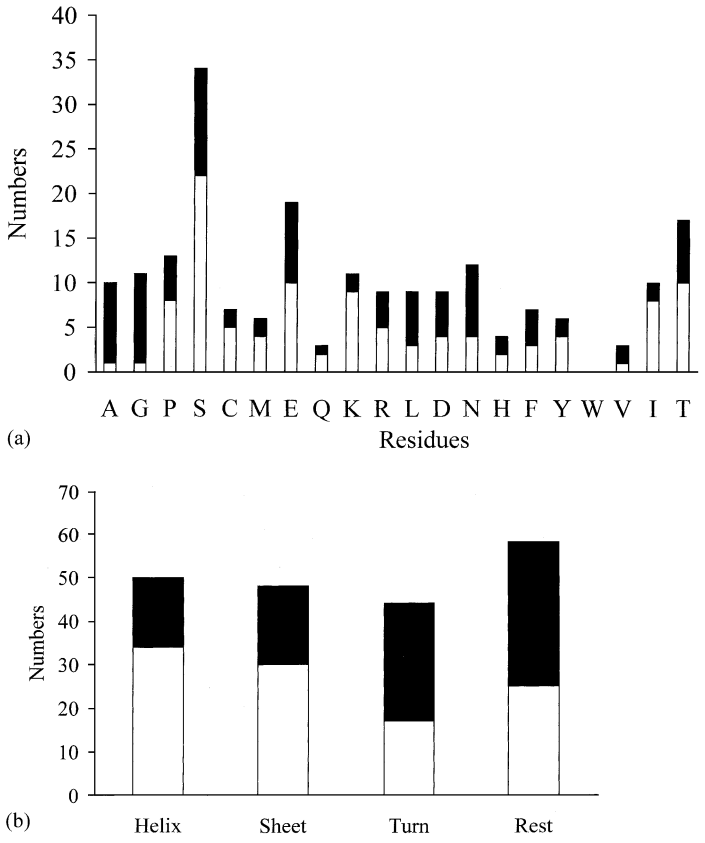

Fig. 45. (a) Histogram of residues with the backbone modelled in more than one conformation and (b) the distribution of the secondary structural elements for all such residues. The bar without shading represents isolated residues in multiple conformation, whereas when more than one consecutive residue have similar features, they are represented in shade.

Ala are likely to be involved when a stretch of (two or more) residues are observed in multiple conformations (shaded part, Fig. 45a). As expected, residues with flexible backbone are exposed. A single, isolated residue having more than one main-chain geometry is more likely to be found in a helix or β-sheet, whereas a continuous stretch of such residues can be encountered more in turns or non-regular regions (Fig. 45b).

17.3. Implications for thermostability

The stability of proteins from thermophilic microorganisms (with optimum temperatures of growth about 60°C) has been studied on the basis of thermodynamics, sequence and the three-dimensional structure to discover strategies for thermal adaptation (Jaenicke and Böhm, 1998; Ladenstein and Antranikian, 1998). Identification of a variety of intrinsically hyperstable enzymes from hyperthermophilic organisms with optimal growth temperatures of 100°C and above (Rees and Adams, 1995) has generated considerable interest for their possible applications in biotechnology (Ludlow and Clark, 1991) and there have been numerous attempts to engineer mesophilic enzymes to be stable at higher temperatures (Lee and Vasmatzis, 1997), maintaining at the same time the conformational flexibility required for the enzyme function (Závodszky et al., 1998). In spite of some striking examples—Van den Burg et al. (1998) have engineered moderately thermostable thermolysin-like protease into a hyperstable enzyme by substituting only eight amino acid residues (out of 319), and Malakauskas and Mayo (1998) achieved a shift in the thermal transition from 83°C for the wild-type protein to more than 99°C for the seven-fold mutant of the *Streptococcal* protein Gβ1—a generally applicable set of rules by which a mesophilic protein could be made thermostable does not seem to exist (Usher et al., 1998).

A pairwise comparison of the three-dimensional structures of homologous proteins with different thermal stabilities has been used to discover determinants that lead to the enhancement of protein thermostability (Querol et al., 1996). On going from lower to higher growth temperatures, numerous differences have been reported (Jaenicke and Böhm, 1998): the clustering of (intra- and/or inter-subunit) ion pairs; improved packing of the hydrophobic core (increased van der Waals interactions); additional networks of hydrogen bonds and enhanced secondary structure propensity; increased helix-dipole stabilization; an increased polar surface area; a decreased number and total volume of cavities; and burying hydrophobic surface area by either tightening interdomain and intersubunit contacts or by increasing the state of association (for some representative references discussing these points, see: Russell et al., 1997; Vogt et al., 1997; Wallon et al., 1997; Auerbach et al., 1998; Facchiano et al., 1998; Karshikoff and Ladenstein, 1998). Though electrostatic interactions involving extensive ion-pair networks seem to be the most common strategy adopted for enhanced thermostability (Vetriani et al., 1998; Xiao and Honig, 1999; Lebbink et al., 1999), different proteins have adapted to different thermal environments by a variety of evolutionary devices. For example, a structural comparison of two glutamate dehydrogenases from the hyperthermophiles *T. maritima* and *P. furiosus* with the enzyme from the mesophilic bacterium *Clostridium symbiosum* has revealed that in the former thermophile the subunit interactions in the enzyme are dominated by ionic interactions realized by large salt bridge networks, whereas in the latter the number of intersubunit ion pairs is reduced and the hydrophobic interactions are increased (Knapp et al., 1997). Recently, Thompson and Eisenberg (1999) have suggested loop deletion as a contributing factor for thermal stability through their

Table 19
Comparison of amino acid compositions (in percentages) of mesophiles and thermophiles (Deckert et al., 1998)[a]

Amino acid	Mesophiles	Thermophiles
Charged residues (DEKRH)	24.11	29.84
Polar/uncharged residues (GSTNQYC)	31.15	26.79
Hydrophobic residues (LMIVWPAF)	44.74	43.36

[a] One-letter abbreviations of the amino acid residues are given in brackets.

comparative analysis of 20 different full genome sequences corresponding to organisms having different optimum temperatures.

A large enough data set from complete sequencing of genomes from different organisms has made it possible to visualize certain trends in amino-acid usage (Table 19). Some correlations seem to hold—compared to mesophiles, genomes of thermophiles encode higher levels of charged amino acids, primarily at the expense of uncharged polar residues (Deckert et al., 1998). Haney et al. (1999) have compared sequences of 115 proteins from *Methanococcus jannaschii* (growth temperature, 85°C) with known sequences of mesophilic *Methanococcus* species. Their analysis has indicated that the composition of Ser, Thr, Asn and Gln is reduced in thermophiles while that of Glu, Arg and Lys increased. Based on an analysis of 70 sequences from six different protein families Menéndez-Arias and Argos (1989) observed that the top two residue exchanges leading to an increased thermal stability are Lys to Arg and Ser to Ala occurring in helices. Mutations of Ser to Ala and Thr to Ala in mesophilic lactate dehydrogenase have been reported to enhance the stability of the enzyme by 20°C or more when compared to the wild type enzyme (Kotik and Zuber, 1993).

From the above it appears that residues like Ser, Thr, Asn are discriminated against in thermophiles. An increased rate of deamidation at higher temperature may provide a rationale for the lower levels of Asn (Jaenicke and Böhm, 1998), but there is no convincing argument for the other two residues. In this connection, results presented in Sections 12.2, 17.1 and 17.2 (Figs. 41b, 44a and 45a) are enlightening (Pal and Chakrabarti, 2001). The polar, uncharged residues which are avoided in thermophiles (Table 19) are found in higher frequencies in disallowed ϕ, ψ regions. Interestingly, the top five replacements that are most biased in number between mesophile and thermophile proteins are Ser → Ala, Lys → Arg, Ser → Lys, Leu → Ile and Asp → Glu (Haney et al., 1999) and all of these involve a change from a residue with a greater propensity of occurrence in disallowed conformations to one with a lower propensity. Due to the greater motion at higher temperature, residues which have a higher propensity to be in a disallowed region may be trapped in such a conformation disrupting the native structure. Additionally, Ser has by far the maximum number of occurrences in multiple conformations of the side chain and the main chain. Although the conformers may be isoenergetic, a change from one state to another may produce enough fluctuation in the protein structure at high temperatures leading to unfolding. Thus a consideration of residues occurring in disallowed regions and in multiple conformations has a bearing on the thermostability of proteins.

18. Prediction of side-chain conformation

Homology modelling for the prediction of protein structure (Blundell et al., 1987; Greer, 1991) involves the recognition of a level of global sequence homology between the structure to be predicted (the target structure) and a protein whose structure is known (the template structure), and then the latter is used to build the backbone atoms of the aligned sequences. Next the main-chain coordinates are predicted for regions where there is no homology to the template structure. Finally, the side-chain conformations are predicted. The simultaneous prediction of all side-chain conformations is a complex combinatorial problem and various strategies have been adopted to address the issue (Bower et al., 1997; Dunbrack, 1999 and references therein). However, they all rely on a decision about the conformational space each side chain is allowed to sample, the energy function for evaluating solutions, and the choice of moves from one possible solution to the next. The dimension of the problem is reduced by incorporating knowledge-based information on the conformations that the various side chains can adopt. Ponder and Richards (1987) compiled from 19 highly refined crystallographic protein structures a side-chain rotamer library representing the most probable combinations of all χ angles for each residue type. By dynamic cluster analysis Tuffery et al. (1991) derived an expanded set of rotamers from 53 protein structures and developed optimization methods to predict side-chain coordinates. Holm and Sander (1991) performed the Monte Carlo optimization with simulated annealing and precalculation of all possible rotamer pairwise interaction energies using the rotamer library of Tuffery et al. (1991) along with a few additional rotamers for aromatic residues. Different aspects of the local environments have also been considered in deriving empirical rules for the prediction of side-chain packing in proteins (Levitt, 1992; Eisenmenger et al., 1993; Laughton, 1994). Due to the interdependence of backbone and side-chain conformations, Dunbrack and Karplus (1993) developed a backbone-dependent rotamer library (instead of looking at ϕ, ψ ranges at different χ_1 angles in Fig. 12, this database enumerates the rotameric probabilities of each amino acid at 20° bins of ϕ, ψ), which has further been expanded (Dunbrack and Cohen, 1997). The use of this library results in considerable improvement in the correct initial placement of the side chain (Bower et al., 1997). However, it seems there are scopes for further improvement with the incorporation of χ preferences that depend on the main-chain secondary structure, the location in specific regions within the secondary structure, or being at chain termini. For example, Pro is almost equally distributed between the g^+ and g^- states, but in helix it is 69% in the g^+ state (Table 6). Likewise, though the overall preference of Trp in the helix is for the t state, at the N-end it prefers to be in the g^- state (Table 13).

Bower et al. (1997) compared the accuracy of their algorithm with two other publicly available softwares, the Monte Carlo sampling program of Holm and Sander (1991,1992) and the mean field theory algorithm described by Koehl and Delarue (1994). For all the methods, Ser is the residue with the worst χ_1 prediction, followed by Glu. Interestingly, Ser is also the residue with the maximum number of occurrences with multiple side-chain conformations, also followed by Glu (Fig. 44a). In Fig. 44d the boundaries of ϕ, ψ angles at three different χ_1 angles for class I residues are shown. Regions common to two χ_1 states are more likely to be occupied by residues that can assume any one of the two χ_1 s, and consequently these are the residues for which χ_1 is susceptible to be misplaced.

19. Conformation in the validation of protein crystal structure

A protein structure derived from experimental data will be subject to many sources of error, both experimental and in the interpretation of results. Some quality-control techniques are, therefore, needed to assess the validity of such models (Laskowski et al., 1998; Dodson et al., 1998; Kleywegt, 2000). For models solved by X-ray crystallography, besides the quality of data, there are at least three categories of criteria to judge the model. The first quantifies the agreement between the model and the experimental diffraction data. The second compares the geometry of the model to the expected distribution derived from small-molecule and protein crystal structures. The third relies on matching nonbonded contacts and residue environments with the expectations derived from the database. Any criterion that has been used explicitly or implicitly during model refinement does not provide a truly independent check on the quality of the model. Normally, during X-ray refinement, geometrical parameters (like bond lengths, bond angles, planarity of different groups, chirality of tetrahedral carbon atoms, etc.) are greatly restrained, while the conformational angles are usually freely rotatable. As a result, the deviation of ϕ, ψ angles from the preferred area of the Ramachandran plot is a useful indicator of the quality of a model.

The program, PROCHECK (Laskowski et al., 1993) divides the Ramachandran plot into four types of area: most favoured, additional allowed, generously allowed and disallowed—the percentages of available area and population in these regions being 11.0 and 81.9, 28.0 and 14.8, 31.0 and 2.0, and 30.0 and 1.3, respectively (Morris et al., 1992). Kleywegt and Jones (1996) have suggested a division of the plot into two areas: core and non-core, the former with 19.7% of the entire plot area accounts for 98% of all non-glycine residues; about 91% of all structures have 10% or fewer outliers. Hooft et al. (1997) have also used the Ramachandran plot to derive an 'objective score' for a protein structure according to where in the plot each of its residue falls. However, as has been mentioned in Section 12, outliers in a Ramachandran plot are not necessarily errors—there are definite patterns involving the residues occurring in the disallowed region and their neighbours.

All the validation tools dealing with conformation do not distinguish between different amino acid residues. But Figs. 10–12 indicate that the allowed ϕ, ψ regions for different classes (Table 5) of residues are quite distinct and also depend on the χ_1 rotameric state. Consequently, the programs can be made more rigorous by comparing the ϕ, ψ values of a model structure with class-based, χ_1-dependent Ramachandran plots.

20. Conclusions

The distribution of ϕ, ψ angles of a given residue lies within the boundary predicted by the Ramachandran plot. However, the introduction of the third dimension, *viz*., the side-chain torsion angle (χ_1), shows that at a given χ_1, the ϕ, ψ distribution is considerably different from the map at another χ_1 and such maps are usually distinct for different residues and can indeed be used for a conformational classification of amino acid residues. This reduces the protein folding alphabet to 7 (or 8, if Ser is considered separately, Table 5) classes of amino acids and can be used in modelling and protein engineering studies for amino acid substitutions causing the minimum

perturbation to local protein fold. This classification reflects the different influence of residues on local protein structures and/or their preferential occurrence in differently folded parts of the molecules. The three-dimensional ϕ, ψ, χ_1 maps (for non-Gly/Ala/Pro residues) have features that can be used to derive parameters which correlate with secondary structural propensities of the residues. The distribution of the ϕ, ψ, χ_1 angles can be influenced by specific interactions between the side-chain and main-chain atoms (notably for Cys, Asp and Asn), by a flanking Pro residue, the secondary structure (and the position in the secondary structure) the residue is located in. The interrelationship between the side-chain and main-chain conformation is exemplified beautifully by Pro residue where a shift of ϕ to a more negative value in going from the *trans* prolyl bond to a *cis* bond causes the puckering of the ring in the *cis* isomer to be predominantly in the DOWN conformation. Non-Gly residues occurring in the normally disallowed region of the Ramachandran plot are clustered in distinct regions. Ser is the residue which is quite conspicuous in its occurrence in the disallowed region or in having multiple main-chain or side-chain conformations. There are some similarities in the types of residues that abound in or avoid the neighbourhood of *cis* peptide bonds and residues in disallowed Ramachandran region. Residues which are discriminated against in thermophilic proteins are found to have the highest propensities to be in disallowed region or in multiple side-chain and main-chain conformations. The ϕ, ψ, χ_1 distribution of the two terminal residues in the polypeptide chain is different from the general distribution. The electrostatic field due to the amino end causes the main-chain torsion angle, ψ, of the first residue to be in an extended conformation, which in turn makes the next residue a more likely location for the start of a β-strand rather than an α-helix. The greater preference for the β-conformation is observed for six locations from the amino end. Identification of systematic features in protein conformations would facilitate protein modelling.

Acknowledgements

We are grateful to the crystallographers who deposited the PDB files (Table 1) on which the work is based, but had to remain anonymous because of the constraint in space. The critical comments of the reviewer are appreciated. Financial support was provided in the form of a grant and a fellowship by the Council of Scientific and Industrial Research and the computational facilities by the Bioinformatics Centre (Department of Biotechnology).

References

Abagyan, R., Totrov, M., 1994. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J. Mol. Biol. 235, 983–1002.

Abola, E.E., Sussman, J.L., Prilusky, J., Manning, N.O., 1997. Protein Data Bank archives of three dimensional macromolecular structures. Methods Enzymol. 277, 556–571.

Allen, F.H., Baalham, C.A., Lommerse, J.P.M., Raithby, P.R., 1998. Carbonyl–carbonyl interactions can be competitive with hydrogen bonds. Acta Crystallogr. B54, 320–329.

Allen, F.H., Kennard, O., 1993. 3D search and research using the Cambridge Structural Database. Chem. Des. Autom. News 8, 31–37.

Altona, C., Sundaralingam, M., 1972. Conformational analysis of a sugar ring in nucleosides and nucleotides. A new description using the concept of pseudorotation. J. Am. Chem. Soc. 94, 8205–8212.

Altschul, S.F., 1991. Amino acid substitution matrices from an information theoretic perspective. J. Mol. Biol. 219, 555–565.

Argos, P., Palau, J., 1982. Amino acid distribution in protein secondary structures. Int. J. Peptide Protein Res. 19, 380–393.

Ashida, T., Kakudo, M., 1974. Conformations of prolyl residues in oligopeptides. Bull. Chem. Soc. Jpn. 47, 1129–1133.

Auerbach, G., Ostendorp, R., Prade, L., Korndörfer, I., Dams, T., Huber, R., Jaenicke, R., 1998. LDH from the hyperthermophilic bacterium *Thermotoga maritima*: the crystal structure at 2.1 Å resolution reveals strategies for intrinsic stabilization. Structure 6, 769–781.

Aurora, R., Creamer, T.P., Srinivasan, R., Rose, G.D., 1997. Local interactions in protein folding: lessons from the α-helix. J. Biol. Chem. 272, 1413–1416.

Aurora, R., Rose, G.D., 1998. Helix capping. Protein Sci. 7, 21–38.

Aurora, R., Srinivasan, R., Rose, G.D., 1994. Rules for α-helix termination by glycine. Science 264, 1126–1130.

Avbelj, F., Moult, J., 1995. Role of electrostatic screening in determining protein main chain conformational preferences. Biochemistry 34, 755–764.

Baker, E.N., Hubbard, R.E., 1984. Hydrogen bonding in globular proteins. Prog. Biophys. Mol. Biol. 44, 97–179.

Balasubramanian, R., Lakshminarayanan, A.V., Sabesan, M.N., Tegoni, G., Venkatesan, K., Ramachandran, G.N., 1971. Studies on the conformation of amino acids. VI. Conformation of the proline ring as observed in crystal structures of amino acids and peptides. Int. J. Protein Res. III, 25–33.

Barlow, D.J., Thornton, J.M., 1988. Helix geometry in proteins. J. Mol. Biol. 201, 601–619.

Benedetti, E., Morelli, G., Nemethy, G., Scheraga, H.A., 1983. Statistical and energetic analysis of side-chain conformations in oligopeptides. Int. J. Peptide Protein Res. 22, 1–15.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. Nucleic Acids Res. 28, 235–242.

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., 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. 112, 535–542.

Betz, S.F., Raleigh, D.P., DeGrado, W.F., 1993. *De novo* protein design: from molten globules to native-like states. Curr. Opin. Struct. Biol. 3, 601–610.

Bhaskaran, R., Ponnuswamy, P.K., 1988. Positional flexibilities of amino acid residues in globular proteins. Int. J. Peptide Protein Res. 32, 241–255.

Bhat, T.N., Sasisekharan, V., Vijayan, M., 1979. An analysis of side-chain conformation in proteins. Int. J. Peptide Protein Res. 13, 170–184.

Blaber, M., Zhang, X.-J., Lindstrom, J.D., Pepiot, S.D., Baase, W.A., Matthews, B.W., 1994. Determination of α-helix propensity within the context of a folded protein. Sites 44 and 131 in bacteriophage T4 lysozyme. J. Mol. Biol. 235, 600–624.

Blaber, M., Zhang, X.-J., Matthews, B.W., 1993. Structural basis of amino acid α helix propensity. Science 260, 1637–1640.

Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E., Thornton, J.M., 1987. Knowledge-based prediction of protein structures and the design of novel molecules. Nature 326, 347–352.

Bower, M.J., Cohen, F.E., Dunbrack, R.L., Jr., 1997. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J. Mol. Biol. 267, 1268–1282.

Brady, G.P., Sharp, K.A., 1997. Entropy in protein folding and in protein-protein interactions. Curr. Opin. Struct. Biol. 7, 215–221.

Branchini, B.R., Nemser, A.R., Zimmer, M., 1998. A computational analysis of the unique protein-induced tight turn that results in posttranslational chromophore formation in green fluorescent protein. J. Am. Chem. Soc. 120, 1–6.

Brandts, J.F., Lin, L.-N., 1986. Proline isomerization studied with proteolytic enzymes. Methods Enzymol. 131, 107–126.

Brooks, C.L. III., Case, D.A., 1993. Simulations of peptide conformational dynamics and thermodynamics. Chem. Rev. 93, 2487–2502.

Brown, J., Klee, W., 1971. Helix-coil transition of the isolated amino terminus of ribonuclease. Biochemistry 10, 470–476.

Bürgi, H.B., Dunitz, J.D., 1983. From crystal statics to chemical dynamics. Acc. Chem. Res. 16, 153–161.
Bürgi, H.B., Lehn, J.M., Wipff, G., 1974. An ab initio study of nucleophilic addition to a carbonyl group. J. Am. Chem. Soc. 96, 1956–1957.
Burley, S.K., Petsko, G.A., 1988. Weakly polar interactions in proteins. Adv. Protein Chem. 39, 125–189.
Carugo, O., Argos, P., 1997. Correlation between side chain mobility and conformation in protein structures. Protein Eng. 10, 777–787.
Chakrabarti, P., 1989. Geometry of interaction of metal ions with sulfur-containing ligands in protein structures. Biochemistry 28, 6081–6085.
Chakrabarti, P., 1991. Does helix dipole have any role in binding metal ions in protein structures? Arch. Biochem. Biophys. 290, 387–390.
Chakrabarti, P., 1993. Anion binding sites in protein structures. J. Mol. Biol. 234, 463–482.
Chakrabarti, P., 1994. An assessment of the effect of the helix dipole in protein structures. Protein Eng. 7, 471–474.
Chakrabarti, P., Chakrabarti, S., 1998. C–H$\cdots$O hydrogen bond involving proline residues in α-helices. J. Mol. Biol. 284, 867–873.
Chakrabarti, P., Pal, D., 1997. An electrophile–nucleophile interaction in metalloprotein structures. Protein Sci. 6, 851–859.
Chakrabarti, P., Pal, D., 1998. Main-chain conformational features at different conformations of the side-chains in proteins. Protein Eng. 11, 631–647.
Chakrabartty, A., Baldwin, R.L., 1995. Stability of α-helices. Adv. Protein Chem. 46, 141–176.
Chan, H.S., 1999. Folding alphabets. Nat. Struct. Biol. 6, 994–996.
Chandrasekaran, R., Ramachandran, G.N., 1970. Studies on the conformation of amino acids. XI. Analysis of the observed side group conformations in proteins. Int. J. Protein Res. 2, 223–233.
Chothia, C., Finkelstein, A.V., 1990. The classification and origins of proteins folding patterns. Annu. Rev. Biochem. 53, 1007–1039.
Chou, P.Y., Fasman, G.D., 1974. Conformational parameters for amino acids in helical, β-sheet and random coil regions calculated from proteins. Biochemistry 13, 211–222.
Chou, P.Y., Fasman, G.D., 1978. Prediction of the secondary structure of proteins from their amino acid sequence. Adv. Enzymol. 47, 45–148.
Cohen, B.I., Presnell, S.R., Cohen, F.E., 1993. Origins of structural diversity within sequentially identical hexapeptides. Protein Sci. 2, 2134–2145.
Creighton, T.E., 1993. Proteins: Structures and Molecular Properties. Freeman, New York (Chapters 1 and 2).
Creamer, T.P., Rose, G.D., 1992. Side-chain entropy opposes α-helix formation but rationalizes experimentally determined helix-forming propensities. Proc. Natl. Acad. Sci. USA 89, 5937–5941.
Creamer, T.P., Rose, G.D., 1994. α-Helix–forming propensities in peptides and proteins. Proteins 19, 85–97.
Cremer, D., Pople, J.A., 1975. A general definition of ring puckering coordinates. J. Am. Chem. Soc. 97, 1354–1358.
D'Aquino, J.A., Gómez, J., Hilser, V.J., Lee, K.H., Amzel, L.M., Freire, E., 1996. The magnitude of the backbone conformational entropy change in protein folding. Proteins 25, 143–156.
Dasgupta, S., Bell, J.A., 1993. Design of helix ends. Amino acid preferences, hydrogen bonding and electrostatic interactions. Int. J. Peptide Protein Res. 41, 499–511.
Dauter, Z., Wilson, K.S., Sieker, L.C., Meyer, J., Moulis, J.M., 1997. Atomic resolution (0.94 Å) structure of *Clostridium acidurici* ferredoxin. Detailed geometry of [4Fe–4S] clusters in a protein. Biochemistry 36, 16065–16073.
Davies, D.R., 1964. A correlation between amino acid composition and protein structure. J. Mol. Biol. 9, 605–609.
Dayhoff, M.O., Schwartz, R.M., Oreutt, B.C., 1978. A model of evolutionary change in proteins. In: Dayhoff, M.O. (Ed.), Atlas of Protein Sequence and Structure, Vol. 5 (suppl. 3). National Biomedical Research Foundation, Washington, pp. 345–352.
Dayhoff, M.O., Barker, W.C., Hunt, L.T., 1983. Establishing homologies in protein sequences. Methods Enzymol. 91, 524–545.
Deane, C.M., Allen, F.H., Taylor, R., Blundell, T.L., 1999. Carbonyl–carbonyl interactions stabilize the partially allowed Ramachandran conformations of asparagine and aspartic acid. Protein Eng. 12, 1025–1028.

Deckert, G., Warren, P.V., Gaasterland, T., Young, W.G., Lenox, A.L., Graham, D.E., Overbeek, R., Snead, M.A., Keller, M., Aujay, M., Huber, R., Feldman, R.A., Short, J.M., Olsen, G.J., Swanson, R.V., 1998. The complete genome of the hyperthermohilic bacterium *Aquifex aeolicus*. Nature 392, 353–358.

De Maeyer, M., Desmet, J., Lasters, I., 1997. All in one: a highly detailed rotamer library inproves both accuracy and speed in the modelling of sidechains by dead-end elimination. Fold. Des. 2, 53–66.

DeTar, D.L.F., Luthra, N.P., 1977. Conformation of proline. J. Am. Chem. Soc. 99, 1232–1244.

Dill, K.A., 1990. Dominant forces in protein folding. Biochemistry 29, 7133–7155.

Dodson, E.J., Davies, G.J., Lamzin, V.S., Murshudov, G.N., Wilson, K.S., 1998. Validation tools: can they indicate the information content of macromolecular crystal structures? Structure 6, 685–690.

Doig, A.J., Sternberg, M.J.E., 1995. Side-chain conformational entropy in protein-folding. Protein Sci. 4, 2247–2251.

Dunbrack Jr., R.L., 1999. Comparative Modeling of CASP3 targets using PSI-BLAST and SCWRL. Proteins Suppl. 3, 81–87.

Dunbrack Jr., R.L., Cohen, F.E., 1997. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci. 6, 1661–1681.

Dunbrack Jr., R.L., Karplus, M., 1993. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J. Mol. Biol. 230, 543–574.

Dunbrack Jr., R.L., Karplus, M., 1994. Conformational analysis of the backbone-dependent rotamer preferences of protein side chains. Nat. Struct. Biol. 1, 334–340.

Dyson, H.J., Rance, M., Houghten, R.A., Lerner, R.A., Wright, P.E., 1988. Folding of immunogenic peptide fragments of proteins in water solution. I. Sequence requirements for the formation of a reverse turn. J. Mol. Biol. 201, 161–200.

Efimov, A.V., 1993. Standard structures in proteins. Prog. Biophys. Mol. Biol. 60, 201–239.

Eisenmenger, F., Argos, P., Abagyan, R., 1993. A method to configure protein side-chains from the main-chain trace in homology modelling. J. Mol. Biol. 231, 849–860.

Engh, R.A., Huber, R., 1991. Accurate bond and angle parameters for X-ray protein structure refinement. Acta Crystallogr. A47, 392–400.

Evans, J.S., Mathiowetz, A.M., Chan, S.I., Goddard, W.A., III, 1995. De novo prediction of polypeptide conformations using dihedral probability grid Monte Carlo methodology. Protein Sci. 4, 1203–1216.

Faber, H.R., Matthews, B.W., 1990. A mutated T4 lysozyme displays five different crystal conformations. Nature 348, 263–266.

Facchiano, A.M., Colonna, G., Ragone, R., 1998. Helix stabilizing factors and stabilization of thermophilic proteins: an X-ray based study. Protein Eng. 11, 753–760.

Fedorov, A.N., Dolgikh, D.A., Chemeris, V.V., Chernov, B.K., Finkelstein, A.V., Schulga, A.A., Alakhov, Y.B., Kirpichnikov, M.P., Ptitsyn, O.B., 1992. *De novo* design, synthesis and study of albebetin, a polypeptide with a predetermined three-dimensional structure. Probing the structure at the nanogram level. J. Mol. Biol. 225, 927–931.

Feher, V.A., Baldwin, E.P., Dahlquist, F.W., 1996. Access of ligands to cavities within the core of a protein is rapid. Nat. Struct. Biol. 3, 516–521.

Finkelstein, A.V., 1995. Predicted β-structure stability parameters under experimental test. Protein Eng. 8, 207–209.

Finkelstein, A.V., Ptitsyn, O.B., 1977. Theory of protein molecule self-organization. I. Thermodynamic parameters of local secondary structures in the unfolded protein chain. Biopolymers 16, 469–495.

Finkelstein, A.V., Ptitsyn, O.B., 1987. Why do globular proteins fit the limited set of folding patterns? Prog. Biophys. Mol. Biol. 50, 171–190.

Frauenfelder, H., Sligar, S.G., Wolynes, P.G., 1991. The energy landscapes and motions of proteins. Science 254, 1598–1603.

Frömmel, C., Preissner, R., 1990. Prediction of prolyl residues in *cis*-conformation in protein structures on the basis of the amino acid sequence. FEBS Lett. 277, 159–163.

Fujimoto, H., 1987. Paired interacting orbitals: a way of looking at chemical interactions. Acc. Chem. Res. 20, 448–453.

Garratt, R.C., Thornton, J.M., Taylor, W.R., 1991. An extension of secondary structure prediction towards the prediction of tertiary structure. FEBS Lett. 280, 141–146.

Gelin, B.R., Karplus, M., 1979. Side-chain torsional potentials: effect of dipeptide, protein, and solvent environment. Biochemistry 18, 1256–1268.

Gerstein, M., Lesk, A.M., Chothia, C., 1994. Structural mechanisms for domain movements in proteins. Biochemistry 33, 6739–6749.

Gibrat, J.-F., Robson, B., Garnier, J., 1991. Influence of the local amino acid sequence upon the zones of the torsional angles ϕ and ψ adopted by residues in proteins. Biochemistry 30, 1578–1586.

Glusker, J.P., Trueblood, K.N., 1985. Crystal Structure Analysis. A Primer. Oxford University Press, New York.

Grathwohl, C., Wüthrich, K., 1981. NMR studies of the rates of proline *cis–trans* isomerization in oligopeptides. Biopolymers 20, 2623–2633.

Gray, T.M., Matthews, B.W., 1984. Intrahelical hydrogen bonding of serine, threonine and cysteine residues within α-helices and its relevance to membrane-bound proteins. J. Mol. Biol. 175, 75–81.

Greer, J., 1991. Comparative modeling of homologous proteins. Methods Enzymol. 202, 239–252.

Gunasekaran, K., Nagarajaram, H.A., Ramakrishnan, C., Balaram, P., 1998. Stereochemical punctuation marks in protein structures: glycine and proline containing helix stop signals. J. Mol. Biol. 275, 917–932.

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

Haney, P.J., Badger, J.H., Buldak, G.L., Reich, C.I., Woese, C.R., Olsen, G.J., 1999. Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species. Proc. Natl. Acad. Sci. USA. 96, 3578–3583.

Heringa, J., Argos, P., 1999. Strain in protein structures as viewed through nonrotameric side chains: I. Their position and interaction. Proteins 37, 30–43.

Hermans, J., 1993. Molecular dynamics simulations of helix and turn propensities in model peptides. Curr. Opin. Struct. Biol. 3, 270–276.

Hermans, J., Anderson, A.G., Yun, R.H., 1992. Differential helix propensity of small apolar side chains studied by molecular dynamics simulations. Biochemistry 31, 5646–5653.

Herzberg, O., Moult, J., 1991. Analysis of the steric strain in the polypeptide backbone of protein molecules. Proteins 11, 223–229.

Hirel, P.H., Schmitter, J., Dessen, P., Fayat, G., Blanquet, S., 1989. Extent of N-terminal methionine excision from *Escherichia coli* proteins is governed by the side-chain length of the penultimate amino acid. Proc. Natl. Acad. Sci. USA 86, 8247–8251.

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

Hol, W.G.J., 1985. The role of the α-helix dipole in protein function and structure. Prog. Biophys. Mol. Biol. 45, 149–195.

Holm, L., Sander, C., 1991. Database algorithm for generating protein backbone and side-chain co-ordinates from a C^{α} trace. Application to model building and detection of co-ordinate errors. J. Mol. Biol. 218, 183–194.

Holm, L., Sander, C., 1992. Fast and simple Monte Carlo algorithm for side chain optimization in proteins: application to model building by homology. Proteins 14, 213–223.

Hooft, R.W.W., Sander, C., Vriend, G., 1997. Objectively judging the quality of a protein structure from a Ramachandran plot. Comput. Appl. Biosci. 13, 425–430.

Horovitz, A., Matthews, J.M., Fersht, A.R., 1992. α-Helix stability in proteins. II. Factors that influence stability at an internal position. J. Mol. Biol. 227, 560–568.

Hubbard, S.J., Gross, K.H., Argos, P., 1994. Intramolecular cavities in globular proteins. Protein Eng. 7, 613–626.

Hue, S.C., Dill, K.A., 1991. Polymer principles in protein structure and stability. Annu. Rev. Biophys. Chem. 20, 447–490.

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

Hutchinson, E.G., Sessions, R.B., Thornton, J.M., Woolfson, D.N., 1998. Determinants of strand register in antiparallel β-sheets of proteins. Protein Sci. 7, 2287–2300.

Hutchinson, E.G., Thornton, J.M., 1994. A revised set of potentials for β-turn formation in proteins. Protein Sci. 3, 2207–2216.

IUPAC–IUB Commission on Biochemical Nomenclature, 1970. Abbreviations and symbols for the description of the conformation of polypeptide chains. Tentative rules. Biochemistry 9, 3471–3479.

Jabs, A., Weiss, M.S., Hilgenfeld, R., 1999. Non-proline *cis* peptide bonds in proteins. J. Mol. Biol. 286, 291–304.

Jaenicke, R., Böhm, G., 1998. The stability of proteins in extreme environments. Curr. Opin. Struct. Biol. 8, 738–748.
James, M.N.G., Sielecki, A.R., 1983. Structure and refinement of penicillopepsin at 1.8 Å resolution. J. Mol. Biol. 163, 299–361.
Janin, J., Wodak, S., Levitt, M., Maigret, B., 1978. Conformation of amino acid side-chains in proteins. J. Mol. Biol. 125, 357–386.
Jeffrey, G.A., Saenger, W., 1991. Hydrogen Bonding in Biological Structures. Springer, New York.
Jia, Z., Vandonselaar, M., Quail, J.W., Delbaere, L.T.J., 1993. Active-centre torsion-angle strain revealed in 1.6 Å-resolution structure of histidine-containing phosphocarrier protein. Nature 361, 94–97.
Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure. Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637.
Kabsch, W., Sander, C., 1984. On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. Proc. Natl. Acad. Sci. USA 81, 1075–1078.
Kang, Y.K., Jhon, J.S., Han, S.J., 1999. Conformational study of Ac–Xaa–Pro–NHMe dipeptides: proline puckering and *trans*/*cis* imide bond. J. Peptide Res. 53, 30–40.
Kang, H.S., Kurochkina, N.A., Lee, B., 1993. Estimation and use of protein backbone angle probabilities. J. Mol. Biol. 229, 448–460.
Karplus, P.A., 1996. Experimentally observed conformation-dependent geometry and hidden strain in proteins. Protein Sci. 5, 1406–1420.
Karplus, P.A., Schulz, G.E., 1985. Prediction of chain flexibility in proteins. Naturwissenschaften 72, 212–213.
Karpusas, M., Baase, W.A., Matsumura, M., Matthews, B.W., 1989. Hydrophobic packing in T4 lysozyme probed by cavity-filling mutants. Proc. Natl. Acad. Sci. USA 86, 8237–8241.
Karshikoff, A., Ladenstein, R., 1998. Proteins from thermophilic and mesophilic organisms essentially do not differ in packing. Protein Eng. 11, 867–872.
Kaul, R., Balaram, P., 1999. Stereochemical control of peptide folding. Bioorg. Med. Chem. 7, 105–117.
Kim, P.S., Baldwin, R.L., 1990. Intermediates in the folding reactions of small proteins. Annu. Rev. Biochem. 59, 631–660.
Kim, C.A., Berg, J.M., 1993. Thermodynamic β-sheet propensities measured using a zinc finger host peptide. Nature 362, 267–270.
Kishan, K.V.R., Zeelen, J.P., Noble, M.E.M., Borchert, T.V., Wierenga, R.K., 1994. Comparison of the structures and the crystal contacts of trypanosomal triosephosphate isomerase in four different crystal forms. Protein Sci. 3, 779–787.
Kleywegt, G.J., 2000. Validation of protein crystal structures. Acta Crystallogr. D56, 249–265.
Kleywegt, G.J., Jones, T.A., 1996. Phi/Psi-chology: Ramachandran revisited. Structure 4, 1395–1400.
Klingler, T.D., Brutlag, D.L., 1994. Discovering structural correlations in α-helices. Protein Sci. 3, 1847–1857.
Knapp, S., de Vos, W.M., Rice, D., Ladenstein, R., 1997. Crystal structure of glutamate dehydrogenase from the hyperthermophilic eubacterium *Thermotoga maritima* at 3.0 Å resolution. J. Mol. Biol. 267, 916–932.
Koehl, P., Delarue, M., 1994. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. J. Mol. Biol. 239, 249–275.
Koehl, P., Levitt, M., 1999. Structure-based conformational preferences of amino acids. Proc. Natl. Acad. Sci. USA 96, 12524–12529.
Kolaskar, A.S., Kulkarni-Kale, U., 1992. Sequence alignment approach to pick up conformationally similar protein fragments. J. Mol. Biol. 223, 1053–1061.
Kossiakoff, A.A., Randal, M., Guenot, J., Eigenbrot, C., 1992. Variability of conformations at crystal contacts in BPTI represent true low-energy structures: correspondence among lattice packing and molecular dynamics structures. Proteins 14, 65–74.
Kotik, M., Zuber, H., 1993. Mutations that significantly change the stability, flexibility and quaternary structure of the L-lactate dehydrogenase from *Bacillus megaterium*. Eur. J. Biochem. 211, 267–280.
Kumar, S., Bansal, M., 1996. Structural and sequence characteristics of long α helices in globular proteins. Biophys. J. 71, 1574–1586.
Kumar, S., Bansal, M., 1998. Dissecting α-helices: position-specific analysis of α-helices in globular proteins. Proteins 31, 460–476.

Kurinov, I., Harrison, R.W., 1995. The influence of temperature on lysozyme crystals. Structure and dynamics of protein and water. Acta Crystallogr. D51, 98–109.

Kuszewski, J., Gronenborn, A.M., Clore, G.M., 1996. Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. Protein Sci. 5, 1067–1080.

Ladenstein, R., Antranikian, G., 1998. Proteins from hyperthermophiles: stability and enzyme catalysis close to the boiling point of water. Adv. Biochem. Eng. Biotech. 61, 37–85.

Laskowski, R.A., MacArthur, M.W., Moss, D.S., Thornton, J.M., 1993. PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. 26, 283–291.

Laskowski, R.A., MacArthur, M.W., Thornton, J.M., 1998. Validation of protein models derived from experiment. Curr. Opin. Struct. Biol. 8, 631–639.

Laughton, C.A., 1994. Prediction of protein side-chain conformations from local three-dimensional homology relationships. J. Mol. Biol. 235, 1088–1097.

Lazardis, T., Archontis, G., Karplus, M., 1995. Enthalpic contribution to protein stability: insights from atom-based calculations and statistical mechanics. Adv. Protein Chem. 47, 231–307.

Lebbink, J.H.G., Knapp, S., van der Oost, J., Rice, D., Ladenstein, R., de Vos, W.M., 1999. Engineering activity and stability of *Thermotoga maritima* glutamate dehydrogenase. II: Construction of a 16-residue ion-pair network at the subunit interface. J. Mol. Biol. 289, 357–369.

Lee, C., Subbiah, S., 1991. Prediction of protein side-chain conformation by packing optimization. J. Mol. Biol. 217, 373–388.

Lee, J.J., Ekker, S.C., von Kessler, D.P., Porter, J.A., Sun, B.I., Beachy, P.A., 1994a. Autoproteolysis in *hedgehog* protein biogenesis. Science 266, 1528–1537.

Lee, B., Vasmatzis, G., 1997. Stabilization of protein structures. Curr. Opin. Biotechnol. 8, 423–428.

Lee, K.H., Xie, D., Freire, E., Amzel, L.M., 1994b. Estimation of changes in side chain configurational entropy in binding and folding: general methods and application to helix formation. Proteins 20, 68–84.

Lesk, A.M., Chothia, C., 1984. Mechanisms of domain closure in proteins. J. Mol. Biol. 174, 175–191.

Levinthal, C., 1968. Are there pathways for protein folding? J. Chim Phys. 65, 44–45.

Levitt, M., 1978. Conformational preferences of amino acids in globular proteins. Biochemistry 17, 4277–4285.

Levitt, M., 1992. Accurate modelling of protein conformation by automatic segment matching. J. Mol. Biol. 226, 507–533.

Lewis, P.N., Momany, F.A., Scheraga, H.A., 1973. Chain reversals in proteins. Biochim. Biophys. Acta 303, 211–229.

Lifson, S., Sander, C., 1979. Antiparallel and parallel β-strands differ in amino acid residue preferences. Nature 282, 109–111.

Lifson, S., Sander, C., 1980. Specific recognition in the tertiary structure of β-sheets of proteins. J. Mol. Biol. 139, 627–639.

Lim, V.I., 1974. Algorithms for prediction of α-helical and β-structural regions of globular proteins. J. Mol. Biol. 88, 873–894.

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

Lovell, S.C., Word, J.M., Richardson, J.S., Richardson, D.C., 2000. The penultimate rotamer library. Proteins 40, 389–408.

Ludlow, J.M., Clark, D.S., 1991. Engineering considerations for the application of extremeophiles in biotechnology. Crit. Rev. Biotechnol. 10, 321–345.

Luque, I., Mayorga, O.L., Freire, E., 1996. Structure-based thermodynamic scale of α-helix propensities in amino acids. Biochemistry 35, 13681–13688.

Lyu, P.C., Liff, M.I., Marky, L.A., Kallenbach, N.R., 1990. Side chain contributions to the stability of α-helical structure in peptides. Science 250, 669–673.

MacArthur, M.W., Thornton, J.M., 1991. Influence of proline residues on protein conformation. J. Mol. Biol. 218, 397–412.

MacArthur, M.W., Thornton, J.M., 1999. Protein side-chain conformation: a systematic variation of χ_1 mean values with resolution—a consequence of multiple rotameric states? Acta Crystallogr. D55, 994–1004.

Maccallum, P.H., Poet, R., Milner-White, E.J., 1995. Coulombic interactions between partially charged main-chain atoms not hydrogen-bonded to each other influence the conformations of α-helices and antiparallel β-sheet. A new method for analysing the forces between hydrogen bonding groups in proteins includes all the Coulombic interactions. J. Mol. Biol. 248, 361–373.

Maigret, B., Perahia, D., Pullman, B., 1970. Molecular orbital calculations on the conformation of polypeptides and proteins. IV. The conformation of the prolyl and hydroxyprolyl residues. J. Theor. Biol. 29, 275–291.

Malakauskas, S.M., Mayo, S.L., 1998. Design, structure and stability of a hyperthermophilic protein variant. Nat. Struct. Biol. 5, 470–475.

Marshall, G.R., Bosshard, H.E., 1972. Angiotensin II. Studies on the biologically active conformation. Circ. Res. 31 (Suppl. II), 143–150.

Matthews, B.W., 1995. Studies on protein stability with T4 lysozyme. Adv. Protein Chem. 46, 249–278.

McDonald, I.K., Thornton, J.M., 1994. Satisfying hydrogen bonding potential in proteins. J. Mol. Biol. 238, 777–793.

McGregor, M.J., Islam, S.A., Sternberg, M.J.E., 1987. Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. J. Mol. Biol. 198, 295–310.

Menéndez-Arias, L., Argos, P., 1989. Engineering protein thermal stability: sequence statistics point to residue substitutions in α-helices. J. Mol. Biol. 206, 397–406.

Milner-White, E.J., 1988. Recurring loop motif in proteins that occurs in right-handed and left-handed forms. Its relationship with alpha-helices and beta-bulge loops. J. Mol. Biol. 199, 503–511.

Milner-White, E.J., Bell, L.H., Maccallum, P.H., 1992. Pyrrolidine ring puckering in *cis* and *trans*-proline residues in proteins and polypeptides. Different puckers are favoured in certain situations. J. Mol. Biol. 228, 725–734.

Minor Jr., D.L., Kim, P.S., 1994a. Measurement of the β-sheet-forming propensities of amino acids. Nature 367, 660–663.

Minor Jr., D.L., Kim, P.S., 1994b. Context is a major determinant of β-sheet propensity. Nature 371, 264–267.

Momany, F.A., McGuire, R.F., Burgess, A.W., Scheraga, H.A., 1975. Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. J. Phys. Chem. 79, 2361–2381.

Morris, A.L., MacArthur, M.W., Hutchinson, E.G., Thornton, J.M., 1992. Stereochemical quality of protein structure coordinates. Proteins 12, 345–364.

Moult, J., 1997. Comparison of database potentials and molecular mechanics force fields. Curr. Opin. Struct. Biol. 7, 194–199.

Muñoz, V., Serrano, L., 1994. Intrinsic secondary structure propensities of the amino acids, using statistical ϕ–ψ matrices: comparison with experimental scales. Proteins 20, 301–311.

Muñoz, V., Serrano, L., 1995. Elucidating the folding problem of helical peptides using empirical parameters. II. Helix macrodipole effects and rational modification of the helical content of natural peptides. J. Mol. Biol. 245, 275–296.

Myers, J.K., Pace, C.N., Scholtz, J.M., 1997. Helix propensities are identical in proteins and peptides. Biochemistry 36, 10923–10929.

Némethy, G., Gibson, K.D., Palmer, K.A., Yoon, C.N., Paterlini, G., Zagari, A., Rumsey, S., Scheraga, H.A., 1992. Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. J. Phys. Chem. 96, 6472–6484.

Némethy, G., Leach, S.J., Scheraga, H.A., 1966. The influence of amino acid side-chains on the free energy of helix-coil transitions. J. Phys. Chem. 70, 998–1004.

Nicholson, H., Söderlind, E., Tronrud, D.E., Matthews, B.W., 1989. Contributions of left-handed helical residues to the structure and stability of bacteriophage T4 lysozyme. J. Mol. Biol. 210, 181–193.

Nicholson, H., Tronrud, D.E., Becktel, W.J., Matthews, B.W., 1992. Analysis of the effectiveness of proline substitutions and glycine replacements in increasing the stability of phage T4 lysozyme. Biopolymers 32, 1431–1441.

Niefind, K., Schomburg, D., 1991. Amino acid similarity coefficients for protein modeling and sequence alignment derived from main-chain folding angles. J. Mol. Biol. 219, 481–497.

O'Neil, K.T., DeGrado, W.F., 1990. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. Science 250, 646–651.

Pace, C.N., Scholtz, J.M., 1998. A helix propensity scale based on experimental studies of peptides and proteins. Biophys. J. 75, 422–427.

Padmanabhan, S., Marqusee, S., Ridgeway, T., Laue T, .M., Baldwin, R.L., 1990. Relative helix-forming tendencies of nonpolar amino acids. Nature 344, 268–270.

Pal, L., Basu, G., 1999. Novel protein structural motifs containing two-turn and longer 3_{10}-helices. Protein Eng. 12, 811–814.

Pal, D., Chakrabarti, P., 1998. Different types of interactions involving cysteine sulfhydryl group in proteins. J. Biomol. Struct. Dyn. 15, 1059–1072.

Pal, D., Chakrabarti, P., 1999a. Graphical representation of the salient conformational features of protein residues. Protein Eng. 12, 523–526.

Pal, D., Chakrabarti, P., 1999b. *Cis* peptide bonds in proteins: residues involved, their conformations, interactions and locations. J. Mol. Biol. 294, 271–288.

Pal, D., Chakrabarti, P., 1999c. Estimates of the loss of main-chain conformational entropy of different residues on protein folding. Proteins 36, 332–339.

Pal, D., Chakrabarti, P., 2000a. Terminal residues in protein chains: residue preference, conformation, and interaction. Biopolymers 53, 467–475.

Pal, D., Chakrabarti, P., 2000b. β-Sheet propensity and its correlation with parameters based on conformation. Acta Crystallogr. D56, 589–594.

Pal, D., Chakrabarti, P., 2000c. Conformational similarity indices between different residues in proteins and α-helix propensities. J. Biomol. Struct. Dyn. 18, 273–280.

Pal, D., Chakrabarti, P., 2001. On the residues in the disallowed region of the Ramachandran map and protein thermostability, in preparation.

Park, S.H., Shalongo, W., Stellwagen, E., 1993. Residue helix parameters obtained from dichroic analysis of peptides of defined sequence. Biochemistry 32, 7048–7053.

Parthasarathy, S., Murthy, M.R.N., 1999. On the correlation between the main-chain and side-chain atomic displacement parameters (B values) in high-resolution protein structures. Acta Crystallogr. D55, 173–180.

Paulus, H., 1998. The chemical basis of protein splicing. Chem. Soc. Rev. 27, 375–386.

Penel, S., Hughes, E., Doig, A.J., 1999. Side-chain structures in the first turn of the α-helix. J. Mol. Biol. 287, 127–143.

Perler, F.B., 1998. Protein splicing of inteins and Hedgehog autoproteolysis: structure, function, and evolution. Cell 92, 1–4.

Petukhov, M., Muñoz, V., Yumoto, N., Yoshikawa, S., Serrano, L., 1998. Position dependence of non-polar amino acid intrinsic helical propensities. J. Mol. Biol. 278, 279–289.

Pickett, S.D., Sternberg, M.J.E., 1993. Empirical scale of side-chain conformational entropy in protein folding. J. Mol. Biol. 231, 825–839.

Piela, L., Némethy, G., Scheraga, H.A., 1987. Proline-induced constraints in α-helices. Biopolymers 26, 1587–1600.

Ponder, J.W., Richards, F.M., 1987. Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193, 775–791.

Ponnuswamy, P.K., Sasisekharan, V., 1970. Studies on the conformation of amino acids. III. Backbone conformation of N- and C-terminal glycyl and alanyl residues. Int. J. Protein Res. II, 47–57.

Ponnuswamy, P.K., Sasisekharan, V., 1971. Studies on the conformation of amino acids. IX. Conformations of butyl, seryl, threonyl, cysteinyl, and valyl residues in a dipeptide unit. Biopolymers 10, 565–582.

Prasad, B.V.V., Balaram, P., 1984. The stereochemistry of peptides containing α-aminoisobutyric acid. Crit. Rev. Biochem. 16, 307–348.

Preissner, R., Bork, P., 1991. On α-helices terminated by glycine. 1. Identification of common structural features. Biochem. Biophys. Res. Commun. 180, 660–665.

Presta, L.G., Rose, G.D., 1988. Helix signals in proteins. Science 240, 1632–1641.

Privalov, P.L., Gill, S.J., 1988. Stability of protein structure and hydrophobic interaction. Adv. Protein Chem. 39, 191–234.

Querol, E., Perez-Pons, J.A., Mozo-Villarias, A., 1996. Analysis of protein conformational characteristics related to thermostability. Protein Eng. 9, 265–271.

Quinn, T.P., Tweedy, N.B., Williams, R.W., Richardson, J.S., Richardson, D.C., 1994. Betadoublet: *de novo* design, synthesis, and characterization of a β-sandwich protein. Proc. Natl. Acad. Sci. USA 91, 8747–8751.

Radzicka, A., Pedersen, L., Wolfenden, R., 1988. Influences of solvent water on protein folding: free energies of solvation of *cis* and *trans* peptides are nearly identical. Biochemistry 27, 4538–4541.

Ragone, P.A., Facchiano, F., Facchiano, A., Facchiano, A.M., Colonna, G., 1989. Flexibility plot of proteins. Protein Eng. 2, 497–504.

Ramachandran, G.N., Chandrasekaran, R., 1972. In: Lande, S. (Ed.), Proceedings of the Second American Peptide Symposium, Cleveland, 1970. Progress in Peptide Research, Vol. II. Gordon and Breach, New York, pp. 195–215.

Ramachandran, G.N., Lakshminarayanan, A.V., Balasubramanian, R., Tegoni, G., 1970. Energy calculations on proline residues. Biochem. Biophys. Acta 221, 165–181.

Ramachandran, G.N., Mitra, A.K., 1976. An explanation for the rare occurrence of *cis* peptide units in proteins and polypeptides. J. Mol. Biol. 107, 85–92.

Ramachandran, G.N., Ramakrishnan, C., Sasisekharan, V., 1963. Stereochemistry of polypeptide chain configurations. J. Mol. Biol. 7, 95–99.

Ramachandran, G.N., Ramakrishnan, C., Venkatachalam, C.M., 1965. Determination of the possible conformations of the residues linked in a polypeptide chain. Biopolymers 3, 591–592.

Ramachandran, G.N., Sasisekharan, V., 1968. Conformation of polypeptides and proteins. Adv. Protein Chem. 23, 283–437.

Ramachandran, G.N., Venkatachalam, C.M., 1968. Stereochemical criteria for polypeptides and proteins. IV. Standard dimensions for the *cis*-peptide unit and conformation of *cis*-polypeptides. Biopolymers 6, 1255–1262.

Ramakrishnan, C., Ramachandran, G.N., 1965. Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units. Biophys. J. 5, 909–933.

Ramakrishnan, C., Srinivasan, N., 1990. Glycyl residues in proteins and peptides: an analysis. Curr. Sci. 59, 851–861.

Ramakrishnan, C., Srinivasan, N., Prashanth, D., 1987. Conformation of glycyl residues in globular proteins. Int. J. Peptide Protein Res. 29, 629–637.

Recsei, P.A., Huynh, Q.K., Snell, E.E., 1983. Conversion of prohistidine decarboxylase to histidine-decarboxylase: peptide chain cleavage by nonhydrolytic serinolysis. Proc. Natl. Acad. Sci. USA 80, 973–977.

Rees, D.C., Adams, M.W.W., 1995. Hyperthermophiles: taking the heat and loving it. Structure 3, 251–254.

Reimer, U., Scherer, G., Drewello, M., Kruber, S., Schutkowski, M., Fischer, G., 1998. Side-chain effects on peptidyl–prolyl *cis*/*trans* isomerisation. J. Mol. Biol. 279, 449–460.

Rejto, P.A., Freer, S.T., 1996. Protein conformational substates from X-ray crystallography. Prog. Biophys. Mol. Biol. 66, 167–196.

Richards, F.M., 1977. Areas, volumes, packing and protein structures. Annu. Rev. Biophys. Bioeng. 6, 151–176.

Richardson, J.S., 1981. The anatomy and taxonomy of protein structure. Adv. Protein Chem. 34, 167–339.

Richardson, J.S., Richardson, D.C., 1988. Amino acid preferences for specific locations at the ends of α helices. Science 240, 1648–1652.

Richardson, J.S., Richardson, D.C., 1989. Principles and patterns of protein conformation. In: Fasman, G.D. (Ed.), Prediction of Protein Structure and the Principles of Protein Conformation. Plenum Press, New York, pp. 1–98.

Riddle, D.S., Santiago, J.V., Bray-Hall, S.T., Doshi, N., Grantcharova, V.P., Yi, Q., Baker, D., 1997. Nat. Struct. Biol. 4, 805–809.

Ringe, D., Petsko, G.A., 1986. Study of protein dynamics by X-ray diffraction. Methods Enzymol. 131, 389–433.

Risler, J.L., Delorme, M.O., Delacroix, H., Henaut, A., 1988. Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. J. Mol. Biol. 204, 1019–1029.

Rooman, M.J., Kocher, J.-P.A., Wodak, S.J., 1992. Extracting information on folding from the amino acid sequence: accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. Biochemistry 31, 10226–10238.

Roterman, I.K., Lambert, M.H., Gibson, K.D., Scheraga, H.A., 1989. A comparison of the CHARMM, AMBER and ECEPP potentials for peptides. II. ϕ–ψ maps for *N*-acetyl alanine *N*′-methyl amide: comparisons, contrasts and simple experimental tests. J. Biomol. Struct. Dyn. 7, 421–453.

Rose, G.D., Gierasch, L.M., Smith, J.A., 1985. Turns in peptides and proteins. Adv. Protein Chem. 37, 1–109.

Russell, R.J.M., Ferguson, J.M.C., Hough, D.W., Danson, M.J., Taylor, G.L., 1997. The crystal structure of citrate synthetase from the hyperthermophilic archae on *Pyrococcus furiosus* at 1.9 Å resolution. Biochemistry 36, 9983–9994.

Salemme, F.R., 1983. Structural properties of protein β-sheets. Prog. Biophys. Mol. Biol. 42, 95–133.
Samanta, U., Pal, D., Chakrabarti, P., 2000. Environment of tryptophan side chains in proteins. Proteins 38, 288–300.
Sankararamakrishnan, R., Vishveshwara, S., 1992. Geometry of proline-containing alpha-helices in proteins. Int. J. Peptide Protein Res. 39, 356–363.
Sasisekharan, V., Ponnuswamy, P.K., 1970. Backbone and side-chain conformations of amino acids and amino acid residues in peptides. Biopolymers 9, 1249–1256.
Sasisekharan, V., Ponnuswamy, P.K., 1971. Studies on the conformation of amino acids. X. Conformations of norvalyl, leucyl and aromatic side groups in a dipeptide unit. Biopolymers 10, 583–592.
Schellman, C., 1980. The α_L-conformation at the ends of helices. In: Jaenicke, R. (Ed.), Protein Folding. Elsevier and North-Holland. New York, pp. 53–61.
Schmid, F.X., Mayr, L.M., Mücke, M., Schönbrunner, E.R., 1993. Prolyl isomerases: role in protein folding. Adv. Protein Chem. 44, 25–66.
Schimmel, P.R., Flory, P.J., 1968. Conformational energies and configurational statistics of copolypeptides containing L-proline. J. Mol. Biol. 34, 105–120.
Schrauber, H., Eisenhaber, F., Argos, P., 1993. Rotamers: to be or not to be? An analysis of amino acid side-chain conformations in globular proteins. J. Mol. Biol. 230, 592–612.
Serrano, L., 2000. The relationship between sequence and structure in elementary folding units. Adv. Protein Chem. 53, 49–85.
Shakhnovich, E.I., Gutin, A.M., 1990. Implications of thermodynamics of protein folding for evolution of primary sequences. Nature 346, 773–775.
Smith, C.K., Withka, J.M., Regan, L., 1994. A thermodynamic scale for the β-sheet forming tendencies of the amino acids. Biochemistry 33, 5510–5517.
Smith, C.K., Regan, L., 1995. Guidelines for protein design: the energetics of β-sheet side chain interactions. Science 270, 980–982.
Smith, D.D.S., Pratt, K.A., Summer, I.G., Henneke, C.M., 1995. Greek key jellyroll protein motif design: expression and characterization of a first-generation molecule. Protein Eng. 8, 13–20.
Smith, L.J., Bolin, K.A., Schwalbe, H., MacArthur, M.W., Thornton, J.M., Dobson, C.M., 1996. Analysis of main chain torsion angles in proteins: prediction of NMR coupling constants for native and random coil conformations. J. Mol. Biol. 255, 494–506.
Smith, C.K., Regan, L., 1997. Construction and design of β-sheets. Acc. Chem. Res. 30, 153–161.
Srinivasan, N., Anuradha, V.S., Ramakrishnan, C., Sowdhamini, R., Balaram, P., 1994. Conformational characteristics of asparaginyl residues in proteins. Int. J. Peptide Protein Res. 44, 112–122.
Srinivasan, R., Rose, G.D., 1999. A physical basis for protein secondary structure. Proc. Natl. Acad. Sci. USA 96, 14258–14263.
Sudarsanam, S., Srinivasan, S., 1997. Sequence-dependent conformational sampling using a database of ϕ_{i+1} and ψ_i angles for predicting polypeptide backbone conformations. Protein Eng. 10, 1155–1162.
Stec, B., Zhou, R., Teeter, M.M., 1995. Full-matrix refinement of the protein crambin at 0.83 Å and 130 K. Acta Crystallogr. D51, 663–681.
Stewart, D.E., Sarkar, A., Wampler, J.E., 1990. Occurrence and role of *cis* peptide bonds in protein structures. J. Mol. Biol. 214, 253–260.
Stites, W.E., Meeker, A.K., Shortle, D., 1994. Evidence for strained interactions between side-chains and the polypeptide backbone. J. Mol. Biol. 235, 27–32.
Stites, W.E., Pranata, J., 1995. Empirical evaluation of the influence of side chains on the conformational entropy of the polypeptide backbone. Proteins 22, 132–140.
Stoddard, B.L., Pietrokovski, S., 1998. Breaking up is hard to do. Nat. Struct. Biol. 5, 3–5.
Summers, N.L., Carlson, W.D., Karplus, M., 1987. Analysis of side-chain orientations in homologous proteins. J. Mol. Biol. 196, 175–198.
Summers, N.L., Karplus, M., 1990. Modeling of globular proteins. A distance-based data search procedure for the construction of insertion/deletion regions and Pro $\leftrightarrow$ non-Pro mutations. J. Mol. Biol. 216, 991–1016.
Sun, S., Thomas, P.D., Dill, K.A., 1995. A simple protein-folding algorithm using a binary code and secondary structure constraints. Protein Eng. 8, 769–778.

Swindells, M.B., MacArthur, M.W., Thornton, J.M., 1995. Intrinsic ϕ, ψ propensities of amino acids, derived from the coil regions of known structures. Nat. Struct. Biol. 2, 596–603.

Thompson, M.J., Eisenberg, D., 1999. Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability. J. Mol. Biol. 290, 595–604.

Thornton, J.M., Chakauya, B.L., 1982. Conformation of terminal regions in proteins. Nature 298, 296–297.

Thornton, J.M., Sibanda, B.L., 1983. Amino and carboxy-terminal regions in globular proteins. J. Mol. Biol. 167, 443–460.

Tomii, K., Kanehisa, M., 1996. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng. 9, 27–36.

Topham, C.M., McLeod, A., Eisenmenger, F., Overington, J.P., Johnson, M.S., Blundell, T.L., 1993. Fragment ranking in modelling of protein structure. Conformationally constrained environmental amino acid substitution tables. J. Mol. Biol. 229, 194–220.

Tuffery, P., Etchebest, C., Hazout, S., 1997. Prediction of protein side chain conformations: a study on the influence of backbone accuracy on conformation stability in the rotamer space. Protein Eng. 10, 361–372.

Tuffery, P., Etchebest, C., Hazout, S., Lavery, R., 1991. A new approach to the rapid determination of protein side-chain conformations. J. Biomol. Struct. Dyn. 8, 1267–1289.

Usher, K.C., De la Cruz, A.F.A., Dahlquist, F.W., Swanson, R.V., Simon, M.I., Remington, S.J., 1998. Crystal structures of Che Y from *Thermotoga maritima* do not support conventional explanations for the structural basis of enhanced thermostability. Protein Sci. 7, 403–412.

Van den Burg, B., Vriend, G., Veltman, O.R., Venema, G., Eijsink, V.G.H, 1998. Engineering an enzyme to resist boiling. Proc. Natl. Acad. Sci. USA 95, 2056–2060.

Vàsquez, M., Némethy, G., Scheraga, H.A., 1994. Conformational energy calculations on polypeptides and proteins. Chem. Rev. 94, 2183–2239.

Venkatachalam, C.M., 1968. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. Biopolymers 6, 1425–1436.

Vetriani, C., Maeder, D.L., Tolliday, N., Yip, K.S.-P., Stillman, T.J., Britton, K.L., Rice, D.W., Klump, H.H., Robb, F.T., 1998. Protein thermostability above 100°C: a key role for ionic interactions. Proc. Natl. Acad. Sci. USA 95, 12300–12305.

Vihinen, M., Torkkila, E., Riikonen, P., 1994. Accuracy of protein flexibility predictions. Proteins 19, 141–149.

Vogt, G., Woell, S., Argos, P., 1997. Protein thermal stability, hydrogen bonds, and ion pairs. J. Mol. Biol. 269, 631–643.

von Heijne, G., 1991. Proline kinks in transmembrane α-helices. J. Mol. Biol. 218, 499–503.

von Heijne, G., Blomberg, C., 1977. The β structure: inter-strand correlations. J. Mol. Biol. 117, 821–824.

Wallon, G., Kryger, G., Lovett, S.T., Oshima, T., Ringe, D., Petsko, G.A., 1997. Crystal structures of *E. coli* and *S. typhimurium* 3-isopropyl MDH and comparison with their thermophilic counterpart from Thermus thermophilus. J. Mol. Biol. 266, 1016–1031.

Walther, D., Argos, P., 1996. Intrahelical side chain-side chain contacts: the consequences of restricted rotameric states and implications for helix engineering and design. Protein Eng. 9, 471–478.

Wang, J., Wang, W., 1999. A computational approach to simplifying the protein folding alphabet. Nat. Struct. Biol. 6, 1033–1038.

Weiss, M.S., Hilgenfeld, R., 1999. A method to detect nonproline *cis* peptide bonds in proteins. Biopolymers 50, 536–544.

Weiss, M.S., Jabs, A., Hilgenfeld, R., 1998. Peptide bonds revisited. Nat. Struct. Biol. 5, 676.

West, M.W., Hecht, M.H., 1995. Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. Protein Sci. 4, 2032–2039.

Whitlow, M., Teeter, M.M., 1986. An empirical examination of potential energy minimization using the well-determined structure of the protein crambin. J. Am. Chem. Soc. 108, 7163–7172.

Wiberg, K.B., Murcko, M.A., 1988. Rotational barriers. 2. Energies of alkane rotamers. An examination of gauche interactions. J. Am. Chem. Soc. 110, 8029–8038.

Williams, R.W., Chang, A., Juretic, D., Loughran, S., 1987. Secondary structure predictions and medium range interactions. Biochem. Biophys. Acta 916, 200–204.

Williams, K.A., Deber, C.M., 1991. Proline residues in transmembrane helices: structural or dynamic role? Biochemistry 30, 8919–8923.

Wilmot, C.M., Thornton, J.M., 1990. β-Turns and their distortions: a proposed new nomenclature. Protein Eng. 3, 479–493.

Wilson, K.S., Butterworth, S., Dauter, Z., Lamzin, V.S., Walsh, M., Wodak, S., Pontius, J., Richelle, J., Vaguine, A., Sander, C., Hooft, R.W.W., Vriend, G., Thornton, J.M., Laskowski, R.A., MacArthur, M.W., Dodson, E.J., Murshudov, G., Oldfield, T.J., Kaptein, R., Rullmann, J.A.C., 1998. Who checks the checkers? Four validation tools applied to eight atomic resolution structures. J. Mol. Biol. 276, 417–436.

Wodak, S.J., Rooman, M.J., 1993. Generating and testing protein folds. Curr. Opin. Struct. Biol. 3, 247–259.

Wolynes, P.G., 1997. As simple as can be? Nat. Struct. Biol. 4, 871–874.

Word, J.M., Lovell, S.C., LaBean, T.H., Taylor, H.C., Zalis, M.E., Presley, B.K., Richardson, J.S., Richardson, D.C., 1999. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. J. Mol. Biol. 285, 1711–1733.

Wouters, M.A., Curmi, P.M.G., 1995. An analysis of side chain interactions and pair correlations within antiparallel β-sheets: the differences between backbone hydrogen-bonded and non-hydrogen bonded residue pairs. Proteins 22, 119–131.

Wright, P.E., Dyson, H.J., Lerner, R.A., 1988. Conformation of peptide fragments of proteins in aqueous solution: implications for initiation of protein folding. Biochemistry 27, 7167–7175.

Wu, W.-J., Raleigh, D.P., 1998. Local control of peptide conformation: stabilization of *cis* proline peptide bonds by aromatic proline interactions. Biopolymers 45, 381–394.

Xiao, L., Honig, B., 1999. Electrostatic contributions to the stability of hyperthermophilic proteins. J. Mol. Biol. 289, 1435–1444.

Xu, Q., Buckley, D., Guan, C., Guo, H.-C., 1999. Structural insights into the mechanism of intramolecular proteolysis. Cell 98, 651–661.

Yang, A.S., Honig, B., 1995. Free energy determinants of secondary structure formation. 1. α-helices. J. Mol. Biol. 252, 351–365.

Yang, J., Spek, E.J., Gong, Y., Zhou, H., Kallenbach, N.R., 1997. The role of context on α-helix stabilization: host-guest analysis in a mixed background peptide model. Protein Sci. 6, 1264–1272.

Yao, J., Feher, V.A., Espejo, B.F., Reymond, M.T., Wright, P.E., Dyson, H.J., 1994. Stabilization of a type VI turn in a family of linear peptides in water solution. J. Mol. Biol. 243, 736–753.

Yun, R.H., Anderson, A., Hermans, J., 1991. Proline in α-helix: stability and conformation studied by dynamics simulation. Proteins 10, 219–228.

Závodszky, P., Kardos, J., Svingor, A., Petsko, G.A., 1998. Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. Proc. Natl. Acad. Sci. USA 95, 7406–7411.

Zhang, X., Wozniak, J.A., Matthews, B.W., 1995. Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozyme. J. Mol. Biol. 250, 527–552.

Zimmerman, S.S., Pottle, M.S., Némethy, G., Scheraga, H.A., 1977. Conformational analysis of the 20 naturally occurring amino acid residues using ECEPP. Macromolecules 10, 1–9.