The study presented in this thesis deals with analysis of protein crystal structures with emphasis on backbone stereochemistry. The investigations involve identification and analysis of common structural features observed among the unrelated protein crystal structures which are available in the Brookhaven Protein Data Bank (PDB). The thesis in essence attempts to provide a rational knowledge base for design of protein structural mimics, protein engineering and structure prediction.

Conformational studies of peptides and proteins, involving theoretical methods and data analysis of protein crystal structures, are the major ongoing projects in the group with which the author is associated. The work reported in various chapters in this thesis can be broadly classified under the groups (i) studies on unusual stereochemistry (Chapter 2 and Chapter 5) (ii) obtaining rules which can be used in predictions (Chapter 3 and Chapter 7) and (iii) analyses on structural motifs (Chapter 4 and Chapter 6).

Chapter 1 of the thesis is of an introductory in nature and briefly discusses (a) conformation of polypeptide chains, (b) structure determination (c) globular proteins mainly in connection with its structure, (d) identification of secondary structures (e) some of the applications of database analysis of protein structures and (f) the selection of representative dataset and characterization.

Chapter 2 summarizes a study on the nature and distribution of Ramachandran disallowed conformations of amino acid residues observed in high resolution protein crystal structures. A dataset consisting of 110 high resolution non-homologous protein crystal structures from the Brookhaven protein data bank (PDB) was examined. The dataset consisted a total of 18,708 non-Gly residues, which were characterized based on their backbone dihedral angles ($\phi$, $\psi$). Residues falling outside the defined 'broadly
allowed limits" on the Ramachandran map were chosen and the reported B-factor value of the α-carbon atoms was used to further select the well defined disallowed conformations. The conformations of the selected 66 disallowed residues clustered in distinct regions of the Ramachandran map indicating that specific φ, ψ distortions are preferred under compulsions imposed by local constraints. The distribution of various amino acid residues in the disallowed residue dataset showed a predominance of small polar/charged residues with bulky hydrophobic residues being infrequent. As a further check for all the 66 cases non-hydrogen van der Waals short contacts in the protein structures were examined. The analysis reveals that short contacts are eliminated in most cases by local distortions of bond angles. An analysis of the conformation of the identified disallowed residues in related protein structures reveals instances of conservation of unusual stereochemistry.

Chapter 3 deals with an analysis on the nature of α-helix stop signals. A dataset of 1057 helices were identified from 250 high resolution (≤2.0Å) non-homologous protein crystal structures. The backbone dihedral angles (φ, ψ) of the terminating residue (T) were found to cluster either in the left-handed helical region (αL, φ=20° to 125° and ψ=−45° to 90°) [469 helices (44%)] or in the extended region (E, φ=180° to −30° and ψ=60° to 180° and −180° to −150°) [459 helices (43%)] of the Ramachandran map. Gly residues were found to have an overwhelming preference to occur as the αL-terminator (T) resulting in the classical Schellman motif, with a strong preference for hydrophobic residues at position 'T-4' and 'T+1'. In the case of E-terminated helices His, Asn, Leu, and Phe were found to occur with high propensity at position 'T'. Quite remarkably Pro residues, with single exception, were absent at position 'T', but had the highest propensity at position 'T+1'. Examination of the frequencies of hydrophobic (h) and polar (p) residues at positions flanking Gly/Pro reveals that Pro residues flanked by polar amino acids have a very strong tendency to terminate helices. Examination of a segment ranging from 'T-4' to 'T+3' appeared to be necessary to determine whether helix
termination or continuation occur at Gly residues. The two types of helix termination ($\alpha_L$ and $\alpha_E$) signals also differed dramatically in their solvent accessibility.

Chapter 4 discusses an analysis on $\beta$-hairpins which have short connecting loops (1-5 residues). $\beta$-hairpins were identified from the dataset of 250 high resolution ($\leq 2.0\AA$), non-homologous protein crystal structures. The conformational preferences of the loop segments have been analyzed with the specific aim of identifying frequently occurring motifs. Type $\Gamma'$ and $\Pi'$ $\beta$-turns were found to have a high propensity for occurrence in 2 residue loops. For 3 and 4 residue loops, the major conformational motif in the linking segments is $\alpha_R-\alpha_R-\alpha_L$ (type I $\beta$-turn followed by a residue in a left-handed helical conformation) and $\alpha_R-\alpha_R-\alpha_R-\alpha_L$ (a $\pi$-turn motif), respectively. The present larger dataset confirms the high occurrences of these motifs which have been identified in earlier analyses [Sibanda, B L and Thornton, J M (1985) Nature 316, 170-174 and Sibanda, B L, Blundell, T L and Thornton, J M (1989) J Mol Biol 206, 759-777]. In addition to type $\Gamma'$ and type $\Pi'$ $\beta$-turns, several examples of type I $\beta$-turn nucleated 2 residue loop hairpins, in spite of having opposing sense of twist to that of type $\Gamma'$ $\beta$-turn have also been observed. Examination of these frequently occurring motifs (flanked by extended conformation [$\beta$]) in the dataset reveals that the motifs $\beta-\alpha_R-\alpha_R-\alpha_L-\beta$ and $\beta$-type $\Gamma'$-$\beta$ have equal propensity and type $\Pi'$ indeed having highest propensity to nucleate $\beta$-hairpins. The larger number of examples in this study allows estimation of the specific amino acid preferences for loop positions in 2, 3 and 4 residue loops. Small polar residues Asn, Asp, Ser, Thr and Gly and Pro in general have a high propensity for the loop positions but they reveal specific positional preferences in these frequently occurring motifs. There are no strong compositional preferences in the strand segments. Several Cys-Cys pairs have been identified at the non-hydrogen bonded positions of $\beta$-hairpins as many as 6 are disulfide bonded pairs.

Chapter 5 briefly describes a study on $\beta$-turn interconversions observed in protein structures. Inspection of molecular models reveals that interconversions between
type I and type II β-turn structures may be readily possible without breaking the intramolecular 4→1 hydrogen bond, by a process which involves an approximately 180° flip of the central peptide unit. Of the 250 high-resolution (≤ 2.0 Å), non-homologous protein crystal structures, 136 proteins had “homologous entries” (alternate structures, site specific mutants or homologous sequences from different sources) in the PDB which have sequence homology > 40% and the structure determined at high resolution (≤ 2 Å). Based on the sequence alignment of the representative protein with each of its “homologous entries”, 55 examples β-turns undergoing conformational interconversions (type I/III ↔ type II or type I/III' ↔ type II') were identified. Examination of the secondary structures at the flanking positions of the 55 examples of β-turns reveals that a significant number of the examples (16) occur in short segments (≤ 6 residues) linking the secondary structures. A further examination reveals that 7 examples occur in the loop region of the β-hairpins indicating the formation of ordered secondary structures on either side of the β-turn does not preclude local conformational dynamics. In β-turns undergoing flips, Pro (11 examples), Lys (9 examples) and Ser (7 examples) were most often found at the 1+1 position. Glycine was found to occur overwhelmingly at the position 1+2 (28 examples) while Ser (7 examples) and Asn (6 examples) were amongst the most frequent residues. In order to estimate the energy barrier for the type I ↔ type II β-turn interconversions, peptide models Ac-Pro-Aib-NHMe and Ac-Pro-Gly-NHMe were chosen. The AM1 level calculation reveals that the path which corresponds to the outward rotation of the central carbonyl group is barrierless (3.2 kcal/mol in the case of Ac-Pro-Aib-NHMe and 2.8 kcal/mol in the case of Ac-Pro-Gly-NHMe) suggesting that concerted flips of central peptide units involving correlated single bond rotation can occur with essentially negligible activation energy barriers.

Chapter 6 describes a study on αL conformation and on the antigenic tip of the V3 loop peptide of HIV-1 gp120. Residues adopting left-handed helical conformation (αL) was identified from the dataset of 250 high resolution (≤ 2.0 Å), non-homologous
\( \alpha_L \) conformation, 1,510 corresponds to Gly residues, representing the majority of the examples and 1,062 corresponds to non-Gly residues. A preliminary classification of the \( \alpha_L \) conformations, which are distributed in various motifs, leaves out only 67 examples as unclassified. A novel multiple turn conformation, which involves \( \alpha_L \), has been observed for a segment GPGRAFY in the crystal structure of a complex of HIV-1 gp120 V3 loop peptide with the Fab fragment of a neutralizing antibody [Ghiara et al., (1994) Science 264, 82-85]. A structural motif has been defined for the peptide segment employing idealized backbone conformations characterized by ranges of virtual \( C^\alpha \) torsion angles and bond angles. A search in the 250 protein crystal structures permitted identification of 64 examples of similar structural motifs. Two major conformational families have been identified, which differ primarily in the conformation at residue 3. The observed conformation at residue 3 in family 1 is left-handed helical (\( \alpha_L \)) and that in family 2 is right-handed helical (\( \alpha_R \)). Of the 30 examples in family 1, 17 examples have Gly residues at position 3. Of the 31 examples in family 2, 10 examples have Asn/Asp at position 3. Computer modeling of the V3 loop tip sequence using the two backbone conformational families as starting points leads to minimum-energy conformations in which antigenically important side-chains occupy similar spatial arrangements.

The Chapter 7 of the thesis deals with a preliminary study on structural analysis of membrane proteins. Our laboratory has recently been interested in studying membrane protein structures and their comparison with water-soluble globular proteins with the goal to predict the conformation of the membrane spanning segments. This chapter briefly introduces observed structural folds and common structural features in membrane proteins, and prediction of membrane protein structures. A dataset consisting of 55 membrane protein sequences, for which experimental information was available, were obtained from the SWISS-PROT sequence database. Examination of the amino acid preferences to occur in the transmembrane segments and in the interior of the water-soluble globular proteins reveals that both have similar characteristics in their nature of
chemical polarity An analysis on amino acid positional preferences in globular and membrane protein helices has also been carried out

Appendix A deals with confirmation of the important conclusions, obtained in the analysis of Ramachandran-disallowed conformation, with a dataset consisting of 48 protein crystal structures determined at much higher resolution (<1.5 Å) These protein structures have been selected from the latest release of PDB (July, 1997)

The study presented in this thesis has implication in the design of structurally and functionally important motifs, protein engineering, and structure predictions

The projection diagram given in this thesis have been prepared using the program ORTEP which was modified to run on IBM PC's by Prof C Ramakrishnan and Viji For plotting the backbone dihedral angles on the Ramachandran map CRPSPLOT a program developed by Prof C Ramakrishnan and Nagarajaram, was used Ribbon diagrams were prepared using the program MOLSCRIPT kindly provided by Dr Per Kraulis, Center for Structural Biochemistry, Karolinska Institute, Sweden All the text materials were typed using the Microsoft Word version 6.0 (Copyright Microsoft Corporation) All the programs required for identification and analysis were developed by the author for the research purpose The programs were written mostly in C and some in FORTRAN to run on IBM PC's and unix work stations