PART II

PROTEIN DATA ANALYSIS
CHAPTER 4

INTRODUCTORY SURVEY ON PROTEIN DATA ANALYSIS

4.1 INTRODUCTION

The molecular structures of several proteins (more than a hundred) have been solved with reasonable accuracy (with errors in the coordinate of the order of 0.1Å) using X-ray crystallographic techniques in the past few decades. Some of the structures solved at a lower resolution are recently taken up for studies at a higher resolution (Cohen et al., 1981; Takano and Dickerson, 1981). The coordinate data are available with Protein Data Bank or as microfische atlas (AMSM, 1978). Prompted by the access to a wealth of structural data on proteins, work on protein data analysis is being pursued in this laboratory during the past few years. This second part of the thesis describes the analyses carried out by the author regarding some aspects of protein conformation. A brief introduction to the present studies is given in this chapter along with a short review of the earlier data analyses in the literature.
4.2 REVIEW OF CONFORMATIONAL PARAMETERS

Though the positional parameters of the atoms in a macromolecule like protein unambiguously specify its entire spatial structure, a proper evaluation of the specific aspects of information contained in them is better achieved through the use of other quantities/parameters developed for the specific purposes. Various parameters derived using the geometry of the peptide backbone have found application in the characterisation of structural and folding features through different representations.

The pair of fundamental dihedral angle parameters $\phi$ and $\psi$ of the peptide chain, describing the rotations around the bonds $N-C^\alpha$ and $C^\alpha-C'$ respectively, was proposed by Ramachandran et al., (1963). The values of these two angles taken consecutively over the entire chain uniquely specify the complete conformation of all backbone atoms. Using a conformational map based on stereochemical criteria, Ramachandran and Sasi-makharan (1968) arrived at the 'allowed' regions of conformation, which broadly correspond to the major secondary structures found in polypeptides. Conclusions drawn from such maps have enabled corrections
to improbable conformations during model building. Empirical $\varphi - \psi$ diagrams based on observed data on proteins, on the other hand, have been found useful in elucidating certain conformational preferences, an example of which may be found in Chapter 7, (Ravichandran and Subramanian, 1981).

The $\varphi - \psi$ parameters, however, are inadequate to efficiently extract information about certain aspects of protein foldings, as for example, the directional relationships between small segments of the chain. So, more specialised parameters were proposed for such purposes. These are either angle parameters defined for various levels of association of the polypeptide units or vector parameters describing the geometry. It is convenient to classify these under three levels, dealing namely with short segments of protein chain, secondly the backbone units and thirdly the side chain atoms.

The protein chain can be studied at a gross level using the system of virtual bonds between the $C^\alpha$ atoms carrying the residues. This concept is widely used in polymer statistics (Brant and Flory, 1965) and an application of this is found in Chapter 6.
of this thesis. For a molecular architecture described
by the virtual bonds of constant length (Pauling and
Corey, 1951), the corresponding bond angle and torsion
angle emerge as suitable parameters. Thus, the virtual
bond angle defined by \( \varepsilon_i = C_{i-1}^\alpha - C_i^\alpha - C_{i+1}^\alpha \) (Srinivasan
et al., 1977) and the torsion angle \( \theta \) defined for a
segment of four C\(^\alpha\) atoms by
\[ \theta_i = C_{i-1}^\alpha - C_i^\alpha - C_{i+1}^\alpha - C_{i+2}^\alpha \]
(Srinivasan et al., 1977) were proposed. The constancy
of \( \theta \) in regular structural regions made it suitable
for delineating such regions (Srinivasan and Rajan,
1978; Srinivasan, 1978). A probability distribution of
\( \theta \) values in several proteins led to the prediction of a
new secondary structural feature called \( \varepsilon \)-helix
(Srinivasan et al., 1976b). The angle \( \varepsilon \) has a restric-
ted range of 0°–180° while \( \theta \) has a full range of
variation from -180° to 180°.

In a study of the gross conformational features
of protein chains, a segment containing four C\(^\alpha\) atoms
serves as the smallest convenient unit. Such a seg-
ment is referred to as the local helical segment. If
a constant value is assumed by a conformational para-
meter like \( \theta \) for successive such segments, it indicates
the presence of regular structures. Quite often,
however, one is interested in knowing the directional
features of adjacent segments. A vector parameter $\mathbf{h}_i$ describing the direction of the axis of a local helical segment at residue $i$ was proposed by Srinivasan et al. (1975) to satisfy this requirement. In an ideal $\alpha$-helix the $\mathbf{h}_i$'s at different residues align themselves parallely. Any deviation from such parallel alignment indicates how much does a given helix deviate from the ideal case. Such deviations are quantitatively assessed by means of the angle between the corresponding axial vectors. An angle parameter $\theta_{ij}$ defined between any two vectors $\mathbf{h}_i$ and $\mathbf{h}_j$ is generally useful in such studies (Rajan et al., 1976). This angle helps to study the spatial orientation of any segment $i$ from another segment $j$ and can serve as a measure of distortions in helical regions. In regular structural regions, $\theta_{ij}$ is a constant. Best helical parameters for the $\alpha$-helix based on myoglobin data were arrived at using this parameter (Rajan and Srinivasan, 1977). Another vector parameter employed in data analysis is the virtual bond $\mathbf{l}_i$ itself between successive C$^\alpha$ atoms. The distribution of the distance between any two C$^\alpha$ atoms at $i$ and $j$, denoted by $l_{ij}$ or of the angle $\theta_{ij}$ discussed earlier serves as a structural fingerprint (Ooi and Nishikawa, 1973; Rajan and Srinivasan, 1977).
The parameters $\phi$ and $\gamma$ belong to the peptide unit and therefore to the second level. The peptide unit has been analysed using the unit vectors along the plane normals denoted by $\vec{n}_i$. The angle between any two normals $\vec{n}_i$ and $\vec{n}_j$ is denoted by $\gamma_{ij}$ (Balasubramanian and Srinivasan, 1976). Although these normals can provide information on the orientations of peptide planes, they are not so convenient to handle. A new dihedral angle parameter between adjacent peptide planes defined using a pair of real bonds and one virtual bond introduced in Chapter 5, is suitable for easy application (Srinivasan and Revichandran, 1982).

4.3 DATA REPRESENTATIONS

4.3.1 Graphical Representations

Representations of the parameters described above either taken singly or as different combinations provide comprehensive methods of characterising structural and conformational features. While many of these characterise the folding patterns, some of these serve as correlation maps. The single parameter representations, called the 'chain plots' are the linear plots
of the value the parameter assumes at each residue site, as a function of the site. The parameters like $\Theta_i$ or $\eta_i^*$ have been plotted like this. The $\Theta$-plots are valuable aids in analysing the sense and the pattern of folding. The twin parameters $\phi_i$ and $\psi_i$ have also been plotted together as a function of $i$ in a variation of such chain plots (Balasubramanian, 1977). The circular chain plot which pack more information such as type of residues etc., is another modification of the chain plot (Srinivasan and Srinivasan, 1977).

In the two-parameter representations, the variation of one of the parameters on the other, as observed in a given protein, is shown on a plane diagram. The energy plot as a function of $\phi, \psi$ values is a typical example. Yet another representation of similar kind in the form of square-matrix plots (sometimes called $i-j$ maps), deal with a single parameter but considered between two residues $i$ and $j$. The $l_{ij}$ maps (Rossmann and Liljas, 1974) pictureise folding patterns and

$^*$ $\eta_i$ denotes the angle between two consecutive vectors $h_i$ and $h_{i+1}$. In the usual notation this will be represented by $\kappa_{i,i+1}$ but for convenience the second subscript is dropped.
symmetry and the $\eta_{ij}$ maps (Rajan and Srinivasan, 1977) describing the segment orientations are useful structural fingerprints. The $\eta_{ij}$ is useful in the analysis of distortions present in helices (Srinivasan et al., 1981) as seen in Chapter 8. Pairs of parameters of different levels like $(\theta_i, \gamma_i)$, $(\phi_i, \theta_i)$, $(\psi_i, \theta_i)$ have been plotted in the nature of correlation maps. A better correlation study is made possible through a 'triple-plot' where three different parameters are studied together by plotting them in a three-dimensional perspective graph (Chandrasekhar, 1982).

The complete three-dimensional structures are in some cases studied through models (Richards, 1968; Lee and Richards, 1971) or stereo pictures in computer graphics incorporating thermal ellipsoids (Diamond, 1978; Johnson, 1976). The structure cartoons (Dickerson and Gies, 1968; Richardson, 1977; Schulz and Schirmer, 1974) provide spatial orientations of secondary structures.

4.3.2 Stereographic Projection

This method is well known in crystallography in the study of interfacial angles of crystals
(Phillips, 1954; Cullity, 1967). It can also be applied to represent clearly the relative orientation of the segment axes in a protein by connecting sequentially the projections of the corresponding $\vec{h}_i$ vectors on a stereogram (Srinivasan et al., 1975). This is particularly powerful in the examination of helical regions (Srinivasan and Rajan, 1978). An application of this technique in bringing out the interesting features of distortions in helices which are otherwise not apparent, is described in Chapter 8 (Srinivasan et al., 1981).

4.4 CONFIGURATIONAL STATISTICS IN PROTEINS

The configurational properties of macromolecular systems like polymer chains can be studied as a statistical average over the entire chain. Two parameters generally useful in such studies and employed in the solution studies of synthetic polymers as well as biopolymers are the radius of gyration and the end-to-end distance. Results from the solution studies describe the chain properties as the average over the different configurations the chain assumes in the solution. But, in the case of proteins, they
are unique in that they take up invariably a specific folded conformation. The 'native' conformational state of the globular protein is specific and completely retained in the solid state. Although for a given protein the conformation is precise enough at the atomic level, the wealth of crystallographic coordinate data on several proteins are available to enable us to extract statistical information. With this in view, a systematic analysis along this line has been recently carried out in this laboratory. The radius of gyration and the end-to-end vector were studied for several proteins. The relation between the mean square radius of gyration and the mean square end-to-end distance in the case of proteins was deduced to be \( \langle h_n^2 \rangle = 3.7 \langle s_n^2 \rangle \) (Srinivasan and Chandrasekhar, 1982). The radius of gyration itself was found to have a one third power dependence on the number of atoms in proteins (as well as in 'medium-sized' molecules). This as well as the results on packing in protein molecules obtained through radius of gyration are discussed by Srinivasan and Vijayalakshmi (1982). An analysis of the end-to-end distances in individual proteins was also taken up using a bootstrap technique to generate segments of various lengths and the conclusions derived are presented in Chapter 6 of this thesis.