

**THEORETICAL STUDIES ON PROTEIN FOLDING:
PROTEIN CRYSTAL STRUCTURE DATA ANALYSIS**

**A THESIS
SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY**

**BY
V. RAMABRAHMAM**

**SCHOOL OF LIFE SCIENCES
UNIVERSITY OF HYDERABAD
HYDERABAD 500 134**


JULY 1983


SCHOOL OF LIFE SCIENCES
UNIVERSITY OF HYDERABAD
HYDERABAD 500 134

Dated: 27th July 1983

CERTIFICATE


This is to certify that Mr. V. Ramabrahmam, has carried out the research work embodied in the present thesis under my supervision and guidance for the full period prescribed under the Ph.D. Ordinances of the University. I recommend his thesis entitled "THEORETICAL STUDIES ON PROTEIN FOLDING: PROTEIN CRYSTAL STRUCTURE DATA ANALYSIS" for submission for the degree of Doctor of Philosophy in this University.


PROF. P.S. RAMAMURTY
Dean
School of Life Sciences


DR. A.S. KOLASKAR
(Supervisor)

D E C L A R A T I O N

I hereby declare that the work presented in this thesis has been carried out by me under the supervision of Dr. A.S. Kolaskar and that this has not been submitted for a degree or diploma of any other University.



V. RAMABRAHMAM

Date: July 27, 1983
Place: Hyderabad

A.S. Kolaskar
DR. A.S. KOLASKAR
(Supervisor)

ACKNOWLEDGEMENTS

I am short of words to express my gratitude to Dr. A.S. Kolaskar for suggesting this interesting topic and for his stimulating, efficient and professionalised guidance. He provided both direction and stabilization for this 'folding'. I am benefited in many ways from our association. My regards to him.

I thank ICRISAT library authorities for allowing me to make use of Microfiche reading facilities. I thank W.A. Khan of our C I L for his assistance during computations.

I owe a lot to my near and dear for comforting me by sharing my moments of disturbance and joy.

Special mention should be made of the assistance rendered by my wife Kameswari and of the help by my friends, Jagannadham, Mitra, M.S.N. Murt and Rāma, during the final phase of thesis preparation..

I am thankful to Mr. T. Rama Rao for his neat and efficient typing of this manuscript. I thank Mr. Giridharan for the neat sketches.

Finally, I am indebted to University Grants Commission for its financial assistance in the form of junior and senior research fellowships. Also my salutations to protein crystallographers, but for whose dedicated, efficient and hard work, this thesis might not have taken shape.


V. RAMABRAHMAM

S Y N O P S I S

Many attempts are being made to predict the three dimensional structure of a protein from its amino acid sequence. Conformational energy calculations and analysis of protein crystal structure data are two principal theoretical approaches. The success achieved in the prediction of three dimensional structure using conformational energy calculations being limited, the efforts were concentrated to retrieve maximum information from available crystal structures of large number of globular proteins and develop an algorithm on a micro computer to predict rough three dimensional structure of globular proteins from amino acid sequence.

As a first step towards this goal, individual (ϕ, ψ) -probability distribution maps were obtained from crystal structure data of globular proteins for the 20 proteinous amino acids. The comparison of these (ϕ, ψ) -probability maps with one another has shown that (ϕ, ψ) -probability distributions are similar for some residues though their side chains exhibit different chemical and physical properties. This similarity seems to be an intrinsic property of amino acid residues and we termed it as "main chain conformational similarity". Using this similarity together with other similarities among amino acids, an attempt is made to understand amino acid replacements taken place during evolution in homologous proteins. Main chain conformational similarity among amino acids is further utilised to postulate a set of obligatory amino acids that might be present in primitive proteins. The (ϕ, ψ) -probability distribution data are also analysed to obtain main chain conformations significantly effected by respective side chains.

In order to get some insight into the near-neighbour interactions, amino acid pair potentials were derived by analysing data on crystallographically observed secondary structures, assuming four-state model. These amino acid pair potentials are different from those obtained by combining two respective single residue potentials. Our analysis has also shown that only about 55% of the observed secondary structures in proteins have inbuilt high tendency to assume the said structure. We have termed these as type I secondary structures. Remaining observed secondary structures, termed type II, seem to have been formed mainly due to tertiary interactions. This seems to be the major reason for limited success in predicting secondary structures of proteins using present-day techniques. Therefore as an alternative, to obtain rough three dimensional structure of a protein from primary structure, we have divided the (ϕ, ψ) -probability map into three regions, namely, region A (ϕ from -140°C to 0°C and ψ from -100°C to 0°C), region B (ϕ from -180°C to 0°C and ψ from 80°C to 180°C) and region C (rest of (ϕ, ψ) -plane). Amino acid pair potentials were computed for regions A and B and are used to predict these regions using a simple algorithm.

Chapter-wise summary of the thesis is given below:

CHAPTER I:

An attempt is made to review the existing literature on theoretical studies of prediction of protein structures, both secondary and tertiary. Mechanism of protein folding is also discussed very briefly to get an overall view of this subject.

CHAPTER II:

In this chapter, the method adopted to obtain sets of residues exhibiting main chain conformational similarity, using (ϕ, ψ) -probability data obtained from crystal structures of 38 different globular proteins is discussed. The results obtained suggest that main chain conformational similarity is as much an intrinsic property of amino acids as their secondary structure preference, chemical nature and other physical properties.

CHAPTER III:

With the assumption that an amino acid mainly consists of two parts, the main chain and the side chain, the (ϕ, ψ) -probability maps were analysed to obtain those main chain conformations of amino acids which are significantly effected by their side chains. The method used to derive these conformations is discussed. The stabilizing or destabilizing effect of respective side chains on main chain conformations is also discussed.

CHAPTER IV:

Information obtained in chapters II and III is not sufficient to develop an algorithm for the prediction of three dimensional structure. Further, since the single residue potentials used in secondary structure predictions are mostly compositional in nature, amino acid pair potentials are derived, in which sequence effect is present to a certain extent. Amino acid pair potentials were found to vary not only with types of amino acid residues in a pair but also with the position they occupy in a pair and the secondary structure in which these pairs occur.

CHAPTER V:

The amino acid pair potentials were used to analyse the crystallographically observed secondary structural segments. This analysis has shown that for only about 55% of observed secondary structures the calculated average potential was maximum for the observed structure. Results obtained are not very different, when single residue potentials derived by us, Chou and Fasman and Garnier and Robson, are used. Reasons for the limited success of present-day secondary structure prediction attempts are discussed in this chapter.

CHAPTER VI:

So, as an alternative to prediction of secondary structures, we have developed a simple algorithm based on probability of occurrence of pairs of amino acids in thickly populated regions of Ramachandran plot, to predict these regions.

APPENDIX I

The method applied to understand the amino acid substitutions occurred during evolution in homologous proteins cytochrome C and haemoglobin, using main chain conformational similarity and other properties. of amino acid residues obtained is presented and the results are discussed.

APPENDIX II

The method developed to derive obligatory amino acids in primitive proteins utilizing main chain conformational similarity among amino acids is discussed and the results obtained are presented.

Thus in short, in this thesis, it has been pointed out that, using a micro computer, one can get large amount of useful information from available crystal structure data, which not only throws light on properties of individual amino acids, but also on observed secondary structures in globular proteins. Further, attempt has been made to show that the derived information can be successfully utilized, in a simple fashion to get a rough three dimensional structure of protein from its primary structure.

LIST OF PUBLICATIONS

1. Conformational similarity among amino acid residues 1. Analysis of proteins crystal structure data. A.S. Kolaskar and V. Ramabrahmam, Int. J. Biolog. Macromolecules, (1981), 3, 171.
2. Side chain characteristic main chain conformations of amino acid residues. A.S. Kolaskar and V. Ramabrahmam. Int. J. Peptide Proteins Res. (1982), 19, 1.
3. Conformational properties of pairs of amino acids. A.S. Kolaskar and V. Ramabrahmam, Int. J. Peptide Proteins Res. (1983), 22, 83.
4. Nature of amino acid substitutions in homologous proteins during evolution. A.S. Kolaskar and V. Ramabrahmam. Int. J. Biol. Macromolecules, (1982), 4, 151.
5. Obligatory amino acids in primitive proteins. A.S. Kolaskar and V. Ramabrahmam, Biosystems (1982), 15, 105.
6. Analysis of crystallographically observed secondary structures. A.S. Kolaskar and V. Ramabrahmam (communicated).
7. A review of prediction of protein structural organisation A.S. Kolaskar and V. Ramabrahmam (in preparation)

C O N T E N T S

	<u>Page Number</u>
Certificate	i
Declaration	ii
Acknowledgements	iii
Synopsis	iv
List of Publications	ix
CHAPTER I: Prediction of protein structural organisation - A brief review	1
CHAPTER II: Conformational similarity among amino acid residues	18
CHAPTER III: Side chain characteristic, main-chain confor- mations of amino acid residues	42
CHAPTER IV: Conformational properties of pairs of amino acids	65
CHAPTER V: Analysis of crystallographically observed secondary structures	86
CHAPTER VI: Prediction of rough structure of proteins using Ramachandran plot	104
APPENDIX I: Nature of amino acid substitutions in homologous proteins during evolution	115
APPENDIX II: Obligatory amino acids in primitive proteins	135
REFERENCES	145

CHAPTER I

**PREDICTION OF PROTEIN STRUCTURAL ORGANISATION -
A BRIEF REVIEW**

I.1 INTRODUCTION

The solution of protein folding problem, being complex and challenging, has attracted the attention of several workers in the last 30 years. Many models of protein structural organisation have been proposed, the first one being an oil drop model by Kauzmann (1959). Since then many review articles have appeared on this subject; based on theoretical and experimental studies (Tranford, 1968; Wetlaufer and Ristow, 1973; Baldwin, 1975; Anfinsen and Scheraga, 1975; Nemethy and Scheraga, 1977; Richards, 1977; Creighton, 1978; Privalov, 1979; Jaenicke, 1980; Ptitsyn and Finkelstein, 1980; Schechter and Goldberger, 1980; Ptitsyn, 1981; Wetlaufer, 1981; Richardson, 1981; Rossmann and Argos, 1981; Ikegami, 1981; Ghelis and Yon, 1982; and Kim and Baldwin, 1982). A study of these articles suggests that the understanding of protein folding and prediction of three dimensional structure of proteins from amino acid sequence are still in a nascent stage, though a large amount of information has been gathered by experimental studies, using various techniques such as electrophoresis (Creighton, 1978; Wetlaufer, 1981; Kim and Baldwin, 1982); X-ray crystallography as compiled by Feldmann (1976); ORD and CD (Brahms and Brahms, 1980); Laser Raman spectroscopy (William and Dunker, 1981) and NMR (Wagner and Wüthrich, 1982).

In this chapter, an attempt has been made to briefly review the methods which have been developed for the theoretical prediction of various levels of protein structural organisation. These theoretical prediction schemes are mostly based on analysis of protein crystal structure data.

In short, we have avoided giving historical background of the subject as this has been done in many of the reviews cited above. Similarly review of literature on experimental studies is not undertaken since the work presented in this thesis is concerned with the analysis of protein crystal structure data. On the other hand this chapter is aimed at providing the necessary background and continuity for the work presented in the succeeding chapters of this thesis.

With this in view we have discussed the advantages and drawbacks of the methods used to predict the different levels of protein structural organisation. Further, we have also discussed the information gained by analysing experimental data at the amino acid residue level. This is considered to be essential because certain properties of individual amino acids such as conformational similarity or specific main chain conformations get affected by individual side chains (details are presented in Chapter II and Chapter III respectively) are useful in understanding the phenomenon of protein folding. This chapter is divided into three major sections.

- (i) Individual properties of amino acids - present status
- (ii) Secondary structure prediction schemes - achievements and drawbacks.
- (iii) Attempts at predicting other structural levels - i.e., topology, domains and tertiary structure.

I.2 INDIVIDUAL RESIDUE PROPERTIES

Hydrodynamic studies of homo- and hetero- polypeptides, ORD and CD studies on a large number of polypeptides and proteins and analysis of

protein crystal structure data have been used by several workers to divide the 20 proteinous amino acids into many categories based on their chemical and physical nature including secondary structure forming capacities. These are listed in Table I.

Among the properties cited in Table I, hydrophobicity, which is mostly entropic in nature, has been assumed by Kauzmann (1959) as the guiding force of protein folding. Though the values determined by Nozaki and Tanford (1971) are compositional in nature, no attempt is made till recently to study the variation of hydrophobicity along the primary structure of a protein. Kyte and Doolittle (1982) have given one such method. Manavalan and Ponnuswamy (1978) looked at this property quite differently and considered hydrophobicity of a residue in a protein as a function of amino acids which are near-neighbours sequentially and from the three dimensional point of view. They have developed an algorithm to determine the hydrophobicity of an amino acid in a particular environment. A critical analysis of various methods to determine hydrophobic character of amino acid residues was given recently by Meirovitch et al. (1980).

Determination of amino acid residue tendencies to be present in various secondary structures, another important amino acid property, has also been widely pursued. Robson and Pain (1974) have used information theory to know the amino acid preference of secondary structures. Tanaka and Scheraga (1976) have derived the conformational properties of amino acids in secondary structures using statistical mechanics. Another method,

TABLE I

Physical and chemical properties of amino acids

Amino acid	Bulkiness ^a	polarity ^a	R _F ^a	P _{k_l} ^a	Hydrophobicity ^b	Refractivity ^d	Chemical nature ^e	Secondary structure preference ^f
Ala	11.50	0.00	9.9	2.34 ^c	0.87	4.34	Aliphatic	α -helix
Arg	14.28	52.00	4.6	1.81 ^c	0.85	26.66	Basic	--
Asp	11.68	49.70	2.8	2.01 ^c	0.66	12.00	Acidic	Reverse turn
Asn	12.82	3.38	5.4	2.02	0.09 ⁻	13.28	Acid Amide	Reverse turn
Cys	13.46	1.48	2.8	1.65	1.52	35.77	Sulphur containing	α -helix
Glu	13.57	49.90	3.2	2.19	0.67	17.26	Acidic	α -helix
Gln	14.45	3.53	9.0	2.17	0.00	17.56	Acid Amide	α -helix
Gly	3.40	0.00	5.6	2.34 ^c	0.10	0.00	Aliphatic	Reverse turn
His	13.69	51.60	8.2	1.82	0.87	21.81	Basic	α -helix
Leu	21.40	0.13	17.6	2.36	2.17	18.78	Aliphatic	α -helix
Ile	21.40	0.13	17.1	2.36	3.15	19.06	Aliphatic	β -sheet
Lys	15.71	49.50	3.5	2.18	1.64	21.29	Basic	α -helix
Met	16.25	1.43	14.9	2.28	1.67	21.64	Sulphur containing	α -helix
Phe	19.80	0.35	18.8	1.83	2.87	29.40	Aromatic	β -sheet
Pro	17.43	1.58	14.8	1.99 ^c	2.77	10.93	Imino Acid	Reverse turn
Ser	9.47	1.67	6.9	2.21 ^c	0.07	6.35	Hydroxy	Reverse turn
Thr	15.77	1.66	9.5	2.10	0.07	11.01	Hydroxy	β -sheet
Trp	21.67	2.10	17.1	2.38	3.77	42.53	Aromatic	β -sheet
Tyr	18.03	1.61	15.0	2.20	2.67	31.53	Aromatic	β -sheet
Val	21.57	0.13	14.3	2.32	1.87	13.92	Aliphatic	β -sheet

a: Zimmerman et al. (1968); b: Mozaki and Tanford (1971); c: From measurements without liquid junction, Cohn and Edsall (1943); e: Stryer (1975); f: Levitt (1978). Three letter amino acid code has been used.

which is mostly statistical in nature with little physical meaning or basis, is the analysis of protein crystal structure data (Chou and Fasman, 1974; Levitt, 1978) and assigning potentials to each amino acid residue in various secondary structures. Argos and Palau (1982) have recently calculated correlation coefficients between physico-chemical properties of 20 proteinous amino acids and their occurrence at a particular position within a secondary structural element.

Chemical properties like aliphaticity, aromaticity or acidity (p^K) have not been directly used in protein folding studies. Only recently, bulkiness or volume of each individual residue has been used in protein structural studies (Richards, 1977). Physical properties such as dipole moment of individual amino acid residues are yet to be used. Hol et al. (1981) have used dipole moment of helices and β -strands to predict their association and thus to gain information regarding their topologies. However, as has been discussed in Chapter II of this thesis, no attempt has been made before to look into conformational similarity or the similarity in restriction on the orientation of main chain atoms. This is observed to be an intrinsic property of amino acid residues which can be used profitably in protein folding studies.

1.3 SECONDARY STRUCTURE

As early as in 1951, Pauling and co-workers had postulated the existence of secondary structures such as the α -helix (Pauling et al., 1951) and β -sheet (Pauling and Corey, 1951). Venkatachalam (1968) had

proposed the existence of β -bends (or chain reversals) from model building studies. The direct evidence for existence of α -helix and β -sheet in protein molecules has come from crystallographic studies of globular proteins such as myoglobin (Kendrew et al., 1960) and Lysozyme (Blake et al., 1965). Simultaneously Scheraga's group has done solution studies of homo-polypeptides and hetero-polypeptides. Based on statistical mechanics, they have also developed models first of helix-coil transitions and then sheet-coil transitions (Anfinsen and Scheraga, 1975). During these years a large amount of data has accumulated from X-ray diffraction studies of single crystals of various globular proteins. Also, many attempts were simultaneously made to predict the secondary structural regions of globular proteins.

There exists two main approaches to predict secondary structures in globular proteins:

- (i) Statistical mechanical or physical treatment of polypeptide chains, and
- (ii) Statistical analysis of protein crystal structure data.

Most of the studies based on statistical mechanics have been carried out by Scheraga and co-workers. This method uses the one dimensional Ising model. This approach was reviewed by Anfinsen and Scheraga (1975) and has been applied to predict secondary structures by Tanaka and Scheraga (1976). Finkelstein and Ptitsyn (1971) calculated the potential of a residue by assuming the hydrophobic effect as done by Nozaki and Tanford (1971) in the presence and absence of an hydrophobic environment, and used them to predict secondary structures.

These two methods give secondary structure predictions with the same order of accuracy as by statistical methods developed by analysing protein crystal structure data. The advantage of these methods, since the basis of their development is physical, is that the analysis of their success or failure will be useful in understanding protein folding process. But, in view of the computational complications associated with these methods they are less popular.

On the other hand methods developed using analysis of crystal structure data of globular proteins are simple though totally empirical. Many secondary structure prediction schemes are available under this class (Ptitsyn and Finkelstein, 1970; Bunting et al. 1972; Kabat and Wu, 1973; Nagano, 1973; Burgess et al., 1974; Lim, 1974; Argos et al., 1976; Chou and Fasman, 1978; and Garnier et al., 1978). Only the last two methods by Chou and Fasman (1978) and Garnier et al. (1978) are popular. These two methods are briefly discussed here.

(a) Chou and Fasman method: In this method crystal structure data of globular proteins is analysed and propensities of each amino acid residue in helical, β -sheet and chain reversal states are calculated using the following simple formula:

$$P_{j,k} = P_{j,k}/P_j \quad \text{..... (1)}$$

$P_{j,k}$ = Conformational propensity (or potential) of j amino acid in k state.

$P_{j,k}$ = Probability of finding the j residue in k state

P_j = Probability of finding j residue in proteins.

These calculated potential values range between 0 and 2 for various amino acids and vary according to their preference for a particular secondary structure.

Nucleation sites are determined by locating clusters of four helical residues out of six, for the helical region and three β -sheet residues out of five for the sheet region. Weightages are given to residues for identifying termination regions of these secondary structural stretches. These weightages are derived by a trial and error method. The strategy used for predicting bends is different.

The prediction scheme of Chou and Fasman has the advantage that it can be executed on a programmable calculator also. Hence, this method is used widely though it is not the most efficient. In fact, as opined by Sternberg and Cohen (1982), this method requires subjective judgement in the assignment of α - and β -structures. This method of Chou and Fasman has been modified by some workers recently (Cid et al., 1980; and Palau et al., 1982).

(b) Garnier et al. method: This is a modified and simplified method of Robson and Pain (1971). This is basically a directional method, which uses single residue potentials derived from information theory, and takes care of the influence, of eight residues before and eight residues after, on the conformation of the residue being decided. A parameter decision constant, quantifies the influence of cooperative effects such as hydrophobic interaction on the formation of secondary structures. Busetta and Hospital (1982) have modified and used this method.

Discussion on secondary structure prediction schemes

The accuracy of secondary structure prediction schemes seems to be limited. Bourgeois et al. (1979) have used six different ways of predicting secondary structures of the lac repressor protein. From this study and from other studies (Schulz et al., 1974) it is clear that, no single method predicts the secondary structures with more than 50-55 per cent accuracy. From this it also seems that though a particular method is good for proteins of certain structural class or for predicting particular type of secondary structure, its value as a general prediction scheme is quite limited.

Analysis presented in Chapter V of this thesis divides crystallographically observed secondary structures into two categories and there the possible reasons for the failure of secondary structure prediction schemes is discussed. Also recently Palau et al. (1982) and Busetta and Hospital (1982), have hinted that the upper limit of prediction accuracy has been reached for secondary structure prediction schemes; which is about 60 percent. In the discussions in Chapter V, of this thesis, the need for development and incorporation of weighting factors that can take care of tertiary interactions in secondary structure prediction schemes is expressed to obtain more accurate secondary structure predictions.

I.4 TOPOLOGY PREDICTIONS

The next step in the hierarchy of protein structural organisation is the associations of secondary structures as topologies.

The topological features of association of various secondary structures have been thoroughly reviewed by Richardson (1981) and some reasons for limited topologies were given by Ptitsyn and Finkelstein (1980). Currently, the following procedures are being adopted for the prediction of topologies.

(a) Combinatorial approach: This approach starts with known secondary structural regions of proteins and tries to pack them in the most agreeable state compared to the native structure. All possible associations are generated and then those that violate stereochemical stipulations are excluded. The structures that survive these filters, include a native-like structure. This approach has been followed to pack all α , all β and α/β structures separately (Cohen et al., 1982). The demerit of this approach is that it takes excessive computer time.

(b) Free energy considerations: This method analyses crystal structures, and estimates of different contributions to the free energy of a folded protein are made. These free energy estimates include: hydrogen bonding between residues present in various secondary structures, side-chain-side-chain interactions present in different secondary structures and hydrophobic interactions between various secondary structural segments (Busetta and Barrans, 1982). Prior knowledge of secondary structural regions is necessary and is obtained by predicting secondary structures by other methods.

(c) Signed distance map method: This method is based on C^α -coordinates and does not require the location of specific regular structures.

The handedness of α -helices and β -strands are represented on a signed distance map. Also the handedness of association of two regular structures can be observed (Braun, 1983).

(d) Use of Dipole moments: Hol et al. (1981) have made use of dipole moments of α -helices and β -sheets and their interactions and hence predicted their arrangements in proteins.

Discussion

Lambhardt (1982), in his experimental studies on Ribonuclease S, demonstrated the necessity of folded β -sheet to observe the folding of S-peptide helix. Also the possibility of conversion of helix to sheet or sheet to helix is provided by Louie and Somorjai (1982) based on differential geometry. Thus, formation of secondary structures and associations as a particular topology do not seem to be two independent processes and one seems to influence the formation of the other.

I.5 DOMAINS

Domains are independent structural units of globular proteins. They form very compact globules by having many internal contacts, but a few contacts with other parts of the polypeptide chain (Schulz and Schirmer, 1979). The following properties can generally be ascribed to a domain (Rossmann and Argos, 1981).

(a) Similar domain structures or their amino acid sequences can be found either within the same polypeptide or in a different molecule.

(b) Domains within a polypeptide are spatially separated from each other, or at least form a compact 'glob' or cluster of residues.

(c) Domains have a specific function, such as binding a nucleotide or polysaccharide.

Rašhin (1981), and Wodak and Janin (1981), have made use of surface area measurements to locate the domains in proteins. The two methods differ in their formulations and parameters used. Signal distance map developed by Braun (1983) can also be used to predict domain region of a protein.

I.6 TERTIARY STRUCTURE PREDICTIONS

Many attempts were made to predict directly the tertiary structure from amino acid sequence avoiding prediction of intermediate levels of protein structure. These can be divided roughly into three categories:

- (a) Energy calculations
- (b) Empirical schemes
- (c) Folding simulations

(a) Energy calculations: It is a general belief that the tertiary structure of a protein corresponds to minimum free energy. Nemethy and Scheraga (1977) have observed that energy calculations with a complete description of the free energy surface of polypeptide chain followed by efficient energy minimization could predict protein structures. Adopting this procedure, many attempts have been made to calculate the tertiary

structure i.e., the assignment of various atomic positions in protein space. All these attempts have not been successful beyond a certain modest level. This moderate success might be due to: (a) computational problems arising when dealing with non-linear free energy function which involves a large number of variations in assigning various positions at atomic level, (b) inadequacies of existing potential functions to model solvent effects, and (c) enormous amount of computational time requirements for energy minimization. Cohen et al. (1982) and Goel et al. (1982) have discussed at some length these reasons for limited success of this approach. The problem of excessive computation time was earlier over-come by Levitt and Warshel (1975) by reducing the atomic representation to two effective centres per residue, the main chain and the side chain. But this could not give accurate prediction when used even on small proteins such as BPTI.

(b) Empirical schemes: One simplified empirical method is a geometrical model which does distance geometry calculations; it needs experimental or theoretical information to determine a partial matrix of inter-atomic distances. This approach attempts to calculate the tertiary structure by imposing a set of constraints, which reflect some aspects of the folded structure. The work of Goel et al. (1982) is a recent application of this method and contains references about other groups attempting predictions on similar lines.

Another empirical method that predicts tertiary structure of a protein is by Prabhakaran and Ponnuswamy (1980). They assume the

protein space to be an ellipsoid and try to predict the amino acid residue preference in that space by taking into consideration their hydrophobicity, solvent contact area and spatial positions. Busetta (1982) has also provided an improved residual representation of proteins.

(c) Folding simulations: As mentioned earlier, the simplified approach of Levitt and Warshel (1975), has a limited use in simulating protein folding. Hagler and Honig (1978) have provided a critical analysis of the failures of Levitt and Warshel (1975) method. They observe that bend regions can be introduced in the appropriate position without the use of artificial torsional potential. A somewhat similar method has been developed by Kuntz et al. (1976). They consider $3N$ cartesian coordinates as independent variables, where N is the number of residues in the protein molecule. Robson and Osguthorpe (1979) performed folding simulations using an angular variable φ , which couples the variation of ϕ and ψ of the same residue. Here the authors stress the role of certain coil-state residue having sufficient flexibility to guide the folding of the protein.

A quite different approach was tried by Gö and co-workers (1980). They simulated the folding process assuming a self-avoiding random walk model on a two dimensional square lattice. They studied folding process per se, by simulating folding by the Monte Carlo process of Metropolis et al. (1953) and allowed transitions and rotations of both single units and sequence of units. Meirovitch and Scheraga (1981) also simulated protein folds based on energetic and geometric criterion.

Rapport and Scheraga (1981) have tried to simulate the folding of polypeptide chains of homopolymers on a high speed array processor. Recently Krigbaum and Lin (1982) tried Monte Carlo simulations of protein folding and the structures were generated with centrosymmetric and local interaction potentials for the protein BPTI. All the studies mentioned till now are theoretical studies of folding phenomenon and are in the initial stage and require further improvements.

1.7 CONCLUSIONS

The present state of art, regarding the problem of protein folding without getting into the mechanism of the process, indicates that the wealth of information available in crystal structures of globular proteins has not been fully explored; except probably in the case of prediction of secondary structures. However, as has been discussed in Chapter V of this thesis, and recent experimental results (Wetlaufer, 1981; Lambhardt, 1982; Galat, 1982) all secondary structures are not formed due to intrinsic preference of amino acid residues, but are also due to favourable tertiary interactions, and therefore, there is a need to understand the properties of amino acid residues at other levels such as their orientations in three dimensional space or their preferences in taking main chain dihedral angles ϕ and ψ .

Secondly, one should try to predict, at least, a rough three dimensional structure of proteins directly, without going through the step of prediction of secondary structures. This predicted rough three dimensional structure should then be refined using information regarding

chemical, biochemical, physical and conformational properties of amino acids and considering the polypeptide chain in the native state, rather than attempting to solve the problem by combursome energy minimisation procedures or mere statistical methods. In other words, sufficient weightage should be given to structural class of the folded polypeptide and functional properties of the protein.

In this thesis, an attempt is made to get information about conformational properties of amino acids and observed secondary structures in globular proteins. Further an attempt has been made to show that the derived information can be successfully utilized in a simple fashion to get a rough three dimensional structure of the protein from its primary structure. It may be further mentioned that all required computations are carried out using a simple INTEL 8080 microprocessor based microcomputer.

CHAPTER II

CONFORMATIONAL SIMILARITY AMONG AMINO ACID RESIDUES

II.1 INTRODUCTION

Amino acid residues are the basic building blocks of protein and the peptide unit is the repeating unit of polypeptide chain. Therefore, energy calculations carried out at the dipeptide level gives an insight into conformational behaviour of amino acid residues. The hard sphere model of Ramachandran has identified allowed and dis-allowed regions in (ϕ, ψ) -plane (Ramachandran, 1962; Sasisekharan, 1962; Ramachandran et al., 1962). Refinement of these studies using empirical potential energy functions at dipeptide level were also undertaken by various groups (Desantis et al., 1965; Brant and Flory, 1965; Kitaigorodsky, 1965; Scott and Scheraga, 1965; Ramachandran et al., 1966).

The above-mentioned studies, as well as several other studies carried out later, have shown that the potential energy maps at the dipeptide level vary with the set of semi-empirical potential energy functions used. This prompted Ramachandran (1973) to discuss at length the inaccuracies in the potential functions used in such studies. It has also been noticed that protein empirical energy map derived by Pohl (1971), using crystal structure data of globular proteins, was different in certain parts of (ϕ, ψ) -plane, when compared with the maps obtained using semi-empirical energy function mentioned above and also with those maps obtained from quantum chemical methods such as CNDO/2, PCILO and ab-initio methods (Pullman and Pullman, 1974). This observation indicates not only the inaccuracies in the potential functions used but, probably, also the limitations of dipeptide as a model in protein folding studies.

Further refinements of potential energy functions have been made to match dipeptide energy map with protein energy map obtained using protein crystal structure data by Kolaskar and Prashant (1979). However, the availability of single crystal structure data at a high resolution, of protein having different functions and three dimensional structure, can be used directly to study physico-chemical properties of individual amino acid residues; in particular their conformational properties in proteins.

Secondary structure propensities of amino acid residues, a conformational property, was derived using protein crystal structure data by several workers, including Chou and Fasman (1974), Tanaka and Scheraga (1976), and Levitt (1978). Recently, Meirovitch et al. (1980) have given the details of various attempts made to derive hydrophobic properties of amino acid residues.

However, the (ϕ, ψ) -distribution data of individual amino acid residues have not been utilized till now to study the conformational property of these residues. Attempts have been made only to represent the (ϕ, ψ) -data (Balasubramanian, 1977) and use the distances between various amino acids to get some idea about folding of globular proteins (Goel et al., 1982). Therefore, the analysis of probability distribution in (ϕ, ψ) -plane of individual amino acid residues obtained from large number of different globular proteins was undertaken to understand amino acid behaviour in taking particular main chain conformation in proteins. The rationale is that, since the proteins considered have either different three dimensional structure or different sequences, the effect on a residue

conformation due to near-neighbour residues or particular environment will be masked in (ϕ, ψ) -probability distribution of a residue to a good extent and, for a sufficiently large data, the observed probability distribution can be considered to be similar to that of a residue in random coil state. In other words the (ϕ, ψ) -probability distributions are representative of main chain conformational property of individual amino acid residues. Therefore, to study this aspect and to get an idea about amino acid tendencies in taking main chain conformations, which can be later used in protein folding studies, we have analysed (ϕ, ψ) -data of amino acid residues from 38 different globular proteins using a simple algorithm. The method developed and results obtained are presented in succeeding sections.

II.2 METHOD

(ϕ, ψ) -data used in this study were obtained from Feldmann (1976). 7567 (ϕ, ψ) -values of the 20 proteinous amino acids were collected. The total number of individual amino acids used in this study is given as a histogram in Fig. 1(a). Comparison of this histogram with those obtained by Doolittle (1981) using a very large protein data base (Fig. 1(b)), indicate similar amino acid distribution as in the data base considered in our study.

The proteins considered, along with the resolution to which crystal structure is solved, are given in Table I. Mostly at these resolutions, the main chain dihedral angle values are accurate only upto $\pm 10^\circ$; hence the grid interval of 20° is chosen and the observed (ϕ, ψ) -values were plotted in (ϕ, ψ) -plane for each residue separately. The

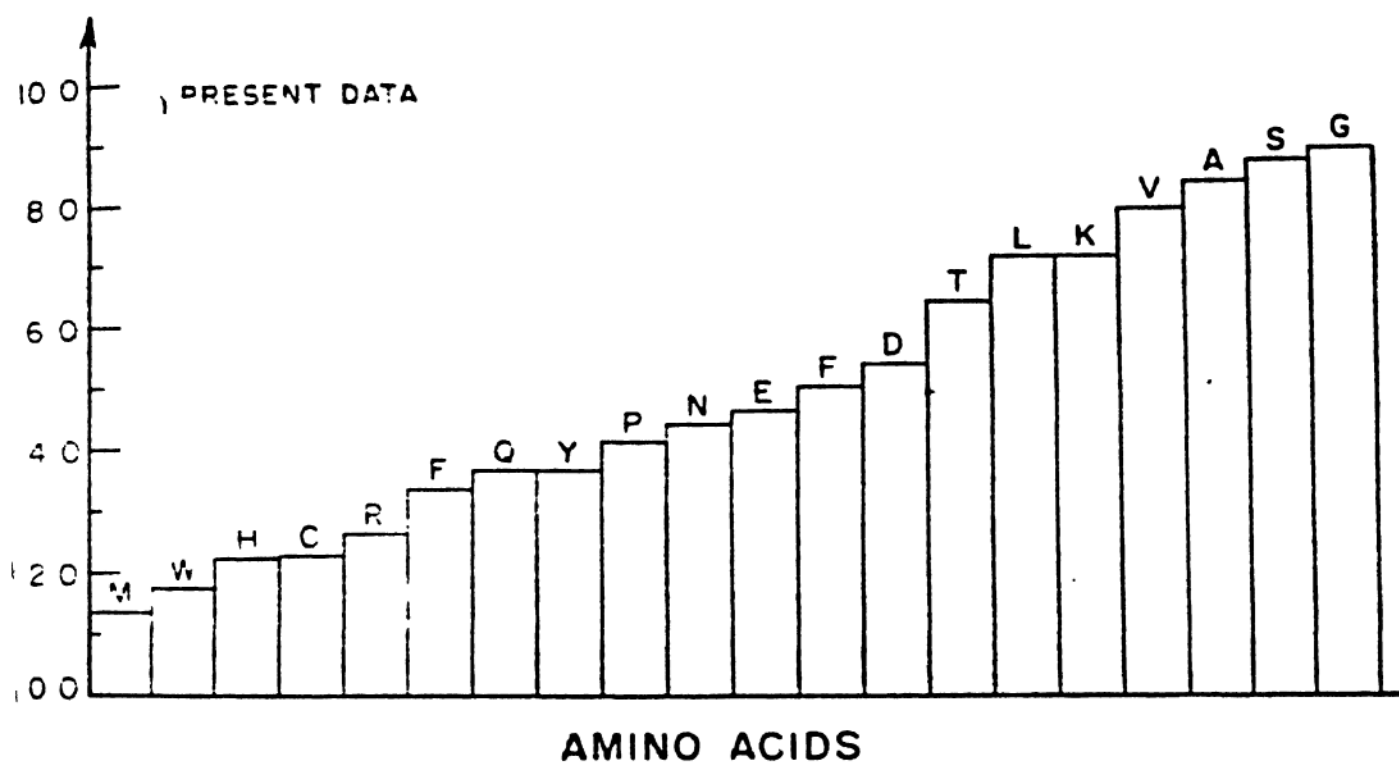


Fig. 1 (a) Distribution of amino acids found in 38 proteins presently considered

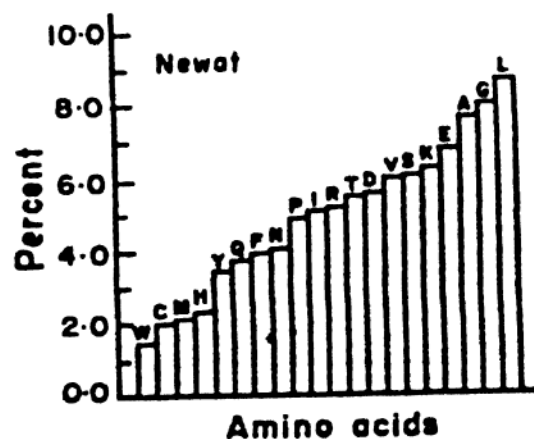
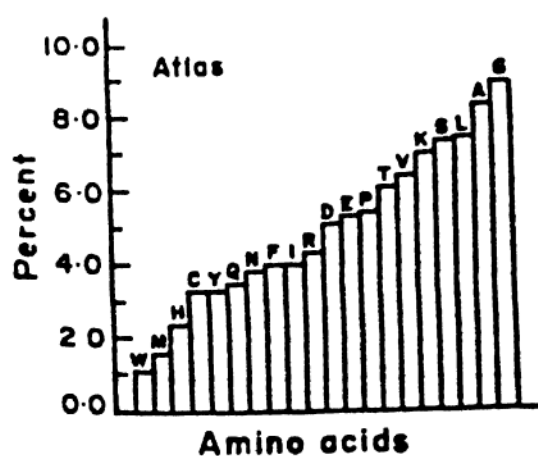


Fig. 1(b) (Left) Distribution of amino acids found in 1081 proteins and peptides listed in Atlas of protein sequences and structure.

(Right) Distribution of amino acids found in 189 peptides and proteins compiled from original literature covering the period 1978 to mid-1980.

TABLE I

List of proteins considered along with the resolution to which
the crystal structure is solved

Name of the protein	Resolution (in Å) to which crystal structure is solved or refined
Lamprey cynamet haemoglobin	2.0
Bovine ferricytochrome b ₅	2.0
Horse deoxy haemoglobin dimer	2.8
Binito ferrocytochrome C	2.3
Tuna ferricytochrome C 'outer'	2.0
Tuna ferricytochrome C 'inner'	2.0
Bacterial Ferricytochrome C ₂	2.0
Bacterial cytochrome C ₅₅₀	2.5
Spermwhale metmyoglobin	1.4
Bacterial rubredoxin	1.54
Bacterial high potential protein	2.0
Bacterial ferredoxin	2.0
Subtilisin BPN'	2.5
Bovine α-chymotrypsin A	2.8
Bovine chymotrypsinogen A	2.5
α-chymotrypsin A	1.9
Bovine trypsin	1.9
Porcine tosyl elastase	2.5
Papain	2.8
Bacterial thermolysin	2.3
Bovine carboxypeptidase A complex	2.0
Bovine trypsin-trypsin inhibitor complex	1.9
Dog fish lactate dehydrogenase complex	2.8
Horse alcohol dehydrogenase complex	2.4
Lobster glyceraldehyde-3-P-dehydrogenase green'	2.9
Bacterial oxidised flavodoxin	1.9
Bovine ribonuclease S complex	2.0

contd.....

Table 1 contd.....

Bacterial nuclease complex	2.0
Human bence-jones protein	2.0
Human immunoglobulin G 'Fab new'	2.0
Jack bean concanavalin A	2.4
Chicken lysozyme	2.0
Chicken triose phosphate isomerase monomer	2.5
Carp calcium-binding protein B	1.85
Human carbonic anhydrase B	2.0
Human carbonic anhydrase C	2.0
Human prealbumin dimer	2.5
Bacterial semiquinone flavodoxin	1.9

number of points in each grid thus obtained was normalised and these normalised maps are given in Fig. 2.

In order to determine similarity among (ϕ, ψ) -probability maps of various residues the (ϕ, ψ) -probability map of each residue was compared grid-wise with (ϕ, ψ) -maps of other residues. The total discrepancy between any of these probability maps is calculated using the simple relation:

$$\Delta P_j = \sum_{k=1}^{324} |P_{\text{ref } k} - P_{jk}|$$

ΔP_j = total discrepancy index between two maps.

$P_{\text{ref } k}$ = normalised probability of reference residue in kth grid. Reference residue changes after all the other 19 residue distributions are compared.

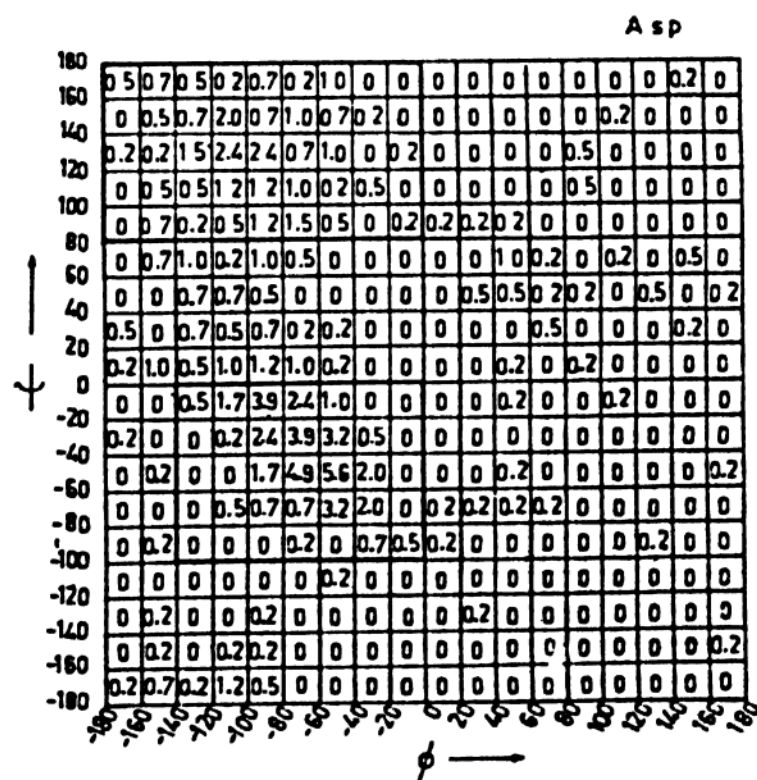
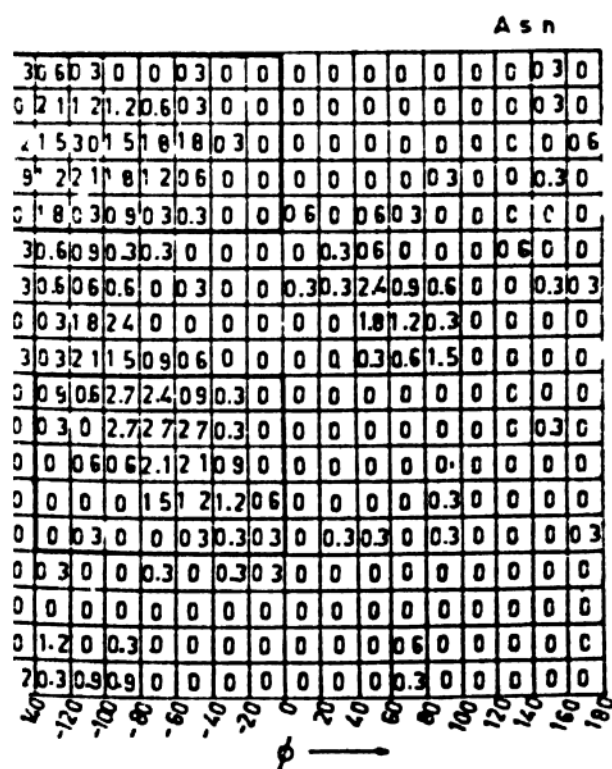
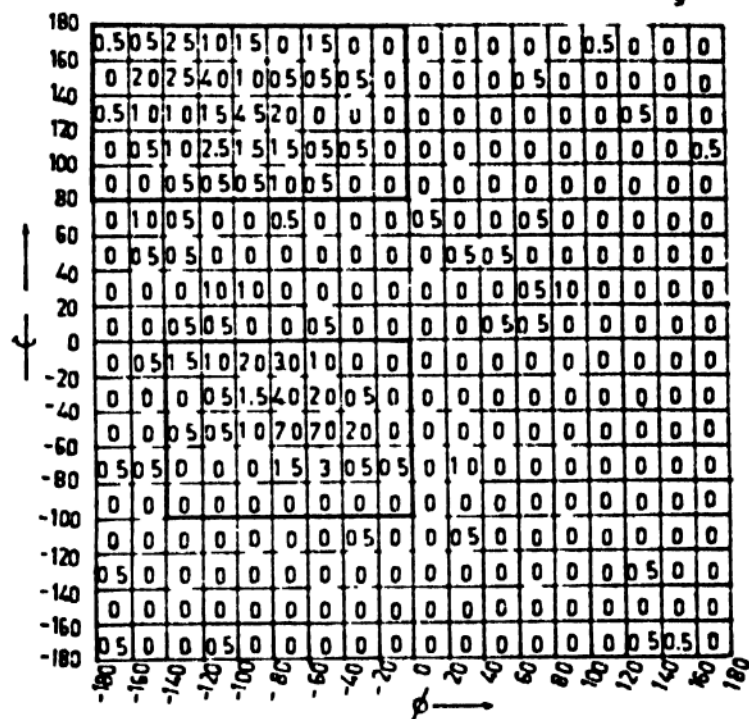
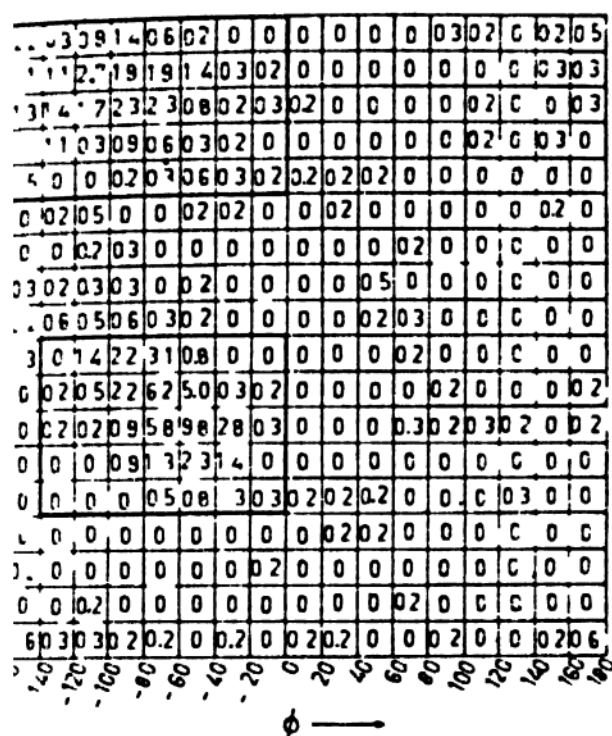
P_{jk} = normalised probability of jth residue in the kth grid.

Total number of grids being 324, k varies from 1 to that number.

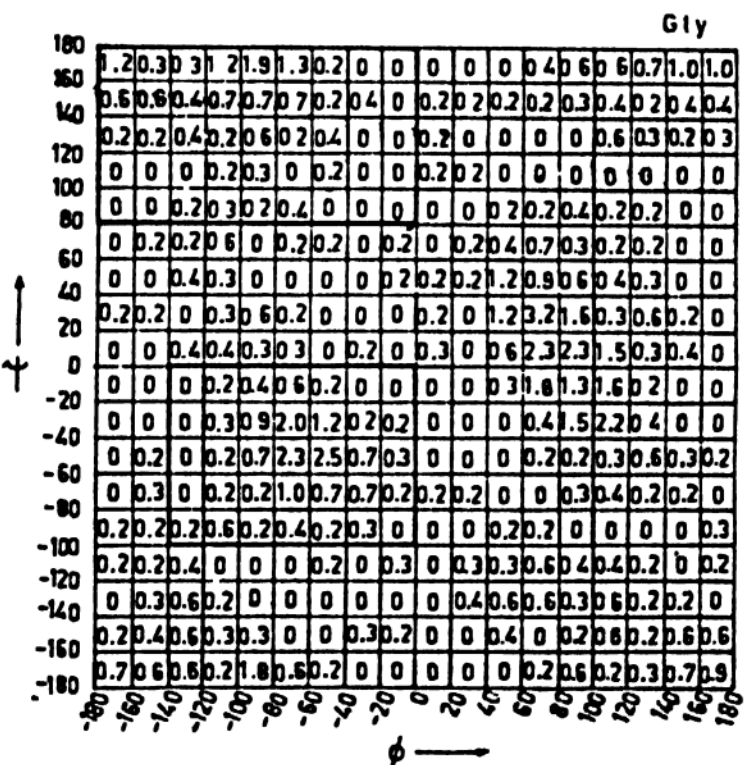
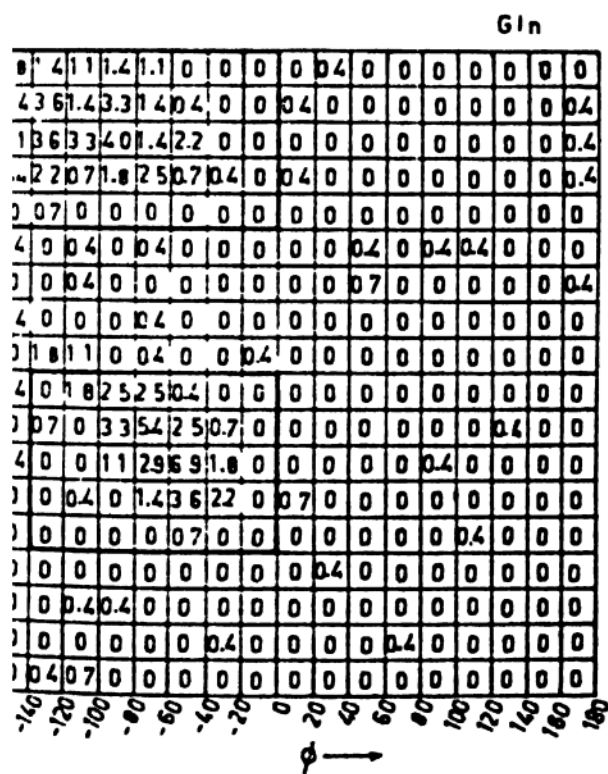
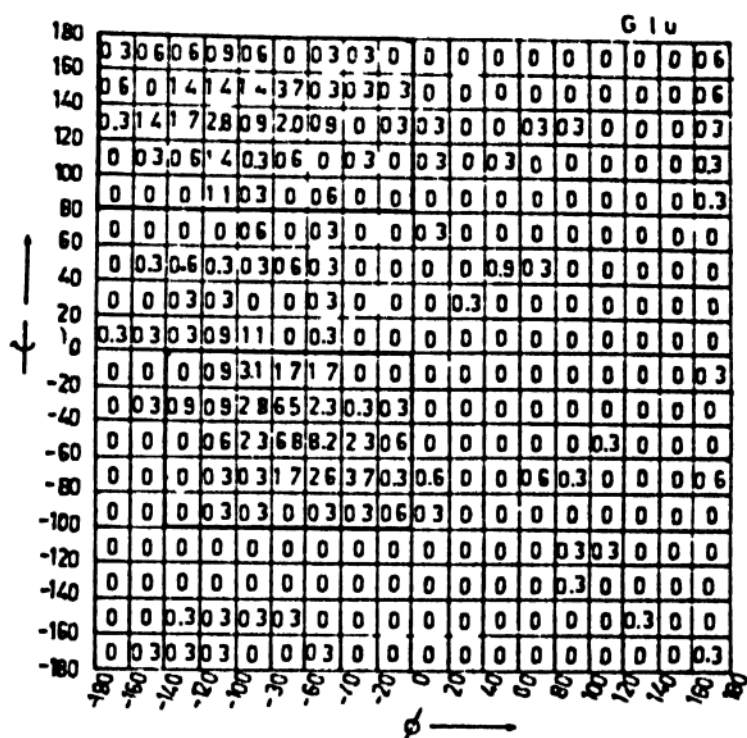
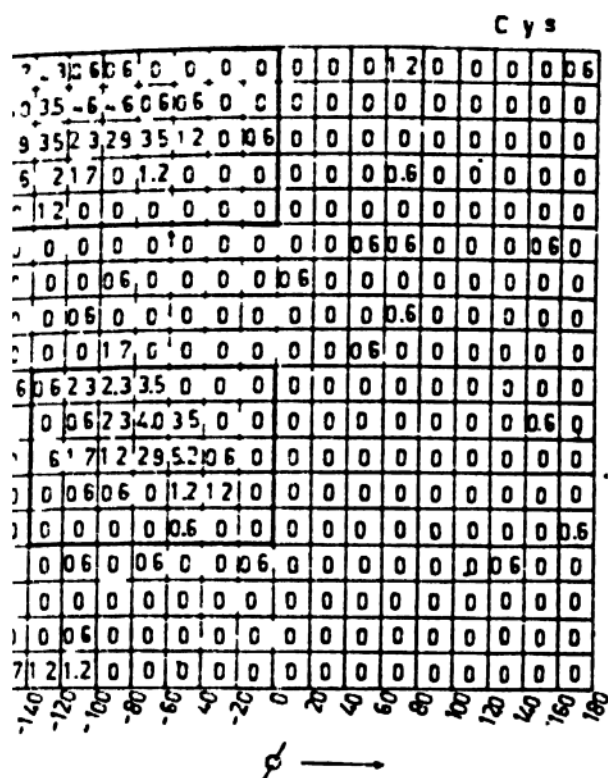
These discrepancy indices are given as percentages in Table II. Standard deviation σ_j associated with each of these discrepancy indices is also computed. These values varied between 2.5 and 5.5 except when compared with Gly and Pro (ϕ, ψ) -distributions which are more than 7 in all cases.

Conformationally similar residues were derived using data in Table II and the associated standard deviations. The ΔP_j minimum along with standard deviation σ_j for each row is used to determine a range for the discrepancy. Any residue whose $\Delta P_j < (\Delta P_{j \text{ min}} + \sigma_{j \text{ min}})$ was considered to be conformationally similar to the residue in the far left column of the row. For

F1 .2. The following figure contains the main chain conformation probability value (approximated to first decimal) as percentages, obtained from crystal structure data of 38 globular proteins. The regions A and B comprise ϕ -variations between -140° and 0° , -180° and 0° respectively and ψ -variations between -100° and 0° and 80° and 180° respectively, are also marked to show that maximum probability distribution of (ϕ, ψ) -map lie in these regions_{A, a}



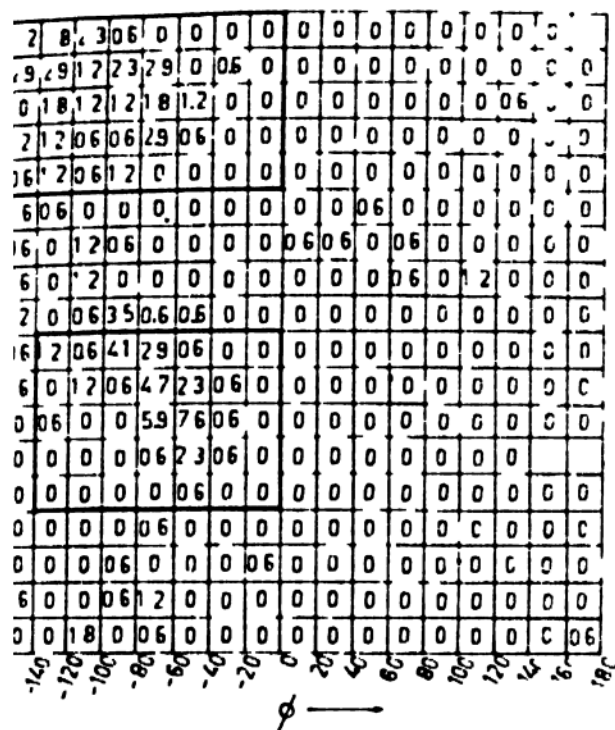
contd.....



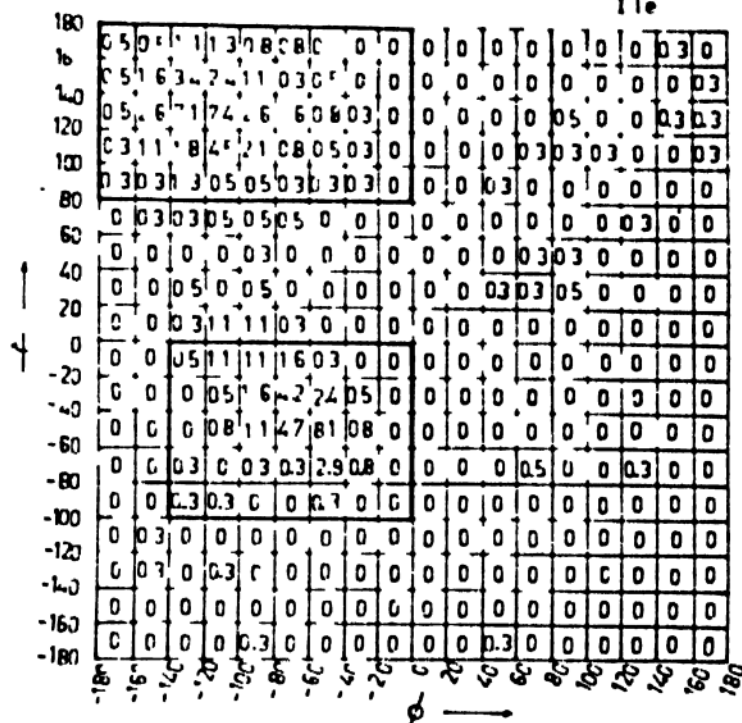
contd.

ig. 2 contd.....

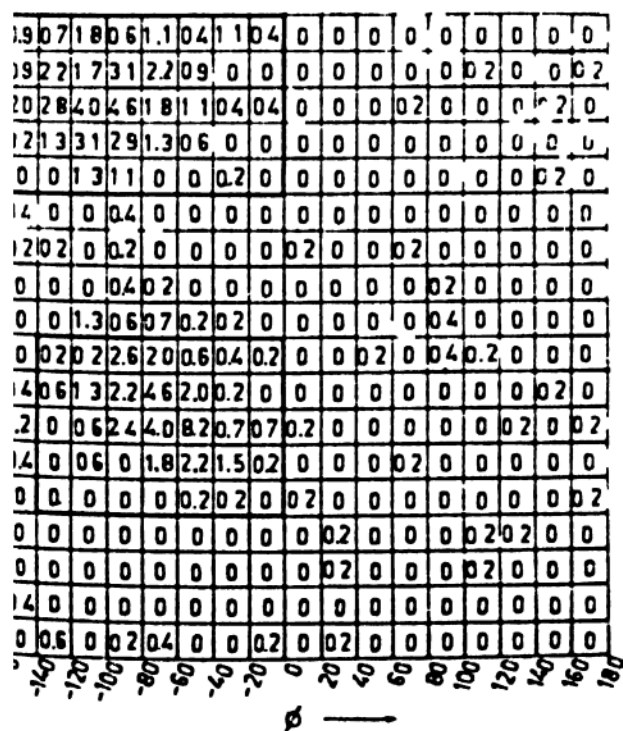
H s



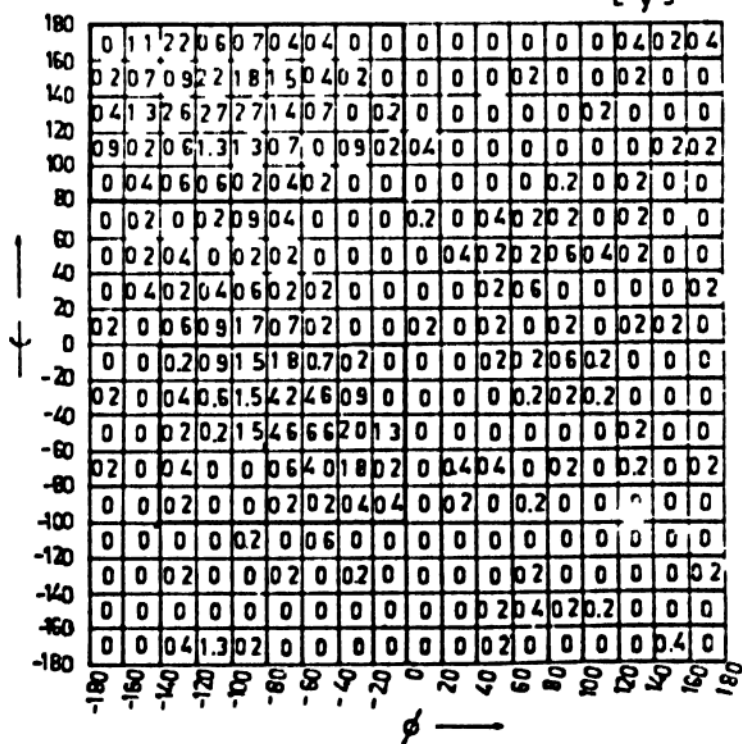
Ile



Leu



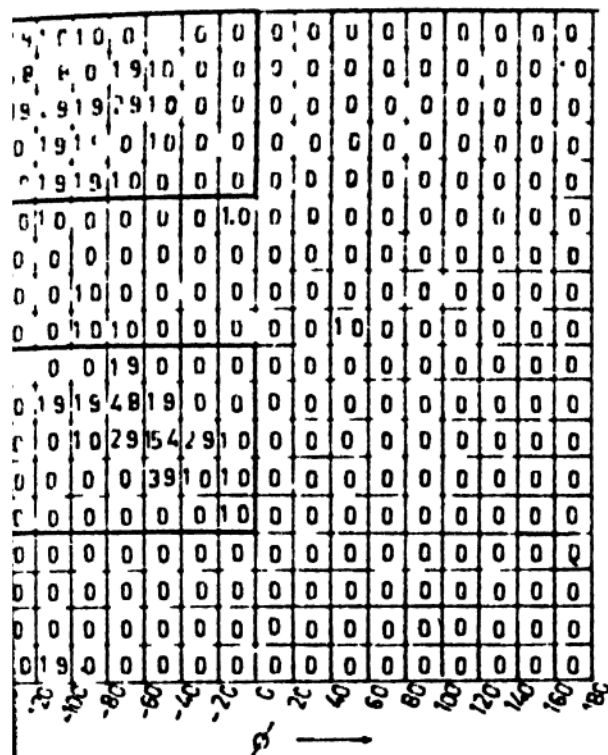
Lys



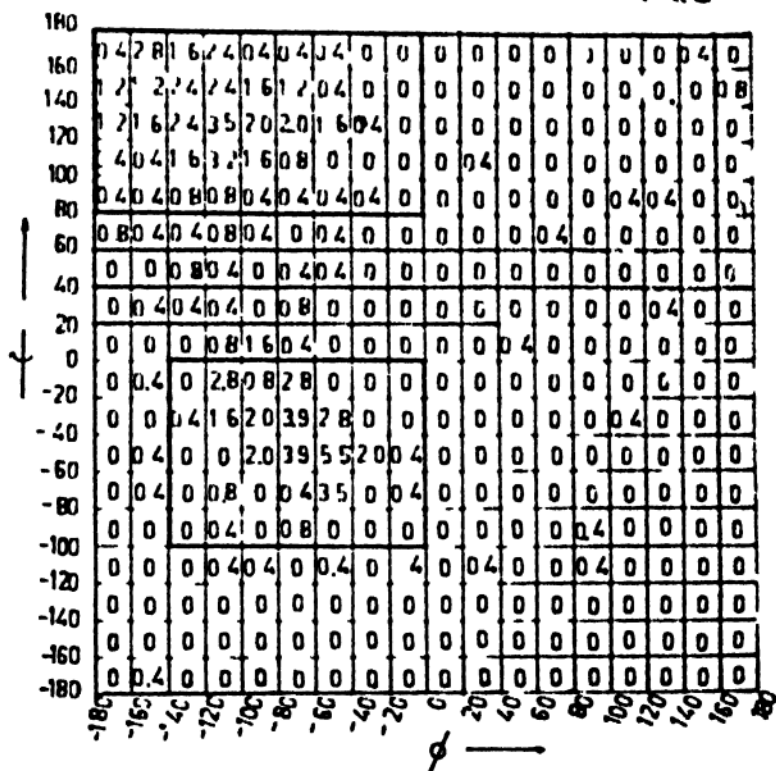
contd.....

2 contd.....

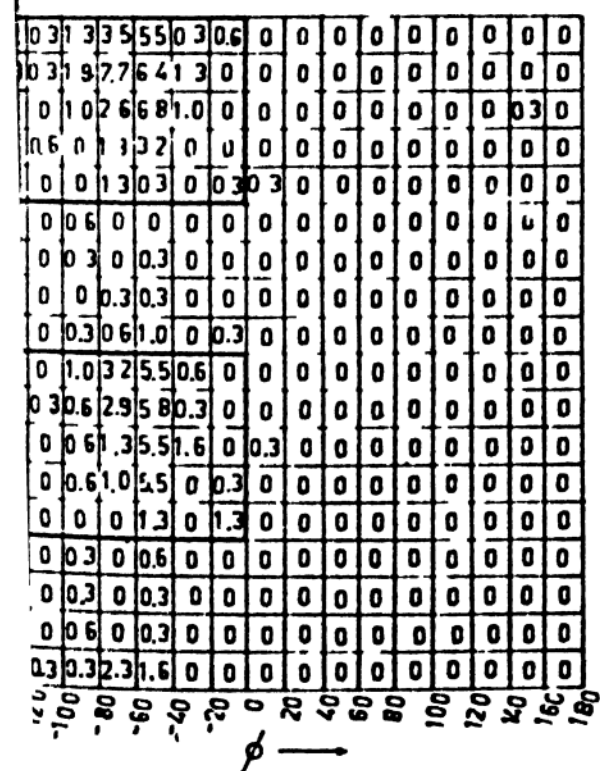
Met



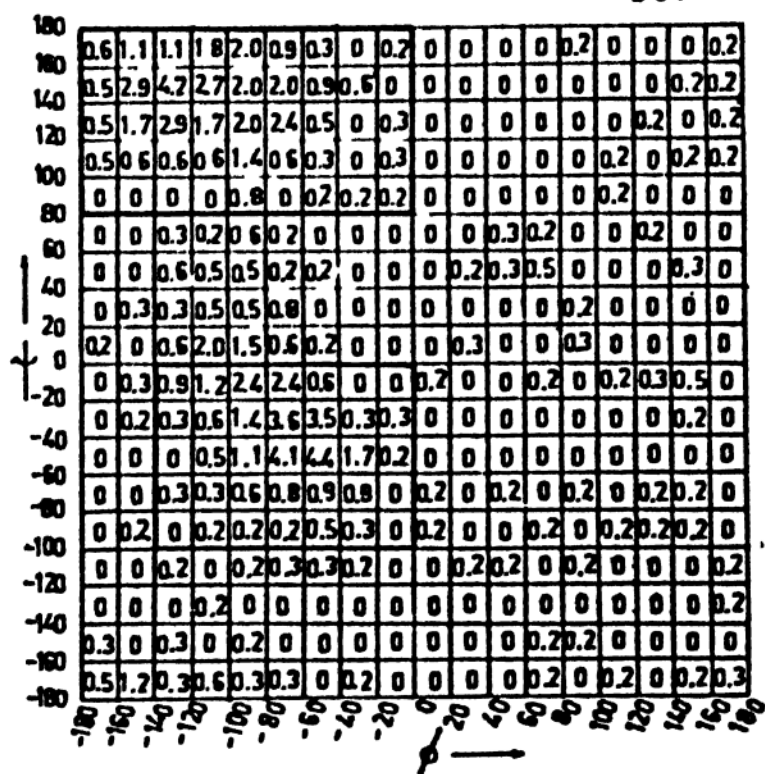
Phe



Pro



Ser



contd.....

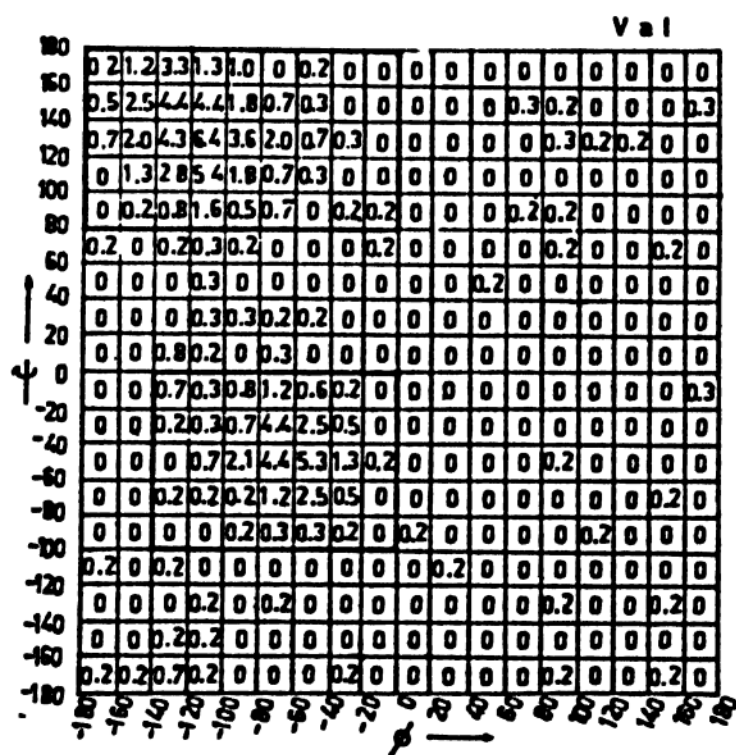
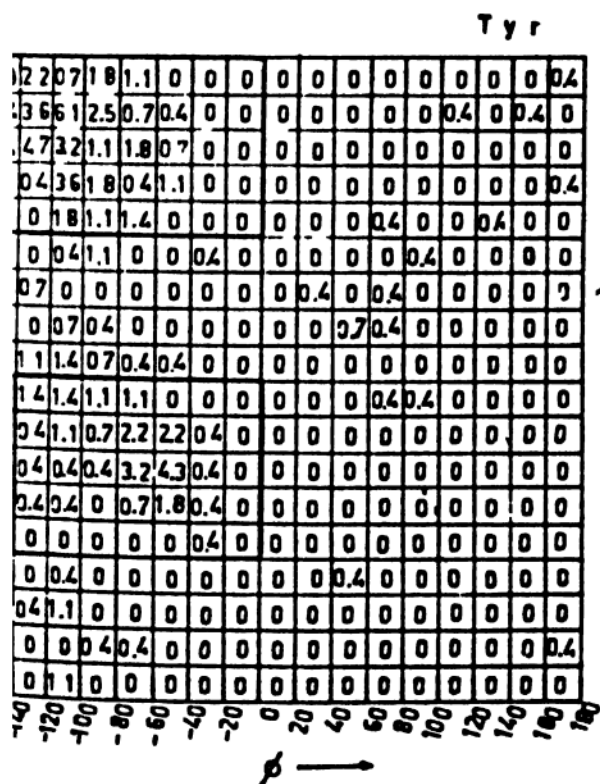
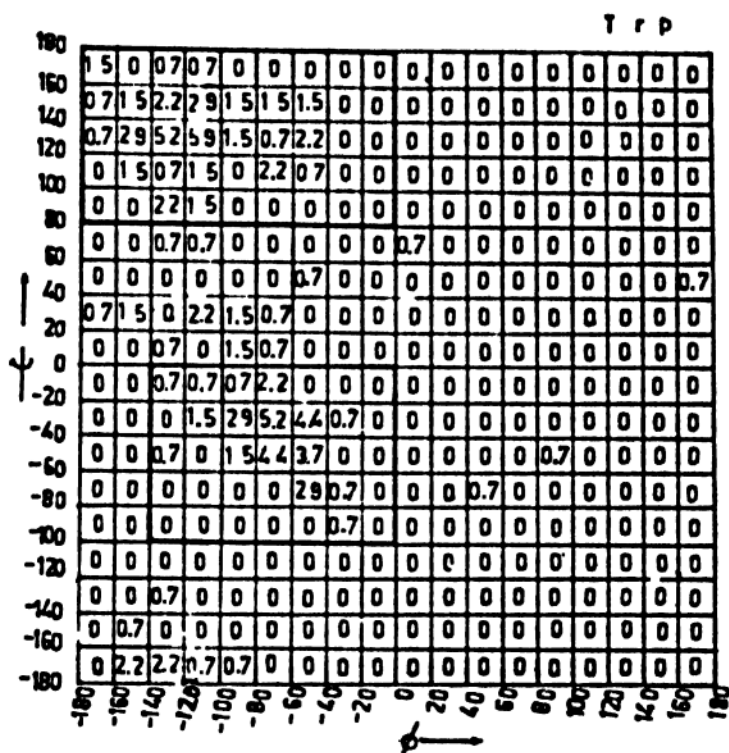


Table II Conformational discrepancy index (%) of amino acid residues with respect to that in the far left column

	Ala	Arg	Asn	Asp	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
<i>N</i>	642	200	336	410	174	352	277	685	171	381	546	547	104	254	313	666	487	136	278	608
Ala	—	37.8	48.2	35.7	41.6	32.2	35.7	60.5	41.2	38.9	35.6	33.1	43.2	38.3	54.8	31.4	38.9	45.6	48.5	40.7
Arg	—	—	45.0	37.1	41.4	41.2	38.2	65.3	38.4	36.5	38.3	36.5	43.2	38.8	61.5	38.9	38.0	47.6	43.3	35.6
Asn	—	—	—	39.7	46.9	45.9	45.4	59.6	47.7	44.3	43.7	44.7	55.7	44.5	65.7	40.1	44.7	46.0	49.3	49.9
Asp	—	—	—	—	44.1	36.6	39.1	62.7	41.4	39.5	39.2	29.7	48.3	38.6	57.5	36.0	38.5	44.8	46.8	44.9
Cys	—	—	—	—	—	48.3	38.7	71.6	40.9	40.1	40.6	41.5	43.2	38.3	66.2	35.3	35.9	40.3	43.5	38.2
Glu	—	—	—	—	—	—	38.3	65.9	42.3	40.7	33.1	36.0	47.6	40.1	59.1	39.7	40.2	48.3	52.7	43.4
Gln	—	—	—	—	—	—	—	67.4	41.1	36.2	31.5	35.1	42.3	35.2	59.2	36.8	33.4	42.1	40.2	35.7
Gly	—	—	—	—	—	—	—	—	70.3	67.9	66.2	61.0	74.0	67.0	76.6	59.1	64.0	71.8	65.7	66.8
His	—	—	—	—	—	—	—	—	—	43.1	39.8	43.5	46.6	43.0	60.9	39.0	39.2	45.5	47.9	45.0
Ile	—	—	—	—	—	—	—	—	—	—	31.2	35.1	41.3	33.2	65.9	36.7	32.6	37.4	37.7	26.0
Leu	—	—	—	—	—	—	—	—	—	—	—	34.3	40.6	33.1	60.5	35.0	32.7	42.6	40.0	33.1
Lys	—	—	—	—	—	—	—	—	—	—	—	—	42.8	34.2	58.1	33.0	35.5	42.6	42.3	38.1
Met	—	—	—	—	—	—	—	—	—	—	—	—	—	40.1	65.0	46.9	42.9	45.1	42.3	38.3
Phe	—	—	—	—	—	—	—	—	—	—	—	—	—	—	61.1	35.6	32.7	40.5	38.5	34.5
Pro	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	58.9	60.7	65.4	66.9	65.7
Ser	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	31.7	43.1	40.8	35.6
Thr	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	39.0	34.6	31.9
Trp	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	45.2	38.9
Tyr	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	33.7
Val	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

N is the total number of points in (ϕ, ψ)-plane for each type of amino acid residue

example, when (ϕ, ψ) -distributions of all 19 residues were compared with that of Ala, $\Delta P_j = 31.4$ was minimum for Ser having $\sigma_j = 3.9$. As can be seen from Table II, only two other residues Glu and Lys have ΔP_j values less than $(\Delta P_{j \text{ ser}} + \sigma_{j \text{ ser}}) = 35.3$. Thus, residues Ser, Glu and Lys are considered conformationally similar to Ala. Such sets of conformationally similar amino acids are given in Table III for each residue.

11.3 RESULTS AND DISCUSSION

A cursory glance at Table II shows that in all cases the least discrepant value of each row is in the range of 25% to 40% the exceptions being Gly and Pro for which the least discrepant value is more than 50%. This indicates that the probability distributions in the (ϕ, ψ) -plane for Gly and Pro are distinct from other residues in proteins. Main chain conformations of Gly and Pro are distinct compared with other residue conformations. This is understandable since Gly and Pro have unique side chains. So, in Table III no residue is shown conformationally similar to either Gly or Pro. For each of the remaining 18 types of residues there is at least one residue whose (ϕ, ψ) -probability map is similar to the one with which it is compared. It should be mentioned here that, since the data points in (ϕ, ψ) -plane considered in present study are less than 200 for amino acids, Arg, Cys, Met, His and Trp, the results obtained for these residues are subject to higher statistical fluctuations and thus should be considered preliminary.

Analysis of data given in Table III reveals that hydrophobic amino acids Ala and Leu are conformationally similar to the acid amino acid Glu.

TABLE III

Conformationally similar residues in whole (ϕ, ψ)-plane

Ser Glu Lys

Val Ala Asp Gln His Ile Leu Lys Phe Ser¹ Thr

Asp Ser

Lys

Ser Phe Thr Val

Ala Leu

Leu Thr

-

Arg Ala Asp Cys Glu Gln Leu

Val

Ile Glu Gln Lys Phe Ser Thr Val

Asp

Val Ala Arg Cys Gln Ile Leu Lys Phe Thr Tyr

Leu Gln Ile Lys Ser Thr Val

-

Ala Cys Leu Lys Thr

Ser Gln Ile Leu Phe Tyr Val

Ile Cys Phe Thr Val

Val Thr

Ile

indicates that the (ϕ, ψ)-distribution in the (ϕ, ψ)-map of the residue in left hand column is distinct from those of the remaining residues. Three letter amino acid code has been used.

The bulkiness of the side chain of these amino acids also varies considerably. Similarly, Ser, a neutral residue having hydroxy side chain was found to be conformationally similar to Ala, a hydrophobic and aliphatic amino acid. Further, Ala prefers α -helix while Ser prefers chain reversals; still one finds Ser has a (ϕ, ψ) -distribution similar to that of Ala. Table III further shows that conformationally similar residues need not have side chains of similar chemical nature and bulkiness. For example, Ile is conformationally similar to Phe. Side chains of these residues are hydrophobic in nature and both prefer β -sheet structure, but Ile is aliphatic while Phe is aromatic.

In order to see whether these amino acids having different chemical and physical properties are conformationally similar in general, we carried out a simple analysis, the results of which are given in Table IV. Examples discussed above and the data in Table IV point out that conformationally similar residues need not be similar in other physico-chemical properties. Thus main-chain conformational similarity obtained from present study is an intrinsic property of amino acids, an observation which was not made in earlier studies.

In Table III one may come across an apparent anomaly, such as Lys is conformationally similar to Ala, and Asp not Ala is conformationally similar to Lys. We would like to point out that these results are internally consistent and this apparent anomaly is due to the change of reference residue during comparison. This is illustrated in Fig. 3. ΔP_j for (Ala, Lys) pair is 33.1% while that for (Lys, Asp) is 29.7%, for (Lys, Asp) pair the value of $\sigma_j = 2.7$ and thus ΔP_j for (Lys, Ala) is

TABLE IV

Details about physical and chemical natures and secondary structural affinities of conformationally similar residues mentioned in Table III

Amino acid	Total number of conformationally similar residues	Number of conformationally similar residues having			
		Similarity in bulkiness	Same chemical nature	Same polarity	Same secondary structural preference
Al	3	1	-	-	2
Ar	11	5	1	2	1
Asp	2	2	-	1	2
Asn	1	-	-	1	-
Cys	4	1	-	1	-
Glu	2	1	-	-	2
Gln	2	-	-	1	1
Gly	-	-	-	-	-
His	7	5	1	1	4
Ile	1	1	1	1	1
Leu	6	2	2	3	3
Lys	1	-	-	1	-
Met	11	4	-	4	4
Phe	7	2	-	3	3
Pro	-	-	-	-	-
Ser	5	2	1	2	-
Thr	7	2	1	2	4
Trp	5	2	1	3	4
Tyr	2	-	-	1	2
Val	1	1	1	1	1

Three letter amino acid code has been used.

outside the range indicating that Ala is not conformationally similar to Lys. However, ΔP_j for (Ala, Lys) is well within the range set for conformationally similar residue to Ala.

The pairs of amino acids which exhibit reciprocity by having similar (ϕ, ψ) -distributions are: Ala-Glu, Ala-Ser, Arg-His, Asp-Lys, Cys-Ser, Glu-Leu, Gln-Leu, Gln-Thr, Ile-Val, Leu-Phe, Leu-Thr, Leu-Ser, Phe-Thr, Ser-Thr and Thr-Tyr. This observation indicates that the side chain effect on main chain conformation for these residues is similar in most of the regions of (ϕ, ψ) -plane.

Conformationally similar residues not only have the discrepancy index lowest, indicating that their overall (ϕ, ψ) -distributions are not quite different from that of the residue with which they are compared but the discrepancy value even at grid level is never more than 2.5 for the particular pair of amino acids. Stray cases in which this does not hold good are: Ala-Ser ($-60^\circ, -80^\circ$); Arg-Val ($100^\circ, -20^\circ$) and Arg-Thr ($-60^\circ, -80^\circ$).

From Table III it can be observed that residues like Ala, Ser, Val, Ile are conformationally similar to more than one residue. This observation is made use of in proposing obligatory amino acids in primitive proteins. This analysis is presented in Appendix II.'

During the above analysis, it was also noted that certain amino acid residues which prefer different secondary structures can show similarity in (ϕ, ψ) -distributions. To get more insight into this aspect,

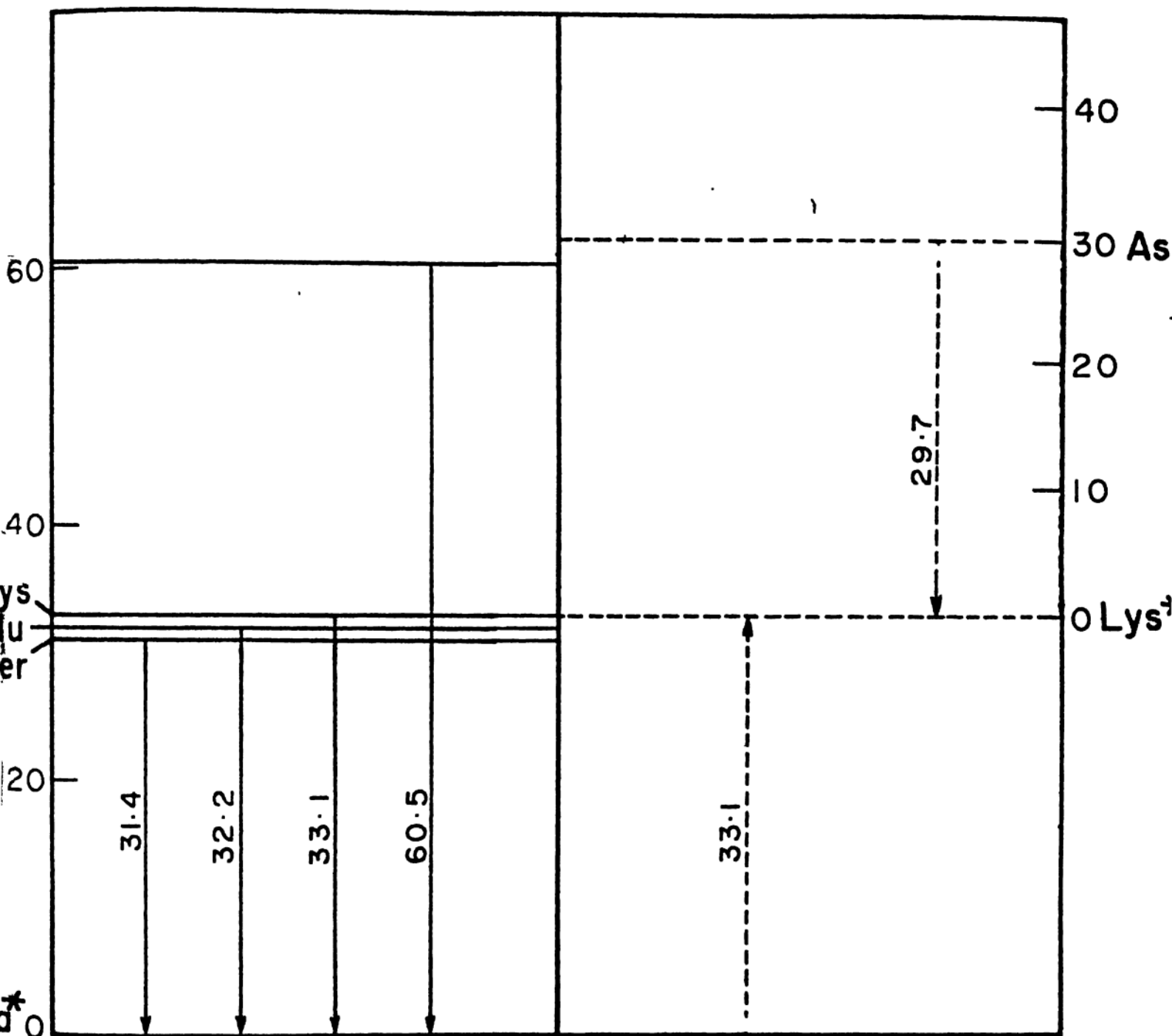


Fig. 3 * indicates reference residue, with whose (ϕ, ψ) -probability distribution the remaining residue (ϕ, ψ) -distributions are compared. Arrows are directed towards reference residue.

two parts of the (ϕ, ψ) -plane, termed region A and region B were considered. These two regions are densely populated and more than 70% of observed conformations of residues, other than Gly and Asn, lie in these regions.

Region A encompasses ϕ and ψ variations in the range -140° to 0° respectively. This region represents the main chain conformations taken by residues in closely packed structures such as α -helix, $^3_{10}$ -helix and chain reversals etc.

Region B is defined as one in which ϕ and ψ vary from -180° to 0° and 80° to 180° respectively. Main chain conformations that fall within this range, are extended in nature and occur in secondary structures such as β -sheet, collagen-type helix and chain reversals. The regions A and B are marked in Fig. 2.

Comparison of (ϕ, ψ) -probability distribution in these regions is carried out after proper normalisation using the procedure discussed above. A set of residues having (ϕ, ψ) -probability distribution similar in region A are given in Table V.A, and those in Region B are given in Table V B for each residue. Examination of the Tables III, V A and V B reveals that a residue conformationally similar to a residue in the whole (ϕ, ψ) -plane need not be similar in region A or region B. Thus, Ala is conformationally similar in whole (ϕ, ψ) -plane to Glu and not similar in region A.

It is interesting to note that using entirely different approach namely that of differential geometry and analysing the parameters

TABLE V A

Conformationally similar residues in region A

Ala	Ser Lys
Arg	His Ala Asp Glu Ile Ser Tyr Val
Asn	Ser Asp Gln
Asp	Ser Lys
Cys	Ser
Glu	Gly Leu
Gln	Asp Ala Glu Leu Lys Thr
Gly	Glu
His	Arg Ala Ile
Ile	Val Ala Arg His Lys
Leu	Glu
Lys	Ala Asp Cys Ile Ser Val
Met	Ile Ala Arg Gln Leu Lys Phe
Phe	Lys Ala Asp Gln Gly Ser Tyr
Pro	Lys Ala Arg Asn Asp Gln Phe Ser Thr Val
Ser	Ala Asp
Thr	Gln Asp
Trp	Ile Ala Arg Asp Gln Gly His Lys Phe Ser Thr Tyr Val
Tyr	Arg His Ile Val
Val	Ile Ala Arg Glu Gly Lys

Three letter amino acid code has been used.

TABLE V B

Conformationally similar residues in region B

Ala	Ser Lys Thr
Arg	Val Asp Lys Thr
Asn	Phe Ile Leu
Asp	Lys
Cys	Ser Lys Thr Val
Glu	Leu Asn Ala Lys Phe Ser Thr
Gln	Leu Lys Phe Thr
Gly	Ser Ala His Phe Thr Tyr
His	Ser Gln Phe
Ile	Val Thr
Leu	Lys Asn Gln Phe Thr
Lys	Leu Asp Phe Ser Val
Met	Tyr
Phe	Thr Asn Ile Leu Lys Tyr Val
Pro	-
Ser	Thr Ala Cys
Thr	Val Ile Phe Ser
Trp	Ile Asn Thr
Tyr	Met Phe Thr Val
Val	Ile Thr

- Indicates that the (ϕ, ψ) -distribution in this region of (ϕ, ψ) -map of the residue in left column is very much distinct from those of the remaining 19 residues. Three letter amino acid code has been used.

κ (curvature) and τ (torsion angle around virtual bond), Rackvosky and Scheraga (1982) have arrived at essentially similar results as obtained in this study. This further confirms that conformational property, namely main-chain conformational similarity, is an intrinsic property of amino acid residues and simple analysis carried out here is sufficient to extract the same.

II.4 CONCLUSIONS

Hitherto protein crystal structure data have been used to obtain similarities in physical properties (Jones, 1975; Meirovitch et al., 1980) and secondary structure affinities (Tanaka and Scheraga, 1976; Levitt, 1978). Analysis presented in this chapter is an addition to this list, and shows the existence of similarity between (ϕ, ψ) -distributions of amino acid residues - termed by us as main chain conformational similarity. Results and discussion presented in this chapter indicate that main chain conformational similarity among amino acid residues is an intrinsic property.

Replacement by conformationally similar residues is expected to alter the overall three dimensional structure of a polypeptide chain minimally, if the side chain can be accommodated. Thus, this property should be used in drug design in addition to other biochemical and chemical properties of side chains of amino acids. Conformational similarity among amino acid residues can be an aid in simplifying the algorithms used in protein folding studies.

C H A P T E R I I I

SIDE CHAIN CHARACTERISTIC MAIN CHAIN CONFORMATIONS OF
AMINO ACID RESIDUES

III,1 INTRODUCTION

In the last chapter (Chapter II), we have discussed the results obtained after comparing normalised (ϕ, ψ) -probability distribution of twenty proteinous amino acids. The (ϕ, ψ) -probability distributions have shown certain discrepancies when compared among themselves. This indicates that the main chain conformations of amino acids residues are affected by side chains. It is also well documented that side chain-side chain interactions are quite important in the formation of native three dimensional structure of polypeptides and proteins. This information is derived from solution, spectroscopic and crystallographic studies of oligopeptides and polypeptides (Anfinsen and Scheraga, 1975). Analysis of crystal structure data of globular proteins have also been carried out by several workers to study side chain-side chain interactions (Krigbaum and Rubin, 1971; Janin et al., 1978; Warme and Margan, 1978; Krigbaum and Komoriya, 1979). However, crystal structure data analysis did not point out main chain conformations characteristic of side chains of amino acid residues, except for giving preferences of amino acid residues towards secondary structures (Chou and Fasaman, 1974; Levitt, 1978). Theoretical studies carried out at the dipeptide level using semi-empirical potential energy functions (Ponnuswamy and Sasisekharan, 1971; Burgess et al., 1974) as well as quantum chemical methods (Pullman and Pullman, 1974), have indicated that the dipeptide energy map changes its shape, with change of side chain group at C^α atom. But, the changes are minor, except in case of Gly and Pro. Therefore, as a logical next step during our study of development of algorithm for

prediction of three dimensional structure of protein was to see the effect of individual side chains on main chain conformations of amino acid residues.

This can be done in two ways. One approach is calculation of potential energy maps for a dipeptide unit having different side chains attached to the C^α atom - as has been done by several groups including those of Desantis et al. (1965); Ramachandran and Sasisekharan (1968); Scheraga (1968) and Flory (1969). From these potential energy values one deducts the contribution to the potential energy by main chain atoms. Thus one can arrive at a set of conformations which are more or less solely stabilised due to the presence of a particular side chain. This approach has its own limitations since the potential functions used by different groups are different and the accuracy of them is limited (Ramachandran, 1973).

The other approach, which we have followed, is purely empirical, using the available (ϕ, ψ) -data from crystal structures of different globular proteins. As mentioned in the previous chapter, each (ϕ, ψ) -probability distribution map represents to a good extent, the probability of conformations taken by a residue in random coil state and thus is a conformational property of the amino acid residue. The environment of any given residue being different in different proteins, the inter-residue interactions are masked in this (ϕ, ψ) -distribution. Therefore, the individual (ϕ, ψ) -probability maps can be used to study the intra-residue interactions. The intra-residue interactions can be divided into three parts: (i) main chain inter-atomic interactions, (ii) side chain inter-atomic interactions, and (iii) main chain-side chain inter-

atomic interactions. Out of these three interaction terms, the main chain inter-atomic interactions are constant for all proteinous amino acids. Therefore, if the other two terms are grouped together, one can find out the effect due to a particular side chain on main chain conformations. In the present analysis a simple algorithm has been developed using (ϕ, ψ) -probability maps of individual residues to derive those conformations which are maximally affected by a specific side chain. The method developed and the results along with discussion are presented in succeeding sections.

III.2 METHOD

A general L-residue probability map is constructed using data of individual (ϕ, ψ) -probability maps given in Fig. 2 of Chapter II, as a first step to derive main chain conformations significantly affected by respective side chains.

III.2.1 Construction of 'general L-residue' map

Probability values P_j^i for the j th grid in (ϕ, ψ) -map of i th residue were used to determine the j th grid probability values P_j^{GR} of general L-residue map using the relation:

$$P_j^{GR} = \sum_{i=1}^{20} P_j^i / 20 \quad \dots\dots (1)$$

The index j varies over all 324 grids of a (ϕ, ψ) -map, the starting grid being $(-180^\circ, -180^\circ)$. The variable ϕ varies over the whole range, at intervals of 20° and then ψ changes by 20° ; thus $j = 18$ corresponds to grid $(-160^\circ, -180^\circ)$. The general L-residue (ϕ, ψ) -map

"GENERAL-L-RESIDUE" PROBABILITY MAP

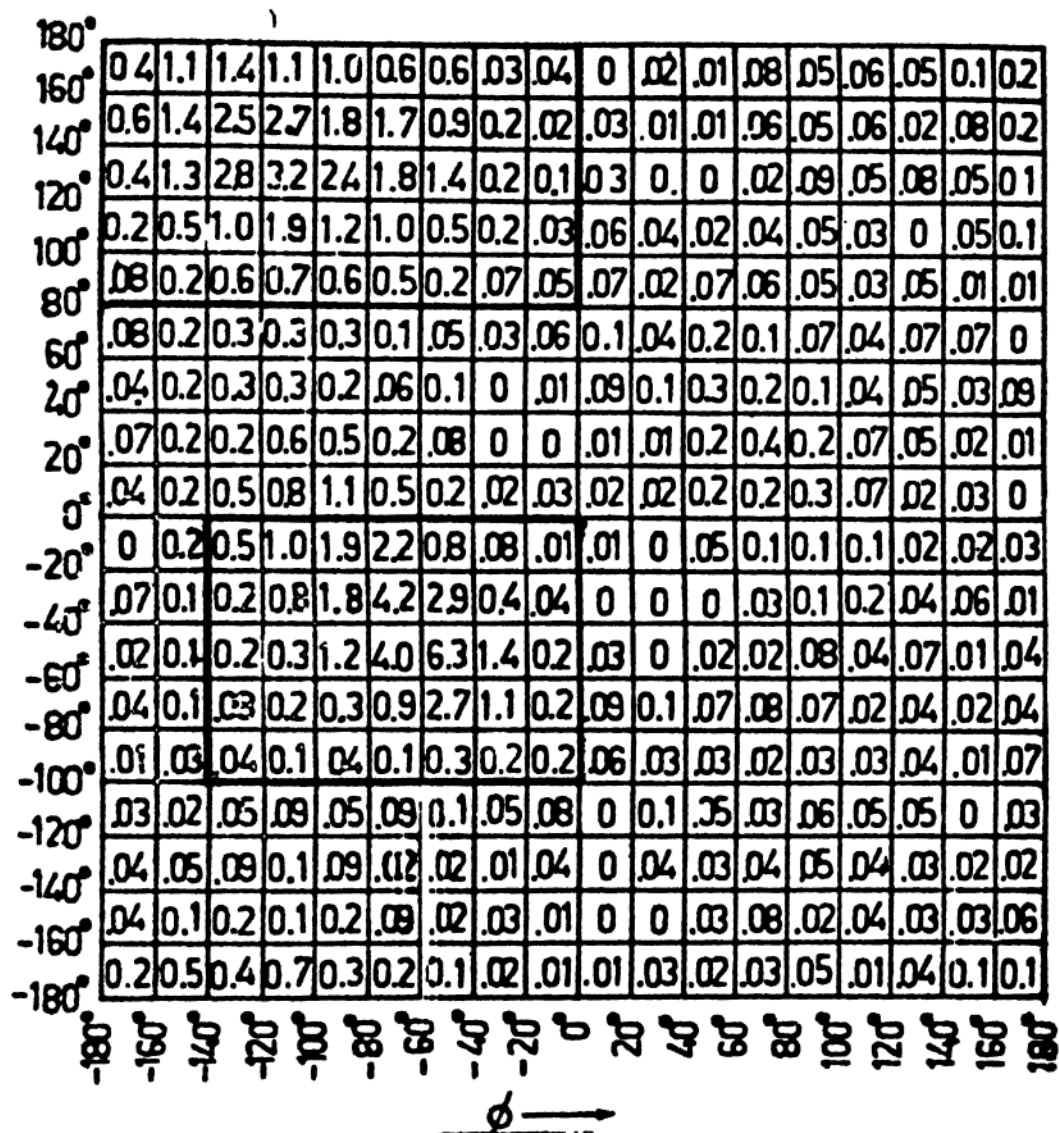


Fig. 1.

Normalised (ϕ, ψ) -probability map obtained from crystal structure data of 38 globular proteins at a grid interval of 20° . Note that the individual normalised (ϕ, ψ) -probability maps of each of the 20 residues were used. The regions A and B comprise ϕ -variations between -140° and 0° , -180° and 0° respectively, and ψ -variations between -100° and 0° and 80° and 180° respectively. These regions are marked in the figure.

This 'general L-residue' map is made up of main chain atoms of usual amino acid residues and a fictitious side chain. The side chain of this residue is the resultant effect of all 20 types of side chains, getting 0.05 weightage from each individual side chain. In other words, this map represents main chain atom interactions with unit weightage and the inter-atomic interactions between individual side chains or interactions among side chain-main chain atoms of any particular residue with low weightage.

III.2.2 Comparison of general L-residue map and individual residue (ϕ, ψ) -probability maps

The normalised (ϕ, ψ) -probability maps of individual residue, excepting that of Gly, were compared grid-wise with the general L-residue map. The comparison with Gly (ϕ, ψ) -map was done after suitably altering the 'general L-residue' map, since Gly (ϕ, ψ) -probability map possesses an inversion symmetry.

The 'difference probability' at grid level is represented by

$$\Delta P_j^i = P_j^i - P_j^{GR} \quad \dots (2)$$

These 'difference probability' maps were computed for each residue. Standard deviation σ_i is also calculated for the difference probability map using the following relation:

$$\overline{\Delta P^i} = \sum_{j=1}^{324} P_j^i / 324 \quad \dots\dots\dots 3(a)$$

$$\sigma_i = \left[\sum_{j=1}^{324} (\Delta P_j^i - \overline{\Delta P^i})^2 / 323 \right]^{1/2} \quad \dots\dots\dots 3(b)$$

Each ΔP_j^i was compared with $2\sigma_i$ and if $|\Delta P_j^i| > 2\sigma_i$ then only ΔP_j^i value is considered to be significantly different from P_j^{GR} value and the corresponding (ϕ_j, ψ_j) -grid was assumed to have a significant contribution from the side chain of the residue and these conformations are termed as 'side chain characteristic conformations'.

III.3 RESULTS AND DISCUSSION

The general L-residue probability map represents not only the probability of conformations taken by main chain atoms, but includes main chain conformations which have similar effect from all twenty side chains. In particular since C^β -atom exists in L-configuration in all residues, except in Gly, its effect on main chain conformations is clearly seen in this distribution. This is the reason why Fig. 1 is termed as the 'general L-residue' (ϕ, ψ) -map and is differentiated from the (ϕ, ψ) -map of either Gly or of any residue having symmetric α -carbon atom. Thus it should be noted that the conformations listed in Table I, as side chain characteristic main chain conformations, are only those which have different influence on main chain atoms from side chain compared to fictitious side chain mentioned above.

ΔP_j^i values are indicators of contribution to intra-residue interactions due to the side chain atoms of a particular residue as these values are obtained by taking out the effect of other influences on main chain conformations in the form of general L-residue probability P_j^{GR} conformations having $|\Delta P_j^i| > 2\sigma_i$ alone can be considered with 95 per cent confidence as the conformations getting significantly influenced

by side chain atoms. Such conformations for each of the 20 amino acid residues together with corresponding ΔP_j^i values are listed in Table I. This table shows that only a small portion of the allowed conformational space is significantly affected by side chain atoms and this effect is different for different side chains which can be intuitively perceived.

It can be seen from Table I that ΔP_j^i values are not always positive. $\Delta P_j^i > 0$ indicates that corresponding main chain conformation is stabilised by interactions due to respective side chain atoms. Similarly, the respective side chain atoms destabilise the main chain conformations if $\Delta P_j^i < 0$. Thus data given in Table I can be categorised into four types namely:

(a) Side chain and main chain stabilised conformations:

$$P_j^i > 0 \quad P_j^{GR} > 0 \quad \Delta P_j^i > 0$$

(b) Side chain destabilised - main chain stabilised conformations:

$$P_j^i > 0 \quad P_j^{GR} \approx 0 \quad \Delta P_j^i > 0$$

(c) Side chain destabilised - main chain stabilised conformations:

$$P_j^i > 0 \quad P_j^{GR} > 0 \quad \Delta P_j^i < 0$$

(d) Side chain destabilised - main chain indifferent conformations:

$$P_j^i \approx 0 \quad P_j^{GR} > 0 \quad \Delta P_j^i < 0$$

P_j^i and P_j^{GR} values are considered positive when their value is greater than $2 \sigma_i$ and as zero when their values are less than $2 \sigma_i$. Main chain

conformations of each residue falling under each of these categories are listed respectively in the four rows (a) to (d) of Table I. The conformations which fall under these four categories are discussed briefly.

Case (a): Side and main stabilised

$$P_j^i > 2 \sigma_i; \quad P_j^{GR} > 2 \sigma_i; \quad \Delta P_j^i > 0$$

Type (a) conformations, given in rows (a) are stabilised by interactions both from main chain atoms ($P_j^{GR} > 0$) and side chain atoms ($\Delta P_j^i > 0$).

Table I shows that residues Asn and Gly do not possess type (a) conformation. This means that the main chain and side chain atom interactions are never in phase for these two residues, when both interactions are contributing significantly.

For residues having type (a) conformation, it is expected that the global energy minimum conformation for their dipeptide map will correspond to one of the type (a) conformations listed in Table I. For example, global energy minimum for Ala and Ile dipeptide maps correspond respectively, to type (a) conformations $(-60^\circ, -60^\circ)$ and $(-120^\circ, 120^\circ)$.

Further, type (a) conformations being energetically quite stable, can act as nucleation sites during the protein folding process. It may be mentioned that nucleation process should start when the residues have taken conformations stabilized due to intra-residue interactions and thus type (a) conformations will probably be the best choice.

TABLE I

"Side chain characteristic" conformations from protein crystal structure data. These conformations are divided into four categories (for details see text) and the corresponding ΔP_j^i values are given beneath. $2\sigma_i$ (σ_i = standard deviation) is also given together with residue. All grid values are in degrees.

Ala $2\sigma_i = 0.76$

(a) Side and main stabilized

(-80, -60)	(-60, -60)	(-40, -60)	(-80, -40)
1.78	3.49	1.43	2.00

(-60, -40)	(-80, -20)	(-160, 160)
2.07	0.89	1.05

(b) Side stabilised - main indifferent -

(c) Side destabilised - main stabilised

(-140, 120)	(-120, 120)	(-140, 140)
-1.38	-1.44	-1.38

(d) Side destabilised - main indifferent

(-120, 100)	(-140, 160)
-1.6	-1.09

Arg $2\sigma_i = 0.74$

(a) Side and main stabilised

(-80, -60)	(-80, -20)	(-100, 120)	(-120, 140)
3.01	0.77	2.11	1.29

(-140, 160)	(-60, 160)
1.10	0.95

(b) Side stabilised - main indifferent

(20, -80)	(-140, -20)	(-160, 60)
0.90	1.04	0.76

(c) Side destabilised - main stabilised

(-60, -40)	(-140, 120)	(-120, 120)	(-100, 140)
-0.92	-1.78	-1.65	-0.81

(d) Side destabilised - main indifferent

Table I contd...

$$\text{Asn } 2\sigma_i = 0.92$$

- (a) Side and main stabilised -
- (b) Side stabilised - main indifferent
 (-140, -160) (-120, 0) (80, 0) (-120, 20)
 1.04 1.29 1.24 1.24
 (-100, 20) (40, 20) (40, 40) (-140, 80)
 1.88 1.56 2.04 1.17
- (c) Side destabilised - main stabilised
 (-60, -80) (-80, -60) (-60, -60) (-80, -40)
 -1.48 -1.90 -4.24 -1.55
 (-140, 120) (-120, 140)
 -1.29 -1.52
- (d) Side destabilised - main indifferent
 (-160, 140) (-80, 140) (-100, 160)
 -1.39 -1.11 -1.02
-

$$\text{Asp } 2\sigma_i = 0.64$$

- (a) Side and main stabilised
 (-40, -80) (-80, -60) (-120, -20) (-100, -20)
 0.89 0.89 0.70 1.96
- (b) Side stabilised - main indifferent
 (-160, 0) (-100, 60) (-100, 80) (-80, 80)
 0.81 0.69 0.66 1.01
- (c) Side destabilised - main stabilised
 (-60, -60) (-120, 100) (-140, 120) (-120, 120)
 -0.72 -0.69 -1.31 -0.71
 (-80, 120) (-140, 140) (-120, 140) (-100, 140)
 -1.07 -1.74 -0.75 -1.08
 ((-80, 140)
 -0.73
- (d) Side destabilised - main indifferent
 (-160, 120) (-160, 140) (-140, 160) (-120, 160)
 -1.04 -0.91 -0.91 -0.89
-

$$\text{Cys } 2\sigma_i = 0.84$$

- (a) Side and main stabilised
 (-120, -20) (-80, -20) (-160, 120) (-80, 120)
 1.29 1.20 1.59 1.65
 ((-160, 140) (-140, 140) (-120, 140) (-100, 140)
 2.63 0.98 1.89 2.79
 (-140, 160)
 2.30
- (b) Side stabilised - main indifferent
 (-160, -180) (-120, -60)
 1.7 1.38

(c) Side destabilised-main stabilised

(-60, -80)	(-80, -60)	(-60, -60)	(-120, 120)
-1.52	-1.11	-1.15	-0.85

(d) Side destabilised - main indifferent

(-80, -80)	(-100, 100)	(-80, 140)
-0.86	-1.23	-1.13

$$\text{Glu } 2\sigma_i = 0.82$$

(a) Side and main stabilised

(-80, -80)	(-40, -80)	(-100, -60)	(-80, -60)
0.85	2.63	1.10	2.83
(-60, -60)	(-40, -60)	(-100, -40)	(-80, -40)
1.91	0.9	1.03	2.31
(-100, -20)	(-60, -20)	(-80, 140)	
1.18	0.95	1.99	

(b) Side stabilised - main indifferent -

(c) Side destabilised - main stabilised

(-140, 120)	(-100, 120)	(-140, 140)	(-120, 140)
-1.07	-1.54	-1.05	-1.29

(d) Side destabilised - main indifferent

(-100, 100)	(-160, 140)	(-140, 160)
-0.95	-1.39	-0.83

$$\text{Gln } 2\sigma_i = 0.68$$

(a) Side and main stabilised

(-60, -80)	(-40, -80)	(-100, -40)	(-80, -40)
0.94	1.10	1.44	1.99
(-120, -20)	(-140, 100)	(-140, 120)	(-100, 120)
0.79	1.16	0.83	1.58
(-60, 120)	(-140, 140)	(-100, 140)	
0.79	1.14	1.44	

(b) Side stabilised - main indifferent

(-140, 0)
1.34

(c) Side destabilised - main stabilised

(-80, -60)	(-120, 100)	(-120, 140)
-1.10	-1.19	-1.26

(d) Side destabilised - main indifferent

(-120, -40)	(-100, 0)	(-160, 140)
-0.77	-1.06	-1.03

Table I contd.....

$$\text{Gly } 2\sigma_i = 1.3$$

- (a) Side and main stabilised —
- (b) Side stabilised - main indifferent
 (-100, -180) (100, -40) (60, -20) (100, -20)
 1.64 1.60 1.47 1.58
 (60, 0) (80, 0) (100, 0) (60, 20)
 2.20 2.20 1.46 2.76
- (c) Side destabilised - main stabilised —
- (d) Side destabilised - main indifferent
 (-120, 160) (80, 40) (60, 60) (80, 60)
 -1.50 -1.70 -2.70 -1.91
 (-140, 120) (-120, 120)
 -1.50 -1.55
-

$$\text{His } 2\sigma_i = 0.86$$

- (a) Side and main stabilised
 (-80, -60) (-60, -60) (-100, -20) (-100, 0)
 1.86 1.28 2.15 2.45
 (-80, 100) (-160, 140) (-80, 140) (-120, 160)
 1.89 1.53 1.22 1.21
- (b) Side stabilised - main indifferent
 (-120, -180) (-80, -160) (-160, 0) (100, 20)
 1.06 1.08 1.01 1.10
 (-120, 40)
 0.91
- (c) Side destabilised - main stabilised
 (-140, 120) (-120, 120) (-100, 120) (-120, 140)
 -1.02 -1.98 -1.22 -1.54
- (d) Side destabilised - main indifferent
 (-100, -60) (-100, -40) (-120, 100) (-160, 120)
 -1.17 -1.22 -1.33 -1.28
 (-60, 140)
 -0.87
-

$$\text{Ile } 2\sigma_i = 0.88$$

- (a) Side and main stabilised
 (-60, -60) (-120, 100) (-160, 120) (-140, 120)
 1.81 2.55 1.34 4.31
 (-120, 120) (-140, 140)
 4.2 0.94
- (b) Side stabilised - main indifferent —
- (c) Side destabilised - main stabilised
 (-100, -20)
 -0.89
- (d) Side destabilised - main indifferent

Table I contd.....-

$$\text{Leu } 2\sigma_i = 0.64$$

- (a) side and main stabilised
- | | | | |
|-------------|-------------|-------------|-------------|
| (-80, -80) | (-100, -60) | (-60, -60) | (-80, -40) |
| 0.97 | 1.21 | 1.92 | 0.72 |
| (-120, 100) | (-100, 100) | (-160, 120) | (-120, 120) |
| 1.20 | 1.70 | 0.73 | 0.88 |
| (-100, 120) | (-100, 140) | (-120, 160) | |
| 2.19 | 1.30 | 0.70 | |
- (b) Side stabilised - main indifferent —
- (c) Side destabilised - main stabilised
- | | | |
|------------|-------------|-------------|
| (-60, -40) | (-120, 140) | (-140, 160) |
| -0.90 | -1.06 | -0.67 |
- (d) Side destabilised - main indifferent
- | | |
|--------------|-------------|
| (-120, -180) | (-120, -20) |
| -0.70 | -0.83 |
-

$$\text{Lys } 2\sigma_i = 0.52$$

- (a) Side and main stabilised
- | | | | |
|--------------|------------|------------|-------------|
| (-120, -180) | (-60, -80) | (-40, -80) | (-80, -60) |
| 0.58 | 1.35 | 0.76 | 0.58 |
| (-40, -60) | (-60, -40) | (-100, 0) | (-140, 160) |
| 0.63 | 1.65 | 0.58 | 2.08 |
- (b) Side stabilised - main indifferent
- | | | | |
|------------|------------|-------------|------------|
| (-20, -60) | (-100, 60) | (-180, 100) | (-40, 100) |
| 1.04 | 0.58 | 0.73 | 0.75 |
- (c) Side destabilised - main stabilised
- | | | | |
|-------------|------------|-------------|-------------|
| (-120, 100) | (-60, 120) | (-160, 140) | (-140, 140) |
| -0.63 | -0.65 | -0.66 | -1.55 |
| (-120, 160) | | | |
| -0.58 | | | |
- (d) Side destabilised - main indifferent
- | |
|------------|
| (-60, 100) |
| -0.56 |
-

$$\text{Met } 2\sigma_i = 1.32$$

- (a) Side and main stabilised
- | | | | |
|-------------|------------|-------------|-------------|
| (-60, -60) | (-40, -60) | (-140, 140) | (-120, 140) |
| 9.06 | 1.51 | 2.34 | 3.06 |
| (-140, 160) | | | |
| 1.49 | | | |
- (b) Side stabilised - main indifferent
- | |
|------------|
| (-100, 80) |
| 1.36 |
- (c) Side destabilised - main stabilised —
- (d) Side destabilised - main indifferent
- | | |
|-------------|-------------|
| (-100, -20) | (-100, 140) |
| -1.94 | -1.81 |

Table I contd.....

Phe $2\sigma_i = 0.58$

- (a) Side and main stabilised
 (-60, -80) (-100, -60) (-40, -60) (-120, -40)
 0.87 0.80 0.59 0.80
 (-120, -20) (-120, 100) (-180, 140) (-160, 160)
 1.74 1.24 0.59 1.62
 (-120, 160)
 1.23
- (b) Side stabilised - main indifferent
 (-80, -100) (-180, 60) (-180, 120)
 0.66 0.71 0.76
- (c) Side destabilised - main stabilised
 (-60, -60) (-100, -20)
 -0.81 -1.15
- (d) Side destabilised - main indifferent
 (-40, -80) (-60, -20) (-100, 160)
 -1.07 -0.76 -0.63

Pro $2\sigma_i = 1.76$

- (a) Side and main stabilised
 (-60, -80) (-60, -40)
 2.79 2.87
- (b) Side stabilised - main indifferent
 (-80, -180) (-60, -20) (-60, 100) (-60, 120)
 2.01 4.71 2.65 5.37
 (-80, 140) (-60, 140) (-80, 160) (-60, 160)
 6.01 5.56 2.89 4.91
- (c) Side destabilised - main stabilised -
- (d) Side destabilised - main indifferent
 (-80, -60) (-140, 120) (-120, 120) (-140, 140)
 -2.7 -2.78 -3.15 -2.15
 (-120, 140)
 -2.38

Ser $2\sigma_i = 0.58$

- (a) Side and main stabilised
 (-120, 0) (-160, 140) (-140, 140) (-120, 160)
 1.16 1.46 1.74 0.67
 (-100, 160)
 0.93
- (b) Side stabilised - main indifferent
 (-160, -180)
 0.70

contd.....

- (c) Side destabilised - main stabilised
 (-60, -80) (-60, -60) (-80, -40) (-120, 100)
 -1.77 -1.97 -0.62 -1.31
 (-120, 120) (-80, 120)
 -1.50 -0.60
- (d) Side destabilised - main indifferent
 (-60, 120) (-120, 80)
 -0.93 -0.68

Thr $2\sigma_i = 0.66$

- (a) Side and main stabilised
 (-120, -40) (-100, -20) (-160, 120) (-140, 120)
 0.87 1.76 0.77 0.92
 (-120, 120) (-120, 140) (-120, 160) (-100, 160)
 2.39 1.40 1.13 1.24
 (-80, 160)
 0.79
- (b) Side stabilised - main indifferent —
- (c) Side destabilised - main stabilised
 (-80, -60) (-60, -60) (-60, -40)
 -2.34 -1.60 -1.07
- (d) Side destabilised - main indifferent
 (-100, -60) (-40, -60)
 -0.76 -0.76

Trp $2\sigma_i = 0.92$

- (a) Side and main stabilised
 (-160, -180) (-100, -40) (-60, -40) (-80, 100)
 1.7 1.13 1.49 1.17
 (-160, 120) (-140, 120) (-120, 120)
 1.66 2.37 2.73
- (b) side stabilised - main indifferent
 (-140, -180) (-160, 20) (-120, 20) (-100, 20)
 1.78 1.26 1.66 0.97
 (-140, 80) (-160, 100) (-180, 160)
 1.59 0.97 1.04
- (c) Side destabilised - main stabilised
 (-60, -60)
 -2.65
- (d) Side destabilised - main indifferent
 (-40, -60) (-100, -20) (-100, 100) (-80, 120)
 -1.38 -1.21 -1.23 -1.06
 (-160, 160) (-100, 160)
 -1.13 -1.02

contd.....

Tyr $2\sigma_i = 0.90$

- (a) Side and main stabilised
 (-120, 100) (-140, 120) (-120, 120) (-180, 140)
 1.69 1.90 1.52 1.93
 (-140, 140) (-120, 140) (-160, 160)
 1.13 3.41 1.74
- (b) Side stabilised - main indifferent
 (-120, -140) (-140, -20) (-120, 80) (-80, 80)
 0.97 0.98 1.11 0.98
- (c) Side destabilised - main stabilised
 (-60, -60) (-80, -40) (-80, -20)
 -2.01 -2.07 -1.14
- (d) Side destabilised - main indifferent
 (-40, -60) (-100, -40) (-160, 120) (-80, 140)
 -1.02 -1.09 -0.92 -0.98
-

Val $2\sigma_i = 0.86$

- (a) Side and main stabilised
 (-100, -60) (-120, 80) (-140, 100) (-120, 100)
 0.97 0.96 1.79 2.52
 (-140, 120) (-120, 120) (-100, 120) (-160, 140)
 -1.50 3.26 1.23 1.07
 (-140, 140) (-120, 140) (-140, 160)
 1.97 1.73 1.89
- (b) Side stabilised - main indifferent —
- (c) Side destabilised - main stabilised
 (-60, -60) (-80, -20)
 -1.06 -1.08
- (d) Side destabilised - main indifferent
 (-100, -40) (-100, -120) (-100, 0) (-80, 140)
 -1.15 -1.12 -1.06 -1.05
-

Further, the analysis of ΔP_j^i values for type (a) conformations of residues, which prefer secondary structures such as α -helix indicate certain interesting features.

- (i) ΔP_j^i values are maximum in α -helical region for α -helix preferers.
- (ii) ΔP_j^i was found maximum for α -helix preferring residues for following main chain conformations; Ala (-60° , -60°); Glu (-80° , -60°); Met (-60° , -60°); Gln (-100° , -40°); His (-100° , -20°), Leu (-60° , -60°) and Lys (-60° , -60°). This indicates that stabilisation from respective side chains for helix preferers is not necessarily maximum for (-60° , -60°)-grid, an ideal α -helix conformation. Similar observations were made for the (-100° , 120°)-grid for β -sheet preferers.

Thus this analysis indicates that this spread of main chain conformations in α -helical or β -sheet regions is due to influence of side chains of amino acid residues.

Case (b): Side stabilised - main indifferent conformations:

$$\Delta P_j^i > 2 \sigma_i; \quad P_j^{GR} < 2 \sigma_i; \quad \Delta P_j^i > 0$$

This type of conformations, given in rows (b) of Table I, have the contribution from main chain atoms in negligible quantity and the stabilisation is solely due to specific side chains. Thus these conformations can be used as a marker to distinguish one residue from the other.

Residues Ala, Glu, Ile, Leu, Thr and Val do not have a single conformation under (b) category. This indicates that the side chain of

these residues do not stabilise main chain conformations in a characteristic fashion. The stabilising effect of these side chains on the main chain conformations is similar to fictitious side chain effect in this case.

Case (c): Side destabilised - main stabilised

$$P_j^i > 2 \sigma_i; \quad P_j^{GR} > 2 \sigma_i; \quad \Delta P_j^i < 0$$

The conformations which fall under category (c) have $P_j^{GR} > P_j^i$ contribution to stabilisation energy, from main chain and side chain atoms are out of phase. In spite of the destabilisation effect from side chain atoms, the main chain atom interactions are strong enough to drive the residue to take these conformations. Residues Gly, Met and Pro do not have a single conformation under this category.

Case (d): Side destabilised - main indifferent

$$P_j^i < 2 \sigma_i; \quad P_j^{GR} > 2 \sigma_i; \quad \Delta P_j^i \sim 0$$

For conformations of this type, the stabilising main chain interactions term gets almost nullified by side chain destabilising interactions term. A conformation of this type, assumed by a particular residue in a polypeptide chain, results from a delicate balance between main chain and side chain atom interactions acting in opposite directions and the effect of neighbouring residues both sequentially and spatially. Thus inter-residue interactions would play quite a significant role in making residues assume this type of conformation.

III.4 GENERAL DISCUSSION

Main chain conformations in allowed regions of (ϕ, ψ) -plane, that do not figure in Table I for respective residues, may be assumed to be neither stabilised nor destabilised and hence have neutral influence from side chain atoms. These conformations are not characteristic of any particular side chain.

Table I further shows that there are a few main chain conformations which can be either stabilised or destabilised by several side chains. For example, grids $(-100^\circ, -20^\circ)$, $(-80^\circ, -60^\circ)$ and $(-60^\circ, -60^\circ)$ which lie in region A of the (ϕ, ψ) -map (see legend of Fig. 1) and $(-140^\circ, 140^\circ)$, $(-120^\circ, 120^\circ)$, $(-140^\circ, 120^\circ)$, $(-120^\circ, 100^\circ)$, $(-120^\circ, 140^\circ)$, and $(-80^\circ, 140^\circ)$, which lie in region B (see legend of Fig. 1) received significant contribution from side chains of ten or more residues. All these grids are not only part of thickly populated regions of (ϕ, ψ) -map but also are representatives of two major secondary structures, α -helix and β -sheet, of globular proteins.

Further one notices that these above-mentioned conformations of region A, are stabilised not only by side chains of α -helix preferring residues but also by side chains of residues which are categorised as either α -helix breakers or indifferent. Thus, indicating that though it is correct to say that all α -helix preferring residues should have at least one conformation among those belonging to region A and the side chains should have stabilising effect on main chain conformations, the converse is not true.

Similar observations are made for conformations, representing β -sheet and which are part of region B. Further, in case of conformations taken in β -sheets, the reasons for significant contribution from the side chains of residues to these extended conformations is difficult to explain. Thus these conformations require more detailed study. However, a point to be noted here is that for β -sheet preferring residues, their respective side chains stabilise one or the other of the conformations of region B, mentioned above.

In contrast to above mentioned conformations where several side chains contribute significantly for a given conformation, one can observe from Table I that a few main chain conformations exist which are characteristic of only one residue. Such conformations are given in Fig. 2.

3.5 CONCLUSIONS

The discussion of the results mentioned above brings out the following points:

- (i) Number of side chain characteristic conformations is far smaller compared to the number of allowed conformations of (ϕ, ψ) -map and vary with the residue.
- (ii) Side chain atoms beyond C^β -atom affect main chain conformations significantly depending on the nature of the side chain.
- (iii) Type (a) conformations, which are stabilised both by main chain and side chain atom interactions, might be the most suitable conformations to be taken by amino acid residues during nucleation of protein folding process.

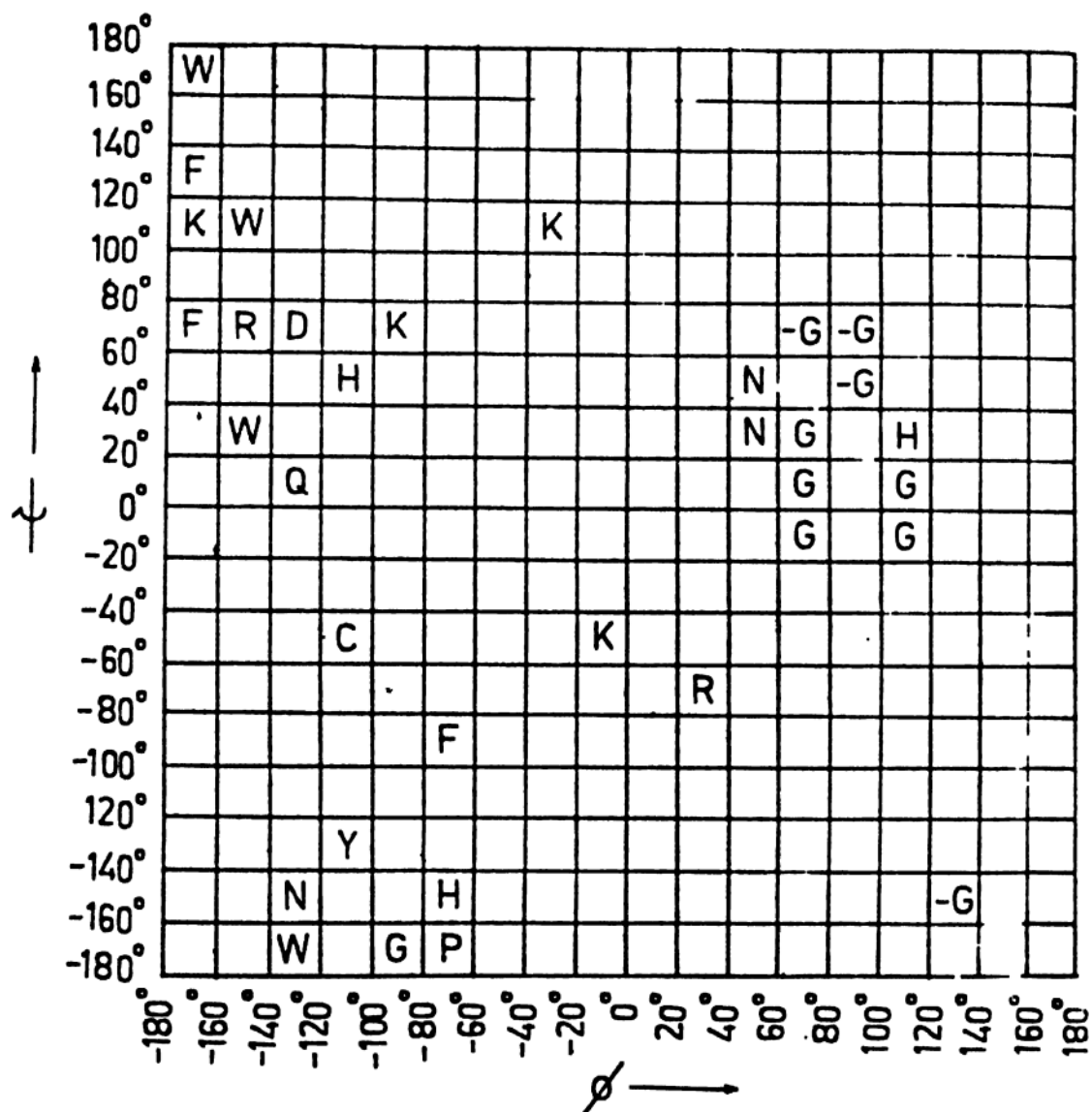


Fig. 2

Residue mentioned in a grid of (ϕ, ψ) -map takes that conformation which has a significant contribution from its side chain. Side chain of any other residue does not contribute significantly for this grid. - sign before a residue indicates that this conformation is destabilised by respective side chain. Single letter amino acid code has been used.

- (iv) A set of main chain conformations which can be used as a marker to distinguish one residue from the other is also obtained.

CHAPTER IV

CONFORMATIONAL PROPERTIES OF PAIRS OF AMINO ACIDS

IV.1 INTRODUCTION

It has been stated in Chapter I that our aim is to predict rough three dimensional structure of proteins and polypeptides from amino acid sequence. The information which we have derived using crystal structure data of a large number of globular proteins and discussed in Chapters II and III, is at single amino acid level. Since we have used statistical methods to derive this information the effect due to particular environment is not considered. In particular, the effect of amino acids in the neighbourhood of a residue is also not taken into account. In order to simulate this effect at least artificially, we thought of using presently available secondary structure prediction schemes, such as Robson and Pain (1971); Chou and Fasman (1978) or Garnier et al. (1978). But these methods have achieved only limited success in predicting secondary structures. These results have prompted us to digress from our main task, to find out the reasons for this limited success. Even recent modifications of these methods and their use in secondary structure prediction (Busetta and Hospital, 1982; Palau et al., 1982) have shown that the prediction accuracy does not exceed 60%.

Secondary structural prediction methods mentioned above have one common factor, namely, the use of single residue potentials in prediction schemes. These single residue potentials are derived without regard to the neighbouring residues or their effect and thus are purely compositional in nature. Additional information available in a given sequence is thus not utilised. The sequence effect has been artificially simulated in these methods to some extent (Chou and Fasman, 1978; Garnier et al., 1978) by giving certain

different weightages to the potentials of sequentially nearer residues. Thus a basic assumption is made in all these methods that neighbourhood interactions are additive in nature. In contrast to this, conformational energy calculations carried out on model di, tri and tetra peptides (Ramachandran and Sasisekharan, 1968) indicate that near neighbour interactions are non-linear. Therefore, we have analysed the crystal structure data from 31 different globular proteins to obtain the potentials for pairs of amino acids in four secondary structural states, helix, extended structure, chain reversals and coil region. These pair potentials are further analysed.

IV.1.1 Earlier work on pair potentials

Ptitsyn and Finkelstein (1971) have derived amino acid pair sequences by considering 9 proteins. Amino acid pair potentials were also derived by Kabat and Wu (1973), Nagano (1973), Periti (1974), Robson and Pain (1974). There is a main difference between our present study and studies mentioned. We considered the pairs formed only by adjacent residues, while in above four studies the pairs are formed with residues which are distant by 1 to 7 residues, along the sequence in both directions, from the residue under consideration.

IV.2 METHOD

Thirty one (31) different globular proteins given in Table I were considered for the present study. The helical and extended structure regions are taken from crystal structure data as reported in original papers

TABLE I

List of proteins considered together with number of residues in various secondary structures (protein-wise)

Name of the protein	Number of residues in helix	Number of residues in extended structure	Number of residues in chain reversals	Number of residues in coil state
1	2	3	4	5
Human haemoglobin α -chain	112	0	15	14
Human haemoglobin β -chain	117	0	20	9
Lamprey haemoglobin	117	0	6	25
Spermwhale metmyoglobin	121	0	25	7
Bacterial ferri cytochrome C ₂	38	0	45	29
Bonito Ferro cytochrome C	34	0	43	26
Bacterial cytochrome C ₅₅₀	58	0	38	39
Bacterial Rubredoxin	0	53	0	0
Jackbean concanavalin A	0	118	56	53
Bovine trypsin	11	97	53	87
Porcine trypsin elastase	11	86	74	84
Human Bence-Jones protein RFI (monomer I)	0	73	13	27
Human Immunoglobulin I	0	134	24	59
Human Immunoglobulin II	0	182	33	25
Human prealbumin (monomer II)	9	53	4	38
Chicken lysozyme	37	12	57	23
Bovine Cymotrypsinogen A	26	100	40	79
Bacterial High Potential Protein	9	17	33	26
Carp calcium binding protein	62	9	18	19
Bovine ferri cytochrome b ₅	52	24	0	9
Bovine ribonuclease S complex	32	43	27	22
Bacterial nuclease S complex	36	48	29	29
Human carbonic anhydrase C	29	107	53	59
Subtilisin BPN'	86	37	72	80

contd.....

	2	3	4	5
Bovine Carboxypeptidase A complex	115	55	81	56
Lobster GPD	109	119	49	57
Chicken triose phosphate isomerase (monomer)	124	50	23	46
Bacterial semiquinone flavodoxin	57	42	21	18
Bacterial thermolysin	110	73	76	67
Dogfish apolactate dehydrogenase Complex	120	79	52	31
Bacterial Ferredoxin	0	54	0	0
Horse alcohol dehydrogenase Complex	96	121	70	87
Papain	60	37	38	79

and compiled by Feldmann (1976) in AMSOM. The chain reversal regions are computed using simple algorithm which is briefly discussed below. Regions of protein sequences other than helix, extended structure and chain reversals are considered as coil region.

IV.2.1 Computation of chain reversal regions

The distances between each C^α -atom of the i^{th} and $(i+3)^{\text{rd}}$ residue have been calculated. Those residues, for which the distance is below 7.0 \AA , have the possibility of being either in helix or in chain reversals. To exclude those residues which lie in the helical region the following criterion has been used. For the cases for which the distance $(C_i^\alpha \dots C_{i+3}^\alpha) = d_i \leq 7.0 \text{ \AA}$, the distances $(C_{i-1}^\alpha \dots C_{i+2}^\alpha) = d_{i-1}$ and $(C_{i+1}^\alpha \dots C_{i+4}^\alpha) = d_{i+1}$ were also computed. If these distances were found to be in the range of $d_i \pm 1.5 \text{ \AA}$, then the residues i to $(i+3)$ were assumed to be in helical region. Under ideal conditions in a helix, $d_{i-1} = d_i = d_{i+1}$. However, inaccuracies in the coordinates of C^α -atoms and distortions in the helices will not give such condition and so, by trial and error, the factor 1.5 \AA has been considered a good estimate. In the present analysis, those chain reversals that are part of either N- or C- terminal of helix are excluded to avoid double counting. Thus the residues, i to $i+3$, for which distance between their C^α -atoms is less than 7 \AA , and that are not part of helical region are taken as chain reversals.

IV.2.2 Computation of pair potentials

Following procedure is adopted to derive the potentials of Pairs formed by 20 amino acids in the four different secondary structural states.

If ACDEF is a stretch representing a particular secondary structural region in protein sequence under consideration, pairs AC, CD, DE and EF were formed from them. The number of times a pair, formed by i^{th} and j^{th} type of residues, occurs in k^{th} secondary structural state, (N_{ijk}) was computed for all pairs in four states. Both i and j vary from 1 to 20 and k from 1 to 4. The total number of each type of pairs N_{ij} , ($N_{ij} = \sum_{k=1}^4 N_{ijk}$) occurring in 31 different globular proteins are given in Table II.

Then P_{ijk} , the potential for a pair of amino acids formed by i^{th} and j^{th} type of amino acids in k^{th} secondary structure was calculated using following relation:

$$S_{ijk} = n_{ijk} / \sum_{i=1}^{20} \sum_{j=1}^{20} N_{ijk},$$

where, $n_{ijk} = N_{ijk} / \sum_{k=1}^4 N_{ijk},$

$$\bar{S}_{ij} = \sum_{k=1}^4 S_{ijk} / 4 ,$$

and $P_{ijk} = S_{ijk} / \bar{S}_{ij} \quad \dots\dots (1)$

The values of P_{ijk} obtained for each type of pair of amino acid residues in each of the four secondary structural states are given in Table II.

IV.2.2 Computation of single residue potentials

The number of times each amino acid residue occurs in each of the four secondary structures (N_{ik}) was calculated. Index i varies from 1 to 20 and k from 1 to 4. Then P_{ik} , the potential of each residue in each of the secondary structures is calculated as:

Number of occurrences of amino acid pairs in the 31 proteins

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	70	8	35	17	16	44	19	20	41	41	6	18	23	21	11	44	26	48	9	18
C	14	1	5	2	0	18	3	6	13	14	1	8	4	9	7	9	4	7	1	4
D	40	6	18	19	15	36	3	26	21	24	3	15	16	4	7	31	17	27	6	13
E	22	5	14	18	16	24	5	9	25	30	6	13	10	13	10	15	15	23	5	7
F	13	4	11	15	7	18	5	10	17	15	4	7	10	3	8	16	19	11	1	8
G	40	14	35	30	13	38	8	34	41	32	3	19	20	17	19	48	39	49	8	22
H	12	8	2	11	9	15	3	5	11	7	3	4	14	0	1	9	10	12	4	6
I	25	5	20	22	11	28	6	20	23	19	4	20	15	11	8	22	15	23	3	11
K	43	6	29	13	15	31	14	23	33	35	13	20	16	3	6	35	25	42	3	28
L	28	4	22	20	12	31	17	30	39	29	7	16	18	23	17	43	32	35	6	10
M	5	0	6	5	5	6	1	5	12	3	2	5	2	3	1	5	3 ^c	11	1	1
N	12	7	10	11	12	20	4	13	14	16	3	14	12	15	8	18	22	19	10	7
P	17	4	22	24	6	25	4	9	14	20	4	12	8	3	5	28	8	24	4	10
Q	24	3	14	8	6	20	8	9	15	12	2	11	9	11	8	15	10	10	4	9
R	7	3	4	6	5	15	3	10	10	20	3	8	6	13	3	20	7	16	2	5
S	46	16	20	22	15	51	9	19	28	37	9	20	14	20	14	50	39	39	14	27
T	32	9	22	19	21	31	10	14	26	28	5	12	23	14	6	30	23	32	8	14
V	54	19	40	22	16	35	18	22	30	42	4	19	14	15	14	42	33	45	6	11
W	8	2	4	4	1	16	2	8	8	4	1	4	2	2	5	7	8	11	1	3
Y	15	5	16	6	3	30	5	8	12	13	2	7	15	9	8	20	20	12	6	10

These pairs can be read by considering the residue, first from vertical column and then from horizontal row. Single letter amino acid code has been used.

following order: (i) helix (h), (ii) extended structure (e), (iii) chain reversal (t) and (iv) coil-region (c).

These pairs can be read by considering the residue, first from verticle column and then from horizontal row.

Single letter amino acid code has been used.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	1.82	0.44	1.24	2.44	2.16	0.87	1.58	1.80	1.67	2.36	2.70	1.28	0.52	1.75	1.16	0.88	0.88	1.56	1.11	1.31
C	0.52	1.95	0.55	0.23	1.01	0.84	0.21	1.14	0.59	1.00	0.71	0.39	1.09	0.00	0.82	0.87	1.24	1.30	0.39	1.38
D	0.65	0.78	1.22	0.75	0.40	1.46	1.00	0.91	0.78	0.32	0.00	1.85	0.58	1.18	0.00	1.11	0.74	0.28	1.86	0.37
E	1.01	0.81	0.99	0.58	0.42	0.83	1.21	0.16	0.97	0.33	0.59	0.48	1.81	1.07	2.02	0.94	1.15	0.86	0.64	0.95
F	0.50	4.00	1.73	1.95	0.00	0.21	0.00	1.29	1.04	0.77	0.00	1.18	0.00	0.41	0.00	1.20	0.96	0.53	4.00	0.00
G	1.58	0.00	0.00	2.05	0.00	1.79	0.00	2.71	0.55	1.91	0.00	0.83	2.61	2.18	1.16	0.42	3.04	1.67	0.00	1.81
H	1.26	0.00	0.00	0.00	0.00	0.71	3.18	0.00	1.74	0.87	0.00	1.99	1.39	0.69	1.34	0.00	0.88	0.00	1.44	
I	0.65	0.00	2.27	0.00	0.00	1.29	0.82	0.00	0.68	0.45	4.00	0.00	0.00	0.72	1.92	1.04	0.00	0.92	0.00	0.75
K	1.03	1.20	1.37	1.76	0.00	0.39	1.46	1.12	1.22	1.55	0.00	0.22	1.09	0.00	0.44	0.40	0.20	1.38	0.68	0.27
L	0.63	1.27	0.62	0.41	0.88	0.81	0.00	1.33	0.55	0.98	0.00	0.46	0.69	1.81	0.94	0.63	1.06	0.73	2.14	1.43
M	1.59	1.01	1.32	0.98	0.47	1.62	0.00	0.94	1.17	0.78	3.18	2.19	1.46	1.44	2.23	2.35	1.69	0.93	0.00	1.36
N	0.75	0.52	0.69	0.85	2.66	1.18	2.54	0.61	1.06	0.68	0.82	1.14	0.76	0.75	0.39	0.61	1.05	0.96	1.18	0.94
P	2.34	0.00	1.33	1.19	1.48	0.81	2.16	1.30	2.15	1.52	1.62	0.54	0.40	1.97	3.03	0.70	1.60	1.42	1.61	2.89
Q	0.57	1.79	0.84	0.63	1.04	0.68	0.00	0.46	0.00	0.57	0.57	0.00	0.84	0.59	0.00	0.73	0.56	1.50	1.69	0.61
R	0.30	0.00	0.89	1.33	0.41	0.81	1.21	0.73	0.77	1.81	1.81	1.81	0.67	0.95	0.64	1.56	0.45	0.80	0.00	0.00
S	0.79	2.21	0.93	0.86	1.07	1.69	0.63	1.51	1.07	0.00	0.00	1.65	2.09	0.49	0.33	1.01	1.39	0.28	0.70	0.50
T	2.02	0.00	0.69	1.19	1.70	0.35	0.82	1.87	1.37	1.45	1.95	0.48	0.86	2.17	1.99	0.93	1.24	1.10	0.00	0.93
V	0.00	2.19	1.82	1.00	1.80	0.73	1.74	1.18	0.48	1.53	2.05	1.51	0.90	0.00	1.58	1.47	1.52	1.94	4.00	1.47
W	0.97	0.00	0.58	1.19	0.00	2.32	0.00	0.63	0.76	0.81	0.00	1.60	0.00	1.83	0.00	0.78	0.35	0.00	0.00	0.78
Y	1.01	1.81	0.90	0.62	0.50	0.60	1.44	0.33	1.39	0.21	0.00	0.42	2.24	0.00	0.43	0.81	0.90	0.96	0.00	0.81
	0.86	0.73	0.54	1.23	1.09	0.38	1.47	0.59	0.96	0.44	0.00	0.84	0.17	0.43	0.55	0.15	0.27	0.61	0.50	0.65
	0.91	0.77	0.79	0.78	0.58	1.20	0.00	1.49	0.19	0.70	2.23	0.44	1.46	0.45	0.96	1.23	0.77	0.80	2.63	0.86
	0.80	1.64	1.08	1.03	1.38	1.27	0.82	0.39	1.47	1.49	1.77	0.70	1.46	1.44	1.53	1.47	1.53	1.14	0.00	1.36
	1.42	0.85	1.59	0.96	0.95	1.15	1.71	1.53	1.38	1.36	0.00	2.01	0.91	1.69	0.96	1.15	1.43	1.45	0.87	1.13
	3.05	0.80	0.00	1.08	0.43	1.73	1.13	0.00	0.83	0.49	2.79	1.11	0.45	0.00	0.00	1.98	1.19	2.23	0.00	0.54
	0.36	0.84	0.00	0.76	1.36	0.78	0.00	2.58	0.59	1.03	0.00	0.00	0.71	0.00	4.00	0.00	0.42	0.87	2.61	1.15
	0.00	2.01	2.63	0.60	0.72	0.83	1.89	0.00	2.34	1.64	0.00	0.00	2.26	0.00	0.00	1.33	0.67	0.54	1.39	1.83
	0.59	0.35	1.37	1.56	1.50	0.65	0.98	1.42	0.24	0.85	1.21	2.89	0.59	0.00	0.00	0.69	1.73	0.56	0.00	0.48
	1.69	0.82	0.93	1.72	0.76	0.77	1.20	2.06	0.85	1.60	2.96	0.95	0.66	0.68	1.38	1.00	0.74	0.33	0.00	0.69
	1.13	1.74	1.96	0.61	1.60	1.49	1.27	0.99	1.25	1.90	1.04	1.41	0.70	2.15	1.45	1.58	1.57	2.10	1.17	1.82
	0.51	0.00	0.62	0.00	0.00	1.08	1.01	0.63	0.57	0.34	0.00	0.64	1.86	0.57	0.77	0.84	0.83	0.56	1.86	0.58
	0.67	1.44	0.49	1.67	1.65	0.67	0.52	0.33	1.33	0.17	0.00	1.00	0.77	0.59	0.40	0.58	0.86	1.01	0.97	0.90
	1.37	0.56	1.28	1.23	1.75	0.81	1.48	1.00	1.28	1.07	1.23	0.55	0.68	2.79	1.20	1.01	1.55	1.49	1.37	0.78
	0.81	0.59	0.54	0.65	0.53	0.43	0.00	1.41	0.45	1.36	0.65	0.78	1.19	0.00	1.27	0.47	0.33	0.98	1.44	1.10
	1.00	1.88	1.07	0.52	0.84	2.22	1.66	0.56	1.43	0.72	0.52	1.23	1.14	0.00	1.01	0.75	0.78	0.47	0.00	1.09
	0.82	0.97	1.11	1.61	0.87	0.53	0.86	1.02	0.84	0.84	1.61	1.44	0.99	1.21	0.52	1.76	1.35	1.06	1.19	1.02
	2.16	0.90	1.23	1.88	0.88	0.97	1.29	1.07	1.69	1.75	1.57	0.48	0.00	0.99	1.76	1.29	0.88	1.66	1.13	0.67
	0.57	0.00	0.16	0.99	0.62	0.64	1.81	2.12	0.94	0.66	1.10	1.02	0.94	1.04	0.93	1.09	0.93	1.87	0.00	1.06
	0.68	1.52	1.81	0.63	1.48	1.02	0.72	0.00	0.50	1.05	0.88	0.81	0.74	0.83	0.74	0.87	0.42	0.19	1.89	1.69
	0.59	1.58	0.80	0.49	1.02	1.37	0.19	0.82	0.87	0.54	0.46	1.69	2.32	1.15	0.58	0.75	1.76	0.29	0.98	0.58

	A	C	D	E	F	G	H	I	K	L	M	H	P	Q	R	S	T	V	W	Y
M	0.86	0.00	1.46	2.44	1.55	1.87	0.00	1.67	1.82	1.29	4.00	1.28	0.00	1.07	4.00	1.47	0.51	1.12	0.00	0.00
	0.90	0.00	0.00	0.86	2.45	0.00	0.00	0.89	0.64	2.71	0.00	0.00	0.00	1.13	0.00	0.00	2.15	1.59	0.00	4.00
	0.00	0.00	0.00	0.00	0.00	1.05	0.00	0.00	1.02	0.00	0.00	2.16	4.00	1.80	0.00	1.24	0.00	0.00	0.00	0.00
	2.24	0.00	2.54	0.71	0.00	1.09	4.00	1.45	0.53	0.00	0.00	0.56	0.00	0.00	0.00	1.29	1.34	1.30	4.00	0.00
	0.58	2.10	0.00	1.09	1.13	0.53	1.00	0.74	0.75	0.72	0.00	0.55	1.16	1.60	0.00	0.18	0.60	1.52	0.84	0.82
	1.22	0.56	2.38	0.38	0.00	0.00	2.12	0.78	1.32	0.00	2.83	0.58	0.61	0.56	0.38	0.58	1.11	1.20	1.33	0.43
N	1.45	0.88	0.63	0.61	1.89	1.78	0.00	2.06	1.26	1.20	0.00	0.93	1.46	0.45	3.00	2.13	1.71	0.95	0.00	2.75
	0.75	0.46	0.98	1.91	0.98	1.69	0.88	0.43	0.66	2.08	1.17	1.93	0.76	1.39	0.62	1.11	0.52	0.33	1.83	0.00
	1.24	0.00	0.34	0.65	1.37	0.00	2.14	0.94	0.75	0.33	1.00	0.57	0.40	1.46	0.00	0.38	0.89	0.86	1.69	0.40
P	0.65	0.80	0.18	0.27	1.44	0.91	0.00	1.00	0.26	1.04	2.12	0.30	0.85	0.00	0.79	0.93	0.47	1.64	0.89	0.84
	1.39	2.54	1.42	2.40	0.00	1.45	0.00	0.00	1.68	1.92	0.00	1.90	2.04	0.00	1.26	1.48	1.49	0.00	1.42	0.67
	0.72	0.66	2.06	0.68	1.19	1.63	1.86	2.06	1.31	0.71	0.88	1.23	0.71	2.54	1.96	1.21	1.16	1.50	0.00	2.09
	1.37	0.00	0.93	2.81	0.00	0.85	1.34	0.86	1.73	1.60	1.95	1.55	0.00	0.91	0.97	0.79	0.60	1.64	0.74	1.31
Q	0.80	1.17	0.49	0.00	3.43	0.71	0.00	0.91	0.78	1.02	2.05	0.41	0.94	1.29	0.51	1.39	0.63	1.29	0.78	2.31
	1.02	1.86	1.96	0.79	0.00	1.70	1.50	0.72	0.83	0.54	0.00	0.00	0.74	1.53	0.82	0.44	2.51	0.00	2.48	0.00
	0.80	0.97	0.61	0.41	0.57	0.74	1.17	1.50	0.65	0.84	0.00	2.03	2.32	0.27	1.70	1.38	0.26	1.07	0.00	0.38
	1.80	1.13	2.04	1.82	1.38	0.24	2.79	0.72	1.56	0.91	4.00	0.45	0.00	1.14	1.07	1.05	0.54	0.70	2.14	0.73
R	0.63	0.00	1.07	0.64	1.46	1.03	0.00	2.67	0.41	1.54	0.00	0.47	0.78	0.90	1.13	0.56	1.14	1.71	0.00	0.77
	0.00	1.89	0.00	1.02	1.16	1.23	0.00	0.61	0.66	0.92	0.00	1.51	0.00	0.96	1.80	1.47	0.91	0.78	0.00	1.23
	1.57	0.98	0.89	0.53	0.00	1.49	1.21	0.00	1.37	0.64	0.00	1.57	3.22	1.00	0.00	0.92	1.41	0.81	1.86	1.27
	1.54	0.88	1.40	1.43	0.24	0.36	0.81	0.78	0.60	0.50	1.09	0.50	1.04	1.02	0.75	0.29	0.26	0.68	1.11	0.26
S	0.94	0.69	0.37	0.00	1.51	1.08	0.43	1.23	0.76	1.60	0.77	0.53	0.28	0.36	0.26	0.83	1.11	1.83	1.46	1.25
	0.82	1.47	1.47	1.60	1.20	1.10	1.36	0.98	1.81	0.85	1.83	1.96	1.31	1.72	1.68	1.44	1.62	0.65	0.46	1.33
	0.71	0.96	0.76	0.97	1.04	1.46	1.41	1.02	0.94	1.05	0.32	1.02	1.37	0.99	1.31	1.44	1.00	0.84	0.97	1.15
	1.42	0.00	0.49	1.23	0.52	0.57	1.45	1.00	0.85	1.21	0.00	1.63	0.90	1.14	0.00	0.60	0.51	0.58	0.50	0.54
T	0.75	3.24	0.69	0.65	1.84	0.97	0.38	1.31	0.60	1.42	1.79	0.69	0.79	0.90	1.51	0.63	1.60	1.72	2.63	1.14
	0.80	0.00	1.38	0.69	0.88	1.35	1.22	1.25	1.19	0.68	0.00	0.55	1.52	0.48	0.00	1.41	0.57	0.78	0.00	0.91
	1.03	0.76	1.43	1.43	0.76	1.10	0.95	0.43	1.36	0.70	2.21	1.14	0.79	1.49	1.49	1.36	1.32	0.91	0.97	1.41
V	2.02	0.57	0.96	0.70	0.49	0.64	0.45	0.88	1.65	1.92	1.05	0.99	0.52	1.44	0.77	0.54	0.94	0.78	0.70	0.70
	0.99	1.81	1.31	2.23	1.81	1.46	2.37	2.22	1.15	0.91	1.11	1.47	0.55	0.51	1.91	2.29	1.24	2.10	1.47	1.48
	0.36	0.96	0.64	0.30	0.41	0.89	0.00	0.29	0.00	0.16	0.00	0.67	1.32	1.21	0.87	0.46	0.59	0.29	0.00	0.59
	0.63	0.66	1.09	0.77	1.28	1.02	1.18	0.61	1.20	1.01	1.84	0.87	1.60	0.84	0.45	0.71	1.23	0.83	1.83	1.23
	2.56	0.00	0.90	0.00	0.00	0.73	2.14	1.81	0.43	1.95	4.00	0.00	0.00	1.95	0.00	0.98	0.55	0.67	0.00	0.00
W	0.54	0.00	0.00	1.15	0.00	0.77	0.00	1.43	1.37	2.05	0.00	3.14	0.00	2.05	0.93	0.52	0.58	2.48	0.00	4.00
	0.00	0.00	1.52	0.00	0.00	0.81	0.00	0.76	1.45	0.00	0.00	0.00	0.00	0.00	0.00	1.65	0.00	0.56	0.00	0.00
	0.89	4.00	1.58	2.85	4.00	1.69	1.86	0.00	0.75	0.00	0.00	0.86	4.00	0.00	3.07	0.86	2.97	0.29	4.00	0.00
Y	0.24	0.79	0.72	0.60	2.62	0.54	0.89	0.44	0.64	2.09	4.00	0.64	0.48	0.46	0.97	0.77	0.17	0.98	0.61	0.72
	1.75	2.51	2.03	1.89	1.38	0.80	0.00	2.81	0.34	0.63	0.00	0.00	0.25	1.94	0.92	1.23	1.63	1.04	1.29	2.67
	1.19	0.00	0.40	1.00	0.00	2.00	0.00	0.75	1.07	0.50	0.00	1.60	0.00	1.60	0.00	1.46	0.65	1.44	0.55	1.03
	0.83	0.69	0.84	0.52	0.00	0.66	3.11	0.00	1.95	0.78	0.00	3.36	1.67	1.60	0.76	1.35	0.75	1.43	1.07	0.00

$$S_{ik} = n_{ik} / \sum_{i=1}^{20} N_{ik},$$

$$\text{where, } n_{ik} = N_{ik} / \sum_{k=1}^4 N_{ik},$$

$$\bar{S}_i = \sum_{k=1}^4 S_{ik} / 4,$$

$$\text{and } P_{ik} = S_{ik} / \bar{S}_i \quad \dots\dots\dots (2)$$

The single residue potentials derived thus are given in Table IV.

4.3 RESULTS AND DISCUSSION

4.3.1 Analysis of single residue potentials

As can be seen from Table IV, potential values for 20 proteinous amino acids are obtained in coil region for the first time. Also in Table IV, single residue potentials derived by Levitt (1978) for each residue in helix, extended structure and chain reversals, are given in brackets. It can be seen that the potential values are of same order except in very few cases. This agreement shows that our data set is comparable with others and is sufficient.

The major difference is in the potential values for Cys when the potential values derived by us and Levitt (1978) are compared. From our computation, Cys has high potential in extended structure and low potential in helical state. But in Levitt's computation, Cys has high potential in helix and low potential in β -sheet.

Table IV further indicates that all the residues take conformations in all the four structural states. The potential values for several residues

TABLE IV

Single residue potentials obtained using
data of proteins mentioned. Values given
from Levitt (19

amino acid residue	Helix	Extended structure	Chain reversals	Coil region
ala	1.46 (1.29)	0.80 (0.90)	0.81 (0.77)	0.93
arg	0.80 (1.11)	1.31 (0.74)	1.10 (0.81)	0.83
asp	0.98 (1.04)	0.91 (0.72)	1.28 (1.41)	0.82
asn	1.47 (1.44)	0.74 (0.75)	0.96 (0.99)	0.83
ile	1.20 (1.07)	1.21 (1.32)	0.75 (0.59)	0.84
lys	0.62 (0.56)	0.88 (0.92)	1.28 (1.64)	1.22
met	1.26 (1.22)	0.77 (1.08)	1.11 (0.68)	0.86
pro	1.08 (0.97)	1.43 (1.45)	0.66 (0.51)	0.82
ser	1.15 (1.23)	1.71 (0.77)	0.97 (0.96)	1.16
thr	1.24 (1.30)	1.06 (1.02)	0.84 (0.58)	0.86
trp	1.50 (1.47)	0.97 (0.97)	0.74 (0.41)	0.79
tyr	0.78 (0.90)	0.77 (0.76)	1.30 (1.28)	1.14
val	0.58 (0.52)	0.80 (0.64)	1.25 (1.91)	1.36
leu	1.16 (1.27)	0.92 (0.80)	1.08 (0.98)	0.84
his	0.99 (0.96)	1.04 (0.99)	1.04 (0.88)	0.93
gln	0.74 (0.82)	0.97 (0.95)	1.30 (1.32)	0.98
glu	0.75 (0.82)	1.08 (1.21)	0.97 (1.04)	1.21
pho	1.03 (0.91)	1.47 (1.49)	0.53 (0.47)	0.98
pro	0.81 (0.99)	1.24 (1.14)	0.72 (0.76)	1.23
tyr	0.74 (0.72)	1.28 (1.25)	1.03 (1.05)	0.95

are statistically not different in all f
 Therefore, we feel that classification of
 preferring or breaking a particular seco
 potential values derived above will be of little practical use. The
 single residue potentials have been computed to compare our results
 on amino acid pair potentials discussed below and thus to achieve
 internal consistency in our results. A point can be mentioned here
 that the amino acid residues Asn and Thr have fairly large potential
 value in the coil state indicating that these two residues do not prefer
 any single regular structure.

IV.3.2 Analysis of amino acid pair potentials

Total number of amino acid pairs used to calculate individual amino
 acid pair potentials (P_{ijk}) are 6028, out of which 1634 pairs are in
 helical state, 1548 pairs in extended structure, 973 pairs in chain
 reversals and 1873 pairs in coil state. There being 400 types of pairs
 of amino acids formed by 20 types of amino acids, certain types of pairs
 have occurred few times not only in particular state but also in all four
 structures put together. The value of the potential for those pairs of
 amino acid residues which have occurred only a few times in our data set will
 have much higher values of standard deviation and further may not give
 true potentials. Therefore, to differentiate those pairs, whose potential
 values can be used, with certain amount of confidence following analysis
 was done.

The mean value of occurrence of a pair (\bar{N}) is calculated using Table II. P_{ijk} values, whose $\sum_{k=1}^4 N_{ijk}$ are $\geq \bar{N}$ ($\bar{N} = 15$) only have been used for further analysis. Other P_{ijk} values given in Table II can at best be used for qualitative discussion.

When one looks at the potential value given in Table III and IV, it would appear that the pairs, formed by amino acid residues which individually have higher potential for a particular secondary structure, also have higher potential for the same secondary structure. But this is not always true. Some pairs have less potential for the secondary structure, inspite of individual residues forming the pairs having higher potentials in that secondary structure. Such pairs are given in Table V(a).

Tables III and IV also illustrate that even though the constituent amino acid residues have less potential for a particular secondary structural state, the pair formed by such residues may have a higher potential in that secondary structural state as can be seen from examples given in Table V(b).

The above observations suggest that the pairs of amino acid residues have different conformational properties from those of their constituent single amino acid residues. However, it does not clearly indicate whether one can reconstitute the pair potentials given in Table III. To clarify this amino acid pair potentials have been computed by using single residue potentials of Table IV, in all four states both by addition and multiplication of these individual potential values.

TABLE V (a)

pairs of amino acids having less potential in the secondary structure in which the constituent residues have high potential. Potential values are given in the brackets

Helix

Ile (1.08), Gln (1.16), Ile-Gln (0.68); Ile (1.08), Val (1.03), Ile-Val (0.33);
 Lys (1.15), Ile (1.08), Lys-Ile (1.00); Leu (1.24), Gln (1.16), Leu-Gln (0.99);
 Gln (1.16), Gln (1.16), Gln-Gln (0.91); Val (1.03), Glu (1.47), Val-Glu (0.70);
 Val (1.03), Phe (1.20), Val-Phe (0.49); Val (1.03), Ile (1.08), Val-Ile (0.88).

Extended structure

Ile (1.43), Ile (1.43), Ile-Ile (0.99); Leu (1.06), Leu (1.06), Leu-Leu (0.66);
 Leu (1.06), Arg (1.04), Leu-Arg (0.93); Leu (1.06), Thr (1.08), Leu-Thr (0.93);
 Val (1.47), Leu (1.06), Val-Leu (0.91);

Chain reversals

Cys (1.10), Gly (1.28), Cys-Gly (0.71); His (1.11), Gly (1.28), His-Gly (0.83);
 Asn (1.30), Gln (1.08), Asn-Gln (0.45); Gln (1.08), Ser (1.30), Gln-Ser (0.44);
 Tyr (1.03), Asp (1.28), Tyr-Asp (0.40); Tyr (1.03), Ser (1.30), Tyr-Ser (0.65).

Coil

Lys (1.16), Gly (1.22); Lys-Gly (0.53); Lys (1.16), Lys (1.16), Lys-Lys (0.84);
 Lys (1.16), Pro (1.36), Lys-Pro (0.99); Aln (1.14), Thr (1.21); Asn-Thr (0.26);
 Thr (1.21), Pro (1.36), Thr-Pro (0.79).

Three letter amino acid code has been used.

TABLE V (b)

Pairs of amino acids having high potential in the secondary structure in which the constituent residues have low potential. Potential values are given in the brackets

Helix

Asp (0.98), Asp (0.98), Asp-Asp (1.37); Asp (0.98), Pro (0.58), Asp-Pro (1.09); Arg (0.99), Ser (0.74), Arg-Ser (1.05); Ser (0.74), Asp (0.98), Ser-Asp (1.40).

Extended structure

Ala (0.80), Pro (0.80), Ala-Pro (1.09); Gly (0.88), Gly (0.88); Gly-Gly (1.20); Gly (0.88), Pro (0.80), Gly-Pro (1.40); Lys (0.71), Pro (0.80); Lys-Pro (1.11); Gln (0.92), Ser (0.97), Gln-Ser (1.39); Ser (0.97), Gly (0.88); Ser-Gly (1.08).

Chain reversals

Ala (0.81), Ile (0.66), Ala-Ile (0.91); Glu (0.96), Glu (0.96), Glu-Glu (1.33); Glu (0.96), Leu (0.84), Glu-Leu (1.81); Lys (0.98), Ala (0.81), Lys-Ala (1.00); Lys (0.98), Lys (0.98), Lys-Lys (1.43); Leu (0.84), Leu (0.84), Leu-Leu (1.05); Thr (0.97), Lys (0.98); Thr-Lys (1.11).

Coil

Ala (0.93), His (0.86), Ala-His (1.21); Ala (0.93), Gln (0.84), Ala-Gln (1.07); Asp (0.82), Phe (0.84), Asp-Phe (2.66); Glu (0.83), Phe (0.84), Glu-Phe (1.07); Ile (0.82), Glu (0.83), Ile-Glu (1.67); Leu (0.86), Gln (0.84), Leu-Gln (1.14); Gln (0.84), Ser (0.98), Gln-Ser (1.38); Ser (0.98), Phe (0.84), Ser-Phe (1.04); Ser (0.98), Leu (0.86), Ser-Leu (1.05); Ser (0.98), Ser (0.98), Ser-Ser (1.44); Ser (0.98), Tyr (0.95), Ser-Tyr (1.15); Val (0.98), Asp (0.82), Val-Asp (1.09); Val (0.98), Phe (0.84), Val-Phe (1.28); Val (0.98), His (0.86), Val-His (1.18); Tyr (0.95), Ser (0.98), Tyr-Ser (1.35).

Three letter amino acid code has been used.

These calculated potential values were compared with observed potentials after proper normalisation for pairs of amino acids given in Table III. The comparison shows that the association effect between residues in the pair which is included in the values of potentials given in Table III, is considerable. In other words, comparing the ratio of potentials calculated for the pair of amino acids with observed potentials in all four states, the variation is considerable indicating that short-range interactions are not incorporated in single residue potentials used presently in protein folding studies.

In order to study whether the potential values P_{ijk} are functions of the position of amino acid residue forming the pairs, we have analysed the pair potentials of those pairs which are formed by a same residues, but existing in different positions in the pair such as AC and CA. The following criterion was used to identify the sets of pairs of amino acids which have different P_{ijk} values.

$$\text{The value, } \bar{P} = P_{ijk} + P_{jik} / 2 \quad \dots (3)$$

when $i \neq j$ is computed

$$\text{And, if } (\bar{P} - 10\% \bar{P}) < (P_{ijk} \text{ and } P_{jik}) < (\bar{P} + 10\% \bar{P}) \quad \dots (4)$$

only then P_{ijk} and P_{jik} values were considered to be the same.

When the above analysis is carried out in all the four states, some sets of pairs satisfied condition (4) in particular secondary structure only. For example, the set of pairs Ala-Glu and Glu-Ala satisfy condition (4) only in helical state. Thus the residues of these pairs show

positional effect when they occur in secondary structures other than helix. The sets of pairs of amino acids together with secondary structures in which the residues of the pair do not exhibit positional effect, are given in Table VI. This observation indicates that some amino acids when they form pairs in particular secondary structures, the association effect is same even if the residues interchange positions in the pair.

The significant difference in P_{ijk} and P_{jik} values, as revealed by above analysis, indicates the side chain of a residue has different effects on tripeptide conformation when it occurs in position one or two of the pair. This has also been shown from conformational energy calculations on tri-peptides (Brown III et al., 1972). The asymmetry in potential values is, not simply due to the statistics applied. This information is completely lost when one is working at single amino acid residue level.

Thus potentials of pairs of amino acids, given in Table III, include the residue-residue interactions at the short-range. The association effect is found to be a function of (i) types of amino acids associated, (ii) position they occupy in the pair, and (iii) the secondary structures in which the pair occurs. Thus the pair potential (P'_{ijk}) obtained by combination of any two residues might be split in the following way:

$$P'_{ijk} = (SS_{ik} \cdot X_{i1}) (SS_{jk} \cdot X_{j2}) \dots (5)$$

X_{i1} and X_{j2} are single amino acid potentials when i^{th} type of residue is in position one and j^{th} type of residue is in position two. It should -

TABLE VI

Pairs of amino acids (of type AC and CA) whose P_{ijk} values satisfy the condition 4 (see text)

In all four secondary structural states	Asp-Lys
In helical extended structure and chain reversal states	Ile-Lys
In helical extended structure and coil states	Val-Lys
In extended structure, chain reversal and coil states	Asn-Ser
In helical and extended structure states	Ala-Asp; Ala-Gly, Ala-Ile; Gly-Tyr
In helical and chain reversal states	Val-Leu
In helical and coil states	Ala-Lys
In extended structure and chain reversal states	Phe-Lys; Thr-Val
In extended structure and coil states	Glu-Thr; Phe-Thr; Gly-Leu; Ser-Tyr
In chain reversal and coil states	Ala-Thr; Glu-Ser
In helical state only	Ala-Glu; Ala-Leu; Asp-Ile; Gly-Lys; Gly-Val
In extended structure state only	Ala-Ser; Asp-Gly; Glu-Gly; Glu-Phe; Phe-Ser; Gly-Ile; Gly-Asn; Gly-Arg; Gly-Ser; Ile-Leu; Ile-Val; Leu-Pro.
In chain reversal state only	Ala-Gln; Asp-Pro; Gly-Pro; Gly-Gln; Gly-Thr; Ile-Ser; Leu-Ser; Ser-Thr
In coil state only	Ala-Tyr; Asp-Leu; Asp-Val Lys-Leu; Lys-Thr; Leu-Arg Ser-Val.

Three letter amino acid code has been used.

be noted that X_{i1} and X_{j2} represent not only the type of amino acid but also the position of residue in the pair. Thus for a residue i , X_{i1} and X_{j2} are assumed to be different. SS_{ik} and SS_{jk} are secondary structural contribution to the association effect.

Based on above assumptions, if the observed potentials of pairs (P_{ijk}) can be partitioned into potentials of single residues according to eq. (5), and the association effect is linear, the ratio of potentials of corresponding pairs in the given secondary structure should satisfy the following equation:

$$\frac{P_{AA}}{P_{CA}} = \frac{P_{AC}}{P_{CC}} = \dots \frac{P_{AY}}{P_{CY}}$$

The ratio will also be equal to:

$$\frac{P_{AA}}{P_{CA}} = \frac{P_{CA}}{P_{CC}} = \dots \frac{P_{YA}}{P_{YC}}$$

where, P_{AA} etc. are P_{ijk} values for particular secondary structure. But these ratios differed considerably even after considering the statistical fluctuations. Thus this analysis confirms that pairs of amino acids are not mere linear combination of two residues.

IV.5 CONCLUSIONS

From this study, we find that amino acids Thr, Lys, Gly, Pro, Asn and Trp have higher potential in coil state. It is also clear from the results and discussion above that single amino acid potentials, though statistically more reliable, will not account for the interactions of adjacent

residues. Our present study indicates that these interactions vary depending upon (i) the position of amino acid residue in an amino acid pair, and (ii) the secondary structure in which these pairs are present. Probably the most important result of our study, which might be intuitively obvious at the outset, is that pairs of amino acids cannot be regarded as a linear combination of constituent amino acids. Even at the secondary structural level, the conformational properties of pairs of amino acids differ from the conformational properties of constituent amino acid residues. Secondary structure prediction algorithms must, therefore, incorporate this aspect for more accurate results.

C H A P T E R V

ANALYSIS OF CRYSTALLOGRAPHICALLY OBSERVED SECONDARY STRUCTURES

V.1 INTRODUCTION

In the previous chapter it is mentioned that algorithms developed to predict secondary structures are not successful beyond a particular limit. One of the reasons for this limited success, we thought, might be the use of single residue potentials, and, therefore, potentials were developed for pairs of amino acid residues. As a continuation of our efforts to understand the reasons for failure of statistical methods used quite commonly to predict secondary structures, the analysis of observed secondary structures of globular proteins was undertaken. This chapter contains the details of this analysis.

V.2 METHOD

Thirty one (31) proteins given in Chapter IV are considered for this analysis. The helical and extended structure regions of these proteins are taken from original papers as compiled by Feldmann (1976) in AMSOM and the chain reversals are computed using the algorithm described in Chapter IV.

The analysis of these observed structures was carried out in the following way. The amino acid sequence of observed secondary structural stretch is considered. If ACDEFGH is the single letter coded amino acid sequence, of that particular stretch, pairs AC, DE and EF were considered first. The pair potentials for respective secondary structures were added up and the totals T_{α_1} , T_{β_1} , T_{t_1} and T_{c_1} in helical, extended structure, chain reversals and coil state are obtained. Then the pairs CD, EF and GH

were considered and similarly total potentials, T_{α_2} , T_{β_2} , T_{t_2} and T_{c_2} were obtained in the four secondary structural states. Then the average potentials, A_{α} , A_{β} , A_t and A_c were obtained for the stretch under consideration. It may be mentioned that residues other than those occurring at the N- and C-terminal are counted twice.

$$\begin{aligned}
 A_{\alpha} &= T_{\alpha_1} + T_{\alpha_2}/2 \\
 A_{\beta} &= T_{\beta_1} + T_{\beta_2}/2 \\
 A_t &= T_{t_1} + T_{t_2}/2 \\
 A_c &= T_{c_1} + T_{c_2}/2
 \end{aligned}
 \qquad \dots\dots (1)$$

Such four average potentials, A_{α} , A_{β} , A_t and A_c were computed representing the four structural states, helix, extended structure, chain reversals and coil state, for each observed secondary structural stretch.

Similarly, average potentials were calculated for all observed secondary structures using a single residue potentials derived by us (Chapter IV, Table IV), Chou and Fasman (1978) and Garnier et al. (1978). The results and the analysis of these results are given in following sections.

V.3 RESULTS AND DISCUSSION

During the analysis of average potentials, for each observed secondary structural stretch in all four secondary structural states, computed using amino acid pair potentials or single residue potentials, one important observation was made. The observation being that the average potential is not always maximum for the observed secondary structural state. We thus termed,

the secondary structural stretches for which the average potential is a maximum for the observed secondary structural state, as type I secondary structures and the secondary structural stretches, for which the average potential is a maximum, for a secondary structural state other than observed state, as type II secondary structures. The observed secondary structures in 31 different globular proteins divided into type I or type II based on an analysis of potential calculated using amino acid pair potentials are given in Table I.

To understand the formation of type I and type II secondary structures, three different aspects were studied. They are: (a) composition of amino acids of these structures, (b) the occurrence of type I and type II structures in different portions of primary structure of protein, and (c) the length of the secondary structural stretch.

(a) Composition of amino acids: Study of the composition of the observed secondary structures was undertaken mainly because it is inherently assumed in all prediction schemes of secondary structures based on statistics that the secondary structural stretches found by crystallographic analysis, always have residues which either prefer or are indifferent to that structure. However, as mentioned above, we have found considerable number of observed secondary structures which have less calculated potential for the state in which they exist. The calculated average potential for the observed secondary structural state for the type II secondary structures can be less, if (i) most of the constituting amino acids have low potential for the observed secondary structures, or (ii) most of the residues present

TABLE I

Secondary structural regions, h (helix), e (extended structure) and t (chain reversals), observed in globular proteins are given below along with their categorisation namely type I and type II (*) secondary structure. Structural type of the protein is given in brackets. The 'others' category also includes proteins which have one domain belonging to one type of structural protein while the other to the other type.

Human Haemoglobin α -chain (α)

h	3-18	20-35	52-71	80-88	94-112	36-43*	118-141*
e	-						
t	73-76	75-78	72-75*	89-92*	113-116*		

Human Haemoglobin β -chain (α)

h	57-76	85-93	99-117	123-146	35-41*	50-56*	
e	-						
t	42-45	43-46	77-80	78-81	118-121	80-83*	94-97*

Lamprey Haemoglobin (α)

h	12-29	45-52	62-66	67-88	92-106	111-127	132-148	30-44*
e	-							
t	55-58	53-56*						

Spermwhale Met-myoglobin (α)

h	1-19	20-35	36-42	51-57	58-77	86-94	100-118	125-148
e	-							
t	43-46	44-47	46-49	78-81	79-82	95-98	121-124	82-85* 119-122*

Bacterial Ferri Cytochrome C₂ (α)

h	2-10	64-71	75-80	96-110				
e	-							
t	12-15	15-18	21-24	32-35	43-46	49-52	84-87	14-17* 26-29*
	35-38*	39-42*	53-56*	54-57*	56-59*			

contd.....

Bonito Ferro Cytochrome C (α)

h	1-11,	61-69	90-103						
e	-								
t	12-15	14-17	15-18	35-38	43-46	50-53	51-54	53-56	70-73
	71-74	75-78	21-24*	26-29*	27-30	32-35*	73-76*		

Bacterial Cytochrome C₅₅₀ (α)

h	55-56	72-80	106-119	5-16*	81-91*				
e	-								
t	33-36	38-41	92-95	22-25*	23-26*	27-30*	40-43*	41-44	45-48*
	121-124*	122-125*	129-132*	130-133*					

Bacterial Rubredoxin (β)

h	-								
c	1-16	35-46	17-27*	28-34*	47-53*				
T	-								

Jack bean concanavalin A (β)

h	-								
e	24-30	46-48	49-56	60-67	73-79	89-97	104-108	124-131	169-179
	4-10*	35-38*	109-115*	139-145*	147-149*	153-157*	208-215*		*
t	14-17	34-37	80-83	117-120	160-163	183-186	201-204	203-206	226-229
	15-18*	31-34*	81-84*, 82-85*	134-137*	216-219*	222-225*	227-230*	229-232*	
	230-233*								

Bovine trypsin (β)

h	235-245*								
e	29-37	41-46	50-55	63-69	80-93	101-108	132-141	179-184	196-202
	212-217	155-166*	224-229*						
t	23-26	48-51	56-59	70-73	95-98	96-99	115-118	167-170	175-178
	186-189	16-19*	25-28*	72-75*	116-119*	119-122*	192-195*		

Table I contd.....

Porcine tosyl elastase (β)

h	244-254*								
e	29-36	44-52	53-58	68-73	141-147	186-193	214-221	222-226	
	86-89*	108-115*	204-210*	233-240*					
t	22-25	59-62	99-102	136-139	157-160	160-163	195-198	196-199	
	227-230	1-4*	8-11*	10-13*	12-15*	61-64*	116-119*	122-125*	134-137*
	154-157*	161-164*	164-167*	166-169*	169-172*	178-181*	229-232*		

Human Bence Jones Protein REI (Monomer I) (β)

h	-								
e	1-7	16-26	31-40	42-50	69-77	84-92	61-67*	95-105*	
t	55-58	8-11*	79-82*	80-83*					

Human Immunoglobulin I (β)

h	-								
e	1-5	7-12	16-25	31-38	83-91	132-140	147-151	202-208	
	41-46*	53-66*	68-75*	94-108*	116-120*	160-169*	173-182*	193-199*	
t	27-30	126-129	127-130	128-131	154-157	155-158	186-189		
	188-191	77-80*							

Human Immunoglobulin II (β)

h	-								
e	1-13	15-25	33-41	43-53	67-75	76-85	92-101	103-114	115-118 155-163
	178-188	198-205	208-216	54-58*	120-129*	141-150*	166-176*		
t	60-63	61-64	63-66	28-31*	29-32*	87-90*	132-135*	135-138*	136-139*
	189-192*	191-194*	193-196*	194-197*					

Human Prealbumin Monomer II (β)

h 75-83
 e 54-56 115-123 11-19* 29-35* 42-49* 67-75* 89-97* 104-112*
 t 1-4

Chicken lysozyme ($\alpha + \beta$)

h 5-15 80-85 24-34* 88-96*
 e 42-46* 50-54* 59-60*
 t 17-20 19-22 36-39 66-69 74-77 79-82 99-102 103-106
 106-109 113-116 115-118 124-127 69-72 20-23* 82-85*
 106-111* 119-122* 122-125*

Bovine chymotrypsinogen A ($\alpha + \beta$)

h 164-173 230-245
 e 28-35 42-49 50-56 101-110 133-141 212-219 63-69* 81-92*
 155-162* 177-186* 193-201* 223-229*
 t 108-111 115-118 4-7* 11-14* 13-16* 16-19* 23-26* 95-98*
 96-99* 116-119* 125-128* 202-205* 203-206*

Bacterial high potential protein ($\alpha + \beta$)

h 12-16 28-31
 e 69-73 48-51* 57-58* 59-64*
 t 3-6 8-11 20-23 37-40 38-41 40-43 42-45 22-25* 23-26*
 43-46* 53-56* 77-80*

contd...

Carp calcium binding protein ($\alpha + \beta$)

h	8-18	26-33	40-51	61-70	99-107	79-90*
e	56-60*	95-98*				
t	2-5	20-23	1-4*	34-37*	35-38*	71-74*

Bovine Ferricytochrome b_5 ($\alpha + \beta$)

h	33-38	42-50	64-74	8-16*	54-62*	80-87*
e	3-7	19-25	28-31*	51-53*	75-79*	
t	-					

Bovine ribonuclease S complex ($\alpha + \beta$)

h	3-13	24-34	50-59						
e	41-48	71-75	105-111	118-124	80-86*	96-104*			
t	16-19	36-39	65-68	87-90	91-94	92-95	17-20*	66-69*	112-115*

Bacterial nuclease S complex ($\alpha + \beta$)

h	99-107	122-134	54-67*						
e	12-19	21-27	30-36	38-41	70-78	89-95	108-113*		
t	3-6	46-49	47-50	49-52	83-86	137-140	139-142	1-4*	115-118*
									117-120*

Human carbonic anhydrase C ($\alpha + \beta$)

h	34-38	155-162	219-229	16-20*					
e	30-33	39-41	45-53	55-62	65-71	87-96	116-124	206-212	
	215-218	229-231	239-241	76-82*	108-110*	125-126*	140-151*		
	171-176*	191-196*	256-259*						
t	7-10	8-11	11-14	12-15	99-102	128-131	129-132	163-166	167-170
	180-183	198-201	234-237	249-252	21-24*	24-27*	104-107*	130-133*	

Substilisin BPN' (α/β)

h	14-20	64-73	103-117	132-147	223-238	268-275	5-10*	242-252*
e	28-32	120-124	148-152	205-209	45-50*	89-94*	213-217*	
t	23-26	36-39	39-42	51-54	56-59	60-63	97-100	98-101
	157-161	159-162	193-196	219-222	263-266	83-86*	85-88*	
	166-169*	168-171*	171-174*	181-184*	187-190*	259-262*	261-264*	

Bovine carboxypeptidase A complex (α/β)

h	14-28	82-88	112-122	215-231	285-306	72-80*	94-103*	173-187*
								254-262*
e	265-271	32-46*	49-53*	60-66*	104-109*	190-196*	200-204*	239-241*
t	8-11	29-32	56-59	89-92	123-126	142-145	148-151	150-153
	153-156	159-162	169-172	206-209	232-235	273-276	275-278	
	277-280	3-6*	4-7*	67-70*	162-165*	242-245*	244-247*	245-248*
	250-253*	278-281*						

Lobster GPD (α/β)

h	9-22	36-46	101-112	147-162	250-264	316-332	201-216 *	
								279-286*
e	1-7	27-32	90-97	115-120	126-128	142-146	168-179	224-233
	236-246	270-274	289-302	303-312	56-61*	62-67*	185-191*	192-194*
t	47-50	78-81	79-82	83-86	84-87	85-88	138-141	163-166
	218-221	219-222	221-224	265-268	48-51*	76-79*	122-125*	129-132*
	130-133*	133-136*						

contd.....

Table I contd.....-

Chickentriose phosphate isomerase Monomer (α/β)

h 79-87 95-102 105-120 130-137 138-154 177-196 197-204 17-31*
44-55* 213-223*

e 5-11 36-42 89-93 122-129 205-209 60-63* 159-167* 227-231*

t 237-240 12-15* 56-59* 57-60* 74-75 * 169-171* 234-237* 243-246*

Bacterial semiquinone flavodoxin (α/β)

h 10-27 66-74 93-107 124-138

e 1-6 30-35 108-110 115-119 48-55* 80-89* 111-114*

t 62-65 39-42* 42-45* 43-46* 56-59* 57-60* 77-80*

Bacterial thermolysin (others)

h 137-150 235-246 260-274 281-296 301-312 67-87* 160-179*

e 27-30 37-46 52-54 60-63 97-106 112-116 3-13* 15-25* 31-36*
55-58* 119-123*

t 107-110 126-129 130-133 132-135 151-154 181-184 209-212
210-213 216-219 217-220 249-252 250-253 276-279 297-300 92-95*
128-131* 133-136* 187-190* 189-192* 190-193* 192-195* 194-197* 197-200*
204-207* 205-208* 207-210* 300-302* 311-314*

Dogfish apolactate dehydrogenase Complex (bthers)

h 2-6 55-70 120-130 165-181 249-263 308-329 33-44* 107-109*
227-236* 237-245*

e 23-28 77-81 92-97 159-161 188-192 281-295 48-53* 134-139* 200-207*
267-279* 298-303*

t 88-91. 140-143 184-187 208-211 213-216 19-22* 84-87*
 86-89* 102-105* 103-106* 142-145* 152-155* 196-199* 209-212*
 216-219* 221-224*

Bacterial ferredoxin (others)

h -

e 1-7 8-19 35-46 20-26* 27-34* 47-54*

t -

Horse alcohol dehydrogenase Complex (others)

h 47-54 229-236 250-259 275-283 353-365 170-187* 202-212* 305-310*
 324-336*

e 34-40 62-65 72-78 128-132 145-146 148-152 193-199 218-224
 287-293 312-318 369-374 9-14* 22-29* 41-45* 68-71* 86-92* 135-138*
 156-160* 138-243* 262-269* 347-352*

t 1-4 102-105 105-108 116-119 123-126 244-247 271-274 295-298 296-299
 319-322 342-345 3-6* 55-58* 80-83* 100-103* 115-118* 139-142* 140-143*
 161-164* 165-168* 338-341*

Papain (others)

h 24-43 49-57 67-78 117-128 137-143*

e 5-7 130-131 169-175 179-183 185-191 206-108 111-112* 162-167*

t 19-22 58-61 61-64 62-65 82-85 96-99 195-198 201-204

8-11* 84-87* 98-101* 198-201* 199-202*

are indifferent to that structure, or (iii) residues having high or low potential for the observed secondary structural state are almost equal in number.

Our analysis on this type II secondary structures has revealed that, amino acids which have less potential for observed secondary structural state are more in most of the type II structures. This analysis suggests that most of type II structures are formed not because of clustering of amino acids which prefer that secondary structures, followed by other types of residues, or because most of the amino acids are indifferent to that structure, but because of tertiary interactions which make these segments assume the secondary structure not necessarily favoured by constituent amino acid residues.

On the other hand, this analysis shows that type I secondary structures are comprised mostly of residues which have high potential for the observed secondary structural state. This can be expected since the calculated average potential is maximum for the observed secondary structural state.

(b) Occurrence in N-, central and C-terminal parts of protein:

Further, it was our interest to see whether type I or type II structures are localised in one particular region of the protein. Therefore, amino acid sequence of each protein is divided into N-, Central C-terminal parts. The frequency of occurrence of type I and type II structures have indicated that type I and type II structures are distributed throughout the primary structure of the protein and are not localised in any particular part of the protein. The results of this analysis are presented in Table II.

TABLE II

Distribution of type I and type II secondary structures

	<u>Helix</u>		<u>Extended structure</u>		<u>Chain reversals</u>	
	Type I	Type II	Type I	Type II	Type I	Type II
<u>Along protein primary structure</u>						
N-terminal	30	14	62	27	58	64
Central part	37	11	40	42	85	74
C-terminal	38	16	45	34	61	51
<u>Lengthwise</u>						
Long	71	24	20	20	-	-
Medium	34	16	115	84	-	-
Short	-	1	12	3	-	-
Average length (\bar{L}):	11.73		7.03			
Standard deviation (σ) :	4.4		2.79			

(c) Length: Type I and Type II helicies and extended structures are further divided as long, medium and short structures. To do such a classification, first the average length (\bar{L}) (i.e., the average number of residues in the stretch) and standard deviation (σ) are calculated. If the number of residues in the stretch is greater than $(\bar{L} + \sigma)$, the stretch is termed 'long'; if the number of residues is less than $(\bar{L} - \sigma)$, the stretch is 'short'; if the number of residues are in between $(\bar{L} + \sigma)$ and $(\bar{L} - \sigma)$, the stretch is of 'medium' length. In Table II, \bar{L} and σ , are given for helix and extended structure. Observation of Table II indicates that roughly 75 per cent of long helicies are of type I, while only 50 per cent of long extended structures are of type I. The percentage of type II extended structures of medium length is 42. Thus like helical structures, extended structures do not have much different percentages for type I and type II structures.

Summary of analysis of Table I is given in Table III. The 31 proteins are grouped into various structural classes to which they belong. Table III suggests that type I helicies are more in number for all structural types. Type II structures are maximum for chain reversals and minimum for helicies.

As an after thought, the influence of four residues before the N- and four residues after the C-terminals of the observed secondary structural stretches on these structures was studied. The average potential values calculated after this addition indicate that the assignment of one secondary structure over the other for a given stretch becomes less easy and thus in very few cases one finds the calculated potential to be maximum for the

TABLE III

Total number of each type of observed secondary structures in different structural types of globular proteins is given along with the division into type I and type II. The 'others' category also includes proteins, which have one domain belonging to one type of structural protein while the other to the other type.

Structural type of protein	Helix		Extended structure			Chain reversals			
	Observed	Type I	Type II	Observed	Type I	Type II	Observed	Type I	Type II
α	42	36	6	-	-	-	64	35	29
β	3	-	3	96	59	37	84	40	44
$\alpha + \beta$	33	22	11	57	30	27	90	51	39
α/β	38	27	11	44	26	18	80	43	37
Others	30	20	10	57	32	25	75	35	40
TOTAL	146	105	41	254	147	107	393	204	189

observed secondary structure. This study indicates that addition of few residues at either N- or C-terminal or both does not improve the results.

V.4 CONCLUSIONS

Analysis presented so far indicates the existence of two types of secondary structures and these have been identified in all the 31 different proteins considered. Since similar results were obtained when either single residue potentials or pair potentials were used, this analysis is not an artifact due to the use of pair potentials.

The first type, i.e., type I secondary structures are those stretches of polypeptide chain which have an in-built tendency to assume the observed secondary structure. Thus type I structures are good candidates to act as nucleation sites during protein folding process.

The type II secondary structures are formed as a result of interactions among amino acid residues that are close not only at primary structure level but also at tertiary structure level. This analysis further showed that type II secondary structures mostly consist of such amino acids which have a low probability to exist in the observed secondary structure.

Secondary structure prediction schemes thus should take into consideration tertiary interactions also, since many of the secondary structures are formed due to tertiary interactions. But presently available crystal structure data seems to be inadequate to derive statistically significant

weightages for tertiary interactions; as these interactions vary with structural class of the protein. Hence direct prediction of tertiary structure without being concerned about the prediction of secondary structures seems to be a better alternative at present. Such an attempt is made in Chapter VI.

C H A P T E R VI

PREDICTION OF ROUGH STRUCTURE OF PROTEIN USING RAMACHANDRAN PLOT

VI.1 INTRODUCTION

It has been discussed in Chapter V that knowledge of tertiary interactions is a prerequisite for accurate prediction of secondary structures. Also it is hinted that direct prediction of tertiary structure avoiding the step of assigning the secondary structures is a better alternative. This we have thought mainly because tertiary interactions can not only force certain stretches of polypeptide chain to take regular structural conformations but can also disrupt some of the already formed regular structures or convert one secondary structure into the other. In other words, we are suggesting that secondary structures formed in the initial stage of protein folding process may not remain in the same state. This also points out that if a helix is formed at an earlier stage of folding, there is a possibility and there are pathways on the energy surface to convert this helix to extended structures and vice-versa. Semi-empirical conformational calculations carried out on dipeptide varying the peptide geometry clearly indicate above possibility (Kolaskar, unpublished); also Louie and Somorjai (1982) have provided a basis for such conversions based on differential geometry.

Conformational energy calculations (Levitt and Warshel, 1975; Nemethy and Scheraga, 1977) and geometrical models (Goel et al., 1982 and references cited therein) are two important attempts for the direct assignment of protein tertiary structure. Success achieved with conformational energy calculations is modest because of complex and

time-consuming computations. Geometrical models too met with only reasonable success (Goel et al., 1982) and also require computers with huge memory for execution. Both methods use external coordinate system and thus one has to carry out the conversion from internal to external coordinate system at every stage.

The prediction of tertiary structure can be less complex if the prediction of only main chain dihedral angles, ϕ and ψ of each residue is undertaken. Thus the parameters to be determined per residue will be reduced to only two per residue. Since ϕ and ψ are internal parameters, no conversion is necessary.

Because the present study is attempted on a micro-computer prediction of (ϕ, ψ) -values at a small grid interval is not undertaken. The allowed (ϕ, ψ) -conformational space of Ramachandran map is divided into three parts: (i) region A (ϕ varies between -140° and 0° , ψ between -100° and 0°); (ii) region B (ϕ varies between -180° and 0° and ψ between 80° and 180°), and (iii) region C, the remaining allowed region of (ϕ, ψ) -plane. It can be noticed by observing (ϕ, ψ) -values of region A, that main chain atoms will be quite close when residues take conformations in this region. Conformations taken by residues in regular structures such as α -helix or bends, are part of this region. Thus region A represents conformations which are densely packed. Region B, on the other hand represents mostly extended conformations. The distances between C_i^α to C_{i+3}^α will be considerably larger than 7 Å in this region as against 5 Å to 7 Å in region A.

In this chapter, an algorithm is developed, which assigns each amino acid, conformation in one of the three parts of (ϕ, ψ) -plane, region A, region B or region C. Thus prediction of rough three dimensional structure based on short range interactions only was attempted. The method and discussion of this study are presented in succeeding sections.

VI.2 METHOD

VI.2.1 Single residue potential calculations

The single residue potentials in the three regions, region A, region B and region C are determined for each residue using a formula similar to eq. 2 used in Chapter IV. Data from 31 proteins listed in Chapter IV are used to derive the single residue potentials and these values are given in Table I, for regions A, B and C, for each of the 20 proteinous amino acid residues. From Table I, it can be seen that the potential values for most of the residues in the two regions, region A and region B, are not very different. Hence, these potential values will not be of much use in developing algorithms to predict the tertiary structure.

VI.2.2. Calculation of amino acid pair potentials

In order to see whether the pairs of amino acid residues provide potentials which are distinguishable in regions A, B and C, using a similar formula as eq. 1 of Chapter IV, the potential values of 400 amino acid pairs, in the three states, were determined. These values are presented in Table II. The amino acid pair potentials given in Table II differ appreciably in three states for a given pair. The pair potentials

TABLE I

Single residue potentials obtained from proteins considered

Amino acid residue	Region A	Region B	Region C
Ala	1.37	0.84	0.71
Cys	0.96	1.20	0.72
Asp	1.67	0.72	1.21
Glu	1.45	0.72	0.80
Phe	1.02	1.13	0.76
Gly	0.48	0.36	2.86
His	1.03	0.96	1.03
Ile	0.95	1.33	0.54
Lys	1.15	0.84	1.00
Leu	1.11	1.18	0.52
Met	1.16	1.18	0.48
Asn	0.78	0.78	1.71
Pro	1.08	1.21	0.54
Gln	1.12	1.09	0.66
Arg	1.11	1.00	0.82
Ser	0.94	1.04	1.04
Thr	0.92	1.20	0.78
Val	0.88	1.46	0.42
Trp	0.93	1.11	0.92
Tyr	0.68	1.39	0.83

Three letter amino acid code has been used.

TABLE II

Potential value for each pair of amino acid residues in the three regions of (ϕ, ψ) -plane in the following order: (i) region-A (a); (ii) region B (b) and (iii) region C (c). These pairs can be formed by considering the residue, first from vertical column and then from horizontal row. Single letter amino acid code has been used.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0.46	0.26	0.40	0.53	0.46	0.13	0.57	0.56	0.40	0.61	0.74	0.21	0.19	0.46	0.38	0.34	0.34	0.48	0.29	0.22
C	0.25	0.46	0.06	0.14	0.08	0.00	0.12	0.24	0.14	0.24	0.00	0.12	0.49	0.15	0.22	0.28	0.53	0.39	0.22	0.26
D	0.17	0.27	0.30	0.16	0.27	0.48	0.00	0.21	0.26	0.16	0.26	0.36	0.32	0.14	0.40	0.31	0.13	0.13	0.20	0.12
E	0.34	0.00	0.28	0.46	0.00	0.08	0.00	0.38	0.58	0.23	1.00	0.22	0.00	0.00	0.00	0.19	0.25	0.00	1.00	0.00
F	0.56	0.00	0.33	0.54	0.00	0.09	0.00	0.22	0.19	0.62	0.00	0.13	0.00	0.45	0.29	0.22	0.58	0.69	0.00	0.83
G	0.10	1.00	0.39	0.00	0.00	0.24	1.00	0.40	0.23	0.16	0.00	0.31	1.00	0.27	0.34	0.32	0.17	0.31	0.00	0.17
H	0.28	0.38	0.31	0.46	0.40	0.05	0.42	0.37	0.32	0.29	0.00	0.10	0.00	0.00	0.24	0.21	0.14	0.27	0.20	0.06
I	0.12	0.22	0.06	0.00	0.07	0.00	0.00	0.19	0.16	0.21	0.63	0.00	0.12	0.39	0.14	0.06	0.33	0.09	0.00	0.37
K	0.37	0.40	0.32	0.36	0.36	0.24	0.58	0.31	0.25	0.28	0.37	0.31	0.42	0.11	0.25	0.27	0.53	0.42	0.21	0.18
L	0.65	0.00	0.43	0.32	0.56	0.19	0.27	0.37	0.55	0.56	0.59	0.48	0.07	0.43	0.85	0.15	0.44	0.32	0.44	0.66
M	0.16	0.00	0.06	0.14	0.28	0.00	0.00	0.22	0.14	0.15	0.00	0.00	0.34	0.17	0.00	0.09	0.26	0.42	0.26	0.00
N	0.19	0.48	0.19	0.19	0.17	0.46	0.73	0.13	0.19	0.18	0.41	0.28	0.15	0.20	0.15	0.53	0.30	0.14	0.30	0.24
O	0.60	0.00	0.13	0.38	0.18	0.18	0.00	0.25	0.45	0.37	0.59	0.15	0.00	0.74	0.38	0.21	0.28	0.24	0.00	0.43
P	0.10	0.36	0.23	0.30	0.21	0.15	0.35	0.39	0.19	0.28	0.00	0.53	0.33	0.00	0.44	0.56	0.49	0.37	0.00	0.17
Q	0.30	0.64	0.23	0.13	0.06	0.27	0.21	0.11	0.19	0.17	0.41	0.32	0.39	0.26	0.18	0.24	0.10	0.16	1.00	0.40
R	0.09	0.10	0.00	0.36	0.28	0.03	0.07	0.13	0.15	0.16	0.17	0.11	0.09	0.03	0.08	0.03	0.11	0.17	0.15	0.15
S	0.06	0.18	0.05	0.07	0.00	0.07	0.00	0.15	0.00	0.11	0.20	0.04	0.36	0.05	0.10	0.07	0.09	0.12	0.09	0.00
T	0.25	0.25	0.35	0.31	0.72	0.15	0.26	0.34	0.38	0.44	0.12	0.22	0.28	0.26	0.32	0.27	0.25	0.30	0.10	0.38
V	0.57	0.42	0.00	0.23	0.14	0.47	0.00	0.00	0.16	0.13	0.35	0.00	0.10	0.00	1.00	0.36	0.34	0.68	0.00	0.14
W	0.17	0.12	0.00	0.00	0.17	0.00	0.00	0.72	0.00	0.15	0.41	0.00	0.11	0.00	0.00	0.14	0.27	0.11	0.58	0.15
Y	0.05	0.14	1.00	0.32	0.69	0.33	0.10	0.28	0.38	0.35	0.24	0.41	0.79	0.00	0.00	0.50	0.39	0.20	0.64	0.20
Z	0.49	0.45	0.21	0.75	0.28	0.06	0.29	0.45	0.17	0.38	0.72	0.33	0.12	0.09	0.41	0.30	0.07	0.13	0.00	0.18
aa	0.26	0.35	0.35	0.19	0.53	0.22	0.51	0.32	0.24	0.51	0.28	0.33	0.49	0.73	0.31	0.55	0.58	0.73	0.24	0.57
bb	0.13	0.21	0.18	0.06	0.19	0.33	0.20	0.09	0.19	0.11	0.00	0.20	0.21	0.18	0.28	0.15	0.35	0.14	0.14	0.10
cc	0.47	0.38	0.33	0.40	0.50	0.20	0.42	0.23	0.28	0.24	0.12	0.10	0.29	0.81	0.19	0.22	0.35	0.28	0.30	0.06
dd	0.16	0.15	0.07	0.19	0.14	0.06	0.08	0.36	0.12	0.37	0.29	0.29	0.25	0.00	0.43	0.15	0.32	0.32	0.70	0.27
ee	0.25	0.09	0.25	0.17	0.17	0.22	0.29	0.19	0.23	0.22	0.21	0.45	0.45	0.19	0.38	0.17	0.22	0.05	0.00	0.28
ff	0.50	0.42	0.43	0.45	0.28	0.18	0.41	0.28	0.36	0.38	0.39	0.13	0.00	0.28	0.51	0.27	0.23	0.39	0.35	0.19
gg	0.27	0.00	0.20	0.32	0.33	0.18	0.34	0.48	0.33	0.26	0.61	0.29	0.62	0.41	0.27	0.32	0.41	0.42	0.41	0.11
hh	0.24	0.58	0.24	0.09	0.39	0.24	0.24	0.24	0.17	0.17	0.00	0.39	0.23	0.19	0.27	0.22	0.27	0.19	0.24	0.40

(TABLE II CONTD..)

A	I	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	P	V	W	Y
0.68	0.00	0.27	0.44	0.36	0.41	1.00	0.40	0.40	0.46	0.00	1.00	0.31	0.06	0.35	1.00	0.44	0.10	0.28	0.00	0.00
0.20	0.00	0.15	0.26	0.64	0.00	0.00	0.46	0.46	0.09	0.77	0.00	0.00	0.00	0.41	0.00	0.26	0.37	0.53	1.00	1.00
0.12	0.00	0.18	0.30	0.00	0.19	0.00	0.14	0.14	0.21	0.23	0.00	0.22	1.00	0.24	0.00	0.30	0.22	0.19	0.00	0.00
0.19	0.61	0.07	0.19	0.24	0.07	0.28	0.23	0.23	0.20	0.29	0.00	0.04	0.10	0.21	0.12	0.13	0.16	0.38	0.21	0.00
0.11	0.18	0.00	0.00	0.00	0.33	0.18	0.18	0.12	0.21	0.46	0.00	0.12	0.08	0.14	0.00	0.10	0.22	0.12	0.00	0.00
0.40	0.21	0.29	0.26	0.21	0.08	0.39	0.37	0.24	0.32	0.54	0.16	0.78	0.49	0.40	0.67	0.37	0.40	0.36	1.00	0.00
0.35	0.19	0.31	0.43	0.14	0.13	0.72	0.28	0.26	0.20	0.25	0.37	0.00	0.40	0.40	0.00	0.17	0.14	0.14	0.59	0.11
0.41	0.67	0.12	0.12	0.16	0.20	0.28	0.53	0.31	0.29	0.58	0.26	0.41	0.00	0.19	0.32	0.48	0.28	0.00	0.51	0.00
0.24	0.13	0.42	0.16	0.29	0.54	0.00	0.19	0.23	0.35	0.17	0.15	0.24	0.00	0.33	0.31	0.38	0.30	0.41	0.38	0.00
0.31	0.19	0.56	0.76	0.19	0.08	0.22	0.24	0.56	0.49	0.46	0.27	0.00	0.26	0.30	0.13	0.27	0.40	0.00	0.12	0.00
0.20	0.23	0.24	0.15	0.43	0.00	0.13	0.42	0.24	0.28	0.54	0.10	0.93	0.61	0.18	0.69	0.21	0.46	0.00	0.55	0.00
0.17	0.00	0.19	0.09	0.38	0.41	0.31	0.33	0.19	0.23	0.00	0.37	0.07	0.12	0.52	0.18	0.25	0.14	1.00	0.33	0.00
0.42	0.74	0.52	0.68	0.49	0.04	0.33	0.11	0.24	0.29	1.00	0.32	0.00	0.33	0.40	0.29	0.15	0.13	1.00	0.44	0.00
0.49	0.00	0.30	0.20	0.00	0.08	0.00	0.65	0.00	0.51	0.00	0.13	1.00	0.38	0.00	0.14	0.53	0.60	0.00	0.26	0.00
0.10	0.26	0.18	0.12	0.51	0.12	0.67	0.23	0.28	0.10	0.00	0.22	0.00	0.29	0.00	0.20	0.32	0.27	0.00	0.30	0.00
0.29	0.37	0.39	0.33	0.21	0.08	0.15	0.20	0.22	0.16	0.00	0.15	0.10	0.16	0.52	0.14	0.13	0.21	0.22	0.19	0.00
0.19	0.14	0.06	0.00	0.49	0.05	0.00	0.41	0.26	0.55	0.00	0.04	0.12	0.00	0.15	0.18	0.26	0.53	0.17	0.29	0.00
0.27	0.30	0.38	0.31	0.29	0.25	0.20	0.24	0.41	0.22	1.00	0.34	0.78	0.34	0.13	0.45	0.38	0.13	0.40	0.24	0.00
0.34	0.00	0.25	0.22	0.08	0.05	0.56	0.26	0.29	0.21	0.19	0.42	0.06	0.26	0.00	0.13	0.17	0.14	0.10	0.06	0.00
0.24	0.57	0.34	0.15	0.40	0.06	0.13	0.61	0.21	0.62	0.67	0.10	0.47	0.31	0.46	0.21	0.44	0.53	0.55	0.34	0.00
0.33	0.43	0.29	0.24	0.26	0.35	0.31	0.13	0.13	0.28	0.17	0.13	0.23	0.48	0.23	0.54	0.21	0.26	0.16	0.07	0.24
0.49	0.21	0.30	0.21	0.18	0.11	0.16	0.18	0.32	0.37	0.16	0.23	0.16	0.54	0.22	0.24	0.17	0.20	0.46	0.16	0.00
0.32	0.49	0.31	0.45	0.70	0.19	0.43	0.55	0.38	0.36	0.37	0.18	0.46	0.32	0.59	0.60	0.55	0.66	0.54	0.67	0.00
0.14	0.29	0.31	0.21	0.12	0.45	0.25	0.05	0.12	0.14	0.00	0.13	0.38	0.14	0.20	0.16	0.12	0.08	0.00	0.17	0.00
0.52	0.00	0.36	0.00	1.00	0.00	0.59	0.52	0.28	0.52	1.00	0.00	0.00	0.59	0.16	0.22	0.00	0.18	1.00	0.00	0.00
0.30	0.63	0.64	0.21	0.00	0.07	0.00	0.30	0.33	0.30	0.00	0.63	0.00	0.00	0.00	0.13	0.48	0.63	0.00	0.44	0.00
0.18	0.37	0.00	0.25	0.00	0.40	0.41	0.18	0.39	0.18	0.00	0.37	1.00	0.41	0.34	0.31	0.21	0.19	0.00	0.00	0.00
0.20	0.13	0.20	0.20	0.00	0.06	0.21	0.08	0.07	0.56	1.00	0.14	0.00	0.09	0.15	0.22	0.05	0.07	0.19	0.27	0.00
0.62	0.82	0.40	0.24	0.24	0.08	0.49	0.45	0.51	0.28	0.00	0.66	0.53	0.43	0.34	0.31	0.43	0.48	0.43	0.73	0.00
0.13	0.00	0.10	0.56	0.14	0.21	0.29	0.00	0.20	0.17	0.00	0.20	0.47	0.19	0.51	0.21	0.25	0.24	0.38	0.00	0.00

are then utilised to develop algorithms to assign conformational state to each of the amino acids, from the amino acid sequence of the protein.

A total of 6015 pairs could be formed from the 31 proteins considered, out of which 1653 pairs lie in region A, 1420 pairs in region B and 2942 pairs in region C. We are aware of the statistical fluctuations associated with these pair potentials. Still these potentials are considered for the present study, because their derivation and use are novel in protein folding studies.

VI.2.3 Prediction of rough three dimensional structure

(a) Consideration of pairs of amino acids: If ACDEF.... is the amino acid sequence of a protein, pairs AC, CD, DE were formed and each pair was assigned the conformational state, for which the pair potential is maximum. This assignment when compared with the observed conformational states, coincided with half of them.

(b) Consideration of quartet - unit weightage: The primary structure of the protein is considered by taking 4 residues at a time along the sequence. For example, say, residues A C D F constitute one such quartet. Pairs AC, CD and EF are formed and corresponding potential values are separately added for the three states by giving all the pairs unit weightage in all the three states.

The average value of the potential values is determined separately for the three states. The central pair CD is assigned the conformational

state, for which the average value is maximum. This algorithm could assign 50-55 per cent of the conformational states correctly.

(c) Consideration of quartet-different weightages: If A_i , C_{i+1} , D_{i+2} , E_{i+3} are four residues in positions 1, 2, 3 and 4 of the quartet, pairs (A_i, C_{i+1}) , (A_i, D_{i+2}) , (A_i, E_{i+3}) , (C_{i+1}, D_{i+2}) , (C_{i+1}, E_{i+3}) and (D_{i+2}, E_{i+3}) were formed. These pairs were given different weightages in different states as given in Table III.

The computed potential for the quartet for each of the three states was calculated with different weightages mentioned in Table III and the central pair (C_{i+1}, D_{i+2}) was assigned that conformational state for which the computed potential is maximum. These studies were carried out on four proteins considering one each from the four structural classes α , β , $\alpha+\beta$ and α/β . This assignment too when compared with observed states, could give correct assignment in the range of 50-55 per cent.

In this assignment, the short-range interactions are considered with due weightage for potentials in respective conformational states. Thus though the weightages were changed suitably, the prediction accuracy has not improved.

VI. 3 DISCUSSION

The above studies indicate that, consideration of short-range interactions could assign 50 per cent of the conformational states correctly. The prediction accuracy is in the same order as for most secondary structure

TABLE III

Details of weightages given to pairs in algorithm (c) (see text)

Region of (\emptyset, γ)-space	(A_i, C_{i+1})	(A_i, D_{i+2})	(A_i, E_{i+3})	(C_{i+1}, D_{i+2})	(C_{i+1}, E_{i+3})	(D_{i+2}, E_{i+3})
<u>Set I</u>						
Region A	Unit	0.5	0.5	Unit	0.5	Unit
Region B	Unit	0.5	0.25	Unit	0.5	Unit
Region C	Unit	0.5	0.5	Unit	0.5	Unit
<u>Set II</u>						
Region A	Unit	0.5	0.5	Unit	0.5	Unit
Region B	Unit	0.5	0	Unit	0.5	Unit
Region C	Unit	0.25	0.25	Unit	0.25	Unit
<u>Set III</u>						
Region A	Unit	0.5	0.75	0.75	0.5	0.5
Region B	Unit	0.5	0	0.5	0.5	0.25
Region C	Unit	0.25	0.5	0.5	0.25	0.5

prediction schemes, which is 55-60 per cent. There is a major difference in these two accuracy percentages. Our assignment is at tertiary structure level while secondary structure prediction schemes predict only secondary structural regions, and thus there remains some regions in protein sequences for which no structure could be assigned. These regions could be predicted only when tertiary interactions are properly incorporated in prediction schemes.

We are aware that our present assignment is inaccurate. But the point to be noted here is that, the analysis presented here is the beginning of another novel way of predicting three dimensional structure of proteins from amino acid sequence. This algorithm can further be improved by combining the information derived in Chapters II and III, about conformationally similar residues and side chain characteristic main chain conformations, respectively. This analysis also suggests that long range interactions are not just additions or linear combinations of short-range interactions. The weightages for tertiary (long-range) interactions have to be developed for proteins depending on the structural class to which the protein belongs and the functional properties of the protein. The salient features of this analysis are:

- (a) Method is simple and novel; though inaccurate at present.
- (b) It gives initial rough range of conformations for residues in a protein molecule, which can be further refined, using other refinement methods and then (ϕ, ψ)-values of each residue can be known.
- (c) A possibility exists in this algorithm to suitably incorporate long and medium range interactions to improve the prediction accuracy of the method.

A P P E N D I X I

NATURE OF AMINO ACID SUBSTITUTIONS IN HOMOLOGOUS PROTEINS
DURING EVOLUTION

INTRODUCTION

Diversity in protein functions is the result of protein evolution. During evolution, mutations have occurred which have either conserved or changed the traits of a whole protein or some parts of the protein. The mutations might also have replaced or retained certain portions of protein structure or just individual amino acids in certain places. This would have affected the chemical or physical properties of the proteins. The comparison of chain folds or sheet topologies of proteins cannot give information about the nature of amino acid substitutions in particular places.

To have some idea of the effect of mutations in proteins - whose function has been conserved - from different species, establishment of individual nature of both the replaced and replacing amino acids is necessary. Each residue position is a site for observation, where conservation or change of different characteristics of amino acid residues present might have occurred. Proteins that have same function in all compared organisms and whose amino acid sequences from many species are available, are suitable for this type of study. These proteins are known homologous proteins.

Phylogenetic trees, which give some information about evolutionary pathway, can be constructed by considering homologous proteins. But recently Felsenstein (1982) critically reviewed the art of phylogenetic tree construction. He observes that currently used numerical methods are influenced by defective criterion and 'unscientific' defence is put forward for both the methods of phylogenetic tree construction and results obtained based on them.

In this study, the two homologous proteins haemoglobin and mitochondrial cytochrome C have been considered and studied to understand the nature of replacement substitutions that have taken place in these proteins during evolution by analysing the chemical, physical and conformational properties (such as conformational similarity) of the amino acid residues. It should be noted that earlier workers have studied only one or the other property of the amino acids and have not looked at in particular the conformational properties though it is known now for sometime that several proteins having different primary structure have essentially similar three dimensional structure. Thus the information regarding the function and utility of the protein is stored not only in amino acid residues but also in over all three dimensional structure of the protein. In other words, conformational properties of amino acids can be one of the traits which one can use even in phylogenetic tree construction. With this in mind we have attempted by using simple method to study various above-mentioned properties which is discussed below.

METHOD

The amino acid sequences of mitochondrial cytochrome C, α and β -chains of haemoglobin for the species considered were taken from the Handbook of Biochemistry and Molecular Biology (1976). The replacement substitutions in these sequences were compared as follows:

For cytochrome C, the reference species ranged from Rhesus monkey to Pacific lamprey, and for haemoglobin from Rhesus monkey to Kangaroo. Replacement in the sequences of other species were studied with respect

to the sequence of the reference species. The amino acid residue in the sequence of the reference species which is replaced is termed as replaced residue, and the residue that occupies same position in the sequence of other species as replacing residue. Each replacement substitution was studied by separately taking into consideration the following properties of amino acid residues.

(a) Secondary structure preference: Levitt's (1978) classification of secondary structural preference of amino acid residues was used. This classification has no overlap and clearly states whether a particular residue favours, is indifferent to or breaks a particular secondary structure. A replacing residue that prefers or is indifferent to the secondary structure preferred by the replaced residue is considered to have similar secondary structural preference. The similar residues are given in Table I.

(b) Conformational similarity: In Chapter II, we have discussed the method used to arrive at conformationally similar residues to the various amino residues (see Table III, Chapter II). This table gives a set of amino acid residues having a high probability to adopt similar main chain conformations, for each amino acid residue. Thus if the replacing residue is conformationally similar to the replaced residue, a minimal change in tertiary¹ structure is expected. Such replacements which have taken place during evolution of proteins will conserve the tertiary structure of the protein.

(c) Composition of condons: Residues that have two bases in common, irrespective of their position in respective condons, were considered to be similar, in this study (see Table II).

TABLE I

Conformational preferences of amino acids for α -helix,
 β -sheet and reverse turns (Levitt, 1978)

Type of secondary structure	Favouring			Indifferent			Breaking		
α -helix	Ala His Lys	Leu Glu Cys	Met Gln	Val Trp Arg	Ile Asp	Phe Asn	Tyr Ser	Thr Pro	Gly
β -sheet	Val, Trp	Ile Tyr	Phe Thr	Ala His Arg	Leu Gly	Met Ser	Glu Asp Cys	Gln Asn	Lys Pro
Reverse turn	Gly Asn	Ser Pro	Asp	Glu Tyr	Gln Thr	Lys Arg	Ala His Phe	Leu Val Trp	Met Ile Cys

Three letter amino acid code has been used.

TABLE II

Amino acid residue	Residues having similar composition of codons											
Ala	Ser	Glu	Asp	Val	Thr	Pro	Gly					
Cys	Ser	Phe	Val	Arg	Gly	Trp						
Asp	Ala	Glu	Val	Gly	Val	His	Asn					
Glu	Ala	Asp	Lys	Gly	Val	Gln						
Phe	Ile	Leu	Val	Ser	Tyr	Cys						
Gly	Ser	Ala	Asp	Val	Glu	Arg						
His	Arg	Asp	Gln	Tyr	Asn	Leu	Pro					
Ile	Val	Leu	Thr	Phe	Lys	Met	Asn	Ser				
Lys	Ile	Thr	Arg	Glu	Gln	Asn	His					
Leu	Ile	Gln	Val	Phe	Pro	His	Arg					
Met	Val	Leu	Ile									
Asn	Asp	Ser	Ile	Thr	His	Tyr	Lys					
Pro	Ala	Ser	Thr	Leu	His	Arg						
Gln	Leu	Pro	Arg	Glu								
Arg	Lys	Thr	Ile	His	Gln	Leu	Ser	Met	Pro	Trp	Cys	
Ser	Ala	Thr	Leu	Phe	Tyr	Cys	Pro					
Thr	Ser	Ile	Asn	Lys	Arg	Ala						
Val	Ile	Leu	Phe	Ala	Asp	Gly	Met					
Trp	Ser	Leu	Arg	Gly								
Tyr	Phe	Ser	Cys	His	Asn	Asp						

Three letter amino acid code has been used.

(d) Chemical nature: Classification of the chemical nature of residues was adopted from Stryer (1975) (see Table III).

(e) Polarity: The 20 residues are divided as hydrophobic, neutral polar or polar, as described by Lehninger (1975) (see Table III).

A simple algorithm was developed to study these features of replacement substitutions in mitochondrial cytochrome C and the α and β -chains of haemoglobin in different species. For example, the value 67 in the third column of the first row of Table IV (a) signifies that when the sequence of human cytochrome C is compared with that of horse cytochrome C (horse is reference species), 8 out of 12 replacements (67 per cent) are such that the replacing residues have similar secondary structure affinity as the replaced residues. The percentages of observed replacement substitutions for various properties of amino acid residues, considered above, are given in Tables IV (a), IV (b), IV (c), IV (d) and IV (e) for cytochrome C. Similar results are given in Tables V (a), V (b), V (c), V (d) and V (e) for α and β -chains of haemoglobin.

RESULTS AND DISCUSSION

The results given in Tables IV (a) and V (a) show that the similarity in secondary structural affinity of the replacing and replaced residues accounts for between 20-100 per cent replacement substitutions; although for most cases the percentage is around 65-75.

Examination of replacement substitutions by conformationally similar residues provided significant information. Similarity of the residues

TABLE III

Residues having (a) same chemical nature (Stryer, 1975) and
(b) same polarity (Lehninger, 1975)

Amino acids	Classification
<u>(a) Chemical nature</u>	
Gly Ala Val Leu Ile	Aliphatic
Ser Thr	Hydroxy
Asp Glu	Acidic
Gln Asn	Acid amides
Lys His Arg	Basic
Phe Tyr Trp	Aromatic
Cys Met	Sulfur containing
Pro	Imino acid
<u>(b) Polarity</u>	
Ala Leu Ile Val Pro Phe Trp Met	Hydrophobic
Ser Thr Tyr Asn Gln Cys Gly	Neutral polar
Asp Glu Lys Arg His	Polar

Three letter amino acid code has been used.

at tertiary structure level explains number of replacements, but not all. In order to understand the reasons for the low percentages in Tables IV (b) and V (b), we have considered the conformation of amino acids as observed from the crystal structures of Tunacytochrome C and human haemoglobin. Analysis of crystal structure data indicates that almost all the replacing residues can adopt a similar main chain conformation as that of the replaced residue. This conclusion is derived from our observation that when their (ϕ, ψ) -maps are compared the replacing and replaced residues have very low discrepancy values between the grids that represent these observed main chain conformations. Thus, the replacing amino acid residues seem to be those which have a minimal effect on the three dimensional structure of the protein.

When the composition of condons of replacing and replaced residues was considered, it could account for 70 per cent of the replacements in most cases. Similarity in chemical nature or polarity accounted for less than 50 per cent replacements (see Tables IV (c), IV (d), IV (e), and V (c), V (d) and V (e), for cytochrome C and haemoglobin respectively).

Since no single feature of the amino acids could account for all the replacements, the data was examined to see whether any of these properties is conserved in a position affected by replacements during evolution. The results of the studies are not affirmative of this for both cytochrome C and haemoglobin.

During evolution, therefore, the emphasis seems to be on conservation of three dimensional structure of homologous proteins rather than conservation of local structure, chemical nature or polarity. The diversity in local properties may reflect the species - characteristic

RAY/SHAWNS SPOLIS

[illegible]

Species studied	REFERENCE SPECIES																											
	Man	Rhesus Monkey	Horse	Donkey	Cow	Camel	Elephant Seal	Dog	Bat	Rabbit	Kangaroo	Chicken	Man	King Penguin	Pekin Duck	Pigeon	Snapping Turtle	Rattle Snake	Bull Frog	Tuna	Bull Frog	Battle Snake	Bull Frog	Tuna	Bonito	Carp	Dog Fish	
Man	100	75	64	82	80	80	92	91	82	78	90	69	69	69	55	75	80	64	67	60	65	69	58	75				
Rhesus Monkey	—	75	73	80	78	78	91	90	80	75	91	67	67	67	50	73	79	67	65	55	60	62	57	70				
Horse	—	—	100	25	40	40	71	67	71	50	71	73	75	75	60	73	73	64	57	61	59	50	65	75				
Donkey	—	—	—	25	40	40	71	67	71	50	63	73	75	75	60	73	73	57	57	59	56	44	65	73				
Cow	—	—	—	—	33	33	80	75	67	60	86	70	73	73	56	70	70	57	50	53	50	44	53	67				
Camel	—	—	—	—	—	—	75	67	60	100	83	70	78	78	57	75	75	68	55	50	47	50	50	60				
Elephant Seal	—	—	—	—	—	—	—	100	100	83	100	70	70	70	63	67	67	62	50	53	50	50	56	53				
Dog	—	—	—	—	—	—	—	—	100	80	100	70	70	70	63	78	67	67	50	47	44	50	53	50				
Bat	—	—	—	—	—	—	—	—	—	80	80	60	60	60	50	63	56	62	46	44	41	33	50	56				
Rabbit	—	—	—	—	—	—	—	—	—	—	83	63	63	63	53	71	78	67	53	50	56	44	47	63				
Kangaroo	—	—	—	—	—	—	—	—	—	—	—	83	80	80	70	82	82	67	62	59	65	67	65	82				
Chicken	—	—	—	—	—	—	—	—	—	—	—	—	100	100	100	75	100	63	45	50	56	45	63	56				
Man	—	—	—	—	—	—	—	—	—	—	—	—	—	100	100	75	100	70	45	50	56	40	60	61				
King Penguin	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100	75	100	70	50	53	59	45	65	58				
Pekin Duck	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	67	100	65	45	50	56	40	59	56				
Pigeon	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	88	67	50	59	65	55	66	68				
Snapping Turtle	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	73	50	59	56	44	6	68				
Rattle Snake	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	54	56	58	55	63	63				
Bull Frog	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	57	64	44	67	67				
Tuna	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	50	50	56	67				
Bonito	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	50	56	67				
Carp	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	46	60				
Dog Fish	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	69				

EXTENSION SPECIES

Species studied	Man	Elephant Monkey	Monkey	Cow	Camel	Elephant Seal	Dog	Bat	Rabbit	Kangaroo	Chicken	Man	King Penguin	Pekin Duck	Pigeon	Snapping Turtle	Rattle Snake	Bull Frog	Tuna	Bonito	Carp	Log Fish	Bottle Turkey
Man	—	0	55	36	27	40	25	27	28	44	40	38	31	38	28	25	33	29	33	33	46	29	45
Elephant Monkey	—	—	30	27	30	27	30	20	50	36	42	33	33	42	20	27	27	35	30	30	38	30	40
Monkey	—	—	—	30	30	29	33	14	50	43	45	43	43	50	30	27	27	30	33	29	30	29	38
Monkey	—	—	—	—	40	14	17	0	53	63	36	33	33	42	20	28	29	43	33	31	36	34	40
Cow	—	—	—	—	33	0	0	0	40	43	40	36	44	45	22	20	24	42	29	25	33	29	33
Camel	—	—	—	—	—	25	33	20	50	67	56	44	56	56	29	25	21	45	31	27	30	31	40
Elephant Seal	—	—	—	—	—	—	0	25	33	38	40	30	40	40	25	22	24	33	24	19	25	28	27
Log	—	—	—	—	—	—	—	33	40	43	40	30	30	40	25	22	24	33	24	17	25	29	29
Bat	—	—	—	—	—	—	—	—	20	25	20	10	20	20	0	15	19	23	17	18	22	22	25
Rabbit	—	—	—	—	—	—	—	—	—	50	38	23	38	38	0	14	17	36	23	23	44	29	41
Kangaroo	—	—	—	—	—	—	—	—	—	—	50	50	40	40	30	27	29	34	41	41	67	45	39
Chicken	—	—	—	—	—	—	—	—	—	—	—	50	100	67	25	25	27	36	31	31	45	32	39
Man	—	—	—	—	—	—	—	—	—	—	—	—	—	33	0	67	25	45	31	31	40	37	44
King Penguin	—	—	—	—	—	—	—	—	—	—	—	—	—	67	25	65	30	42	33	35	45	35	42
Pekin Duck	—	—	—	—	—	—	—	—	—	—	—	—	—	—	33	43	24	27	25	25	30	29	33
Pigeon	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	50	22	33	29	29	36	32	32
Snapping Turtle	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	27	40	35	31	33	37	37
Rattle Snake	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	29	28	31	30	23	33
Bull Frog	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	38	36	56	40	43
Tuna	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0	17	32	22	22
Bonito	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	17	21	13
Carp	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	15	108
Log Fish	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	31

Percentages of replacements by residues similar in secondary structural affinity

129

TABLE V (b)

Percentages of replacements by conformationally similar residues

Species studied	Haemoglobin α -chain										Haemoglobin β -chain									
	Reference species										Reference species									
	Man	Rhesus mon-key	Mouse NB	Rabbit	Dog	Sheep	Pig	Horse Slow	Kangaroo	Species studied	Man	Rhesus mon-key	Mouse C57BL	Rabbit	Dog	Sheep	Pig	Horse	Elephant	Kangaroo
Man	-	25	42	44	30	9	17	17	32	Man	-	50	29	36	20	37	47	32	27	34
Rhesus Monkey	-	-	44	44	38	5	25	25	30	Rhesus Monkey	-	-	21	25	20	32	41	29	27	33
Mouse NB	-	-	-	38	36	24	33	38	37	Mouse C57BL	-	-	-	31	30	44	48	42	30	32
Rabbit	-	-	-	-	32	37	50	52	35	Rabbit	-	-	-	-	33	48	64	44	36	43
Dog	-	-	-	-	-	20	30	33	36	Dog	-	-	-	-	-	46	53	30	33	40
Sheep	-	-	-	-	-	-	27	25	20	Sheep	-	-	-	-	-	-	33	18	30	42
Pig	-	-	-	-	-	-	-	23	27	Pig	-	-	-	-	-	-	-	20	24	50
Horse Slow	-	-	-	-	-	-	-	-	27	Horse	-	-	-	-	-	-	-	-	24	40
										Elephant	-	-	-	-	-	-	-	-	-	36

TABLE V (c)

Percentages of replacements by residues similar in composition of codons

Species studied	Haemoglobin α -chain										Haemoglobin β -chain									
	Reference species										Reference species									
	Man	Rhesus mon-key	Mouse NB	Rabbit	Dog	Sheep	Pig	Horse slow	Kangaroo	Species studied	Man	Rhesus mon-key	Mouse C57BL	Rabbit	Dog	Sheep	Pig	Horse	Elephant	Kangaroo
Man	-	75	84	80	87	77	75	72	71	Man	-	50	79	79	73	63	73	68	62	55
Rhesus Monkey	-		78	76	83	68	67	75	74	Rhesus Monkey	-	-	68	88	80	57	71	68	67	56
Mouse NB	-	-	-	72	88	76	76	79	73	Mouse C57BL	-	-	-	79	80	72	86	83	70	57
Rabbit	-	-	-	-	79	80	90	68	73	Rabbit	-	-	-	-	81	70	71	60	75	54
Dog	-	-	-	-	-	90	83	78	73	Dog	-	-	-	-	-	61	80	80	70	49
Sheep	-	-	-	-	-	-	73	65	70	Sheep	-	-	-	-	-	-	50	54	67	53
Pig	-	-	-	-	-	-	-	69	77	Pig	-	-	-	-	-	-	-	73	71	60
Horse Slow	-	-	-	-	-	-	-	-	73	Horse	-	-	-	-	-	-	-	-	71	56
										Elephant	-	-	-	-	-	-	-	-	-	64

TABLE V (d)

Percentages of replacements by residues of same chemical nature

Species studied	Haemoglobin α -chain										Haemoglobin β -chain									
	Reference species										Reference species									
	Man	Rhesus mon-key	Mouse NB	Rabbit	Dog	Sheep	Pig	Horse Slow	Kangaroo	Species studied	Man	Rhesus mon-key	Mouse C57BL	Rabbit	Dog	Sheep	Pig	Horse	Elephant	Kangaroo
Man	-	75	58	40	35	32	33	39	36	Man	-	38	50	36	40	37	33	36	12	40
Rhesus Monkey	-	-	56	40	38	22	33	38	30	Rhesus Monkey	-	-	46	38	33	36	35	39	20	42
Mouse NB	-	-	-	41	44	40	57	50	40	Mouse C57BL	-	-	-	48	43	50	57	42	45	39
Rabbit	-	-	-	-	39	37	40	36	30	Rabbit	-	-	-	-	48	44	36	28	14	41
Dog	-	-	-	-	-	30	26	30	30	Dog	-	-	-	-	-	46	40	47	26	34
Sheep	-	-	-	-	-	-	33	25	27	Sheep	-	-	-	-	-	-	39	36	33	40
Pig	-	-	-	-	-	-	-	23	39	Pig	-	-	-	-	-	-	-	40	29	47
Horse Slow	-	-	-	-	-	-	-	-	30	Horse	-	-	-	-	-	-	-	-	32	44
										Elephant	-	-	-	-	-	-	-	-	-	28

TABLE V (e)

Percentages of replacements by residues of same polarity

Species studied	Haemoglobin α -chain										Haemoglobin β -chain									
	Reference species										Reference species									
	Man	Rhesus Mon-key	Mouse NB	Rabbit	Dog	Sheep	Pig	Horse Slow	Kangaroo	Species studied	Man	Rhesus Mon-key	Mouse C57BL	Rabbit	Dog	Sheep	Pig	Horse	Elephant	Kangaroo
Man	-	25	58	48	39	36	42	45	43	Man	-	63	57	43	60	52	47	44	46	47
Rhesus Monkey	-	-	50	56	38	37	42	57	41	Rhesus Monkey	-	-	61	56	73	54	53	46	50	47
Mouse NB	-	-	-	41	36	28	38	33	33	Mouse C57BL	-	-	-	62	50	53	62	50	41	50
Rabbit	-	-	-	-	46	50	55	48	41	Rabbit	-	-	-	-	57	44	36	52	32	49
Dog	-	-	-	-	-	30	35	33	49	Dog	-	-	-	-	-	54	60	53	44	51
Sheep	-	-	-	-	-	-	40	25	27	Sheep	-	-	-	-	-	-	56	46	42	54
Pig	-	-	-	-	-	-	-	23	39	Pig	-	-	-	-	-	-	-	40	43	50
Horse Slow	-	-	-	-	-	-	-	-	37	Horse	-	-	-	-	-	-	-	-	44	42
										Elephant	-	-	-	-	-	-	-	-	-	41

functions of these homologous proteins in addition to their common functions.

Thus, only partial explanation of replacements between sequences from any two species by considering individual properties of amino acid residues and non-conservation of any of these properties in replacement substitutions during evolution shows that this type of studies alone cannot provide enough information to establish an evolutionary pathway of proteins.

This study also points to the need for such studies, to aid in understanding whether micro-evolution and macro-evolution have taken place in phase, i.e., to establish whether changes observed in proteins and nucleic acids correspond one to one with morphological changes of species.

APPENDIX II

OBLIGATORY AMINO ACIDS IN PRIMITIVE PROTEINS

INTRODUCTION

Genes of present day living systems can code for a maximum of 20 types of amino acids. In other words, polypeptides and proteins synthesised using the ribosomal machinery can, during synthesis, have at most 20 types of amino acids, termed 'proteinous' amino acids, though more than 250 types of amino acids occur in natural polypeptides (Mooz, 1976). Thus a natural and fundamental question that arises is why 20 amino acids only were selected as coded amino acids and whether all these 20 amino acids were present in proteins of primitive biosystems. The first part of the question, namely, the selection of 20 amino acids as coded amino acids, has received some attention recently and reasons for their existence in proteins have been suggested. (Rohlfing and Saunders, 1978; Weber and Miller, 1981). However, the latter part of the question; the types of amino acids present in primitive proteins is yet to be raised. In this Appendix we suggest that out of 20 proteinous amino acids only Ser, Val, Leu, Asp, Pro, and Gly were obligatory and were present in proteins of primitive biosystems. The study of the composition, structure and function of such 'primitive proteins' is the subject of this Appendix.

METHOD

In order to get some idea about the composition of these primitive proteins, one should find out the traits that might have remained almost invariant during evolution. One such trait seems to be the topology of domains of proteins, as only a restricted number of topologies of domains have been observed when crystal structure data of a large number of

different globular proteins were analysed (Ptitsyn and Finkelstein, 1980). Domains are those regions of proteins which form very compact globules by having many internal contacts but a few contacts with other parts of the chain (Schulz and Schirmer, 1979). In general a single polypeptide chain protein can consist of two to three domains, each of about 70-80 amino acids (Rashin, 1981). The restricted number of observed topologies of domains might have been due to the availability of a restricted number of amino acids for synthesis of primitive proteins, which we assume consist of single domains, in contrast to proteins of more evolved systems which have multiple domains. If primitive proteins consist of a few types of amino acids and single domains in nature, then the topologies of these primitive proteins will be limited in number due to physical forces as argued by Ptitsyn and Finkelstein (1980). The conformational property of amino acids which determines the topology of domains and three dimensional structure of proteins, can be used to find out the 'obligatory' amino acids in primitive proteins. Hence, making use of the conformational similarity among amino acids, an intrinsic property in three dimensions, derived in Chapter II, a set of a minimum number of amino acids, which can represent conformationally the remaining proteinous amino acids, is derived. The derivation is presented below.

Derivation of obligatory amino acids of primitive proteins: From Table III of Chapter II one can observe that residues such as Ala, Ile, Ser Val are conformationally similar to more than one residue. Analysis of this Table reveals that at least six amino acids are necessary to represent all 20 proteinous amino acids conformationally. The possible sets of six amino acids which can represent all 20 amino acids are given in Table I.

The derivation of this Table is discussed briefly below by taking the example of set II:

Ser occurs in the right column of Table III of Chapter II for Ala, Arg, Asn, Cys, Phe and Thr, indicating that the (\emptyset, ψ) -probability distribution of Ser is similar to the (\emptyset, ψ) -distribution of above-mentioned residues. Thus it can represent any one of these residues with minimal conformational discrepancy. Similarly, Val represents Arg, Cys, Ile, Met, Phe, Thr, Trp and Tyr while Leu occurs in the right column of the above Table for Arg, Gln, His, Met, Phe, Ser and Thr. Asp represents conformationally Asn, Lys, His and Arg. The remaining two of this set of six amino acids are Pro and Gly which have unique (\emptyset, ψ) -probability distributions.

It can be seen that some residues of this set III represent more than one of remaining proteinous amino acids. For example, Arg can be represented by any one of Ser, Val and Leu. Sets I and II of Table I are derived in a similar fashion. Three amino acid residues, Gly, Pro and Asp are common to these three sets and only the first three amino acid residues are different. Therefore, we have looked at the possibility of synthesis of these amino acids in the laboratory. Though the synthesis of Ala, Val or Ala, Ile is not difficult, the synthesis of Gln or Thr is definitely not easy. However, this is not the case with Ser, Val or Leu. This prompts us to suggest that Ser, Val, Leu, Asp, Gly and Pro were obligatory amino acids in proteins by the start of biotic evolution.

TABLE I

Minimum number of amino acids which can represent conformationally
all 20 proteinous amino acids

Set	Amino acids					
I	Ala	Val	Gln	Asp	Pro	Gly
II	Ala	Ile	Thr	Asp	Pro	Gly
III	Ser	Val	Leu	Asp	Pro	Gly

Residues of Set III are suggested as obligatory amino acids
in primitive proteins.

DISCUSSION

The set which we have proposed consists of a single, acidic amino acid, Asp, in place of the two, Asp and Glu, which occur in proteins. The presence of Asp rather than Glu is not surprising because even the chemical nature of Asp is more simple as compared to Glu which contains two $-CH_2$ groups in β and γ positions.. The presence of an acidic amino acid is essential so as to prevent the protein or polypeptide from interacting with hydrated electrons (Scott, 1981). Thus, the necessity for acidic amino acids was enormous at the dawn of biotic evolution but one cannot visualise such an important role for basic amino acids. Therefore, the absence of basic amino acids from the proposed set does not seem to be surprising.

In the proposed set there are three amino acids which are either polar or neutral-polar while the remaining three are hydrophobic in nature. Thus, these hydrophobic amino acids must have formed the interior of the globule of the protein while polar and neutral-polar amino acids might have acted to shield the protein from the effect of hydrated electrons which were present because of the aqueous environment. It is interesting to note that 50% of these obligatory amino acids, Val, Leu and Pro are hydrophobic and roughly the same percentage of hydrophobic amino acids is present in the present set of proteinous amino acids. The proposed set does not contain any aromatic or sulphur containing amino acids. Although aromatic and sulphur containing amino acids have a very important and crucial role in proteins and enzymes of evolved biosystems, their function seems to be specific. Similarly, the biosynthesis of aromatic amino acids suggests that they have been incorporated at a later stage as suggested by Wong (1981).

The proposed set consists of amino acids which prefer α -helix, β -sheet and chain reversals. For example Leu is an α -helix preferer, Val prefers β -sheet and Ser, Asp, Gly and Pro are known to occur frequently in chain reversals (Levitt, 1978; Kolaskar et al., 1980). All proteins which have enzymatic activity are known to have α -helix as one of the secondary structures. Thus, the presence of Leu seems to be essential.

Ser is known to be not only a β -bend preferer but also a site of the attachment of polysaccharide chains. Pro seems to be present in primitive proteins mainly because of its characteristic side chain property which gives rigidity to main chain conformation of the polypeptide chain. Gly not only plays an exactly opposite role to that of Pro, namely, providing flexibility to main chain conformation but also takes conformations which other amino acids can not take because their side chains are in L-configuration. In other words it avoids the need for the presence of amino acids in D-configuration. This has been very clearly shown from the crystal structure studies on insulin (Hodgkin, D., private communication). Thus, each of the amino acids suggested in the above set might have played an important role in primitive proteins which are assumed to be multifunctional and single domain in nature.

The experiments in which primordial conditions were simulated to study synthesis of amino acids have shown that less than half of the 20 proteinous amino acids were produced in more than trace amounts. But it is worth noting that the amino acids Ser, Val, Leu, Asp, Gly and Pro (Set III) are synthesized in non-negligible quantities (Dose, 1976; Wong, 1981). We have avoided comparing our set of obligatory amino acids

with amino acids present in extra-terrestrial matter, or lunar dust, since there are no evidences for existence of life there.

It is further interesting to note that though we have arrived at the proposed set by taking into consideration only conformational properties of the amino acid residues, all but one (Asp) of the amino acids of Set III have four or more codons in the present genetic code and thus the presence of third letter for these amino acids seems to be immaterial. In other words, the doublet code for Gly, Val, Leu, Ser and Pro will be GG, GU, CU, UC and CC, respectively, where only the first two letters from the triple letter code are considered. In the case of Asp there are two codons in the present genetic code and the doublet code for Asp can be considered as GA. Thus, it appears quite plausible that the contemporary triplet code has evolved from a doublet code (Jukes, 1973).

One may check the validity of our arguments in the following fashion:

(i) Recently an algorithm has appeared which can be used to find out domains in the proteins (Rashin, 1981). Consider one such domain, say that of T4 lysozyme. In this domain, which consists of 74 residues, represent all amino acids by the proposed six amino acids using the property of conformational similarity. The sequence of this domain is given in Fig. 1. The topology of the polypeptide chain, the sequence of which is given in Fig. 1(b) will not be very much different from that of the observed topology of the first domain of T4 lysozyme.

((Even if it is different one can find out the topology of this polypeptide chain and compare it with the topologies of the domain known from crystal structures of globular protein). The knowledge of topology will give an idea about the possible functions of the computer-simulated polypeptide chain, which consists of only six types of amino acids, and thus about the range of choice of various substrates. The measured K_a -values for various substrates will indicate the multifunctional nature of this polypeptide chain and also its role in the primitive biosystems. The measurements of other physical constants will indicate the thermodynamical stability of this globule.

(ii) The same experiment of measurement of various properties of the polypeptide chain containing the proposed six amino acid residues can be carried out as follows. One can repeat in the laboratory Fox's type of experiment (1965) taking only the proposed six amino acids instead of 18 amino acids and then carry out the characterization of the synthesized polymers.

Thus, in short, we have discussed in this Appendix that during evolution the topology of the domains remained nearly invariant and the restriction on these topologies is because of the presence of few types of amino acids in primitive proteins, which are single domains in nature. Using only main chain conformational similarities among amino acids, we have arrived at a set of six amino acids and have argued that these are obligatory amino acids at the dawn of evolution of proteins.

- Anfinsen, C.B. and Scheraga, H.A. (1975), *Adv. Prot. Chem.* 29, 205.
- Argos, P., Schwarz, J. and Schwarz, J. (1976) *Biochim et. Biophys. Acta*, 439, 261.
- Argos, P. and Palau, J. (1982), *Int. J. Peptide Prot. Res.* 19, 380.
- Balasubramanian, R. (1977), *Nature*, 266, 856.
- Baldwin, R.L. (1975), *Ann. Rev. Biochem.* 44, 453.
- Blake, C.C.F., Koenig, D.F., Mair, G.A., North, A.C.T., Phyllips, D.C. and Sharma, V.R. (1965), *Nature*, 206, 757.
- Bourgeios, S., Jernigan, R.L., Szu, S.C., Kabat, E.A. and Wu, T.T. (1979), *Biopolymers*, 18, 2625.
- Brahms, S. and Brahms, J. (1980), *J. Mol. Biol.*, 138, 149.
- Brant, D.A. and Flory, P.J. (1965), *J. Am. Chem. Soc.* 87, 2791.
- Braun, W. (1983), *J. Mol. Biol.* 163, 613.
- Brown III, F.R., Hopfinger, A.J. and Blout, E.R. (1972), *J. Mol. Biol.*, 63, 101.
- Bunting, J.R., Athey, T.W. and Cathou, R.E. (1972), *Biochim. et. Biophys. Acta*, 285, 60.
- Burgess, A.W., Ponnuswamy, P.K. and Scheraga, H.A. (1974), *Israel J. Chem.*, 12, 239.
- Burgess, A.W., Scheraga, H.A. (1975), *Proc. Natl. Acad. Sci. U.S.A* 72, 1221.
- Busetta, B. and Hospital, M. (1982), *Biochemica et Biophysica Acta*, 701, 111.
- Busetta, B. and Barrans, Y. (1982), *Biochemica et Biophysica Acta* 709, 73.
- Chou, P.Y. and Fasman, G.D. (1974), *Biochemistry*, 13, 211.
- Chou, P.Y. and Fasman, G.D. (1978), in *Advances in Enzymology* (Meister, A., ed.) Vol. 47, pp. 45, John Wiley and Sons, New York.
- Cid, H., Campos, M. and Aruigada, E. (1980), *FEBS Letters*, 111, 56.

- Cohn, E.J. and Edsall, J.T. (1943), *Proteins, Aminoacids and Peptides*, New York, Reinhold.
- Cohen, T.E., Sternberg, M.J.E. and Taylor, W.R. (1982), *J. Mol. Biol.* 156, 821.
- Creighton, T.E. (1978), *Prog. Biophys. Mol. Biol.* 33, 231.
- Desantis, P., Giglio, E., Liquori, A.M. and Ripamonti, A. (1965) *Nature*, 206, 456.
- Doolittle, R.F. (1981), *Science*, 214, 149.
- Dose, K. (1976), *Protein Structure and Function* (J.L. Fox, Z. Deyl and A. Blazej, Eds.) Marcel Dekker Inc. New York and Basel, pp. 149-184.
- Feldmann, R. (1976) *Atlas of Macromolecular Structure on Microfiche*, 1st Edn. Tracer Jitco Inc., Rockville, MD 20852.
- Felsenstein, J. (1982) *Quart. Rev. Biol.* 57, 379.
- Finkelstein, A.V. and Ptitsyn, O.B. (1971), *J. Mol. Biol.* 62, 613.
- Flory, P.J. (1969), *Statistical Mechanics of Chain Molecules*, Wiley Interscience, New York.
- Fox, S.W. (1965), *Nature*, 205, 328.
- Galat, A. (1982) *Int. J. Biochem.* 14, 883.
- Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) *J. Mol. Biol.* 120, 97.
- Ghelis, C., Yon, J. (1982), *Protein Folding*. New York Academic, In press.
- Go. N., Abe, H., Mizuno, H., Taketomi, H. In "Protein Folding", Jaenicke, R., Ed., Elsevier/North Holland Biomedical Press- Amsterdam (1980), p. 167.
- Goel, N.S., Rouyanian, B., and Sanati, M. (1982), *J. Theor. Biol.* 99, 705.
- Hagler, A.T., Honig, B. (1978), *Proc. Natl. Acad. Sci. U.S.A.* 75, 554.
- Handbook of Biochemistry and Molecular Biology* (1976) *Proteins III*, (Eds. G.D. Fasman) CRC Press.
- Hol, W.G.J., Halie, L.M. and Sander, C. (1981) *Nature*, 294, 532.

- Ikegami, A. (1981), Adv. Chem. Phys. 46, 363.
- Jaenicke, R. Ed. (1980), Protein Folding, Proc. 28th Conf. German Biochem. Soc. Amsterdam, Elsevier/North Holland Biomed. Press. 587, pp.
- Janin, J., Wodak, S., Levitt, M. and Maigret, B. (1978) J. Theor. Biol. 125, 357.
- Jones, D.D. (1975), J. Theor. Biol. 50, 167.
- Jukes, T.H. (1973), Nature, 246, 22.
- Kabat, E.A. and Wu, T.T. (1973), Proc. Nat. Acad. Sci. U.S.A. 70, 1473.
- Kabat, E.A. and Wu, T.T. (1973), Biopolymers, 12, 751.
- Kauzmann, W. (1959) Adv. Prot. Chem. 14, 1.
- Kendrew, J.C., Dickerson, R.E., Strandberg, E.E., Hart, R.G., Davies, D.R., Phillips, D.C. and Shore, V.C. (1960), Nature, 185, 422.
- Kitaigorodsky, A.I. (1965) Acta, Cryst, 18, 585.
- Kolaskar, A.S. and Prashanth, D. (1979), Int. J. Peptide Protein Res. 14, 88.
- Kolaskar, A.S., Ramabrahmam, V. and Soman, K.V. (1980), Int. J. Pept. Protein Res. 16, 1.
- Kolaskar, A.S. and Ramabrahmam, V. (1981), Int. J. Biolog. Macromol. 3, 171.
- Krigbaum, W.R. and Rubin, B.H. (1971), Biochem. Biophys. Acta, 229, 368.
- Krigbaum, W.R. and Komoriya, A. (1979) Biochim. Biophys. Acta 576, 204.
- Krigbaum, W.R. and Lin, S.F. (1982) Macromolecules, 15, 1135.
- Kuntz, I.D., Crippen, G.M., Kollman, P.A., Kimelman, D.J. (1976) J. Mol. Biol. 106, 983.
- Kyte, J. and Doolittle, R.F. (1982), J. Mol. Biol. 157, 105.
- Lambhardt, A.M. (1982), J. Mol. Biol. 157, 357.
- Lehinger, A.L. (1975) in 'Biochemistry', 2nd Ed. Worth, New York.
- Levitt, M., Warshel, A. (1975), Nature (London), 253, 694.

- Levitt, M. (1978), *Biochemistry*, 17, 4277.
- Lim, V.I. (1974), *J. Mol. Biol.* 88, 857.
- Louie, A.H. and Somorjai, R.L. (1982), *J. Theor. Biol.* 98, 189.
- Manavalan, P. and Ponnuswamy, P.K. (1978), *Nature (London)*, 275, 673.
- McMeekin, T.L., Groves, M.L. and Hipp, N.J. (1964), In *Amino Acids and Serum Protein* (J.A. Stekol, Ed.) p. 54, Washington, D.C. American Chemical Society.
- Meirovitch, H., Rackovsky, S. and Scheraga, H.A. (1980), *Macromolecules*, 13, 1398.
- Meirovitch, H. and Scheraga, H.A. (1981), *Macromolecules*, 14, 1250.
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953), *J. Chem. Phys.* 21, 1087.
- Mooz, E.D. (1976) in 'Handbook of Biochemistry and Molecular Biology, Proteins' (G.D. Fasman, Ed.), Vol. 1 (Chemical Rubber Co. Press, Cleveland) p. 111.
- Nagano, K. (1973), *J. Mol. Biol.* 75, 401.
- Nemethy, G. and Scheraga, H.A. (1977), *Q. Rev. Biophys.* 10, 239.
- Nozaki, Y. and Tanford, C. (1971), *J. Biol. Chem.* 246, 2211.
- Palau, J., Argos, P. and Puigdomenech, P. (1982), *Int. J. Peptide Protein Res.* 19, 394.
- Pauling, L., Corey, R.B. and Branson, H.R. (1951), *Proc. Natl. Acad. Sci. U.S.A.* 37, 205.
- Pauling, L. and Corey, R.B. (1951), *Proc. Natl. Acad. Sci. U.S.A.* 37, 241.
- Periti, P.F. (1974), *Boll. Chim. Farm.* 113, 187.
- Pohl, F.M. (1971), *Nature New Biol.* 234, 277.
- Ponnuswamy, P.K. and Sasisekharan, V. (1971), *Biopolymers*, 10, 565.
- Prabhakaran, M. and Ponnuswamy, P.K. (1980), *J. Theor. Biol.* 87, 623.
- Privalov, P.L. (1979), *Adv. Protein Chem.* 33, 167.
- Ptitsyn, O.B. and Finkelstein, A.V. (1970), *Biophysica*, 15, 785.
- Ptitsyn, O.B. and Finkelstein, A.V. (1980), *Q. Rev. Biophys.* 13, 339.
- Ptitsyn, O.B. (1981), *FEBS Lett.* 131, 197.

- Pullman, B. and Pullman, A. (1974), *Adv. Protein Chem.* 28, 347.
- Rackovsky, S. and Scheraga, H.A. (1982), *Macromolecules*, 15, 1340.
- Ramachandran, G.N. (1962), in "Collagen" (N. Ramanathan, Ed.) p. 3, Wiley (Interscience), New York.
- Ramachandran, G.N., Ramakrishnan, C. and Sasisekharan, V. (1963) *J. Mol. Biol.* 7, 95.
- Ramachandran, G.N., Venkatachalam, C.M. and Krimm, S. (1966), *Biophysics*, J. 6, 849.
- Ramachandran, G.N. and Sasisekharan, V. (1968), *Adv. Protein Chem.* 23, 283.
- Ramachandran, G.N. in "Conformation of Biological Molecules and Polymers" (Ed., Bergman, E.D. and Pullman, B.) Academic Press, Jerusalem, 1973, pp. 1-13.
- Rapaport, D.C. and Scheraga, H.A. (1981), *Macromolecules*, 14, 1238.
- Rashin A.A. (1981), *Nature*, 291, 85.
- Richards, F.M. (1977), *Ann. Rev. Biophys. Bioeng.* 6, 151.
- Richardson, J.S. (1981), *Adv. Protein Chem.* 34, 167.
- Robson, B. and Pain, R.H. (1971), *J. Mol. Biol.* 58, 237.
- Robson B. and Pain, R.H. (1974), *Biochem. J.* 141, 853.
- Robson, B. and Suzaki, E. (1976), *J. Mol. Biol.* 107, 327.
- Robson, B., Osguthorpe, D.J. (1979), *J. Mol. Biol.* 132, 19.
- Rohlfing, D.L. and Saunders, M.A. (1978), *J. Theor. Biol.* 71, 487.
- Rossmann, M.G. and Argos, P. (1981), *Ann. Rev. Biochem.* 50, 497.
- Sasisekharan, V. (1962) in "Collagen" (V. Ramanathan, Ed.) p. 39 Wiley (Interscience), New York.
- Scheraga, H.A. (1968), *Adv. Phy. Org. Chem.* 6, 103.
- Schulz, G.E., Barry, C.D., Friedman, J., Chou, P.Y., Fosman, G.D., Finkelstein, A.V., Lim, V.I., Ptitsyn, O.B., Kabat, E.A., Wu, T.T., Levitt, M., Robson, B. and Nagano, K. (1974), *Nature*, 250, 140.
- Schulz, G.E. and Schirmer, R.H. (1979), *Principles of Protein Structure* (Springer Verlag, New York, Heidelberg, Berlin).
- Scott, J. (1981), *New Sci.* 89, 153.

- Scott, R.A. and Scheraga, H.A. (1965), J. Chem. Phys. 42, 2209.
- Sternberg, M.J.E. and Cohen, F.E. (1982), Int. J. Biolog. Macromolecules, 4, 137.
- Stryer, L. (1975) in "Biochemistry" (W.H. Freeman, San Francisco)
- Tanaka, S. and Scheraga, H.A. (1976), Macromolecules, 9, 168.
- Tanford, C. (1968), Adv. Protein Chem. 23, 121.
- Tanford, C. (1970), Adv. Protein Chem. 24, 1.
- Thomas, K.A. and Schechter, A.N. (1980) in "Biological Regulation and Development", Ed. R.F. Goldberger, 2, 43, New York: Plenum.'
- Venkatachalam, C.M. (1968), Biopolymers, 6, 1425.
- Wagner, G. and Wuthrich, K. (1982), J. Mol. Biol. 155, 247.
- Warne, P.K. and Morgon, R.S. (1978), J. Mol. Biol. 118, 273.
- Warne, P.K. and Morgon, R.S. (1978), J. Mol. Biol. 118, 289.
- Weber, A.L. and Miller, S. (1981), J. Mol. Evol. 17, 273.
- Wetlaufer, D.B. and Ristow, S. (1973), Ann. Rev. Biochem. 42, 135.
- Wetlaufer, D.B. (1981), Adv. Protein.Chem. 34, 61.
- William, R.W. and Dunker, A.K. (1981), J. Mol. Biol. 152, 783.
- Wodak, S.J. and Janin, J. (1981), Biochemistry, 20, 6544.
- Wong, J.T. (1981), Trends Biochem. Sci., 6, 33.
- Zimmerman, J.M., Eliezer, N. and Simha, R. (1968), J. Theor. Biol. 21, 170.

CHAPTER 5 -	REGULATION OF NITRATE ASSIMILATION IN CYANO- BACTERIUM ANABAENA CYCADEAE: PARENT AND GLUTAMINE AUXOTROPHIC STRAINS	
5.1.	INTRODUCTION	87-89
5.2.	MATERIALS AND METHODS	90-94
5.3.	RESULTS	94-99
5.4.	DISCUSSION	99-104
CHAPTER 6 -	GENERAL DISCUSSION	105-111
	R E F E R E N C E S	112-126

--:0:--