THEORETICAL STUDIES ON PROTEIN SEQUENCES

A
Thesis
Submitted
for the degree of
Doctor of Philosophy

by

Meeta Rani

Department of Biochemistry
School of Life Sciences
University of Hyderabad

Hyderabad- 500 134, (India)

DECEMBER 1994

Dedicated

to

Mummy & Daddy

Nirmal Bhaiyya & Deepa Bhabhi

Seema & Enrico.

Shrikant & Ripal

Contents

STATEMENT

I hereby declare that the matter embodied in this thesis entitled "Theoretical Studies on Protein Sequences" is a result of investigations carried out by me in the Department of Biochemistry, School of Life Sciences, University of Hyderabad, Hyderabad- 500 134, India, under the supervision of **Dr. Chanchal K Mitra** and this work has not been submitted for any degree or diploma at any other university.

In keeping with the general practice of reporting scientific observations, due acknowledgements have been made where ever the works described are based on findings of other investigators.

Meeta Rani
December 16th 1994

Meeta Rani

Enrollment No: PL-5942.

UNIVERSITY OF HYDERABAD

CERTIFICATE

Certified that the work embodied in this thesis has been carried out by Ms. Meeta Rani under my guidance for the full period prescribed under the Ph.D. ordinance of this University and the same has not been submitted elsewhere for any degree or diploma of any other university.

I recommend her thesis entitled "Theoretical Studies on Protein Sequences" for submission for the degree of Doctor of Philosophy to the University.

Chanchal K Mitra
(Supervisor)

Head
Department of Biochemistry

16/12/94
Dean
School of Life Sciences

Acknowledgements

I am deeply indebted to my teacher and research supervisor, Dr Chanchal K Mitra for having taught me to learn. He has made me realise that "there is no alternative to hard work". Under his deft guidance I learnt the importance of discipline and systematic hard work. I am thankful to him for introducing me to a very interesting problem for my Ph.D. and also for the interesting method to try solving it. It has been a rewarding experience working with Sir.

I wish to express my gratitude to Prof. N C Subramanyam, Dean, School of Life Sciences, and Prof. T Suryanarayana Rao, Head, Department of Biochemistry, for providing all the necessities to carry out the research.

I am thankful to Dr. C Suguna, Bioinformatics, CCMB, Hyderabad, for providing the protein data-base for our research.

I wish to extend my thanks to Prof.G Govil, Chemical Physics Group, TIFR, Bombay, for making arrangements for me to do some very invaluable reference at the TIFR library. I am also thankful to Prof. G Krishnamurthy, Chemical Physics Group, for his kind help during my visit to TIFR.

I am indebted to my friends, Shanta and Bidisha for the moral support they have always extended to me and for the encouragement I have received from them.

My thanks are due to my labmates Surendar and Sulekha for providing a friendly and pleasant atmosphere.

I extend my thanks to all the faculty who in many direct and indirect ways are responsible for the completion of this work.

I am indebted to my parents, brothers and sister for all the encouragement they have extended to me in all my endeavours.

I gratefully acknowledge the financial assistance provided to me by the University Grants Commission, New Delhi, as JRF and SRF.

Preface

Studying proteins has been a very fascinating experience for scientists. Most of the landmarks in Biochemistry and Molecular Biology revolve around protein studies in some way or the other. While the triumphs of scientists like D C Phillips, J F Kendrews, William Astbury and G N Ramachandran, in protein structure determination; the verification of the protein folding dogma by C B Anfinsen and the explanation of antibody diversity by Tonegawa are encouraging, still it cannot be denied that proteins are still very mysterious molecules and a lot remains concealed than revealed regarding their structure and function.

Several elegant methods have been developed over a few decades to study various aspects of proteins. While the only changes are improvements in the existing methods, we have introduced a new method to study proteins using
i) fractal methods ii) harmonic analysis of the protein sequences.

While we had already started our work, there have been reports about the analysis of DNA by fractal methods (Voss, Peng etc.). Regarding proteins there have been reports on the basis of experiments that some proteins occupy a 3-D space of fractional dimensionality (Stapleton et al, 1980) and also about fractal properties of protein surfaces. However, to the best of our knowledge no one has carried out fractal studies on protein sequences in our manner. This thesis is based on the results of the work done by us and most of our results have been published already.

We are aware that treading on untread path has its own problems. It needs a great deal of suggestions and criticisms for improvements. This work may be viewed in this light and we would welcome all such comments for improvements.

Meeta Rani
December 12th, 1994

Chapter 1

Proteins and their structures

1 Introduction

The most conspicuous attribute of living organisms is that they are complicated and highly organised. The cells, of which they are composed of, possess intricate internal structures containing many kinds of complex molecules. The majority of these biomolecules are macromolecules like proteins, nucleic acids and polysaccharides. Of these, proteins are the most abundant molecules, constituting 50% of the dry weight and 15% of the wet weight. They are found in every part of the cell as they fundamental in all aspects of cell structure and function.

Chemically, proteins are nitrogenous complex organic compounds with molecular weights ranging from a few hundred to several millions. All proteins on hydrolysis produce a group of smaller compounds that can be identified as amino acids. Analysis has shown that proteins are polymers made up of these amino acids joined head to tail. Proteins with upto 20 amino acids are called peptides (e.g., oxytocin and vasopressin). Proteins with upto 100 residues are considered as small proteins (e.g., insulin). Medium sized proteins (myoglobin, cytochrome c) contain between 100 to 300 residues and those having more than 300 residues are large proteins (serum albumin, 550 residues approximately). Some proteins can be very large having thousands of residues (myosin 1800 residues approximately).

1.1 Functional diversity of proteins

Proteins are highly versatile molecules performing several different biological functions some of which are described below.

i) Catalysis of biochemical reactions

Life forms make use of chemical reactions to supply themselves continually with chemical energy. These reactions cannot occur fast enough under physiological conditions (aqueous solution, 37° C, pH 7 and atmospheric pressure) to sustain life. Some proteins known as enzymes act as catalysts and their action increase the rate of these reactions by several orders of magnitude in the organisms under the same physiological conditions. The enzymes are highly specific in nature and they represent the largest class of proteins. Nearly 2000 different kinds of enzymes are known, each catalysing a different kind of biochemical reaction. Many enzymes can be "controlled" or regulated by other molecules providing a mechanism to control a set of biochemical processes.

ii) Transport

Proteins transport a variety of particles ranging from macromolecules to electrons. The serum albumin binds free fatty acids tightly and transports these molecules between adipose (fatty) tissue and other tissues and organs in vertebrates. Lipoproteins of blood plasma transport lipids between intestine, liver and adipose tissues. Haemoglobin transports oxygen from lungs to tissues. Proteins of the electron transport chain guide the flow of electrons in the vital process of photosynthesis. Some proteins control transfer of small molecules across the cell membrane.

Hi) Storage of food

Proteins also serve to store food as in ovalbumin of egg-white for the growing chick embryo; casein of milk for the new born calf and gliadin of wheat for germination of the wheat seed.

iv) Structural elements

Some proteins are simply structural. Collagen is the major extracellular structural protein in connective tissue and bone of the vertebrates. Elastin is found in the yellow elastic tissue. Keratin provides the material for hair, nails, hoof, *etc.* of mammals.

v) Toxic proteins

Some seed proteins are highly toxic, like ricin from castor seed and gossypin from cotton seed, probably to keep away predators. Some microbes produce toxins that cause disease in higher organisms like the cholera toxin and diphtheria toxins. Snake venom is also a toxic protein and is used for defensive and offensive purpose.

vi) Muscle contraction

Actin and myosin are crucial components of muscles and other systems for converting chemical energy into mechanical energy and allow contraction of muscles. Continuous beating of the heart is possible by the action of these proteins.

vii) Defence and protection

Immunoglobulins (antibodies) function in the immune systems of the higher organisms to defend against the intruders *(i.e.,* foreign antigens) by combining

with and neutralising them. Thrombin and fibrinogen participate in the blood clotting mechanism to prevent the loss of blood from the vascular system in vertebrates.

viii) Hormone action

As hormones, proteins transmit information between specific cells and organs. For instance, insulin regulates the blood sugar level and somatotropin regulates growth.

ix) Regulation of gene expression

Some proteins called transcription factors control gene expression by binding to specific sequences of nucleic acids thereby turning genes on and off. These DNA-binding proteins have special DNA binding regions on them that are built from a limited number of well defined structural motifs like zinc fingers, leucine zippers and helix turn helix motifs. For example, the glucocorticoid receptor is member of a family of nuclear transcription factors that also includes the thyroid hormone receptor, the retinoic acid receptor, the vitamin D3 receptor and different steroid hormone receptors. All members of this family contain a highly conserved DNA-binding domain that consists of about 70 residues and that binds to activating elements of DNA, called hormone-response elements.

x) Signal transduction

A class of proteins called receptors are plasma membrane proteins that bind specific molecules, such as growth factors, hormones, or neuro-transmitters, and then transmit a signal to the cell's interior that causes the cell to respond in a specific manner. These responses are usually cascades of enzymatic reactions that give rise to many different effects within the cell, including changes in gene-expression. Interferences with these receptor-signalling systems can have drastic consequences like loss of cellular growth control leading to cancer.

1.2 Amino acids - the structural units of proteins

It is extraordinary that all proteins including those having intense biological and toxic effects are built from the same 20 amino acids which individually have little or no biological activity. The specific biological activity, on the other hand is determined by the 3-dimensional conformation of the protein which in turn is determined by the specific sequence of the amino acids in its polypeptide chain(s).

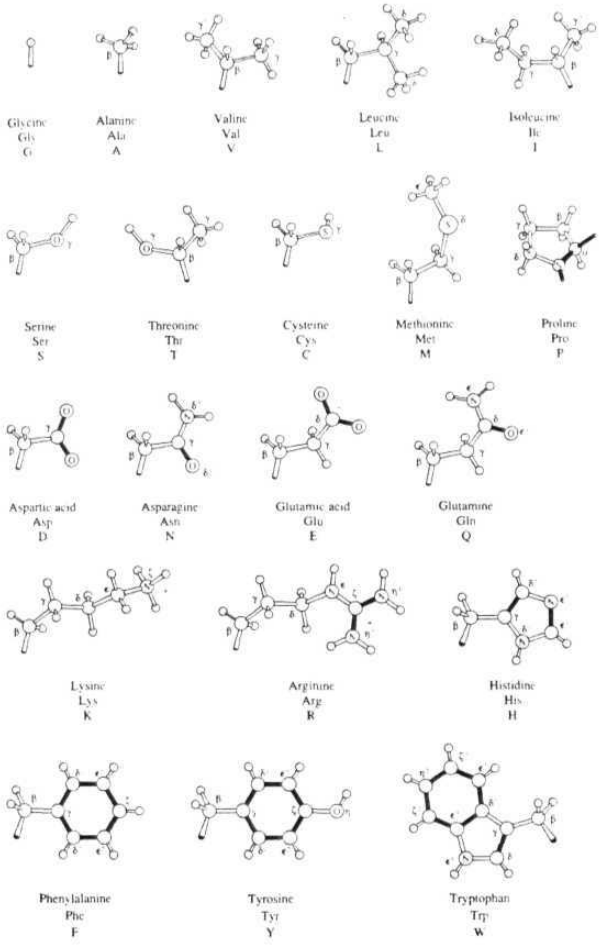

Fig. 1.1: The side-chains of the twenty amino acids that occur naturally in proteins. Double bonds and partial double bonds are black. Below the name of the amino acid are the three-letter and the one-letter codes commonly used. Isoleucine and threonine have asymmetric centres in their side-chains and here only the biologically used isomer is illustrated.

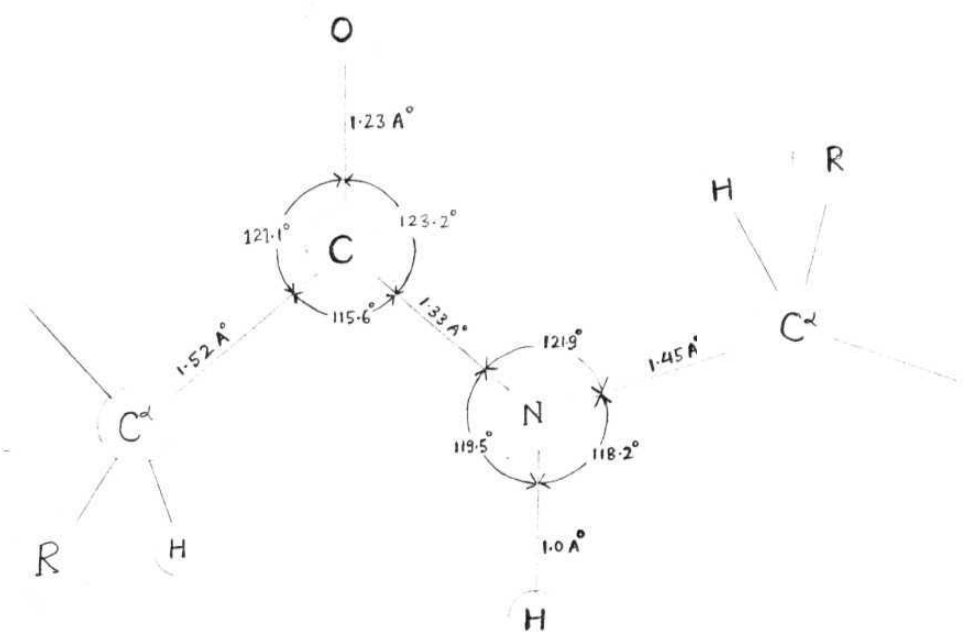

Fig 1.2: The geometry of the peptide backbone, with a peptide bond, showing all the atoms between two C^{α} atoms of adjacent residues. The peptide bond is shaded dark. The dimensions given are averages observed crystallographically in amino acids ans small peptides. (G N Ramachandran et al., Biochim. Biophys. Acta 359, 288-302, 1974).

Amino acids are derivatives of carboxylic acids. These amino acids are of 20 kinds and have the general structure as $R\text{-}CH(NH_2)\text{-}COOH$. R, is called the side-chain. The different amino acids differ only in their side-chains. The various side-chains are described in figure 1.1. The central carbon atom in the amino acids is asymmetric (except glycine) as a result of which there are two possible isomers of the amino acids, the D and L form. Only the L-isomer is seen in natural proteins (except in some proteins of some lower organisms). The amino acids in a protein sequence are linked to the successive amino acid by amide (peptide) linkages to form long linear chains. The peptide linkage is formed by condensation of two amino acids as shown in figure 1.2. Generally 50-3000 such amino acids are linked to form a protein molecule. However, the average number of amino acids in a protein is 250, (Mitra *et al,* 1994). The polypeptide backbone is a repetition of the basic units. Proteins differ from each other only on the basis of their lengths (number of amino acids) and the specific sequence of these amino acids.

1.3 Hierarchies of protein structure

The sequence of the amino acid residues along the covalent backbone of the polypeptide chain is called the primary structure. The secondary structure refers to a regular recurring arrangement in space of the polypeptide chains along one dimension. α–helices and p-sheet are the most common secondary structures. A region lacking any definite geometry of the secondary structure is referred to as a random coil or a coil. Tertiary structure refers to how the polypeptide chain is bent or folded in 3 dimension, to form the compact, tightly folded structure of the globular proteins. The importance of the tertiary structure is due to the fact that the tertiary structure is responsible for the biological activity of the protein. The independently folding unit in a tertiary structure is called a domain, which may be regarded as a supersecondary structure. These domains are usually combinations of α-helices and p-sheets (sometimes random coils also) organised in specific ways like four helix bundles, Greek key motifs, p-hairpins a/p barrels, *etc.* Unrelated proteins may possess similar structural domains. The quaternary structure may be seen in some oligomeric proteins where two or more polypeptide chains interact to be biologically functional. These individual chains are called subunits and they may be similar or dissimilar.

The folded, biologically active form of the protein at physiological conditions is called the native state. The primary structure of a protein sequence has the complete information for the folding pattern of the protein (Anfinsen,1973). Hence, in principle, it should be possible to derive the spatial structure from the primary sequence, without the x-ray analysis. For this statistical analysis of the primary sequences is required to be done *(vide infra).*

1.4 Role of mutations in protein diversity and evolution

A protein sequence differs from another sequence only in the number and kind of amino acid residues and their sequence of arrangement. In principle, sufficient number of changes in the sequence of any protein can change it into another protein. Any change in the primary sequence of a protein is called a mutation. A mutation is defined as an insertion, a deletion or a substitution depending upon whether the change is due to an addition, deletion or replacement of amino acid (at a specified site) in the protein sequence. The central question of protein evolution is how mutational change in amino acid sequence leads to change in the structure and stability, and thereby change in protein function. Protein mutation is a very slow process and occurs as a result of mutation in the gene coding for the protein. The process of evolution depends upon the interplay between two factors. The first is the opportunity provided by the random genetic events *(i.e.,* random mutations) and the other is the improvement in fitness as tested out by natural selection, where the protein is assessed in its natural environment in the organism. It is found quite often that an old protein structure or a piece of it is used in a new and unexpected way without destroying the existing useful properties of the molecule. Thus evolution is quite conservative in its structural aspects. A mutation in any of such conserved sites may be responsible for aberrant proteins. However, mutations are responsible for the evolution and diversity of proteins as well as for aberrant proteins. Some mutations increase the degeneracy of native states and quite often a mutation will be neutral, having no effect on the native state. But the practical effect of such increased degeneracy of mutations on native state of proteins would be decreased enzyme activity, since some of the mutation induced alternative native structures may not be active due to fatal changes in function though the native state is maintained. In conclusion we may say that a mutation may be beneficial, harmful, or neutral depending upon its site of occurrence.

1.7 Statistical analysis of primary sequences

With the accumulation of a large number of known protein sequences, in various databases (over 31 thousand sequences, in a recent release of Swiss prot protein sequence databank at the time of writing this thesis) and availability of good computing facilities it has become possible to study the primary sequences in detail.

A key question is whether protein sequences are random polymers of the 20 amino acid residues or there are some unknown rules governing their structures. Several protein scientists have studied this problem applying theoretical as well as practical approaches. Sorm and Knichal (1962) have applied a theoretical approach in order to find whether all structural similarities of peptidic fragments contained in the primary sequences are real or not while Williams *et al* (1961) used an experimental approach for the same. It has been found that even model structures drawn at random may frequently

exhibit regularities (symmetrical arrangement, repetition of sequences with interchange of individual elements, *etc.)* They argue that only a comprehensive mathematical treatment of all observed regularities within the structure of an individual protein and also between different proteins will make it possible to judge the significance of such regularities. Vonderviszt *et al* (1986) have shown that the abundances of the individual amino acids do not determine the frequency of di- and tri- peptide residues. The pair-frequency distribution of amino acids is highly asymmetrical and every amino acid has a characteristic sequential residue environment in proteins. Cserzo and Simon (1989); Meeta Rani (1990) have demonstrated that the pair preference characters are different for different pairs. (For a detailed account of the theoretical analysis, please refer to section 2.2 in the next chapter).

Studying all the possible preferences and range of preferences (over which these preferences are exerted) requires a rigorous analysis of primary sequences. The spectral analysis of the sequences is useful in such cases. We have applied a new method using the concept of fractals in order to study the primary sequences. The spectral analysis is capable of confirming the presence of fractal properties if they exist. Hence they have been also been applied independently in order to confirm our results based on fractal methods and make the work more complete. The details of the methods are described in the next chapters.

Chapter 2

Methods for studying proteins and their structures

Several elegant methods have been devised to study the properties, structure and function of proteins. The properties normally studied and determined are molecular weight, isoelectric points, sedimentation coefficient, number of subunits or chains, number of residues, the sequence of constituent amino acid residues and kinetics of the protein in case it is an enzyme and determination of 3-D structure. Of these, one of the most important task is the determination of the 3-dimensional structure. X-ray crystallography and NMR spectroscopy are widely used for this purpose although they are laborious and time-consuming. However, with the advent of computers in biological studies and accumulation of a large number of sequences and structure data in various data-bases theoretical studies on protein sequences, computer simulations and modelling have taken a great speed. With computer usage and programming there has also been borrowal, understanding and usage of concepts from mathematics and physics. One such concept has been that of "fractals". Fractal analysis is a relatively recent and useful analytical tool and has very successfully and satisfactorily been used for analysing natural phenomena, natural objects, *etc.* (Voss, 1988).

The first half of this chapter deals with the various methods, both experimental and theoretical, in studying protein structures. The later half of the chapter gives a brief account of a few reports of theoretical studies done on proteins and also introduces our method.

2.1 Methods for studying proteins

The final goal of all studies on protein structures and sequences is to be able to assign a specific role to each protein and the role of its primary sequences in that. It is also interesting to study how the evolution of various proteins is linked to the evolution of an organism. Apart from this another major aim of these studies is to be able to completely understand the principles of protein structure, function and genetics in order to design and engineer new proteins for the betterment of mankind by curing genetic disorders and defects due to aberrant proteins. Several methods have evolved to study the various aspects of proteins, *i.e.,* their physical properties and structure. Some of these methods have been briefly described.

2.1.1 Isolation and purification

A protein should be available in sufficient amounts for analysis. Hence isolation and purification are a must in order to carry out all the other tests and characterisations. The general sequence of analysis is as follows: Isolation

and purification, determination of molecular weight, isoelectric points, sedimentation coefficient, number of subunits or chains, number of residues, the sequence of constituent amino acid residues, enzyme kinetics in case of an enzyme, otherwise the relevant assay of protein activity and finally the determination of the 3-dimensional structure.

The methods used for isolation and purification are generally biochemical methods like isoelectric focusing and chromatography. For a detailed idea on protein purification methods please refer to Bailey (1976) or Haschemeyer and Haschemeyer (1973).

2.1.2 Molecular weight determination

Estimation of molecular weight (M_r) of a polypeptide is of central importance to the characterisation of proteins. Molecular weight determination by SDS-polyacrylamide gel electrophoresis is simple and requires very little sample material. The weight determined should be regarded as approximate. Molecular weight determination of polypeptides by SDS-polyacrylamide gel electrophoresis (PAGE) was first introduced empirically by Shapiro *et al* (1967) and was confirmed and extended by Weber and Osborn (1969) and Dunker and Rueckert (1969). The relative mobilities of the polypeptides are related to their molecular weights. Under appropriate conditions all reduced polypeptides bind the same amount of SDS on a weight basis (1.4 g SDS per g of polypeptide). Viscosity analysis suggested that the reduced polypeptide-SDS complex formed rod-like particles, with lengths proportional to the molecular weight of the polypeptides. This is not seen in case of unreduced polypeptides containing intact disulphide bonds as they do not bind the optimum amount of SDS and have different hydrodynamic volumes.

The amount of SDS bound by a protein is dependent on the structure of the protein (reduced or unreduced), the temperature and the ionic strength of the solution. The mobility of the SDS-polypeptide complex is dependent on its size and not necessarily on size or molecular weight of the original polypeptide. The electrophoretic mobility is proportional to the molecular weight of the polypetide only when th charge/mass ratio of all the SDS complexes are same. Hence only empirical molecular weight can be determined which can be later confirmed by gel filtration, sedimentation equilibrium or sequencing of the protein.

2.1.3 Immunochemical techniques

Next it can be compared with known proteins by immunochemical techniques. The extent to which antibodies (in some animal) produced against a known protein cross reacts with the antibodies against the test protein is an index of conformational similarity with the known protein.

2.1.4 Study of enzyme kinetics

If the protein is an enzyme, then the kinetic constants of the protein can be determined. Enzymes are one of the most remarkable biomolecules known for their extraordinary specificity and catalytic power that are far greater than those of man-made catalysts.

Enzyme kinetics can be studied by following the enzymatic reaction. In order to study the mechanism of the reaction, various useful parameters like order of the reaction, the specific reaction rate, the free energy of activation, Michaelis-Menten constant K M, the maximum initial velocity Vmax, *etc.* can be determined. Also the mechanism of inhibition by various inhibitors of the reaction can be studied.

2.1.5 Protein sequencing

The next stage is sequencing, *i.e.,* determination of the sequence of amino acid residues. This method was first employed successfully be Fredrick Sanger (1961, 1963) to determine the sequences of the two polypeptide chains of insulin. The steps involved in the process are described below in brief.

i) If the protein contains more than one polypeptide chain, the individual chains are first separated and purified.

ii) All the disulphide groups are reduced and the resulting sulfhydryl groups are alkylated to prevent reformation of the disulphide linkages.

iii) A sample of each polypeptide chain is subjected to total hydrolysis, and its amino acid composition is determined.

iv) On another sample of the polypeptide chain the N-terminal and C-terminal residues are identified.

v) The intact polypeptide chain is cleaved into a series of smaller peptides by enzymatic or chemical hydrolysis.

vi) The peptide fragments resulting from step v are separated and their amino acid sequence and composition is determined.

vii) Another sample of the original polypeptide chain is partially hydrolysed by a second procedure to fragment the chain at points other than those cleaved by the first partial hydrolysis. The peptide fragments are separated and their amino acid composition and sequence determined (as in steps v and vi).

viii) By comparing the amino acid sequences of the two sets of peptide fragments, particularly where the fragments from the first partial hydrolysis overlap the cleavage points in the second, the peptide fragments can be placed in proper order to yield the complete amino acid sequence.

ix) The positions of the disulphide bonds and the amide groups in the original polypeptide chain are determined.

2.1.6 Determination of 3-D structure

After the sequencing is done one can go ahead to study the 3-dimensional structure by x-ray diffraction of the crystal protein. The common methods of determination of the 3-dimensional structure are by NMR (Nuclear Magnetic Resonance) and X-ray crystallography. The knowledge of the sequence of the amino acid residues is a must for X-ray crystallography. The general topology of the polypeptide chain in solution can be determined by NMR though the structure obtained in this way is not as detailed and accurate as that obtained by X-ray crystallography, but NMR has the advantage of using a protein in solution rather than in a crystal lattice. The description of methods of 3-D structure determination is given.

2.1.6.1 NMR method

This is a technique for detecting atoms that have nuclei that possess a magnetic moment. These are atoms in which either the protons or the neutrons or both are odd in number. ^{1}H atom (one proton), ^{13}C and ^{15}N are some such isotopes of hydrogen, carbon and nitrogen respectively. Since these nuclei possess both spin and charge the nuclei behave like small magnets. In the presence of an external magnetic field, these nuclei can exist in two states: either in low energy state with their nuclear spin aligned parallel with the field or in a higher energy state with their nuclear spin aligned anti-parallel to the field. In order to change from the low energy to the high energy state the nuclei absorbs appropriate quantum of energy. In a magnetic field of several thousand gauss, such nuclei absorb in the radiowave region of the electromagnetic spectrum giving rise to a phenomenon called nuclear magnetic resonance. The exact frequency of the absorbed radiation from each nucleus depends on the molecular environment of the nucleus and is different for each atom unless they are chemically equivalent as the three hydrogen nuclei (protons) on the CH3 side chain of an alanine residue. These frequencies are obtained relative to a reference signal and are called chemical shifts. The nature, duration and combination of RF pulses can be varied and different molecular properties can be probed by selecting the appropriate combination of pulses.

In principle, it is possible to obtain a unique signal for each hydrogen atom except in case of those which are chemically equivalent. However, this problem has been bypassed by using 2D NMR spectroscopy. A COSY (correlated spectroscopy) experiment gives peaks between hydrogen atoms that are covalently connected through one or two other atoms, for example, the hydrogen atoms connected to nitrogen or carbon atoms within the same amino acid residue. A NOE (nuclear Overhauser spectroscopy) spectrum, on

the other hand gives peaks between pairs of hydrogen atoms that are close together in space even if they are from amino acid residues that are quite distant in the primary sequence.

The assignment of the observed peaks to hydrogen atoms in specific residues along polypeptide chain is done and a list of distance constraints is obtained. However, usually a set of possible structures rather than a unique structure is obtained. These structures are simply the different structures that are compatible with the data obtained. Notwithstanding these limitations about 60 protein structures have been determined by NMR methods

2.1.6.2 X-ray crystallography

The spacing of regularly repeating atomic or molecular units in crystals can be determined by studying the angles and intensities at which X-rays of a given wavelength are scattered or diffracted by the electrons that surround each atom. Atoms with the highest electron density, such as heavy metal atoms, diffract X-rays most and atoms with the lowest electron density (hydrogen atom) diffract X-rays least. X-ray analysis of crystal of salts like NaCl is relatively simple as only two different kinds of atoms are involved and they are regularly spaced. In principle X-ray analysis of very large and complex organic molecules like proteins is possible but the mathematical analysis of the diffraction patterns is very complex because large number of atoms in the molecule may yield thousands of diffraction spots.

In the early 1930s Astbury in England carried out the first pioneering x-ray studies of proteins. He found that hair, wool and other proteins of the a-keratin class are not fully extended but are twisted as they showed a periodicity of 0.5 to 0.55 nm along their long axes, while p-keratins were fully extended as they showed a repeat unit of 0.7 nm. When hair and wool were stretched after steaming they got extended and showed a diffraction pattern similar to the p-keratin class.

Pauling and Corey in the United States recorded the x-ray diffraction patterns of crystals of amino acids and of simple dipeptides and tripeptides and from them deduced the precise structure of the peptide bond. They used precisely constructed models to study all the possible ways of twisting or coiling the backbone of the polypeptide chain along one axis. Pauling predicted the structure of the α-helix which was soon confirmed by x-ray diffraction patterns. The first important breakthrough from X-ray studies came by J.C. Kendrew, (1961) who determined the 3-D structure on the myoglobin molecule (from sperm whale). The diffraction pattern of the crystalline myoglobin contained about 25,000 reflections from 2500 atoms. The analysis took place in two stages. In the first stage 400 diffraction spots were analysed. These results indicated how the polypeptide chain backbone is folded in the myoglobin molecule. In the second stage of analysis about 10,000 spots were studied and the sequence of the amino acids on the basis of the R-groups could be

determined. This tallied with the already known sequence of the protein. However, every protein has its own unique X-ray diffraction pattern and no truly general method for describing protein structures has developed. The determination of the 3-dimensional structure from sequence is still an empirical art based on various semiquantitative observations.

2.2 Theoretical studies and methods for studying proteins.

With the advent of computers and availability of sufficiently large data-bases of primary sequences, secondary structures, 3-D structure data as well as several other information, theoretical studies on protein sequences have become easier and as well as necessary to guide further experimental research. The theoretical studies on proteins have been performed on all the above mentioned data-bases and here a brief account is given on the various methods.

Simon (1986) has compared proteins as general crystals on the basis of several similarities: The protein native state corresponds to the crystal structure; the denatured state to the gaseous state and the molten globule to the liquid. The process of refolding is compared to crystallisation from a supersaturated solution or a supercooled liquid. Just as impurities in crystals decrease their melting temperatures, protein unfolding may be brought about due to replacement of an amino acid, chemical modification and ligand binding. Also, as in crystals, the conformation of the terminal amino acid is determined by the conformation of the previous amino acids.

Vonderviszt *et al* (1986) *et al* have shown that each amino acid residue has a characteristic residue environment within the polypeptide chains. Kolaskar and Ramabrahamam (1983); Cserzo and Simon (1989); Meeta Rani (1990) have shown that the di- and tri-peptide frequencies do not follow simple probability but have intrinsic preferences and repulsions.

Several other scientists have developed scoring matrices for comparing amino acids as well as protein sequences. Fitch and Margoliash (1967) have utilised the sequences of the protein cytochrome c to construct phylogenetic trees to trace the evolution of the organisms.

Regarding study of secondary structures, Blout *et al* (1960) have done pioneering studies and categorised amino acids as helix makers or helix breakers depending upon the influence on the helix. With the availability of protein tertiary structures determined from x-ray crystallographic data, several analyses involving the amino acid composition of protein secondary structures (α-helices, β-sheets p-turns, *etc.)* have resulted. For instance, Scheraga and co-workers (1968, 1969a, 1969b) used a data-base containing four proteins to ascertain the nature of helix-making and helix-breaking residues and a seven protein sample to indicate an asymmetric dipeptide distribution at helix-coil boundaries and examination of the helical and helix-coil transition region

(1975, 1976). Chou and Fasman (1974a, 1974b) analysed the amino acid conformational preference parameters for a given secondary structure and the residue compositions of the four positions of the b-turn regions (1977). Rose (1978) noted that non-hydrophobic character of turn-forming residues allows their prediction from protein primary structures. Lifson and Sander (1979) have examined the amino acid composition of parallel and anti-parallel p sheets.

Zhang and Chou (1992) have proposed a method called maximum component coefficient method, based on which the structural folding class *(i.e.,* whether it is all-α, all-β, $\alpha+\beta$ or a/p) of a protein can be determined from its amino acid sequence (but not the exact 3-D fold). Subsequent reports by Chou and Zhang (1993), Chou and Zhang (1993) claim improvement of method in self-consistency and extrapolativeness. Mao, Chou and Zhang (1994) have related the geometrical description of proteins to their amino acid composition but admit that the amino acid composition alone cannot determine the protein folding type in all cases.

Several methods have been devised to carry out theoretical studies on the 3-D structure prediction based on X-ray data and several other data. In principle, the atomic positions of amino acid residues can be calculated using *ab initio* quantum mechanical methods. In proteins however one must use classical approaches because of the large number of atoms that must be considered. Therefore an empirical force-field is generally used. The force-field function describes the potential energy of the system as a function of the positions of the atoms. The parameters of the equations are derived primarily from the results of *ab initio* quantum mechanics, spectroscopic data and crystallographic data. The force-field is calibrated by fine tuning of these parameters to reproduce structures and energy trends in various model systems. A central idea is that biologically relevant conformations are of low energy. Energy minimisation is accomplished by adjusting the positions of the atoms such that the change in potential energy is zero. For a detail information of the various methods of energy minimisation please refer to Allinger (1976). Unfortunately these minimisation routines search only the local configurational space of a molecule because the system moves downhill over the energy surface and stops at the first minimum energy conformation. The simulated annealing approach is being used to overcome this problem but not much success has been achieved in the prediction of the 3-D conformation of the protein based on the sequence alone.

Other popular methods include knowledge-based methods. They refer to studies in which portions of the sequence of protein of unknown structure are taken from a data-base of known structures. Kolaskar and Ramabrahmam (1983) obtained the conformational properties (potentials) of all the 400 pairs of amino acids from the crystal structure data of globular proteins. Analysis of these potentials showed that tripeptides of amino acid are not linear combination of the constituent amino acids. They suggested that the built in information from the tripeptide may be used in protein folding studies.

Applications of the studies of Ponder and Richards (1987) and work from Blundell's group (1987) are straightforward examples of this approach. Ponder and Richards have constructed a library of the distributions of χ angles found in high resolution crystal structures. Blundell and co-workers have compiled a similar data-base consisting of most of the probable side-chain conformations for each amino acid in different types of secondary structures. These approaches are rapid for determining starting conformations, but the resulting structures will generally need further refinement by conventional force-field methods. These methods are expensive methods considering requirement of greater computation time yet lack of sufficient accuracy and universality.

The 3-D profile method aims at assigning a sequence of amino acids to known protein folds (Bowie ef *al* 1991). In this method, a profile in the form of a table of values is computed from the co-ordinates of the 3-D structure. This profile measures the compatibility of amino acid sequences with the 3-D structure. High scoring sequences are likely to exhibit the fold of the profiled structure. Wilmanns and Eisenberg (1993) have described a modified 3-D profile algorithm that characterises the local environment in terms of the statistical preferences of the profiled residue for neighbours of the specific residue types, main chain conformations or secondary structure.

Proteins possess independently folding structural units called domains. While Vonderviszt and Simon (1986) have tried to predict domain boundaries from amino acid sequences, Zehfus and Rose (1986) have used the property of "compactness" (Z: compactness is defined as the ratio of the volume occupied to the area) to recognise continuous domains from X-ray data of proteins. Using a better algorithm, Zehfus (1994) has been able to identify discontinuous compact protein domains. Compact units in proteins are significant structural elements as the identification of compact domains in proteins provides an interesting new focus for protein folding experiments

2.2.1 Limitations of the popular methods

A method is successful if i) it is universal ii) it is robust iii) it is simple and does not require iot of computational time.

Current computational approaches to predicting the structure of a protein from its amino acid sequence are either based on energy minimisation methods or statistical methods. The former has limitations due to the local minima problem *{vide supra).* Although the simulated annealing approach is a powerful tool in overcoming this problem, still using these methods for prediction of the 3-D conformation has not materialised.

The statistical methods have reduced accuracy as reflected by large statistical fluctuations. Nevertheless, the statistical method has the merit of simplicity and convenience in application and have widely been used by biochemists as

described in the previous section. Besides, the results predicted by statistical method although less accurate, might reduce the scope of searching the conformational space or provide useful information for heuristic approaches.

Therefore, we conclude that "at present", there is no single method sufficiently accurate for relating protein 3-D structure from its sequence. However, combination of several methods like statistical methods and energy minimisation methods and the insight derived from the knowledge and incorporation of the results obtained from the past experiments of hundreds of scientists can go a long way in this endeavour.

2.2.2 Introduction to our method

It must be gratefully acknowledged that a lot of hard work is done by a scientist before he makes available the sequence of some protein. It is by the joint and simultaneous effort of thousands of scientists that we now have the sequence of several tens of thousands of sequences. We are thankful to all of them and we have used fractal and other statistical methods to characterise the immense wealth of information hidden in the data-base of sequences of proteins that are so easily available now. The computer is a very powerful and efficient tool for studying such huge data-bases. Complicated mathematical analysis becomes fairly manageable with a computer.

It is known that Euclidean geometry can describe only objects made by man (say in factories). It is not very useful in understanding and describing a variety of objects, phenomena and processes in nature. Hence mathematicians and scientists in several fields (including us) have been inspired by a relatively new yet convincing form of geometry called fractal geometry (it will be dealt with in more detail in the next chapter) and they have tried to explain several phenomena using fractal geometry. We have tried this method to study protein sequences occurring in nature.

However in order to confirm our results due to fractal analysis we have also carried out spectral analysis on protein sequences considering them as time-series. Spectral analysis is a classical concept and can be applied to deduce regularities and patterns that may not be very obvious otherwise. The spectral and fractal methods, though independent, are expected to supplement each other.

In case of all the tests we have carried out on the protein sequences, we have repeated the tests on simulated random sequences having same amino acid composition and length. And throughout, in all the tests the natural sequences markedly differed from the random sequences.

In the next chapter we give an introduction to the concept of fractals, spectral methods and time-series that have been described in their relevance to our approach. Our methodology along with flow-charts has also been described.

Chapter 3

Our methodology

3.1 Introduction

In this chapter an introduction to our methodology is given which is divided into four major sections. First section is on fractals, the second on fourier series and its spectra and the third on time-series. All these methods have been used by us for studying protein sequences hence they have been described in that perspective. The fourth section describes the overall layout of our methodology with flow-charts.

Section I: Fractals

3.1.1 Symmetry and patterns in nature

The patterns and symmetry of natural objects are difficult to be described succinctly by Euclidean geometry but can beautifully and reasonably accurately be described by a new form of geometry developed by Benoit B Mandelbrot (1983). This form of geometry he named "fractal geometry". While Euclidean geometry can describe only objects made by man (say in factories) fractal geometry has been very useful in understanding and describing a variety of objects, phenomena and processes in nature. In the words of Mandelbrot, "In the last decade, fractal geometry and its concepts have become central tools in most of the natural sciences: physics, chemistry, biology, geology, meteorology, and material science. Fractal images appear complex, yet they arise from simple rules. The computer rendering of fractal shapes leave no doubt about their relevance to nature and fractal geometry now plays a central role in the realistic rendering and modelling of natural phenomena in computer graphics". After the discovery of fractal geometry, strange but true, it appears that almost everything in nature has a fractal nature. In the words of Barnsley there are "fractals everywhere" and it is no accident.

3.1.2 Mathematical treatment of fractals

Mandelbrot defines fractals as sets for which the Hausdorff-Besicovitch dimension (Wallman 1941; Rogers 1970) strictly exceeds the topological dimension D_T. Topological dimension is rather simple in concept and corresponds to Euclidean dimension in classical geometry. According to Barnsley a fractal may be described as a "geometrically complicated subset of a geometrically simple set of points in space". In simple words, just like a set of points constitute a line, surface or space or any other object, fractals too are a set of points arranged according to some given rule or equation. Hence,

they are also geometrical entities and their shape depends upon the arrangement of points as characterised by the given rule or equation. Naturally, different rules or equations give rise to different fractals.

Nature exhibits a higher and also an altogether different level of complexity than it is possible to describe by Euclidean geometry and there are infinite number of distinct scales of lengths of natural patterns. Many patterns of nature are very irregular and fragmented (compared to standard Euclidean geometry) and this irregularity is characteristic and present at all levels of the object showing the pattern. Hence fractals are a family of shapes describing the irregular and fragmented patterns around us. The most useful fractals involve chance and both their regularities and irregularities are statistical. Also their shapes are scaling implying that their degree of irregularity or fragmentation is identical at all scales. The concept of fractal dimensions (Hausdorff-Besicovitch dimension) plays a central role in this. However it is useful to understand the characteristics of a fractal with simple examples. The length of arteries in the human body is very large and depends on how accurately it is measured. Similar is the case regarding the length of the branches of a tree. The coastlines of various countries are very large, and are similarly dependent on the degree of precision of measurement.

3.1.3 Fractal dimension

The fractal dimension gives an idea of the manner in which the fractal object occupies the metric space to which it belongs (Barnsley, 1988). Some fractal sets are curves or surfaces and others are disconnected dusts and yet others are so oddly shaped that there are not appropriate terms for them either in the sciences or the arts.

Fractal dimensions are numbers associated with fractals which give an idea about how densely a fractal object occupies the metric space to which it belongs. Dimension in the classical geometry always has an integral value. For instance, a point, line, plane and 3-D space have dimensions of 0, 1,2 and 3 respectively and are called their topological dimensions (denoted by D_T) Classical geometry cannot attach any meaning to an object with a fractional value of dimension but fractal geometry can and call them- fractals.

There are several kinds of dimensions associated with **a** geometrical object. In case of objects described by Euclidean geometry, all the *useful* dimensions coincide hence it is called a dimensionally concordant set, while in fractal sets all the *useful* dimensions do not coincide and it is called dimensionally discordant. Basic fractals are dimensionally discordant. On the basis of this fact fractals can be treated more rigorously. The topological dimension D_T is always an integer (Brouwer 1975; Menger 1943). The Hausdorff-Besicovitch dimension D need not be an integer. For all of Euclid D= D_T but in case of fractals the Hausdorff-Besicovitch dimension strictly exceeds the topological dimension. Hence, fractals follow the Szpilrajn inequality, *i.e.,* D >= D_T.

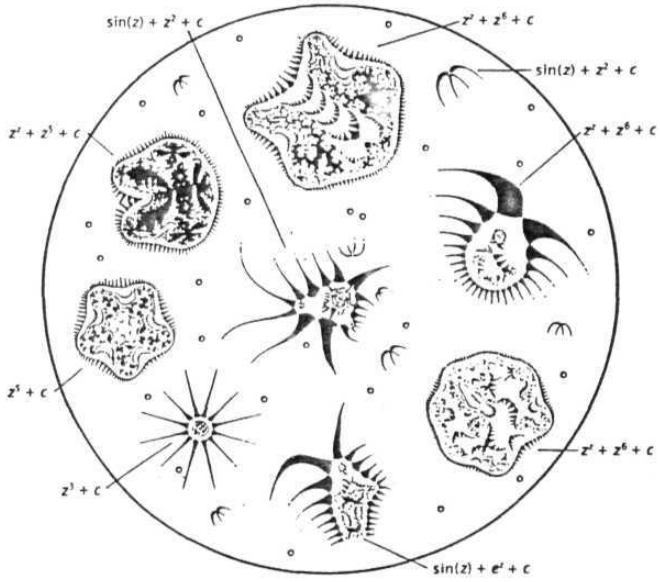

Fig. 3.1 *microscopic view of some biomorphs and their respective generating functions*

The term "fractal set" is rigorously defined but "natural fractal" designates a natural pattern that is usefully represented by a fractal set, e.g., Brownian curves are fractal sets and physical Brownian motion is a natural fractal.

Those sets in which the topological dimension $D_T = 0$, the value of fractal dimension D is greater than 0 (but less than 1), *i.e.,* D lies between 0 and 1. *E.g.,* the original cantor set. Here, D= log2/log3 = 0.6309, a value between 0 and 1. Natural fractals which can be represented by this set are clouds and dusts.

Fractal sets in which $D_T = 1$ and D is greater than 1 (usually between 1 and 2) are fractal curves and fractal lines. *E.g.,* the original Koch's curve. In this case, D = log4/log3 = 1.261. Natural fractals which can be represented by this set are coastlines of islands and circular DNA found in bacteria and plasmids.

However, a fractal may have an integral value of D so long as $D > D_T$, *i.e.,* the so called Szpilrajn inequality is obeyed. Consider a Brownian motion. When a Brownian trajectory is examined increasingly closely, its length increases without bound. The Brownian motion's trail is topologically a curve of dimension 1 but fractally it is of dimension 2 as it fills up the whole plane to which it belongs, hence fractals may have integral dimensions as long as the Hausdorff-Besicovitch dimension exceeds the topological dimension.

Fractal dimensions are important because they can be defined in connection with the real world data and can be measured approximately by means of experiments. For example, one can measure the fractal dimension of the coastline of an island. Fractal dimensions can be attached to clouds, trees, coastlines, feathers, network of neurons in the body, dust in the air at any instant, the distribution of frequencies of light emitted by a flower, the wrinkled surface of the sea during a storm, *etc.* These numbers allow us to compare sets in the real world with the laboratory fractals such as attractors of IFS (iterated function systems).

An Iterated function system (IFS) is a mathematical equation which is iterated (*i.e.*, repeated) over and over again. According to Barnsley (1988), with the help of IFS theory, we can describe a cloud as clearly as an architect can describe a house. IFS theory concerns deterministic geometry: it is an extension of classical geometry. It uses classical geometrical entities like affine transformations, scalings, rotations and congruences to express relations between parts or generalised geometrical objects, namely fractal subsets of Euclidean plane. Using only these relations IFS theory defines and conveys intricate structures (Kawaguchi, 1982; Smith, 1984; Oppenheimer, 1986).

Fractal dimensions of several natural fractals have been calculated.

Table I: A few natural fractals and their fractal dimensions

	natural fractals	fractal dimensions
1	coastlines of islands	1.2
2	membrane dimensions over a wide range	2.17
3	mitochondrial membrane (outer)	2.09
4	mitochondrial membrane (inner)	2.53
5	endoplasmic reticulum	1.72
6	folded surface of mammalian brain	2.73-2.79

3.1.4 Characteristics of fractals

i) Szpilrajn inequality

As already mentioned the Hausdorff dimension D strictly exceeds the topological dimension D_T in case of fractals. This principle is also called the Szpilrajn inequality. Usually, D is fractional but not always so (as in case of Brownian motion).

ii) Scaling behaviour (self-similarity)

The fractals show an invariance of shape or form at all scales. When a small portion of the fractal is enlarged it resembles the whole fractal. This scale invariance may be of a regular or an irregular form and it is called self-similarity. Self-similarity may be geometrically exact or statistically apparent. The self-similarity seen in all natural fractals is generally statistical self-similarity. The branching pattern of trees, neuronal networks and anatomical structures of several organs and organelles in plants and animals show self-similarity.

Another form of scaling behaviour called self-affinity also exists. In this case, there exists a statistical self-similarity but only when the x and y axes are magnified by different extents. Self-affinity has been shown in fractional Brownian motion. Fractional Brownian motion are special cases of Brownian motion (*vide infra).*

Hi) Infinite length and area

Since a fractal possesses detail at all scales its length increases with the size of ruler that is being used to measure (e.g., coastlines), hence it is said to have infinite length. Similarity a fractal surface due to the highly wrinkled or folded area has an infinite area.

3.1.5 Non-linear dynamical systems and the birth of fractals

Though fractals developed independently of non-linear dynamics the connections between them have been established. Basically any system that evolves with time is called a dynamical system. Such systems occur in all branches of science and in virtually every aspect of life. Even the simplest mathematical expressions, when interpreted as dynamical systems yield fractals, which is why fractals are so ubiquitous in nature.

The field of dynamical systems deals with physical and mathematical processes. The basic goal of this field is to predict the eventual outcome of the evolving process and the prediction of a process evolving in time on the basis of its past history. While some dynamical systems are predictable others are not. Even the simplest of dynamical systems depending upon only one variable may yield highly unpredictable and essentially random behaviour. Even the simplest of deterministic dynamical systems may behave unpredictably or randomly leading to a condition called chaos.

A dynamic process can be compared to an iterative process as both evolve with time. An iterative process is that which repeats over and over again. With each iteration the variables or parameters of the dynamical process assume different (or similar) values called iterates. The successive iterates of a point or a number is called the orbit (or trajectory) of that point. Some orbits are stable and others are unstable. Stable orbits are those in which with a slight change in the initial input, the resulting orbit behaves more or less similarly (all orbits of a function like $c(x) = \cos(x)$ are stable).

However, even the simplest of dynamical systems possess orbits which are unstable. An unstable orbit is one for which, arbitrarily close to the given initial input, there is another possible input whose orbit is vastly different from the original orbit, hence small changes in the initial input cause highly unpredictable behaviour. It is of great importance to understand the set of all points in a given dynamical system whose orbits are unstable. The set of unstable orbits in a given dynamical system may be large or small. It has been observed that many simple dynamical systems possess large sets of initial conditions whose orbits are unstable. The set of all points whose orbit are unstable is called the "chaotic set" (Devaney, 1988). Small changes of parameter may radically alter the structure of the chaotic set. Very often the set of points whose orbits are unstable form a fractal. So these fractals are formed by a precise rule: they are simply the chaotic set of a dynamical system.

Many orbits under iteration tend to land up on a particular set called an attractor as it attracts trajectories. The simplest kind of attractor is a fixed point. It describes a system such as a damped pendulum- that always evolves to a single state. The next most complicated attractor is the limit cycle. It corresponds to a system- such as an ideal frictionless pendulum- that evolves to a periodic state. Other attractors are simply called strange

attractors. They describe systems that are neither static nor periodic. The system described by a strange attractor is chaotic. The strange attractor has all the attributes of a fractal. The chaotic dynamics occur on the attractor set itself, *i.e.,* the strange attractor is the set of chaotic orbits or trajectories and forms fractals.

3.1.6 Brownian motion and fractal Brownian motion

Most useful fractals are those which allow chance. One such case is the Brownian motion. Brownian motion is a special case of fractals as it is simple, unstructured and shows lesser complexity. Most of the useful fractals are careful modifications of the Brownian motion. When a Brownian trajectory is examined increasingly closely , its length increases without bound. The Brownian motion's trail is topologically a curve of dimension 1. However, it is fractally of dimension 2 as it fills up the whole plane to which it belongs (in case of two dimensional random walk).

3.1.7 Fractals in nature and biology

We have just seen that fractal structures are often remnants of chaotic non-linear dynamics. Whenever a chaotic process has shaped an environment (the seashore, the atmosphere, geologic faults), fractals like coastlines, clouds and rock formations respectively are left behind.

Nature, and biology in particular abounds in fractals. Fractals are seen in the shape of clouds, meandering pattern of rivers, form of mountains, and in the anatomical structures of animals and plants. For instance, the branching pattern of trees, blood vessels, neuronal networks, the shape of corals, the folds of the intestinal villi, convolutions in the brain of mammals, shape of corals, the arrangement of the elements of vascular bundles, the convolutions in the brain of mammals, *etc.* Various intricate shapes of micro-organisms which cannot be easily described in classical geometry can be generated by iterated function systems (IFS) on a computer. Figure 3.1 shows shapes generated using IFS on a computer which very closely resemble morphology of microbes (Dewdney, 1989). An Iterated function systems (IFS) is a mathematical equation which is iterated over and over again. It uses classical geometrical entities like affine transformations, scalings rotations and congruences to express relations between parts or generalised geometrical objects, namely fractal subsets of Euclidean plane. Using only these relations IFS theory defines and conveys intricate structures (Kawaguchi, 1982; Smith, 1984; Oppenheimer, 1986).

3.1.8 Importance of fractal structures in the human body

In the human body, although the fractal anatomies serve apparently disparate functions in different organ systems, several common anatomical and

physiological themes emerge. Fractal branches or folds greatly amplify the surface area for absorption (as in the intestine, and in the convolutions of the brain), distribution and collection (in blood vessels, bile ducts and bronchial tubes) and information processing (by the nerves). Fractal structures partly by virtue of their redundancy and irregularity are robust and resistant to injury. These fractal structures in the human body result due to chaotic dynamics which offer many functional advantages. Chaotic systems operate under a wide range of conditions and are therefore adaptable and flexible. This plasticity allows systems to cope with the exigencies of an unpredictable and changing environment.

3.1.9 Proteins as fractals

In the previous chapter we have seen how difficult it has been to accurately describe protein structures using the classical methods just as it has been in case of fractals. This cannot be assumed to be sufficient condition that proteins might possess fractal properties, but it is worth trying to find out in case it is so. Proteins possess both regularities and irregularities. The extent of these elements at all structural levels as well as the advantages they offer are described. In fractals, both the regularities as well as the irregularities are characteristic. It seems likely that proteins possess fractal properties hence we can use fractal techniques to study them. Interestingly, Stapleton *et al* (1980) have performed electron spin relaxation measurements on low-spin Fe^{3+} on frozen aqueous solutions and crystals of some iron-containing proteins and concluded that they occupy a space of fractal dimensionality.

3.1.10 Symmetry in protein sequences:

All biological systems evolve and proteins have not been an exception. The evolution of any system can and must be given in terms of its invariant elements. A classical method is the formulation of differential equations, *i.e.,* a method of defining change in terms of what remains unchanged or invariant. Invariance is intimately related to the concept of symmetry which must not be understood in terms of pure geometrical connotation but in a wider sense meaning order within a structure, whether in space or time or ***in*** *abstracto.* In spite of its innumerable uses in physics and chemistry, the concept of symmetry has not been utilised widely and systematically in biology. The idea of relating symmetry and function is relatively new in biology. However, more of symmetry does not imply more order. For instance, lower organisms like coelenterates, radiolaria and some microbes exhibit a high degree of symmetry whereas man and mammals have only bilateral symmetry. Even though extreme complexity exists in biological order, it does express itself in some very simple and obvious symmetrical elements for instance in the structure of its proteins and genes. (Monod J, excerpts from Nobel Symposium 11)

a) Primary structure

All proteins are made up of the same 20 amino acids. In the primary sequence one does not find any significant translational symmetry except in case of proteins where 2 or more regions have arisen due to gene duplication. This produces pseudosymmetric structures with internal symmetry axes and these structures themselves have special functional advantages as they can combine the stability of a single chain with some of the structural flexibility of a multiple-subunit enzyme made of nearly identical units. The sequence similarities between different proteins can be found in case of proteins of the same family. Apart from these one finds several di and tripeptide sequences which exist in significant numbers.

b) Secondary structure

Symmetry at this level is most marked. The secondary structure consists of alpha helices, beta sheets, random coils. The helices are right handed regular helices. Some amino acids have a propensity to be present in the helix (helix formers) while some others break a helix by being present in it. Different side-chains have been found to have weak but definite preferences either for or against being in α-helices. Thus alanine, glutamate, leucine and methionine are good helix formers and proline, glycine, tyrosine and serine are very poor. Such preferences were central to all early attempts to predict secondary structure from amino acid preferences.

c) Tertiary structure

The tertiary structure of proteins possess neat folded patterns of alpha helices, beta sheets, coils, *etc.* but these structure cannot be described succinctly by Euclidean geometry and they are best described as sausage-shaped (at low resolution). But at higher resolutions, *i.e.,* atomic resolution, the structure looks chaotic (in the true sense!). However, in a statistical sense the tertiary structures too have symmetry. A small number of folding patterns describe in outline most of the known protein structures as simple combinations of a few geometric arrangements have been found to occur frequently in protein tertiary structures. These units are called super-secondary structures or motifs like, four bundle helices, helix-loop-helix, p-hairpin, greek key motif, p-a-p motif, *etc.* and based on the presence of these motifs, proteins whose 3-D structure is known, are generally categorised as all-α (cytochrome c, haemoglobin and myoglobin), all-β (elastase, α-chymotrypsin and concanavalin A), $\alpha+\beta$ (lysozyme and papain) or a/p (most enzymes like adenylate cyclase, all glycolytic enzymes, flavodoxin, *etc.)* folding classes (Levitt and Chothia, 1976; Richardson, 1981; Richardson, 1985a; Richardson, 1985b). This implies that folding patterns arises from

intrinsic general properties of polypeptide chains of secondary structures that form globular proteins. In proteins with quite different functions and no detectable evolutionary relationships, these features can be very similar. Such proteins share the same folding pattern and these folding patterns are simple and elegant.

d) Quaternary structure

For carrying out a biochemical process sometimes several similar or dissimilar independent folded units of proteins interact together. These groups of proteins form the quaternary structure. Since quite often similar units are present, they are arranged symmetrically. Oligomeric proteins possess several advantages as cooperative effects, extra stability and more number of clefts for ligand binding hence making the protein more efficient.

3.1.11 Analyses of protein sequences by fractal methods (using data bases)

Fractal methods have been used to analyse DNA sequences by a large number of scientists. Several of them report the fractal properties in DNA (Peng *et al,* 1992; Voss, 1992). Since proteins are linear translations of the DNA molecules one cannot be surprised if proteins too have fractal properties. A mini review by Nandy (1994) gives an account of some such attempts to characterise long DNA sequences by fractal methods. Soloyev *et al,* (1993) have classified functional regions of DNA and RNA based on fractal representation of sets (FRS). They have introduced a new measure of similarity for dividing sequences of different protein classes, such as α-actin, β-actin, globulins, a and p-interferons, *etc.* They claim that the FRS of exon and intron sequences are distinguished and the new measure can be used for searching functional regions in new genomic sequences.

However, in case of proteins for the first time Mitra and Meeta Rani (1993) have shown that the distribution of amino acid residues in proteins follow a fractal pattern and they are probably self-affine (Meeta Rani and Mitra, 1994). This also implies fractal properties in the higher structural levels of proteins.

Section II: fourier series and its spectra

3.2 Fourier series and their spectra

Though spectra and associated functions arose namely as measures of the closeness of the correlation between a time-series and certain harmonic terms they are also associated as transforms of the autocorrelation function.

Fourier considered the expansion of functions in series of harmonic terms of

the type

$$f(x) = \sum_{r=1}^{\infty} a_r \cos(r \cdot x) + \frac{1}{2} \cdot b_0 + \sum_{r=1}^{\infty} b_r \sin(r \cdot x) \quad \textbf{(1)}$$

Notwithstanding the cyclical character of the terms, a wide class of non-cyclical functions can be represented in this way over a limited range. It is sufficient that, in the range — π to + π, f(x) be single-valued, continuous except for a finite number of discontinuities, and have only a finite number of maxima and minima, for such an expansion to be valid. Such a series as shown above is called a fourier series. It has the attractive property that successive terms are orthogonal. For the interval – *n* to + π

$$\int \cos(r \cdot x)\cos(s \cdot x)dx \left\{= 0 \; if \; r \neq s\right\} and \left\{= \pi \; if \; r = s\right\}$$
$$\int \sin(r \cdot x)\sin(s \cdot x)dx \left\{= 0 \; if \; r \neq s\right\} and \left\{= \pi \; if \; r = s\right\}$$
$$\int \cos(r \cdot x)\sin(s \cdot x)dx = 0 \; for \; all \; r \; and \; s \quad (2)$$

On multiplying (1) by cos(rx) and by sin(rx) and integrating, it is found, for the interval – π to + ***n***

$$a_r = \frac{1}{\pi}\int f(x)\cos(r \cdot x)dx \quad (3)$$

$$b_r = \frac{1}{\pi}\int f(x)\sin(r \cdot x)dx \quad (4)$$

and the series (1) may also be written in the form

$$f(x) = \sum_{r=0}^{\infty} C_r \sin(r \cdot x + \phi_r) \quad (5)$$

where ϕ_r is the phase angle.

It appears, then, that a function may be expanded in a series of sines and cosines, the successive terms in (1) having periods 2π, 2π/2, 2π/3, *etc.* and the corresponding angular frequencies being, 1,2,3 with cycle frequencies 1/2π, 2/2π, 3/2π. More generally, when f(x) is defined over the interval 2L, the angular frequencies are typified by πr/L. Thus there is one fundamental frequency π/L and the others are integral multiples of it. Such a representation would be rather artificial it is known that f(x) was the sum of harmonic components with incommensurable frequencies. Hence a more general harmonic series may be considered as follows:

$$f(x) = \sum_{j=0}^{\infty} a_j \cos(\alpha_j \cdot x) + \sum_{j=0}^{\infty} b_j \sin(\alpha_j \cdot x) \quad \textbf{(6)}$$

where the a's can have any real values which leaves no simple way of evaluating a; and bj as in (3) and (4).

However, though the classical approach was to look for concealed harmonics the modern approach is to regard the spectrum as a characteristic of the time-series whether it is truly a sum of harmonics or not.

Section III: Time-series and protein sequences

3.3.1 Introduction to time-series

Time-series is the recording of events according to a horizontal axis in which equal intervals correspond to equal spaces of time. Though the characteristic feature of a time-series is that observations occur in temporal order, it can be extended to spatial situations equally well and the horizontal axis is the space or distance axis. The implication is that we are interested in the relationship of values from one term to the next, in serial correlation along the series.

When we have several series to consider as a multi-variable complex, there will arise a dimension of complexity beyond that of multivariate analysis. In the multivariate analysis we are concerned with the relationships or inter-relationships among variables regardless of the order in which they are presented to us. Whereas with a multi-variable time-series we have to consider, in addition, the correlations and cross-correlations among the series when one or more lead or lag behind each other.

The purpose of time-series analysis:

i) A simple series is taken and a simple system of a mathematical kind is constructed, which describes the behaviour of the series in a concise way.

ii) One can try to explain the behaviour of the series in terms of other variables and relate the observations to some structural rules of behaviour, *i.e.,* a model can be set up as a hypothesis to account for the observations.

iii) One can use the resultant analysis from i) or ii) to forecast the behaviour of the series in future. From i) we work on the assumption that even when we are unaware of the basic mechanism which is generating the series, there is sufficient momentum in the system to ensure that mostly the future behaviour will be like the past. From ii) we have, more insights into the underlying causation and can make projections into the future more confidently.

iv) From ii) we may require to control the system, either by throwing up

warning signals of untoward events that lie ahead or by examining what might happen if we alter some of the parameters in the model.

v) More generally, we may have to consider the joint progress through time/space of a number of variables, *i.e.,* our variable may be a vector of observations. In such a case we are approaching from the statistical angle, the more general subject of mathematical model building.

3.3.2 Analogy of a protein primary structure to a time-series

Protein sequences may be considered as examples of time-series due to two reasons. First, because an amino acid residue which is added earlier in time to the growing polypeptide chain occupies an earlier position in the sequence Hence the successive temporal addition of the amino acids is reflected in the positional sequence of residues in the protein sequences. Secondly, the techniques used to study time-series can be extended to spatial situations also, wherein one or more than one time like dimension is involved (Kendall 1976), as in the case of primary structures of proteins where the horizontal axis is the positional axis instead of the time-axis.

3.3.3 Analysis of protein sequences by techniques of time-series

We have considered the positional distribution (up to 200 positions) of each amino acid residue as a single time-series. Thus we have twenty series of positional distributions, one corresponding to each amino acid residue. Since proteins contain all the twenty amino acids, we also need to understand the position correlation between different kinds of residue. Since a multivariate time-series like situation is seen, these twenty series can be treated as a case of multiple time-series. By studying each series separately we can understand how each residue is correlated positionally to itself within the protein sequences.

3.3.4 Correlation analysis of primary sequences

In order to understand the positional correlations of the amino acids with respect to themselves as well as the other residues, correlation analysis was carried out by performing the serial correlation tests which can measures the dependencies between successive terms in time-series. Autocorrelation and cross-correlation functions are used for the serial correlation determination. The array of correlations of different orders obtained from the distributions are converted by fourier transformation into a spectrum. From such a spectrum for any given pair we can understand the nature of relationship between them. The positional distribution (normalised frequency) of an amino acid residue along protein chains (of the database) is taken as a series of independent values, to begin with. Any correlation can be detected and determined by the serial correlation tests.

Section IV: Layout of our methods

3.4 Flow charts

Programs for all our calculations were written in turbo-pascal version 6.0 and most of our graphs have been plotted using the plotting package sigmaplot version 4.01. The programs were run on an IBM compatible 486. The source code (in pascal) for our programs are not presented. Hence a descriptive flow-chart (instead of a schematic one) is presented. More details are given in the respective chapters.

3.4.1.1 *Section I: Analysis of the primary sequences of proteins and Fractal nature of the positional distribution of amino acid residues*

1).The probability of the 20 amino acids, the 400 pairs & 8000 triplets in data-base calculated as well as their normalised values are calculated.

2).The individual positional frequencies of the 20 residues upto 200 positions are calculated (noisy graphs seen).

3).Normalisation of the above frequencies performed (noises persist still). Noises could be due to:

i) statistical fluctuations

ii) could be intrinsic (neighbourhood attractions or repulsions among residues)

4). Application of Box counting test on the normalised positional distribution of each residue and calculation of their respective fractal dimension was performed. (The values of fractal dimensions lay in between 1.2 to 1.3)

5).Exactly the same test was performed on simulated 5034 random sequences with same amino acid composition. In random sequences the fractal dimensions of the distribution of all the residues was one.

3.4.1.2 *Section II: Autocorrelation analysis: discovery of long range correlations in the distributions*

1) Positional distribution of a residue in proteins is compared to a time-series. (In a time-series analysis one studies the occurrences of an event in time. Instead of time we consider the positions.). The analysis required:

a).Calculation of mean and variance of each series (positional distribution of a residue).

b) Calculation of autocovariances of order k. (k varies from 0 to 99).

c). Calculation of autocorrelations of order k as (autocovariance of order k) / variance of the series obtained.

d) Fourier transformation of the autocorrelations to obtain the spectrum for

each residue.

e) The spectra is nothing but the spectral densities on y-axis plotted against angular frequencies on the x-axis. (both on a logarithmic scale)

f) Deductions from the spectrum:

1) High peaks correspond to strong or intense correlations.

2) Presence of these high peaks in low frequency regions indicate strong long range correlations.

3) Presence of high peaks in high frequency regions indicate strong short range correlations.

4) Presence of high peaks in medium frequency regions indicate strong medium range correlations.

5) It was found that the spectra showed predominant high peaks in low frequency regions indicating strong long range correlations. It also showed peaks corresponding to short and medium range correlations.

6) For comparison, we generated random sequences (5034 number) having the same amino acid composition as the data-base.

7) Random sequences totally lacked long range correlations as clear from absence of the peaks in the low frequency regions

8) Hence real sequences were distinct from random.

9) The spectral exponent and scaling parameter was calculated from the spectra. On the basis of the values obtained we modelled the distribution of amino acids as fractional Brownian motion individual fractals as multi-fractals.

3.4.2 *Periodicities in protein sequences*

1) We treated protein sequences as time-series (justification in text)

2) Since the inter-relationship between various residues was being studied a multiple time-series is obtained.

3) Autocorrelation and cross-correlations are calculated to obtain coherence spectra.

4) One can obtain the periodicity from the coherence spectra.

5) We obtained the coherence spectra for the 400 pairs, 20 for homo-pairs and 380 for hetero-pairs.

6) In case of hetero-pairs coherences for both residues of a pair are symmetric hence only 190 spectra are required.

7) In homo-pairs the coherences are equal to one at all frequencies hence the squares of their spectral densities are used. The squares are used so that the order of magnitude of the spectra are same for both hetero and homo-pairs.

8) From the spectra we calculated the most common frequencies and tabulated them. Hence a knowledge-base of frequent periodicities of amino acid pairs are obtained.

9) The presence of the residue pairs in the knowledge-base at **the** corresponding periodicity was calculated and compared to their occurrence in 10,008 simulated sequences having same amino acid composition and lengths. The comparison was made by assigning scores to sequences on the basis of the presence of the pairs in the knowledge-base at the corresponding periodicity only. Real sequences always scored higher than random ones.

10) The construction of the knowledge base was based on only sequences having at least 200 residues. Hence we tested it exclusively on sequences of some protein families having less than 200 residues. For comparison we simulated 10 random sequences (with same length and amino acid compositions) for each sequence and considered the mean weight due to preferences. It was found that the real sequences always had higher weights than the random ones.

11) Deviation from randomness was tried to be related to higher values of weight due to pair preferences. On the basis we tried to predict the evolutionary' status of the 5 protein families with respect to each other.

3.4.3 *Entropy of protein sequences*

1) The distribution of pairs in protein sequences show preferences and repulsions; This has been tested and proved by us in the previous two chapters. The spectra (chapter 4, section 11) showed long, medium as well as short range correlations in protein sequences. While we used the presence of long range correlations to distinguish them from random sequences (because though random sequences lacked long range correlations, they did show short range correlations and sometimes medium range correlations). William *et al* have shown that random sequences can show short pseudosymmetric patterns.

2) However if a very large data-base is examined even on the basis of short range correlations one can try to distinguish between the pseudosymmetry of random sequences and true and meaningful patterns in real sequences. This is because the short patterns are often parts of long patterns and when they are made to overlap sequentially the long patterns may be reconstructed.

3) Boltzmann and Planck have said that entropy is a function of probability. The order of a system is associated with a number of possible states. When a system exists in a single state the probability of finding it in that state is one. As the number of states increase while the probability of finding it in that state decrease from one to zero, its entropy increases from one to large numbers. Hence the entropy of real sequences (which exist against the disordering effects of entropy) is expected to be higher than random sequences.

4) A relation between entropy and order has been proposed by Shannon and Weaver (1949) and Szilard (1925). To be informative, a message must have a pattern of order (DNA and protein sequences are informational molecules and have definite patterns). Edsall has proposed that production of information results in negative entropy (increase in order). This why protein sequences are expected to have higher entropy than their random counterparts of same lengths.

5) Using Boltzmann's formula (entropy S =R ln W, where W is the probability and ln is the natural logarithm function and R is the universal gas constant), we calculated the entropies of all the 400 pairs when they separated by 0.9 residues between them and stored the values in a 20x2()x 10 array. Analysis was done on a protein data-base having over 30,000 sequences and 10 million residues.

6) Entropy is an extensive property therefore the combined entropy (or mixing entropy) is

$$S_{AB} = S_A + S_B = \ln(Prob_A) + \ln(Prob_B) - \ln(Prob_{AB}),$$

where $Prob_{AB} = Prob_A \cdot Prob_B$

Hence the entropy of a sequence could be calculated based on

i) the types of residue pairs in the sequence

ii) the probabilities of these residue pairs in the data-base

7) The entropies of all the sequences were calculated and simultaneouly a random sequence having same amino acid concentration and length was simulated and its entropy was calculated.

A scoring weight due to entropy was assigned to the sequences for comparison. It was found that the entropy of the real sequences was mostly lower than their random counterparts. The difference was more marked for longer sequences. Both the weight and weights per residue in case of random sequences were higher.

8) Hence on the basis of entropy (in the statistical sense) we could distinguish between real and random sequences.

On the basis of our results we conclude that proteins may be regarded as a special subset of polypeptides and can be distinguished from other on the basis of

i) amino acid pair preferences in the sequences (chapter four).

ii) specific sequence patterns as shown by calculating periodicities in the sequences (chapter five).

iii) entropy of the protein sequences (chapter six).

3.4.4 *Summary of Results*

1 We have demonstrated that the positional distribution of any given amino acid in protein sequences are non-random and their distribution shows a fractal nature as found by applying the box-counting method for demonstration of fractal nature and the fractal dimension of the distribution of any given residue was found to be inversely proportional to its natural abundance. (Mitra C K and Meeta Rani)

2. Spectrum analysis of the distributions showed long range correlations. The spectral exponent and the scaling parameter suggest that the distributions are fractals exhibiting self-affine fractional Brownian motion. (Meeta Rani and

Mitra C K , unpublished results)

3. The protein sequences have been modelled as multi-fractals and the distribution of residue in them as self-affine fBm.

4. A knowledge base of dominant pairs with their periodicities has been constructed. The values have been tested and confirmed in the whole data base as well as sequences of some families.

5. A technique has been developed to calculate the entropies of protein sequences on the basis of short and medium range correlations. This has been extrapolated to be able to distinguish a natural sequence from a non-existing sequence. (Naturally occurring protein sequences are only minute fraction of the astronomical number of possibilities)

6 Finally we have shown that protein sequences are a special subset of the set of polypeptides (Mitra *et al*, 1994). This set shows fractal behaviour both in the individual sequences as well as the distribution of residues.

Hence this thesis reports a new approach of studying protein sequences and attempts to provide a model based on the results.

Chapter 4

Studying the positional distribution of the amino acids in proteins

4 Introduction

Since any given random sequence of polypeptide does not form a biologically meaningful protein (though it may have a unique 3-D structure), the secret of biological activity of proteins lies in their specific primary structures, *i.e.,* the specific sequential distribution of amino acids. Hence an important question is how many kinds of sequences can actually exist and how many completely different amino acid sequences might give a similar 3-dimensional structure for an average sized domain of a protein. It is estimated that between 10^{10} to 10^{12} proteins of different sequences exist in nature (Lehninger, 1986). This is a minute fraction of the sequence space *(i.e.,* the total number of theoretical possibilities). This means that from the sequence space, *i.e.,* the universal set of polypeptides, we obtain a small subset of polypeptides which can fold into stable compact structures; from this subset we get a still smaller set of polypeptides which have active sites and this subset constitutes the biologically meaningful proteins.

Unravelling the amino acid sequences of different proteins and relating each sequence to the properties and functions of the protein and the phylogenetics of an organism are one of the major objectives of contemporary biochemistry. Understanding the nature of primary sequences of proteins is useful in understanding the higher levels of structure, which decide their function, which ultimately run the machinery of the living cell. With the accumulation of a large number of known protein sequences in various data-bases (over 31000 sequences, in the most recent release of Swiss prot protein sequence databank) and good computing facilities it has become possible to study them extensively and try to tap the information in them and we have also tried to do that by devising and using somewhat unconventional methodologies

This chapter is divided into two sections. In section-I, we study the positional distribution of amino acids and find that it has a fractal nature. In section-II, we perform autocorrelation analysis of the distributions and discover long range correlations and self-affinity in them and on the basis of these results we have proposed a model for protein sequences.

4.1 Nature of the positional distributions

Proteins are linear biopolymers made up of twenty different amino acids. The linear sequence of amino acids in a protein molecule is called its primary structure. The primary structure folds to form a 3-dimensional, roughly spherical structure, but with a definite pattern that is called its tertiary structure. The primary structure of a protein sequence has the complete information for the folding pattern of the protein (Anfinsen,1973). As mentioned earlier, protein sequences are a minute fraction of the total number of theoretically possible sequences. Whether naturally occurring protein sequences are random polypeptides or there are some general rules yet undiscovered governing their sequences, has been an intriguing question since long. Both elements of randomness as well as deterministic properties have been discovered in the protein sequences and structures. Dose (1976) pointed out that the composition and primary structure of thermal protenoids differ from the composition of the reactant amino acid mixture indicating a non-random distribution of amino acids in the protenoids. Kolaskar and Ramabrahamam (1983) have shown that the distribution of pairs of amino acids cannot be regarded as linear combinations of frequencies of occurrences of the constituent individual amino acids. Ptitsyn (1984) suggested the presence of random elements by remarking that natural proteins are mainly random sequences which have been edited during the course of evolution to impart biological functionality, while others demonstrated deterministic (Vonderviszt et al, 1986) and regular elements (Cserzo and Simon,1986) as well.

Meeta Rani (1990) has shown that statistically various amino acid pairs as well as triplets are not present in the proteins of a data-base as expected by simple probability. The probability of a pair P(A,B) was calculated as

$$P_{(A,B)} = \frac{Total\ number\ of\ pairs\ AB}{Total\ residues - Total\ sequences}$$

and the probability of a triplet was calculated as

$$P_{(A,B,C)} = \frac{Total\ number\ of\ triplets\ ABC}{Total\ residues - 2 \times Total\ sequences}$$

We calculated the probabilities of the 400 pairs and the 8000 triplets as well as their normalised probabilities in the data-base. (Swiss Prot protein sequence databank, release 10.0, March 1989; vide infra). The normalised probabilities were obtained by dividing the pair or triplet by their individual probabilities of the members of the pair or triplet, i.e.,

Table I : Probabilities *1000 of the 400 pairs in the data-base.

	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
Ala	8.082	1.328	3.864	4.834	2.832	5.966	1.773	4.047	4.125	7.523	1.67	2.881	3.794	3.058	3.937	5.222	4.301	5.369	0.982	2.137
	(1.331)	(0.909)	(0.951)	(1.009)	(0.926)	(1.049)	(0.995)	(0.980)	(0.916)	(1.066)	(0.864)	(0.851)	(0.935)	(0.958)	(0.970)	(0.958)	(0.943)	(1.059)	(0.935)	(0.856)
Cys	1.299	0.607	0.991	0.891	0.772	1.466	0.552	0.856	1.019	1.573	0.334	0.883	1.091	0.835	1.015	1.428	1.07	1.148	0.266	0.672
	(0.889)	(1.725)	(1.014)	(0.772)	(1.049)	(1.071)	(1.286)	(0.862)	(0.940)	(0.926)	(0.719)	(1.084)	(1.117)	(1.087)	(1.040)	(1.088)	(0.974)	(0.941)	(1.055)	(1.119)
Asp	3.899	0.942	2.976	3.402	2.285	3.748	1.009	3.146	3.013	5.054	1.079	2.254	2.764	1.767	2.477	3.425	2.684	3.476	0.761	1.915
	(0.960)	(0.964)	(1.095)	(1.061)	(1.116)	(0.985)	(0.846)	(1.139)	(1.000)	(1.070)	(0.834)	(0.995)	(1.018)	(0.827)	(0.912)	(0.939)	(0.879)	(1.024)	(1.084)	(1.147)
Glu	5.018	1.015	3.599	5.428	2.195	3.888	1.253	3.432	4.207	5.611	1.367	2.791	2.38	2.593	3.435	3.426	3.268	3.858	0.803	1.897
	(1.047)	(0.880)	(1.122)	(1.435)	(0.909)	(0.866)	(0.891)	(1.053)	(1.184)	(1.007)	(0.896)	(1.045)	(0.743)	(1.029)	(1.072)	(0.796)	(0.908)	(0.964)	(0.969)	(0.963)
Phe	2.565	0.775	2.308	2.152	1.676	3.03	0.922	2.066	2.086	3.708	0.774	1.78	2.011	1.564	1.932	3.171	2.424	2.495	0.489	1.358
	(0.838)	(1.052)	(1.127)	(0.891)	(1.087)	(1.057)	(1.027)	(0.993)	(0.919)	(1.042)	(0.794)	(1.044)	(0.983)	(0.972)	(0.945)	(1.154)	(1.055)	(0.976)	(0.925)	(1.080)
Gly	5.462	1.335	3.818	4.312	3.023	6.234	1.69	3.939	4.586	5.889	1.472	3.145	3.54	2.928	3.814	5.233	4.399	4.749	0.99	2.463
	(0.960)	(0.975)	(1.003)	(0.960)	(1.054)	(1.169)	(1.012)	(1.018)	(1.087)	(0.890)	(0.813)	(0.992)	(0.931)	(0.978)	(1.003)	(1.025)	(1.029)	(0.999)	(1.007)	(1.053)
His	1.637	0.557	0.94	1.148	1.172	1.81	0.907	1.163	1.138	2.199	0.425	0.928	1.391	0.96	1.195	1.557	1.208	1.386	0.374	0.774
	(0.919)	(1.299)	(0.788)	(0.816)	(1.305)	(1.084)	(1.733)	(0.960)	(0.861)	(1.061)	(0.749)	(0.934)	(1.168)	(1.024)	(1.003)	(0.973)	(0.902)	(0.931)	(1.214)	(1.056)
Ile	3.973	1.044	2.946	3.048	2.047	3.434	1.264	3.075	3.104	4.606	1.067	2.612	2.809	2.157	2.73	3.895	3.445	3.368	0.676	1.711
	(0.962)	(1.051)	(1.067)	(0.935)	(0.984)	(0.888)	(1.043)	(1.096)	(1.014)	(0.960)	(0.812)	(1.135)	(1.018)	(0.993)	(0.990)	(1.051)	(1.111)	(0.977)	(0.947)	(1.008)
Lys	4.51	1.026	3.099	4.085	2.014	3.773	1.355	3.261	4.431	4.961	1.141	2.701	2.701	2.134	3.067	3.487	3.348	3.82	0.662	2.088
	(1.002)	(0.947)	(1.029)	(1.150)	(0.888)	(0.894)	(1.025)	(1.065)	(1.327)	(0.948)	(0.796)	(1.076)	(0.898)	(0.901)	(1.019)	(0.862)	(0.990)	(1.016)	(0.851)	(1.128)
Leu	7.238	1.497	4.636	5.628	3.424	6.129	2.039	4.568	5.409	8.969	1.884	3.847	4.729	3.858	4.914	6.799	5.717	5.681	1.067	2.591
	(1.025)	(0.881)	(0.982)	(1.010)	(0.962)	(0.926)	(0.984)	(0.952)	(1.033)	(1.093)	(0.838)	(0.977)	(1.002)	(1.039)	(1.041)	(1.073)	(1.078)	(0.963)	(0.875)	(0.893)
Met	2.372	0.353	1.387	1.599	0.938	1.686	0.507	1.214	1.741	2.063	0.629	1.192	1.115	0.885	1.329	1.804	1.526	1.54	0.289	0.668
	(1.227)	(0.758)	(1.072)	(1.048)	(0.962)	(0.931)	(0.893)	(0.924)	(1.215)	(0.918)	(1.022)	(1.106)	(0.863)	(0.870)	(1.029)	(1.039)	(1.050)	(0.954)	(0.863)	(0.840)
Asn	3.081	0.861	2.064	2.396	1.841	3.206	0.899	2.585	2.52	3.839	0.953	2.403	2.628	1.687	2.041	2.98	2.404	2.787	0.656	1.598
	(0.910)	(1.057)	(0.911)	(0.897)	(1.079)	(1.011)	(0.905)	(1.123)	(1.004)	(0.975)	(0.884)	(1.273)	(1.162)	(0.947)	(0.902)	(0.981)	(0.945)	(0.986)	(1.121)	(1.149)
Pro	4.087	0.843	2.595	3.595	1.854	4.672	1.177	2.276	2.671	4.195	0.866	2	3.685	2.24	2.58	3.861	3.051	3.518	0.754	1.55
	(1.007)	(0.864)	(0.956)	(1.123)	(0.907)	(1.229)	(0.988)	(0.825)	(0.888)	(0.889)	(0.670)	(0.884)	(1.359)	(1.050)	(0.952)	(1.060)	(1.001)	(1.038)	(1.076)	(0.929)
Gln	3.28	0.752	1.791	2.614	1.376	2.896	0.916	2.157	2.335	3.827	0.886	1.717	2.207	2.866	2.287	2.56	2.188	2.573	0.558	1.238
	(1.027)	(0.978)	(0.838)	(1.037)	(0.855)	(0.968)	(0.977)	(0.993)	(0.986)	(1.030)	(0.872)	(0.964)	(1.034)	(1.705)	(1.071)	(0.893)	(0.912)	(0.965)	(1.011)	(0.943)
Arg	3.917	1.015	2.735	3.202	2.161	3.633	1.274	2.74	2.892	4.861	1.105	2.195	2.621	2.221	3.811	3.288	2.643	3.308	0.735	1.701
	(0.965)	(1.039)	(1.007)	(1.000)	(1.057)	(0.956)	(1.070)	(0.993)	(0.961)	(1.030)	(0.855)	(0.970)	(0.967)	(1.041)	(1.405)	(0.903)	(0.867)	(0.976)	(1.048)	(1.020)
Ser	5.011	1.421	3.497	3.858	2.877	5.656	1.612	3.499	3.576	6.564	1.325	2.908	3.736	2.811	3.557	6.161	4.255	4.336	0.979	2.191
	(0.919)	(1.083)	(0.959)	(0.897)	(1.047)	(1.107)	(1.007)	(0.944)	(0.885)	(1.036)	(0.763)	(0.957)	(1.026)	(0.980)	(0.976)	(1.259)	(1.039)	(0.952)	(1.039)	(0.978)
Thr	4.467	1.213	2.892	3.181	2.428	4.39	1.32	3.285	3.013	5.438	1.202	2.383	3.395	2.232	2.64	4.135	3.802	4.272	0.867	1.954
	(0.979)	(1.104)	(0.947)	(0.883)	(1.056)	(1.027)	(0.986)	(1.059)	(0.891)	(1.025)	(0.827)	(0.937)	(1.114)	(0.930)	(0.866)	(1.009)	(1.109)	(1.121)	(1.099)	(1.042)
Val	5.127	1.276	3.742	3.923	2.531	4.186	1.471	3.434	3.774	6.099	1.355	2.825	3.389	2.394	3.192	4.617	4.279	4.664	0.872	1.992
	(1.011)	(1.046)	(1.103)	(0.980)	(0.990)	(0.881)	(0.988)	(0.996)	(1.003)	(1.034)	(0.839)	(0.999)	(1.000)	(0.897)	(0.942)	(1.014)	(1.123)	(1.101)	(0.994)	(0.955)
Trp	0.876	0.273	0.715	0.791	0.577	1.016	0.326	0.744	0.806	1.25	0.362	0.653	0.512	0.603	0.694	0.879	0.848	0.841	0.25	0.476
	(0.835)	(1.080)	(1.019)	(0.955)	(1.090)	(1.033)	(1.059)	(1.043)	(1.035)	(1.024)	(1.083)	(1.115)	(0.730)	(1.091)	(0.990)	(0.932)	(1.076)	(0.959)	(1.377)	(1.103)
Tyr	1.997	0.778	1.727	1.696	1.418	2.371	0.714	1.69	1.749	2.83	0.661	1.588	1.74	1.39	1.731	2.276	1.918	1.985	0.507	1.306
	(0.800)	(1.295)	(1.034)	(0.861)	(1.127)	(1.014)	(0.975)	(0.996)	(0.945)	(0.975)	(0.832)	(1.141)	(1.043)	(1.059)	(1.038)	(1.016)	(1.023)	(0.952)	(1.174)	(1.273)

normalised probability of a pair AB is given as

$$\frac{\Pr obability(A,B)}{\Pr obability(A) \bullet \Pr obability(B)}$$

In Table I we have presented the probabilities of occurrence of the 400 pairs as well as their normalised probabilities. The normalised probabilities are in parentheses just below the probabilities. A value close to or equal to one indicates a nearly random distribution of the amino acid residue. A value deviating considerably from one indicates intrinsic preferences or repulsions. A value greater than one indicates preferences or attractions between the members of the pairs and a value less than one indicates repulsions i.e., the pair is rarely seen. In most cases attractions and repulsions were seen.

The probabilities and normalised probabilities in case of the triplets have not been shown here due to their large number (16000 values).

A far greater deterministic element is introduced in native sequences by the forces of evolution and by the requirement of biological functionality. A sequence which is not meaningful in a biological system will either be eliminated or will never be found in nature. We also note that such deterministic forces necessarily must be "imperfect" because they allow for diversities and mutations. The next question that naturally arises is of the extent to which these two forces - the random and deterministic- influence the protein sequences. Recalling Ptitsyn's view that "natural proteins are mainly random sequences which have been edited during the course of evolution to impart biological functionality" and if it is so then it is expected that the more evolved proteins should have accumulated relatively more of such "editions" than the lesser evolved ones, and this probably explains the present non-random nature. We may therefore conclude that all naturally occurring proteins have random as well as deterministic elements. The more evolved proteins would have resulted from deterministic forces to a larger extent than from random and their evolutionary status can be correlated to this.

4.1.1 Proteins as random fractals

If we are to have a model for protein sequences in order to study them qualitatively, the "random fractal model" appears most appropriate. The most common fractals, the geometrical fractals, have no random elements present and are generated by repeated application of some simple geometric rules. Natural fractals, on the other hand, always have some random elements present and are therefore different from classical geometric fractals (Saupe, 1988). Therefore protein sequences have been considered by us as examples of random fractals, or more precisely as constrained random fractals, to stress the presence of deterministic elements (i.e., pair preferences, requirement of biological functionality, etc.). For instance, if a specific family of proteins is considered, then the random elements are responsible for the diversities and mutations within the family and the

deterministic elements impart to them the characteristic features by which they are recognisably members of the same family (i.e., same fractals) and can be characterised by their fractal dimension. By modelling protein sequences as random fractals it is possible to study them by using conventional fractal techniques. We have used the fractal dimensions as a parameter for the study of the twenty different amino acid residues naturally occurring in proteins.

4.1.2 Methodology

Our study of the distribution of amino acids in protein sequences is with a difference. Instead of taking a single protein or a group of proteins and studying their sequences (to explain their biological properties, activity, etc.) we have studied the positional distribution of each residue separately in the whole data-base of protein sequences and how they are effected by the distribution of other residues. This is expected to give the average affinities of amino acid pairs in the protein sequences. All our analyses have been done on the Swiss-Prot Protein Sequence Databank (SWISS-PROT). In the present analysis, the Release 10.0, March 1989 has been used as it was the most recent version available to us at that time. Subsequent analysis has been carried on version 26, 1994.

4.1.3 Protein data-bases

Protein data-bases contain the primary sequences of all the proteins which have been sequenced so far. There has been an explosion in this information in the last 4-5 years. Significant information regarding properties of a sequence which qualify it to be a biologically acceptable protein is expected to be hidden in them. Systematic and careful analysis is expected to throw some light in this matter.

We have selected the SWISS-PROT data-base for all our analysis. The SWISS-PROT is a composite data-base of protein sequences and is built by merging entries from PIR with translated entries from EMBL using a set of automated tools for maintenance and annotation. Some sequence data also comes directly from published literature. SWISS-PROT has been selected as it aims at

1) minimal redundancy
2) maximal annotation
3) integration with other data-bases

Nevertheless, no data base can be truly unbiased due to several reasons. The data-base contains all protein sequences that were available in the literature at that time. Naturally it contains some homologous proteins too. The question here arises is whether this is going to introduce any perceptible

bias in our analysis. Extracting a suitably curtailed dataset from the entire databank is a subjective exercise. A basic question at this point is whether the databank used can be considered as a random sample (of all known and unknown sequences) without replacement. This task has been attempted by Dolittle (1981). When a particular protein is sequenced, it is selected based on several considerations:

1) general interest in the protein
2) biological significance of the protein
3) ease of purification of the protein

and a host of other factors. General interest in a protein gives rise to certain homologies that are present in the data-bases. However their total number is relatively small and we do not expect that this will bias the data-base significantly. It is rather obvious that homologies that are present as a result of evolution cannot be considered as a source of bias in the data-base. The major cause of bias in the data-base we have used comes from fragments that should have been excluded. We feel, however, that screening of the data-base may introduce more bias than it is supposed to remove. Hence we have used the data-base without any modifications.

4.1.4 Formulae and Calculations

The modal value of the sequence length is close to 200 (about 250 actually; please see figure 4.1). The lengths of the sequences in the data-base were counted and plotted against their frequencies. The lengths were plotted on a logarithmic scale on the x-axis and the frequencies on the y-axis. A log scale was chosen so that a large range of sequence lengths could be clearly visualised. The minimum length was less than 10 residues while the maximum was greater than 4000 residues. The modal value was close to 250. Although equal class intervals in the logarithmic plot are unequal in a linear scale, nevertheless the position of the mode is not affected.

We have considered all sequences that had at least 200 residues. There were 5034 such sequences (which is slightly more than half of sequences of the data-base). Choosing a sequence length greater than 200 would reduce the number of sequences considered to less than half the sequences in the data-base. Hence, we settled on a length of 200. All residues beyond the 200th position were ignored.

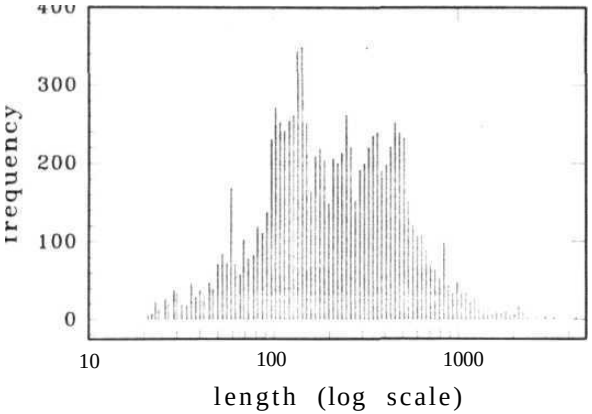

Fig 4.1: The distribution of lengths of the sequences in the data-base (10,008 sequences). On the x-axis are plotted the mid-points of the class intervals to which the sequences belong and on the y axis are plotted the frequencies of the corresponding classes A total of 100 classes we made The x axis has been plotted on a logarithmic scale and the class intervals are equal in logarithmic scale

The positional distribution of the twenty different residues (indicated by their standard one letter codes) were computed as follows:

$$P_j^i = \sum_k A_{jk}^i$$

If A_{jk}^i indicates the presence (1) or absence (0) of a residue of type *i* at the jth position for the kth sequence, the positional distribution is given as

$$P^i = \sum_{jk} A_{jk}^i$$

A_{jk}^i = 1 if residue / is present in the kth sequence at the jth position and A_{jk}^i = 0 otherwise. The resulting P_j^i array is stored in a 20x200 matrix and is called the positional distribution matrix.

This P_j^i matrix was normalised by dividing each element by the total number of sequences. The resulting array was stored in the same *Pj* array and was denoted as the normalised positional distribution matrix.

The total number of a residue of type *i* was obtained as

$$P^i = \sum_{jk} A_{jk}^i$$

where, j=1..200 positions, k=1..5034 sequences and the probability of finding a residue of the type *i* was obtained as

$$P(i) = P^i / (200 \bullet 5034)$$

These probabilities are shown in Table-II and also in figure 4.2 as a histogram

for visual comparison. This also gives us the abundances of various amino acid residues. These values agree with the values reported in the literature (Dolittle 1981; Vonderviszt 1986).

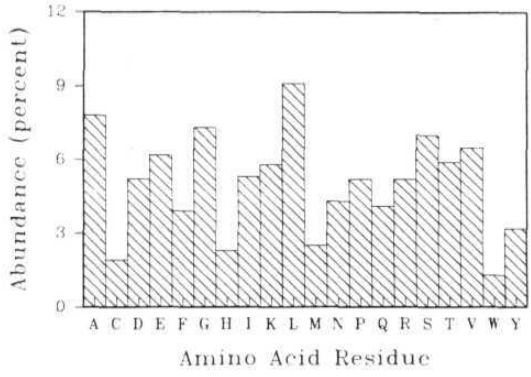

Fig 4.2: The probability of occurrence in percentages of the twenty amino acids in the sequences of the data-base (10,008 sequences and 2,932,613 residues). The amino acids are indicated by their one-letter codes The values obtained agree well with the literature values commonly available

Table II: Percent probabilities of the amino acid residues in the data-base

Amino acid	Probability	Amino acid	Probability	Amino acid	Probability
Ala	7.791	Asx	0.017	Cys	1.875
Asp	5.213	Glu	6.150	Phe	3.927
Gly	7.301	His	2.287	Ile	5.298
Lys	5.778	Leu	9.060	Met	2.481
Asn	4.344	Pro	5.207	Gln	4.099
Arg	5.208	Ser	6.996	Thr	5.855
Val	6.509	Trp	1.347	Unk	0.036
Tyr	3.203	Glx	0.017		

Each element in row /of the normalised positional distribution matrix was divided by $P(i)$ to obtain the relative normalised positional distribution and was stored in the same matrix P_j' where

$$P_j' = P_j' / P(i)$$

The relative normalised positional distribution matrix was used for computation of fractal dimensions for various residues. The P_j' values all lie in the neighbourhood of unity. A value >1 suggesting that a residue is preferred (over the average) at the particular position j and a value < 1 suggesting that

the particular residue is not favoured at that particular position. This however is to be understood in a relative sense; if a particular residue is strongly favoured at any position, the preference for other residues automatically is reduced at the same position.

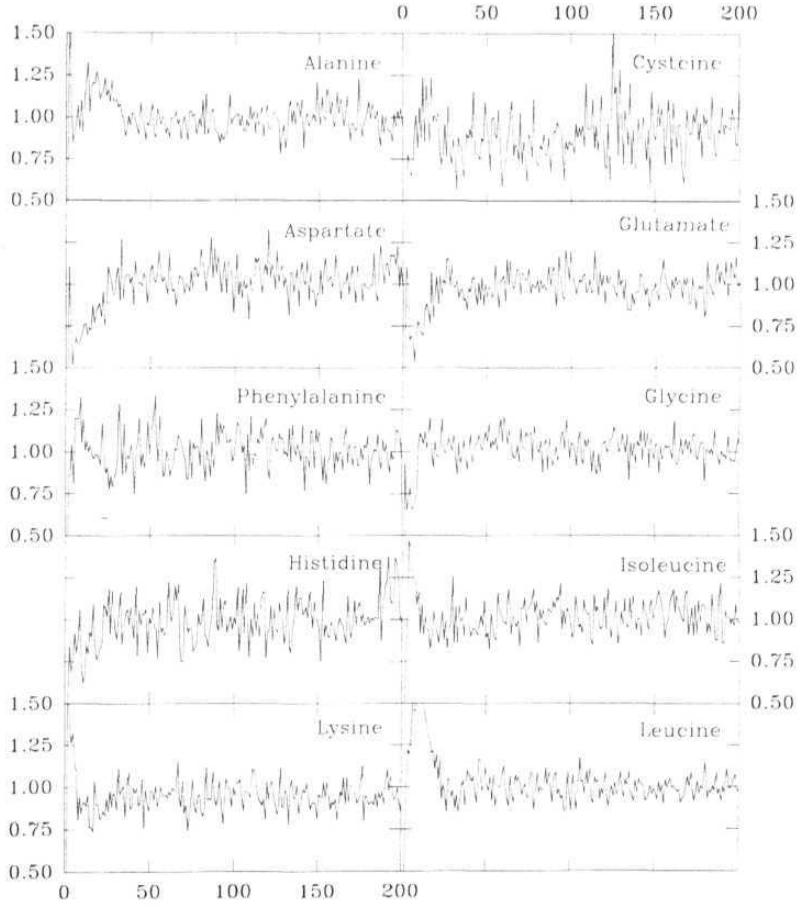

Fig 4.3(a) The positional distribution of first 10 amino acid residues (alanine to leucine). On the x-axis, are represented the positions from 1 to 200 and y-axis represents the normalised frequencies. The graphs appear noisy and resemble white noise However several residues show a greater positional preference at positions between 4 to 25 (Ala, Cys, Phe, He, Lys and Leu). At other positions these residues show a distribution of almost random appearance.

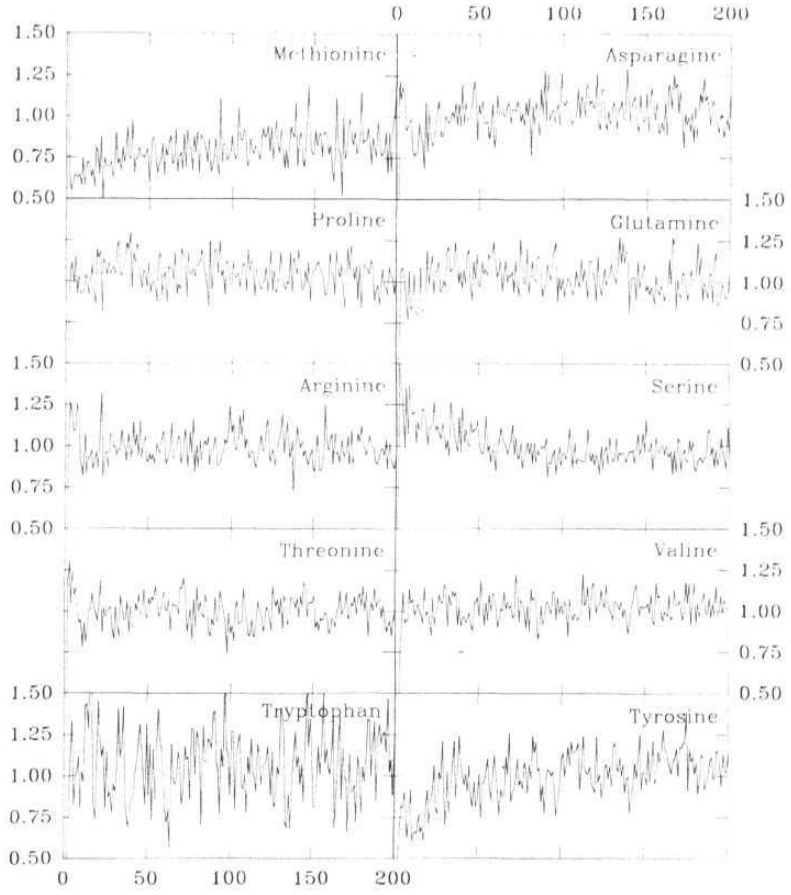

Fig 4.3(b) The positional frequencies of the next 10 amino acids (Methionine to Tyrosine). On the x-axis we have the positional distribution of the residues up to 200 positions. On the y-axis, we have the normalised frequencies at the particular positions. The method of normalisation is described in the text While valine shows less noise, tryptophan shows a highly noisy distribution.

For comparison, 5034 random chains were simulated by a Monte Carlo technique, using the probabilities (abundances) of various amino acids as given in figure 4.2; these sequences were also analysed using exactly the same procedure.

4.1.5 Box-counting algorithm

We have followed the box counting algorithm to determine the fractal

dimensions of the relative normalised positional distribution curves for the twenty amino acid residues. Although details of the principles are available in literature, a brief description has been made here.

Consider a set of points in our case, we consider the relative normalised positional frequencies of a residue) distributed in the plane. Divide the plane into 4^n (as illustrated in figure 4.4) into a number of square grids (taken for computational convenience as 41, 42, 43, 4^4 and 4^5) and count the number of boxes that include at least one point. The limiting slope of the straight line relating

ln(number of boxes containing a point) to n . ln(2),

where n is the order of subdivision (i.e., 1, 2, 3, 4 or 5) - gives the fractal dimension D of the set. Since our set is finite, subdivisions beyond 4^5 is not physically meaningful and hence were not carried out. In addition, instead of actually calculating the limiting slope, we measured the least square slope for computational reasons. Mathematically,

$$\lim_{n\to\infty} \frac{\log(boxcount)}{n \cdot \log 2}$$

where n is the order of subdivision or iteration and 'boxcount at iteration n' is the number of boxes that contain at least one point at a given iteration. Although we have used a least squares technique for calculating the limiting slopes, the correlation coefficients were very high (> 99%) indicating that within a physically meaningful scale, our values are meaningful. The dimensions calculated this way are fractional in nature as expected and are often called box dimensions. The box dimensions calculated for the random sequences are all equal to one. This is expected since in the absence of preferences, the distribution graphs are straight lines parallel to the position axis and a straight line has a dimension of unity.

The fractal dimensions obtained as above are shown in figure 4.5, for the twenty different amino acid residues.

The relative normalised positional distributions were studied by simple statistical tests for randomness. A simple test, the number of turning points (Kendall 1976), reveals that the distributions are random. Although this test cannot detect several kinds of non-randomness, it is useful to demonstrate the fractal nature of the distributions. Further proof of the fractal nature of the distribution is given by the scaling behaviour.

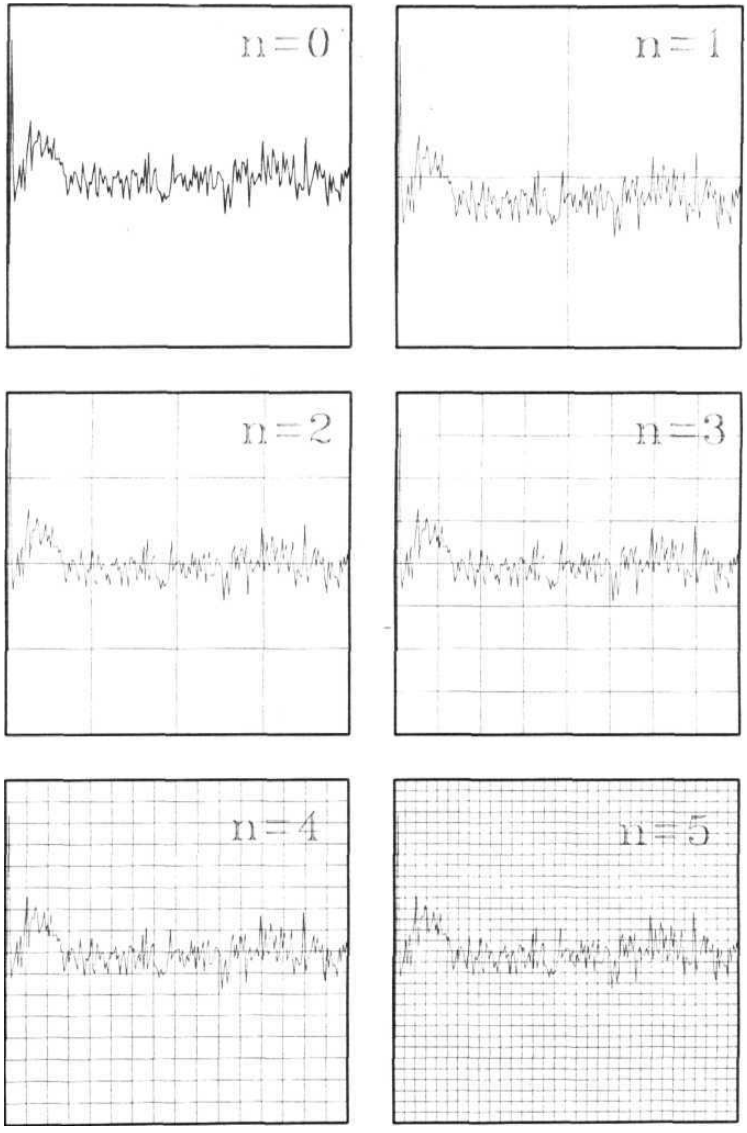

Fig 4.4: Illustration of the box counting algorithm with the example of the positional distribution of alanine. The plane containing the curve is divided into 4^n boxes of equal sizes. n is the number of iterations and boxcount at nth iteration is the number of non-empyt boxes at a given value of n.(Also see table-III for the boxcount values of all residues)

Table III: Boxcount values and the fractal dimensions of the distributions of the amino acids as determined by the box counting algorithm.

A.A	boxcount values at iteration nos					log(boxcount) at iteration nos:					D
	1	2	3	4	5	1	2	3	4	5	
Ala	4	10	19	50	114	1.38	230	2.94	3.91	473	1.199
Cys	4	10	26	78	**140**	**1.38**	2.30	3.26	4.36	4.94	1.322
Glu	4	9	21	55	124	1.38	2.20	3.04	**4.01**	482	1.220
Asp	4	9	18	52	114	**1.38**	2.20	289	3.95	4.73	1.252
Phe	4	9	22	59	129	**1.38**	2.20	309	4.08	4.86	1.273
Gly	4	9	18	48	116	**1.38**	2.20	2.89	387	**4.75**	1.213
His	4	9	20	61	137	**1.38**	2.20	3.00	4.11	4.92	1.296
Ile	4	9	20	59	124	**1.38**	2.20	3.00	4.08	4.82	1.262
Lys	4	10	21	55	119	**1.38**	230	3.04	**4.01**	4.78	1.225
Leu	4	10	20	49	110	1.38	2.30	3.00	390	4.70	1.185
Met	4	8	22	55	**125**	1.38	2.08	3.09	**401**	483	1.271
Asn	4	9	22	46	**124**	**1.38**	2.20	3.09	4.03	482	1.255
Pro	4	9	18	55	**132**	**1.38**	2.20	2.89	**4.01**	488	1.270
Gln	4	9	21	56	133	1.38	2.20	3.04	4.03	490	1.275
Arg	4	9	19	54	124	1.38	2.20	2.94	3.99	4.80	1.249
Ser	4	10	20	53	113	**1.38**	2.30	3.00	398	4.73	1.205
Thr	4	9	19	58	**124**	**1.38**	220	2.94	4.06	482	1.260
Val	4	9	18	52	112	**1.38**	2.20	2.89	3.95	**472**	1.214
Trp	4	13	36	88	154	1.38	2.56	3.58	448	504	1.329
Tyr	4	9	24	62	139	138	2.20	3.18	**4.13**	493	1 302

4.1.6 Fractal nature of distributions

we plot log(1-Pxx / Px) against Dx, the fractal dimension for residue x, where Pxx is the probability of finding a homodipeptide (xx) and Px is the probability of finding a single residue x, we obtain a straight line (figure 4.6). The straight line has a slope of 0.250 and an intercept of -0.341. The correlation coefficient is almost 90%. This suggests that we can write an empirical relation,

$$\log(1 - P_{xx}/P_x) = 0.250 \bullet D - 0.341$$

$$or,$$

$$1 - P_{xx}/P_x = 10^{(0.250 \cdot D - 0.341)}$$

$$= 0.4560 \bullet 1.778^{D}$$

$$or,$$

$$P_{xx}/P_x = 1 - 0.4560 \bullet (1.778)^{D}$$

This gives us a simple relationship between dipeptide and single residue frequencies, based on their fractal dimensions. This also shows the qualitative dependence of the fractal dimension on the abundance of the various residues.

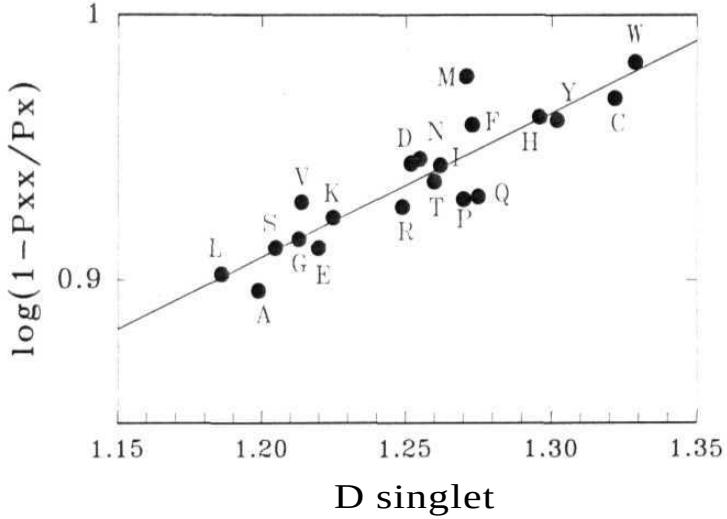

Fig 4.6. The plot of log (1-Pxx/Px), where Px is the probability of finding a singlet (x) and Pxx is the probability of finding a doublet (xx) where x is any residue and xx is its corresponding homodipeptide, as a function of Dx, where Dx is the fractal dimension of x as determined by the box counting algorithm. The points show a correlation of almost 89.9%

4.1.7 Results and discussions

Our results show that the distribution of a given residue along a polypeptide chain is fractal in nature. The fractal dimension D of the distribution is different for different residues but lies in the range of 1.18 to 1.35 (average 1.2), a range common in various naturally occurring fractal objects of Euclidean dimension of 1(Voss, 1988). We also observe a qualitative negative correlation between the probabilities of occurrence of various residues and their fractal dimensions. For example, Leucine (L) has the highest abundance (9.061%) and lowest fractal dimension (1.185) while tryptophan has the lowest abundance (1.347%) and highest fractal dimension (1.342).

The origin of this correlation is not clear at this moment. The abundances of various residues are determined by the genetic code and it is necessary to study the fractal nature of the nucleic acid. While these results had been communicated there have been reports of the fractal nature of DNA sequences (Voss, 1992). But the negative correlation is certainly not due to statistical effects, as studies using pseudo random sequences show a fractal

dimension of one when calculated using the same program, for all the residues. Thus the fractal dimension is most likely determined by intrinsic pair preferences.

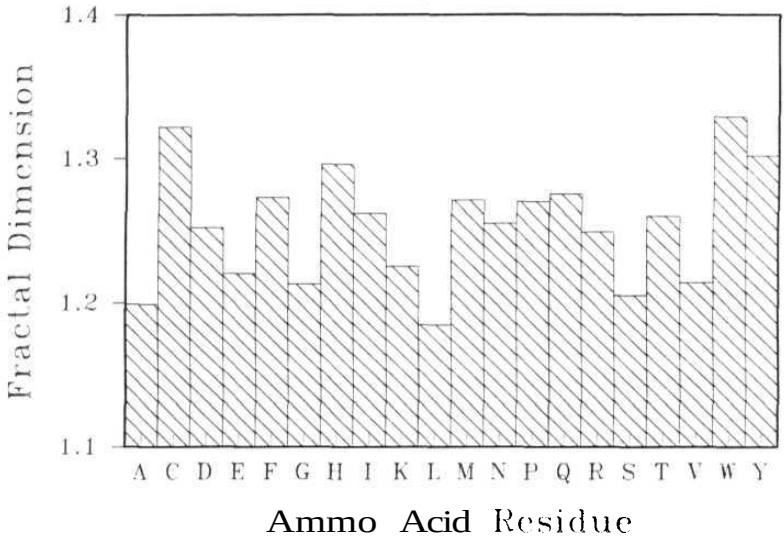

Fig 4 5: The Fractal dimensions determined by box counting algorithm, of the positional distributions of the 20 residues along 200 positions in 5034 proteins of the data-base. The amino acids are indicated by their one letter codes

The distribution graphs do not show any obvious end effects, except perhaps for methionine, which is abnormal. This is because many of the sequences in the data-base have been determined from cDNA sequences and the start codon and the methionine codon happen to be the same. However this materially does not effect the fractal dimensions determined for the positional distribution of methionine.

In the determination of the fractal dimensions of various residues, the graphs were "normalised" so that the relative abundances do not affect the general pattern of distributions. We have also shown, using pseudo-random sequences, that the abundances do not play any direct role in the fractal dimensions determined. Therefore we propose that the fractal dimensions so determined are indicative of the "intrinsic preferences" of various residues. This is apparent from the following justification:

A residue having a low box dimension (e.g., leucine) tends to be distributed more evenly (randomly) and has a "relatively" low preference for itself (1.093 times that expected for a completely random distribution). In contrast, tryptophan has a high box dimension and therefore does not distribute itself

evenly but has high preferences and repulsions. For example it has a high preference for itself (1.377 times than that expected in the case of a random distribution).

We therefore conclude that the fractal dimensions as determined above are indicative of the intrinsic preferences in a general sense and can be used to quantify such preferences for a residue in a native sequence.

The process of renaturation i.e., the formation of the correct 3-D structure from the random one suggests that 3-D structural information of the protein is already present in the protein in the primary sequence (Anfinsen, 1973). Primary sequences of proteins have been studied in great detail with the major objective of predicting the 3-D folding pattern of the final protein by several scientists (please refer to section 2.2). Regarding the secondary structure prediction the Chou-Fasman algorithm or its variants have been the most successful. But still no clear insight is available on the protein folding mechanism. Based on the ideas presented here it appears that fractal structure of primary sequences cause a fractal structure of the 3-D structure of the proteins also. Analysing proteins at all levels of structures by fractal techniques might throw more light on the mysterious relation between the primary sequence and the 3-D structure of the proteins. A point that we would like to stress is that by using the techniques of fractal interpolation it is statistically possible to predict sequences which are biologically meaningful. Since the doublet (homo-di-peptide) and singlet probabilities of the residues are seen to be correlated, it is intuitively obvious that such a possibility exists.

Summary : Statistical studies on positional distribution of amino acid residues in protein sequences of a large data-base showed noises. There was reason to believe that they were non-random noises (probably due to neighbourhood preferences and repulsions). Application of the box-counting method demonstrated fractal natures of these distributions showing a fractal dimension while simulated sequences showed a dimension of one as in case of a distribution free of any neighbourhood preference or repulsion. Fractal dimensions showed an inverse relation with respect to their relative abundances. The fractal dimensions were found to lie in the range 1.18 to 1.32 with an average value of 1.25.

SECTION-II

4.2 Autocorrelation analysis: discovery of long range correlations in the distributions

Having found that the positional distribution of amino acid residues in protein sequences follow fractal patterns, we decided to study the general patterns in them. It was expected that we would be able to distinguish at least in principle, the "allowed" primary structures in natural proteins. Short range

correlations have been observed by a number of scientists (Cserzo and Simon, Vonderviszt et al) and these mostly correspond to regular secondary structures of proteins. We attempted to find more universal features and probably these would be reflected in the form of long range correlations which probably affect protein folding.

As are interested only in the general differences between proteins (natural polypeptides) and random polypeptide sequences, our results are expected to show only their average properties.

Methodology

For our analysis, we have considered the distribution of amino acid residues along the sequential positions on protein chains as time-series and have studied them using techniques applied to time-series. (See 3.3 and 3.4 in the previous chapters). Hence a protein becomes an example of a multiple time-series as it is formed as a result of the joint progress of the twenty residues along its various positions. However in the present case we consider only the distribution of the residues individually.

4.2.1 Protein sequences as time-series:

A time-series is an event which changes with time or alternatively a time-series is the recording of events according to a horizontal axis along which equal intervals correspond to equal intervals of time. The techniques used to study time-series can be extended to spatial situations also, as in the case of primary structures of proteins where the horizontal axis is the positional axis instead of the time-axis.

4.2.2 Positional distributions of amino acids : multiple time-series.

We have considered the positional distribution of each amino acid residue as a single time-series. Thus we have twenty series of positional distributions, one corresponding to each amino acid residue. By studying each series separately we can understand how each residue is correlated positionally to itself within the protein sequences. Since proteins contain all the twenty amino acids, we also need to understand the position correlation between different kinds of residue. Thus these twenty series can be treated as a case of multiple time-series.

The sequence of the steps for analysis for each series is as follows:

1) Calculating mean and variance of each series.

2) Calculating the autocovariances (order 0 to 99)

3) Calculating the autocorrelations of order 'k' as autocovariance of

order 'k' / variance of the series.

4) The fourier transformation of the autocorrelations to obtain spectra
5) Analysis of spectra:
6) Corresponding the peaks to correlations and their periodicities.
7) Studying the spectra for general trends and patterns of correlations.
8) Applying exactly the same tests on random sequences (obtained by Monte Carlo simulations) and comparing them with real sequences.
9) Proposing a model to distinguish between the real and random sequences on the basis of the results obtained by autocorrelation analysis.

4.2.3 The serial correlation test : calculation of autocorrelation

A series of random numbers has no correlations i.e., it is uncorrelated from point to point. The positional distribution (normalised frequency) of an amino acid residue along protein chains (of the data-base) is taken as a series of independent values, to begin with. Any correlation can be detected and determined by the serial correlation tests. Autocorrelation and cross-correlation functions are used for the serial correlation determination. The autocorrelations gives the intra-dependancies and cross-correlations gives the interdependencies between two different series in case of multiple time-series.

In the present calculations we are studying only the positional autocorrelations. The correlation tests are performed on the positional distributions of the residues with respect to themselves. The total number of residues of type r, where r takes values from 1 to 20 and denotes the ordinality of the one-letter code of the amino acid residue and t denotes the positions from 1 to 200 along the protein sequences is denoted by U_t^r

Here *U* is a 20x200 matrix and each row in this matrix denotes **a** positional distribution for one of the twenty amino acids. Considering any residue r the autocorrelation is calculated from their autocovariances. The autocorrelations are calculated as follows:

The mean frequency = $E(U_t^r) = \mu^r$ (1)

The variance = $E(U_t^r - \mu^r)^2 = \sigma_r^2 = \mathrm{var}\, U^r$ (2)

The mean frequency is essentially the observed proportion in the data-base of the amino acid residue.

Let k denote a positional lag in the series U_i^r. The positional lag is allowed to vary from 0 to 99 (hence a maximum periodicity of 100 can be detected).

The kth autocovariance = $E((U_i^r - \mu^r) * (U_{i+k}^r - \mu^r)) = \gamma_k^r$ (3)

and kth correlation $\rho_k^r = \rho_{-k}^r = \gamma_k^r / \sigma_r^2$ (4)

The fourier transformation of the autocorrelations yields the spectral densities.

Spectral density is defined as $w^r(\alpha) = \sum_{k=-\infty} \rho_k^r \bullet e^{i k \alpha} = 1 + 2 \bullet \sum_{k=-\infty} \rho_k^r \bullet \cos(\alpha \cdot k)$ (5)

The x-axis denotes frequencies (in terms of the angular frequency a) and is plotted as $\log(\alpha)$. On the y-axis, the spectral densities are plotted as $\log(w(\alpha))$ vs $\log(\alpha)$ is called the spectrum.

4.2.4 Studying the spectrum

The spectrum reveals a good deal about the relationship between various positions occupied by a residue. The characteristics of this spectrum are related to several important parameters of the positional correlation of the residue being studied. The high peaks always refer to strong correlations and lower peaks refer to weak or no correlation depending upon the relative heights of the peaks in the spectrum. The position on x-axis corresponding to the peak in the graph refers to the periodicity of the residue. Hence presence of high peaks in the low frequency regions indicate long range correlations and their presence in higher frequency regions indicate short range correlations.

Since the graphs were noisy the true peaks were not clear. In order to find the true peaks we smoothened the curves using a spectral window of order nine (rectangular function). The smoothening process was as prescribed by Daniell (1946). Essentially, it involves taking an average of the nine values falling within the window and assigning the average to the middle point in the smoothened spectrum. The window is moved and average of all the points are thus obtained. A program was written in turbo-pascal for smoothening the spectra. The smoothened graphs were then analysed. As already mentioned the highest peaks referred to the most common frequency of occurrence. Broad peaks refer to a range of occurring frequencies. As already mentioned the peaks in the low frequency regions indicated long range correlations, those in high frequency region indicated short range correlations and those in medium frequency region indicated intermediate range correlations. The negative slope of this graph is called the spectral exponent and is indicated as P.

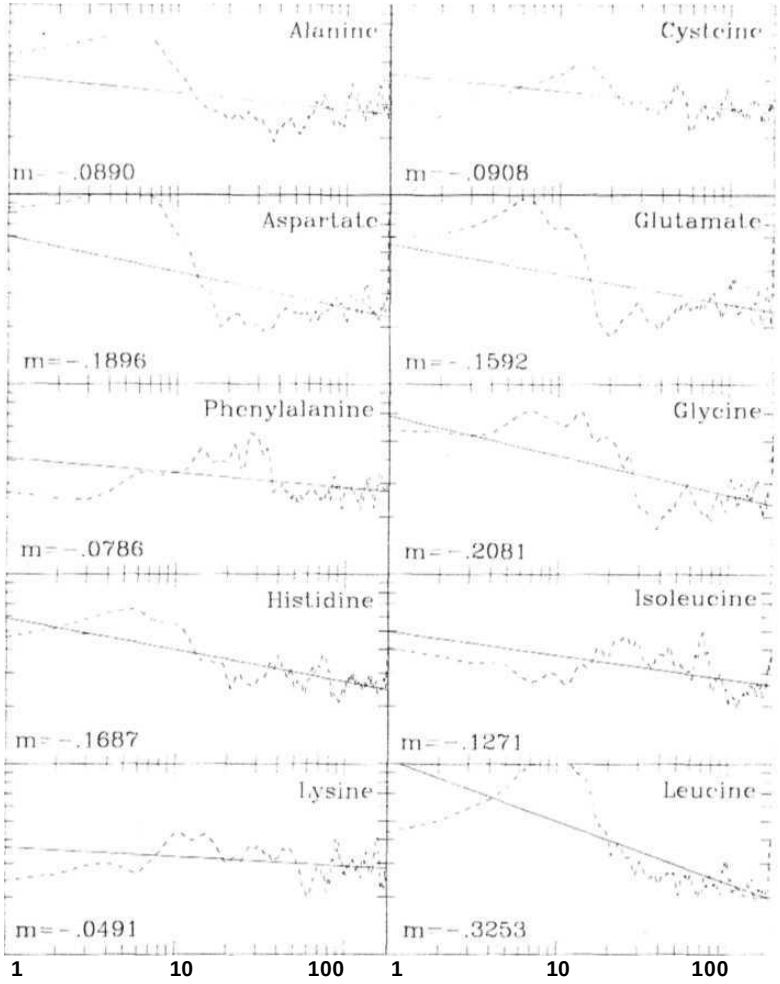

Fig 4.7(a) Real sequences: The logarithms of the spectral densities of the positional frequency distributions of the first 10 amino acids (alanine to leucine). The frequency data have been calculated from 5034 sequences of a data-base. The angular frequency is plotted on the x-axis on a logarithmic scale. The slope of the graph is designated 'm' and is indicated on the left hand side (bottom).

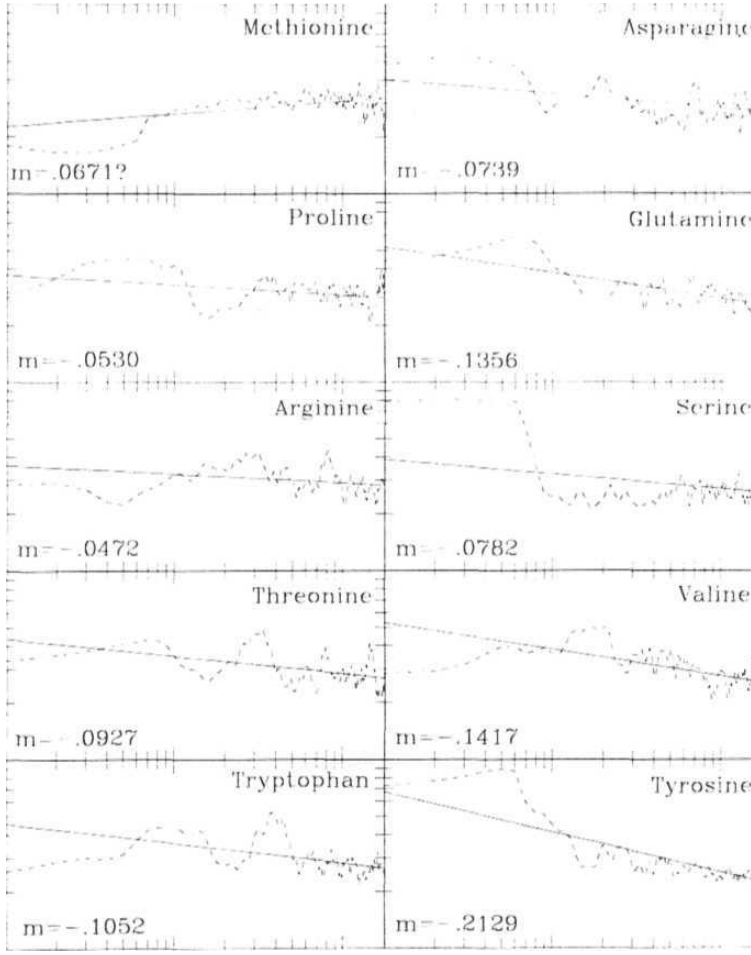

Fig 4.7(b): Real sequences: The logarithms of the spectral densities of the positional frequency distributions of the 10 amino acids (methionine to tyrosine). The frequency data has been calculated from 5034 sequences of the data-base. The angular frequency is plotted on the x-axis on a logarithmic scale. The slope of the graph is designated 'm' and is indicated by on the left hand side (bottom). The name of the amino acid is also indicated on the right hand side (top).

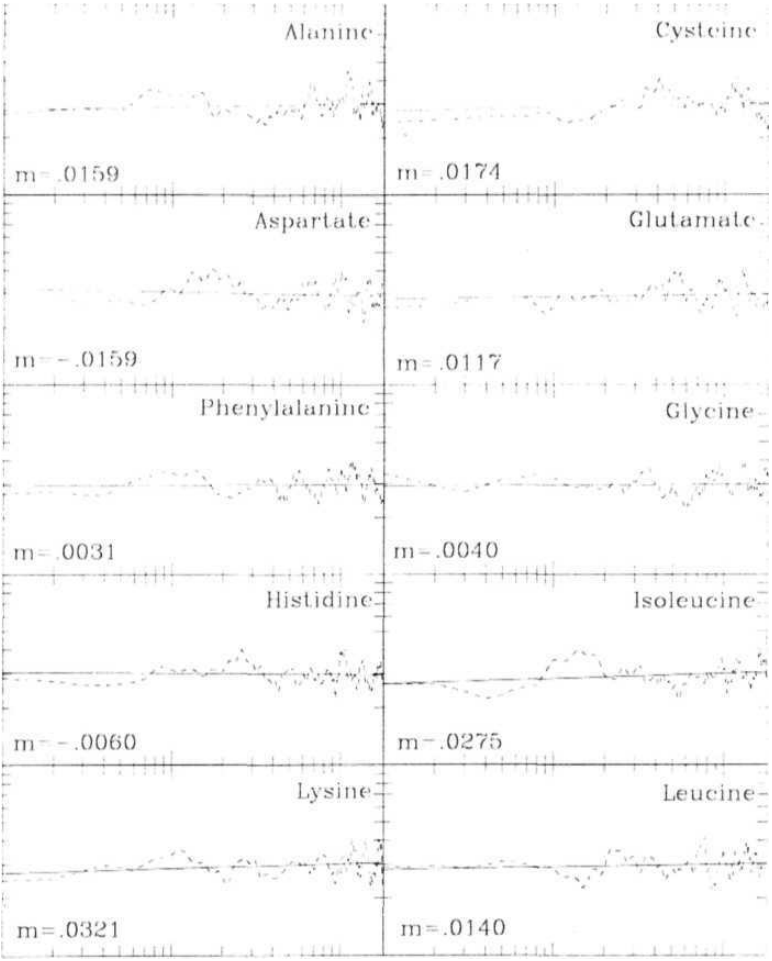

Fig: 4.8 (a): Random sequences: The logarithms of the spectral densities of the positional frequency distributions of the first 10 amino acids (alanine to leucine). The frequency data have been calculated from 5034 simulated random sequences having the same amino acid composition as that of the data-base. The angular frequency is plotted on the **x-axis** on a logarithmic scale. The slope of the graph is designated '**m**' and is indicated on the left hand side (bottom).

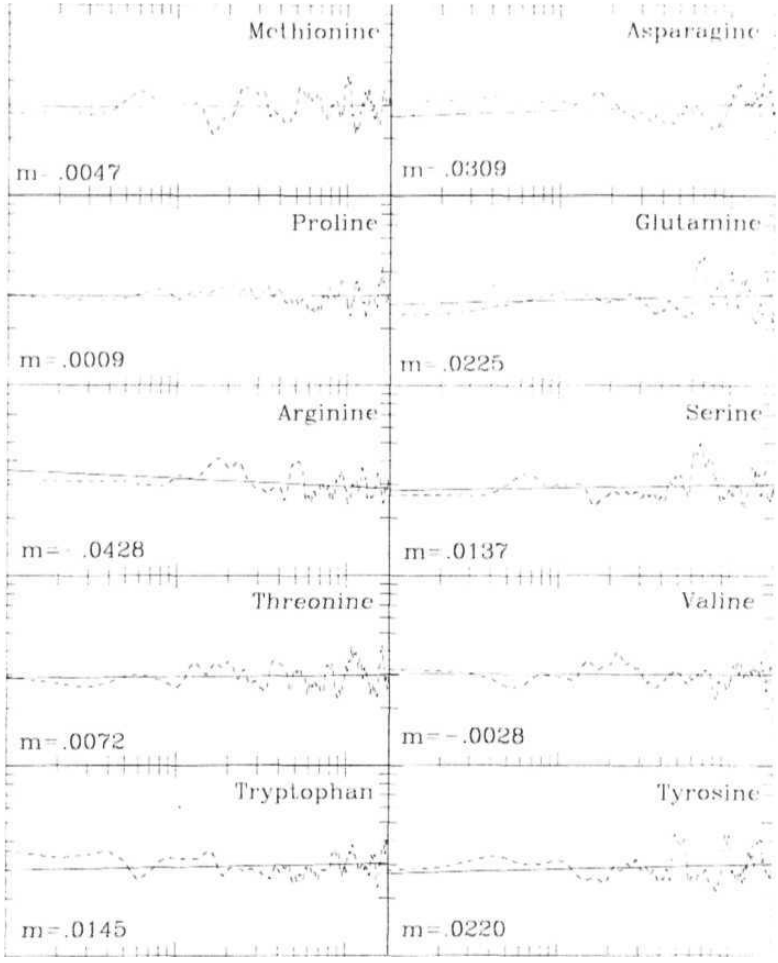

Fig 4.8(b): Random sequences: The logarithm of the spectral densities of the positional frequency distributions of the 10 amino acids (methionine to tyrosine). The frequency data has been calculated from 5034 sequences of the data-base. The angular frequency is plotted on the x-axis on a logarithmic scale. The slope of the graph is designated 'm' and is indicated by on the left hand side (bottom). The name of the amino acid is also indicated on the right hand side (top).

4.2.5 The spectral exponent β and the scaling parameter H

In a typical case, one does not expect to find only one periodicity (or frequency). In such cases multiple peaks are seen. In our case a large number of peaks are observed with varying intensities. The spectra of a

completely random case contains all frequencies in equal amounts and p is expected to be a horizontal straight line. In all other cases, the frequencies are present in unequal amounts and the spectral exponent p measures the distribution of various frequencies. In other words p measures the decrease of the intensity with frequency on a double logarithmic plot. The spectral density varies inversely as the spectral exponent p.

Graphically, the spectral exponent p is the negative slope of the graph of the spectrum plotted as log(spectral density) vs log(frequency). The spectral density p is related to the scaling parameter H. The scaling parameter reflects how the distribution function changes when the independent variable is scaled (i.e., by multiplying with a given constant r). The scaling parameter H characterises the scaling behaviour of fractal traces, random processes, noises etc. Defined explicitly, if we consider a distribution function $y = f(x)$, then for a scale r for x (i.e., x:=r.x) y scales $r^H.y$ (i.e., $y := r^H.y$). For a white noise H is 0.5. Qualitatively, it gives an idea about the correlations in space (or time, as the case may be). In a random walk, the mean distance travelled scales as the square root of the number of steps taken. H lies between 0 and 1. H and p are related by the simple equation H=(1-p/2) for a distribution with a Euclidean dimension of unity. So H=0 and H=1, correspond to the cases when p=1 and p= -1. Thus both p and H provide important information about positional correlations of amino acid residues in protein sequences. One can find nature of distributions, whether they are linear, random or fractal.

The scaling parameter H of a random Brownian motion which has no correlation from point to point is always 0.5. If H > 0.5 then it indicates a positive correlation for the increments of the noise and if H < 0.5 then it indicates a negative correlation between the increments of the noise. The value of H for the positional distributions of the residues were lower than 0.5 hence indicates negative correlations. On the basis of non random distribution patterns with the values of the spectral exponent as well as scaling parameter lying in the range of a fractal object we have modelled the distribution of residues as fractional Brownian motions. A fractional Brownian motion or fBm differs from a ordinary Brownian motion as it contains some correlations. The nature and degree of correlation is described by the spectral exponent and the scaling parameter.

4.2.6 Results

The graphs of the twenty spectra one for each amino acid are presented in figure 4.7(a) and 4.7(b). The graphs of the spectra for the amino acids as in the case of random sequences are presented in figure 4.8(a) and figure 4.8(b). Small values of p suggest that both low and high frequencies are equally probable (as seen in case of all residues except methionine). As already mentioned earlier, methionine showed an anomalous distribution at the first position due to the fact that the start codon matches the codon for methionine. (This also effected all calculations based on the positional

distributions of methionine). High values indicate that high frequencies are far less common compared to low frequencies.

In case of the real sequences the spectral exponent b was found to lie between 0.33 and 0.05 for all residues (except methionine, which had a p value of 0.-07). b has a significant non-zero value compared to the random sequences. The average value of p turns out to be 0.13 (excluding methionine). Presence of high peaks in the low frequency regions indicates long range autocorrelations. This is also apparent from the graphs 4.7(a) and 4.7(b).

In case of simulated random sequences p values are in the range -0.043 to 0.03. Since the values are for random sequences at the expected value is zero and the average slope turns out to be 0.007 (as expected). Also these spectra showed a lack of any long range correlation in any residue distribution as indicated by lack of high peaks in the low frequency regions. This is clear from the graphs 4.8(a) and 4.8(b).

The scaling parameter H, in case of all the residues for real sequences was found to be in the range of 0.48 to 0.34, with an average of 0.45 (including methionine). These values of H also indicate long range correlations. The scaling parameter in case of random sequences was in the range of 0.048 to 0.52 (with an average of 0.5), as expected in case of random noises. For a completely random sequence, H is expected to be 0.5.

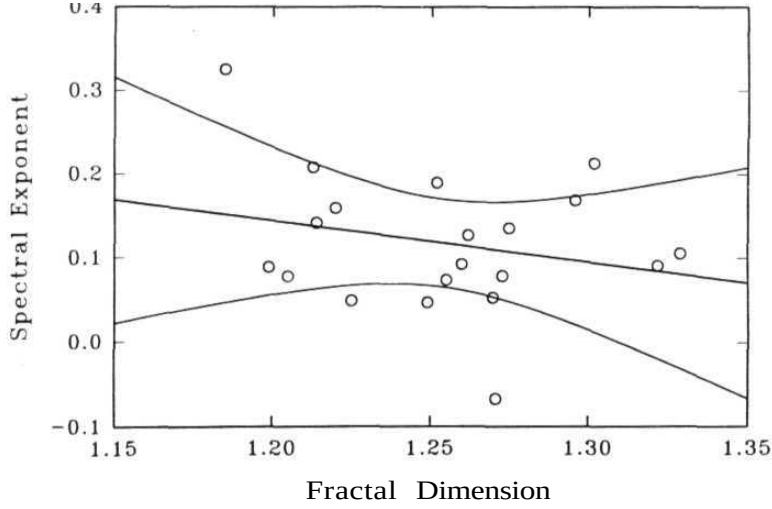

Fig 4.9: A plot of the spectral exponent b against the fractal dimension D of the positional distribution of the of the 20 amino acids. The straight line shows the least square regression line and the two curved lines represent the 99% confidence interval based on D. We do not expect a perfect correlation between the two considering the limited sample size used in our analysis. Methionine has negative spectral exponent due to abnormal distribution (see text)

The fBm characteristics, i.e., β and H and D for the positional distribution of all the twenty residues are presented in table IV.

Table IV fBm characteristics of positional distributions of residues in proteins

No	Res	P	H	D	No	Res	P	H	D
1	Ala	.0890	0.4555	1.199	2	Cys	.0908	0.4546	1.322
3	Asp	.1896	0.4052	1.252	4	Glu	.1592	0.4204	1.220
5	Phe	.0786	0.4607	1.273	6	Gly	.2081	0.3960	**1.213**
7	His	.1687	0.4157	1.296	8	Ile	.1271	0.4365	1.262
9	Lys	.0491	0.4755	1.225	10	Leu	.3253	0.3374	1.185
11	Met	-.0671	0.5336	1.271	12	Asn	.0739	0.4631	1.255
13	Pro	.0530	0.4735	1.270	14	Gln	.1356	0.4322	1.275
15	Arg	.0472	0.4764	1.249	16	Ser	.0782	0.4609	1.205
17	Thr	.0927	0.4537	1.260	18	Val	.1417	0.4292	1.214
19	Trp	.1052	0.4474	1.329	20	Tyr	.2129	0.3936	1.302

4.2.7 Discussions and Conclusions

4.2.7.1 Fractal geometry: the geometry of nature?

Fractals are geometrical objects showing self-similarity (Mandelbrot, 1982) at all magnifications (or scales), i.e., they appear similar at all magnifications. The self-similarity may be geometrically exact or statistically apparent. but self-affine fractals differ from both of these. Self-affine fractals show statistical self-similarity only when the x and y axes are magnified by different scales. The importance of fractal geometry is due to the fact that it has become a very convenient and popular method for analysing objects , processes and phenomena occurring in nature. This is because fractal geometry is capable of providing the language and formalism for studying processes like diffusion limited aggregation, condensation of matter on microscopic scale, etc. (Orbach, 1986); for describing biological phenomena like patterns existing in long DNA sequences (Voss,1992; Peng et al, 1992; Nandy 1994), branching patterns in the network of neurons, arteries, bronchioles, trees, symmetry and patterns of plants and flowers etc.(Oppenheimer, 1986). The origin of fractal patterns in nature is due to chaotic dynamics of non-linear detrministic systems.

4.2.7.2.1 Modelling proteins as fractals

Based on our results we have modelled the pattern of positional distribution of each amino acid in proteins as fractional Brownian motion and any individual protein sequence as multi-fractals.

4.2.7.2.2 Brownian motion and Fractional Brownian motion

In the usual Brownian motion or random walk, the sum of the independent increments or steps leads to a variation that scales as the square root of the total number of steps. Thus H = 0.5 corresponds to a normal Brownian motion. A Brownian motion where H <> 0.5 is called a fractional Brownian motion or an fBm (Voss, 1988). An fBm trace is characterised by the scaling parameter H, spectral exponent β, and the fractal dimension D.

4.2.7.2.3 Relationship between H and p with the fBm trace

An fBm trace is characterised by the scaling parameter H, spectral exponent p, and the fractal dimension D. An fBm trace repeats statistically only when the x and y co-ordinates are magnified by different amounts. If x is magnified by a factor r (x becomes rx) then y must be magnified by a factor r^H , (i.e., y becomes $r^H y$). This non-uniform scaling where shapes are statistically invariant under transformations that scale different co-ordinates by different amounts is called self-affinity mentioned below.

4.2.7.2.4 **Self-affine** fractals

Self-affine processes are not truly random but have some correlations. A statistically self-affine fractional Brownian function vH, provides a good model for many natural scaling processes and shapes. As a function of one variable (E=1), it is a good model for noises, random processes, music etc. We have modelled the distribution of residues in the protein sequences as fBm. The characteristics of these fBm namely the scaling parameter and the spectral exponent have been calculated.

4.2.7.2.5 Positional distribution of residues as self-affine fractals

Based on the studies and results we find that the distribution of the amino acids in protein sequences can be modelled as fractional Brownian motion. Hence, an fBm model has been proposed. The characteristics of an fBm relate to the fractional Brownian motion integrated over an interval of time/space. Differenciating such an fBm w.r.t. time/space gives what is called fractional Gaussian noises. Similarly, the fBm characteristics of the distribution of a residue relate to the average statistical properties of the

residue in the protein sequences. Differenciating such an fBm would give us the distribution of a residue in an individual protein sequence. Hence distribution of a residue in an individual sequence is a fractional Gaussian noise and its distribution in the complete set of natural proteins is a fractional Brownian motion.

4.2.7.2.6 Proteins as multi-fractals

Since an individual protein sequence consists of several residues (the distribution of each being a fractal) an individual protein sequence can be considered as a multi-fractal. Shapes and measures requiring more than one fractal dimension are known as multi-fractals. If a sequence contains all the twenty residues then it requires all the twenty fractal dimensions (one for each residue) to describe it. Proteins like collagen having fewer kinds of residues would be multi-fractals requiring lesser number of fractal dimensions. Different sequences of a given family, say myoglobin for example, have similar primary sequences, hence based on our model, they are multi-fractals of one kind. After characterising this family of multi-fractal, one can in principle, predict unknown sequences (yet undiscovered) but existing in nature, by computer simulations (using the multi-fractal parameters i.e., dimensions of the distribution of the twenty residues already characterised for the protein family).

Also one can, in principle, study the evolution of protein families by observing how the multi-fractal (i.e., primary sequence of a protein) slowly changes with change in the parameters of the multi-fractals and becomes another protein. It may also be extended to the tertiary structures of proteins.

Summary: Treating the sequential positional distribution of amino acid residues in proteins as time series and subsequent analysis, we found long range correlations in them. Further analysis showed them to exhibit self-affinity and they have been modelled as fBm. Proteins themselves hence can be modelled as multi-fractals as they contain all the residues. Multi-fractals are fractals which require more than one dimension to describe them. Hence a protein would require the dimensions of the positional distribution of all the 20 amino acids to describe it ; each of the 20 positional distribution being a self-affine fractal of the fBm kind.

Chapter 5

Spectral analyses of positional distribution of amino acids in proteins

5.1 Introduction

As mentioned in the previous chapter, protein sequences have been considered as time-series and their analysis by techniques used in time-series analysis has been attempted. Protein sequences may be considered as a time-series because a residue which is added earlier in time to the growing polypeptide chain, occupies an earlier position in the sequence. Hence the successive temporal addition of a residue is reflected in sequential position of residues of in the protein sequences. Besides, since the techniques used to study time-series can be extended to spatial situations also, therefore proteins can be studied by time-series considering the horizontal axis as the positional axis instead of the time-axis. The correlations between successive terms can be measured by the serial correlation test. The array of correlations of different orders obtained from the distributions are converted by fourier transformation into a spectrum. From such a spectrum for any given pair we can understand the nature of relationship between them.

While in the previous chapter only the nature of the correlations (only autocorrelations) were studied, here we attempt the complete positional correlation analysis irrespective of the kind of residue pairs. The idea here is to decipher general patterns in the sequences of proteins which can probably distinguish them from random sequences and based on the correlations calculated (from the spectral analysis) to construct a knowledge-base of common periodicities in protein sequences in general and test their ubiquity in real sequences of the data-base and the frequency of their occurrences in simulated sequences.

5.2 Data-base used

We have used the Swiss-Prot protein sequence databank (SPPS databank) release 10.0 March 1989, as mentioned in the previous chapter (section 4.1.3). Only those sequences which had at least 200 residues were used for the analysis (reasons mentioned earlier). The autocorrelation and cross-correlation analysis have been performed only on them and these correlations were determined by using 5034 sequences and sequences having residues less than 200 had not been used to determine them at all. However the results obtained were verified later using the entire data-base. No explicit attempts were made to selectively filter or condition the data-base for the elimination of homologous sequences and families *etc.* (reasons mentioned in the previous chapter).

5.3 Positional distributions of twenty amino acids

In the previous chapter, we have considered the positional distribution of each amino acid residue as a single time-series. Thus we have twenty series of positional distributions, one corresponding to each amino acid residue. By studying each series individually we can understand how each residue is correlated positionally to itself within the protein sequences. Since proteins contain all the twenty amino acids, it is required to understand the position correlation between different kinds of residues *i.e.,* the inter-relationship between the twenty series must also be studied. Thus treating the 20 series together as a case of multiple time-series is very appropriate. For further detail on multiple time-series please refer to Kendall (1976,1983).

5.4 Methodology

The methodology adopted in the complete correlation analysis (both autocorrelations and cross correlations) is similar but also differs in some ways to that applied for the autocorrelations in the previous chapter. Hence the methodology has been described again.

Correlation analysis can be used to detect and calculate correlations within a series of numbers. The autocorrelation analysis carried out in the previous chapter showed that the residues showed long range autocorrelations. It has also been described correlation analysis can be used to detect and calculate correlations within a series of number which are not random. The autocorrelations measure the intradependencies, *i.e.,* how the distribution of a residue is correlated to itself with respect to all the other positions (200 in our case) of the sequences. On the other hand, cross-correlations measure the interdependencies between residues, *i.e.,* how the distribution of a residue is correlated with the distribution of all other residues with respect to all possible positions. Hence for every residue there is 1 series of autocorrelation with 200 different positions and 19 series of cross correlations and each of the 19 series having 200 different positions. Hence in all we have 400 (20x20) pairs; 20 homo-dipeptide pairs and 380 hetero-dipeptide pairs. These autocorrelations as well as cross correlations are converted into spectral densities by fourier transformation. The periodicities from the spectra are calculated and again verified using the whole data-base as well as some protein families which had not been used in the calculations.

5.4.1 Autocorrelation and cross-correlation analysis

The correlation analysis was performed on the positional distribution of the residues with respect to themselves as well as other residues in order to detect the periodicities inherent in the protein sequences. The total number of residues of type *i,* where *i* takes values from 1 to 20 and denotes the ordinality of one letter code of the amino acid residue (A to Z, excluding B, J, O, U, X

and Z, which do not code for the 20 amino acids) and / denotes the positions from 1 to 200 along the protein sequence, is denoted by $u_{i,l}$. u is a 20 x 200 matrix and each row of this matrix denotes a positional distribution for one of the twenty amino acids. Considering any two residues i and j, the correlation between them can be calculated from their covariances. The correlations as mentioned above are of two kinds. When /=/ (both are same amino acid residues) the covariances are called autocovariance and corresponding correlations are called autocorrelations. When $/ \neq j$ (the two amino acid residues are distinct) the covariances are called cross-covariances and the corresponding correlations are called cross-correlations. While the autocorrelations give a measure of the self-preferences among the amino acid residues the cross-correlations give a measure of the preferences between various residues for each other, *i.e.,* the autocorrelations obtained are an index of the intra-dependencies and the cross-correlations of the various inter-dependencies among the positional distributions of the residues in protein chains.

5.4.2 Formulae used

The correlation between the residue /and; are calculated as follows:

The mean of the series for the i th residue

$$\mu_i = E(u_{i,l}) = \sum_l u_{i,l} / 200 \qquad \textbf{(1)}$$

Similarly mean of the series for the jth residue

$$\mu_j = E(u_{j,l}) = \sum_l u_{j,l} / 200 \qquad (2)$$

These frequencies are essentially the observed proportions of the respective amino acid residues in the data-base.

Let s denote a positional lag between the two series $u_{i,l}$ and $u_{j,l}$. *i.e.*, we compare $u_{i,l}$ and $u_{j,l+s}$ always in our calculations. The positional lag s is allowed to vary from 0 to 99 (hence a maximum periodicity of 100 can be detected). The maximum value of positional lag s is chosen as 99. This is appropriate considering the lengths of our series, *i.e.*, 200. In any series of finite length we can detect periodicities or recurrence only upto half the length of the series.

The sth covariance $c(i,j,s)$ between residues /and; is defined as

$$c(i,j,s) = E[(u_{i,l} - \mu_i)*(u_{j,l+s} - \mu_j)] \qquad (3)$$

The sth covariance $c(i,j,s)$ between residues i and j is the correlation when they are separated by s number of residues along the chain. The sth order correlation $r(i,j,s)$, between residues i and j, derived from the sth covariance,

c(i,j,s) is as follows:

$$r(i,j,s) = c(i,j,s) / (\mathrm{var}\,U_i * \mathrm{var}\,U_j)^{1/2} \quad \textbf{(4)}$$

where $varU_i$ and $varU_j$ *are* the variances of the *i*th and *j*th residue's positional distribution. This can be rewritten as,

$$r(i,j,s) = \frac{\frac{1}{n-s}\sum_{l=1}^{n-s}\left(u_{i,l} - \frac{1}{n-s}\sum_{l=1}^{n-s}u_{i,l}\right) * \left(u_{j,l+s} - \frac{1}{n-s}\sum_{l=1}^{n-s}u_{j,l+s}\right)}{\left[\frac{1}{n-s}\sum_{l=1}^{n-s}\left(u_{i,l} - \frac{1}{n-s}\sum_{l=1}^{n-s}u_{i,l}\right)^2 * \frac{1}{n-s}\sum_{l=1}^{n-s}\left(u_{j,l+s} - \frac{1}{n-s}\sum_{l=1}^{n-s}u_{j,l+s}\right)^2\right]^{1/2}} \quad (5)$$

In the above equation n denotes the total number of data points, *i.e.,* 200 in the present case and the symbols *c(ij,s)* and *r(i,j,s) are* used to denote the cross covariance and cross correlation of the sample, whereas *y(i,j,s)* and ρ(*i,j,s*) denote the cross covariance and cross correlation of the population, *i.e.,* the complete series.

5.4.3 The correlogram and the spectra

Let *i* and *j* denote a pair of residues. When *i=j,* let such a pair be called a homo-pair else it be called a hetero-pair. Out of the 400 possible pairs 380 are hetero-pairs and 20 are homo-pairs. We obtained the array of 100 correlations *i.e.,* from 0 to 99, for each pair *i, j.* The array of correlations plotted against the position (*i.e.*, the order) is called a correlogram. The correlations are transformed by fourier transformation into spectral densities. The array of spectral densities plotted against the frequency (angular frequencies) is called a spectrum. The spectrum and the correlogram uniquely determine each other (related by fourier techniques). All our studies are based on the spectra obtained as described above.

For any pair of series say u_i and u_j and positional lag of s number of residues varying from -∝ to +∝ we obtain a set of correlations *r(i,j,s).* Then the spectral density is defined over the range of angles from 0 to *n.* In actual computation s was taken from 0 to 99 only. The angle a is in radians. The spectral densities are obtained by fourier transformations of the correlations ranging from -∝ to +∝ and is as follows:

$$w(i,j,a) = \sum \rho(i,j,s) \cdot e^{i \cdot s \cdot a}$$

In the above expression e=2.7321 and i= $\sqrt{-1}$.

$w(i,j,a)$ is also written alternatively and equivalently as $w_{ij}(a)$. Whereas the summation over s theoretically runs from - ∞ to + ∞, because of the symmetric nature of the autocorrelations, we can consider only half of them, *i.e.,* from 0 to +∞. In addition, we can use a finite sum, in this case up to s=99, as an approximation.

Spectra of homo-pairs

We plotted the squares of the spectral densities (intensities) to obtain the spectra The amplitudes of these were compared to obtain the five highest peaks, as there were often several significantly high peaks with only small differences in heights. The position on x-axis corresponding to the peaks gives the angular frequencies. The periodicities (i.e., linear periodicity) were obtained by dividing 360 by these angular frequencies. All these linear periodicity values are presented in Table I. The first two columns contain the three and one letter code of the residues. In the third column we have shown the probabilities of these residues in the data-base. The remaining five columns show the linear periodicity of the respective amino acid in the data-base as deduced from the spectrum.

Another quantity 'coherence' is used to calculate the spectra of hetero-pairs. It cannot be used for homo-pairs as in case of homo-pairs the coherence is always equal to unity at all angles and no peaks are seen.

Spectra of hetero-pairs

Since we are also deaiing with cross-correlations apart from autocorrelations it is necessary to construct the coherence spectra for analysing the relationship of the hetero-pairs.

Since the cross-correlations are not symmetric, *i.e., p(i,j,s)* $\neq \rho(j,i,s)$, but as $\rho(i,j,s) = \rho(j,i,-s)$ therefore the expression obtained for $w(i,j,a)$ becomes,

$$w(i,j,a) = 1 + \sum\{\rho(i,j,s)*\cos(s\cdot a) + \rho(i,j,-s)*\cos(s\cdot a)\} + \sum\{\rho(i,j,s)*\sin(s\cdot a) - \rho(j,i,-s)*\sin(s\cdot a)\}$$

$$= c(a) + \mathrm{i}\cdot q(a)$$

The quantity $c(a)$ is called co-spectrum and $q(a)$ is called quadrature spectrum. The sum of their squares *i.e.*, $c^2 + q^2$, is called amplitude of the spectrum. The standardised quantity "coherence", *i.e., C[i,j,a] is* defined aswhere w_i and w_j are the spectral densities of u_i and u_j respectively. (The difference between the capital C and the lowercase c are to be noted). The

$$C(i,j,a) = \frac{\left[c^2(a)+q^2(a)\right]}{\left[w_i(a)\cdot w_j(a)\right]}$$

$$= \frac{\left|w_{i,j}(a)\right|^2}{w_i(a)\cdot w_j(a)}$$

coherence spectra are symmetrical, *i.e., C[i,j,a] = C[j,i,a].* Coherence spectra of each pair has 180 values, one corresponding to each angle in one step. There are 380 hetero-pairs *(i.e.,* when $i \neq j$) and due to symmetry of the coherence spectra we have considered only 190 spectra. For each of these pairs, the maximum amplitude of the coherences *(i.e.,* the highest peak) and the angle corresponding to the maximum amplitude was found using a program written in turbo-pascal. We obtained 190 values (the maximum values of the coherences of each of the 190 pair).

These 190 coherences (peaks corresponding to the highest intensities of coherence spectra of each symmetric hetero-pair) were arranged in a decreasing order of magnitude, using a sort program. The 50 highest values, *i.e.,* the 50 most intense peaks out of the 190 were selected for further study. (The remaining peaks with smaller values indicate smaller correlations). The value of the most intense peak was noted down and also the corresponding angular frequency on the horizontal axis. These positions correspond to the highest occurrences or frequencies of the respective pair *i,j.* Corresponding linear periods were obtained by division of 360 by these angular frequencies. The periods give the periodicity, *i.e.,* the number of amino acids separating the residues *i* and *j.* The length of the series being 200, periodicities greater than 100 cannot be detected, hence periods showing values greater than 100 were ignored. (which reduced the number of hetero-pairs from 50 to 44). The periods for the 44 hetero-pairs (as calculated from the coherence spectra) are presented in Table II. The amino acid pairs are presented in the second column. The respective preference values of the pairs are in the third column. Preference is determined by dividing the observed proportions of occurrence by theoretical probability. A value greater than one indicates attractions and less than one indicates repulsions. All those pairs which showed attractions and with preference value greater than 5% over the value of one have been listed here. The pairs numbered 1 (Gly-Pro) to 13 (Trp-His) showed greater than 10% of preference. The periodicity of their occurrences (as determined by the coherence spectra) are shown in the next column titled F.

5.4.4 Verification of periodicities deduced from spectra on the whole data-base

Next the data-base was searched directly to see the accuracy of the period calculated from the coherence spectra. For every pair of residue / and *j* separated by a distance of s, the frequency of occurrence in the whole of the

data-base was actually found out by direct counting. The observed proportion of finding the pair was found out by multiplying the observed proportions for individual residues *i* and *j (i.e.,* the product of μ_i and μ_j). That is the percentage occurrence obtained for each pair was divided by the frequency of the constituent residue to get the preference for the pair. This quantity is called the pair preference. This was done for all the pairs and the values are presented in Table I for self-preferences and Table II for cross-preferences. It is to be noted that these values are essentially average preferences and have been obtained from the whole of the data-base. As one can see in the tables, in case of self-preferences, for each residue up to 5 different periods have been selected from which 80 significant preference values have been chosen; in case of hetero-pairs, though there were 190 coherence spectra only 44 significant cross-preference values were selected. This showed that of self preferences among residues is more common than cross-preferences. Also the intensity of preference is higher in case of self-preferences than in case of cross-preferences. The numbers in Table I and Table II reflect the total number of pairs of residues in the data-base. For this computation all sequences in the data-base were used. Hence the total number of residues that were finally used in the determination of periodicities were much higher. Since all the residues were considered and the residues beyond 200 were not ignored, the total number of residues that were considered is much larger than the number that were used to arrive at the periodicities.

5.4.4.1 A scoring "weight " due to pairs from knowledge-base constructed from results of spectral analysis of protein sequences.

Each of these 124 (80+44) pairs were searched in the whole data-base systematically and their frequencies were calculated. The frequencies were normalised and a value close to one was obtained. If this value was greater than one it indicated a preference and if it was less than one it indicated a repulsion. All these pairs showed a preference value greater than one, which indicated that the periodicity of the pairs in the rest of the data-base was similar *(i.e.,* as in the experimental data-base). To quantify these results we associated a "weight" with each pair based on these values of self-preferences and cross-preferences. The "weight" was defined as the natural logarithm of the observed preference values. The weight of a sequence was the sum of the weights due to the constituent pairs that occur in the sequence (those pairs that didn't exist in our table of self-preference and cross-preference did not contribute to the weight of the sequences; weights due to them may be considered small or negligible compared to the weight of the sequence).

Table I: Periods for various residues

Residue		P(Res)	1st	2nd	3rd	4th	5th
Ala	A	7.791	18	23	10	16	15
Cys	C	1.875	12	36	90	60	2
Asp	D	5.213	5	3	2	8	
Glu	E	6.150	8	6	10	3	
Phe	F	3.927	5	4	2	3	
Gly	G	7.301	9	20	18	4	
His	H	2.287	4	28	2	3	
Ile	I	5.298	4	2	5		
Lys	K	5.778	45	3	4		
Leu	L	9.060	10	2	24	3	
Met	M	2.481	180	7	2	5	4
Asn	N	4.344	6	2	9	4	
Pro	P	5.207	9	4	3	12	
Gln	Q	4.099	4	5	2	3	
Arg	R	5.208	3	6	7	2	
Ser	S	6.996	2	3	4		
Thr	T	5.855	4	2	5	180	
Val	V	6.509	40	12	4	17	3
Trp	W	1.347	7	3	12	5	90
Tyr	Y	3.203	7	4	5	2	

The pair preferences (also referred in the text as self-preferences) of the homo-pairs and the respective common periodicities are shown in the Table I. This was done for all the 20 homo-pairs. The percentage occurrence obtained for each pair was divided by the frequency of the constituent residue to get the preference for the pair. The periodicities were obtained from the spectrum as already mentioned earlier. In case of homo-pairs several high peaks were seen and upto five peaks have been considered.

Table II:Hetero pair preferences (attractions)

No	A A pairs	Pref.	F	*No*	A.A pairs	Pref.	F
1	Gly-pro	1.270	3	*2*	Trp-Cys	1.262	51
3	Phe-Trp	1.227	31	*4*	Cys-Tyr	1.168	**16**
5	Tyr-Trp	1.157	40	*6*	Leu-His	1.156	17
7	Lys-Glu	1.133	11	*8*	Cys-Phe	1.118	19
9	Phe-Ile	**1.114**	30	*10*	Tyr-Cys	1.109	16
11	Tyr-His	1.107	1	*12*	Glu-Lys	1.103	11
13	Trp-His	1.102	7	*14*	Pro-Gin	1.093	7
15	Ile-Lys	1.090	73	*16*	Ile-Trp	1.078	46
17	Trp-Lys	1.074	46	*18*	Leu-Trp	1.073	10
19	Leu-Ile	1.073	17	*20*	Asn-Asp	1.072	1
21	Ser-Asn	1.071	1	*22*	Gly-His	1.071	10
23	His-Ile	1.068	1	*24*	Phe-His	1.067	46
25	Lys-Ile	1.066	73	*26*	Thr-Asn	1.066	19
27	Ala-Pro	1.065	4	*28*	His-Trp	1.065	7
29	Asp-Glu	1.062	**2**	*30*	Asn-Ile	1.062	12
31	Trp-Pro	1.062	38	*32*	Trp-Asn	1.060	38
33	Gln-Glu	1.060	**2**	*34*	His-Tyr	1.060	1
35	Lys-Asp	1.058	7	*36*	Trp-Tyr	1.057	40
37	Trp-Phe	1.055	31	*38*	Glu-Ile	1.055	2
39	Tyr-Asn	1.054	18	*40*	Ile-Tyr	1.053	**44**
41	Asn-Glu	1.053	2	*42*	Ala-Gly	1.052	7
43	Glu-Asp	1.050	**2**	*44*	Tyr-Phe	1.050	**18**

The weight of each sequence in the data-base was calculated as follows: The 80 self-preferences and 44 cross-preferences were fed in a program to calculate the weight of each sequence in the data-base, based on the presence of these pairs at the expected periodicities in the data-base (periodicity here refers to the number of amino acids separating the two members of the pair within a sequence). Finally, the total weight of the sequence was obtained. Simultaneously, a random sequence of the same length as the sequence being analysed was generated and its weight was also calculated in exactly the same manner using our table of preferences. This process was repeated for all sequences in the data-base as well as for

the respective random sequences generated. The weights of the natural as well as the random sequences were compared to check the extent of randomness and significant non-randomness if any in the natural sequences. Finally the weights of the sequences were plotted against the lengths (in log scale). It was found that the weights of the real sequences were higher than the random ones.

These preferences were determined by using only 5034 sequences. Sequences having residues less than 200 had not been used to determine them at all. Since the observed patterns are expected to be valid in general for all sequences, using our table of preferences and corresponding periods, we have extended this technique to calculate the weights of sequences of a few protein families having less than 200 residues. This would confirm the presence (or otherwise) of these pairs with high correlation (at their corresponding periodicity) in sequences other than those which had been used to determine the correlations.

Table III: Weight per residue of a few selected protein families

Protein family	No of sequences	Length of sequences	Wt/res (real seq)	Wt/res (random seq)
Histone (H3)	24	134-136	0.058	0.047 ± 0.0022
Histone (H4)	10	102	0.055	0.046 ± 0.0026
Heat Shock	18	143-174	0.049	0.047 ± 0.0024
Calmodulin	15	138-162	0.048	0.047 ± 0.0016
Myoglobins	68	146-154	0.061	0.048 ± 0.0011
Cytochrome c1	7	104-113	0.054	0.047 ±0.0030
Cytochrome c2	11	102-107	0.053	0.047 ± 0.0036

c1- plant source c2- animal source

We have chosen sequences of the following families namely, calmodulin, cytochrome c, heat shock proteins, histones and myoglobins. All sequences of a given family present in the data-base were considered except fragments *(i.e.,* partially determined sequences). The weight per residue of each family and also the mean weight per residue of the sequences of the family was calculated. Equal number (as in the protein family considered) of random sequences corresponding to each actual sequence (of same length) had been generated and the mean weight per residue and standard error in case of random sequences were calculated. The results are presented in Table III.

5.4.5 Results and discussions

The distribution of lengths of protein sequences in the data-base was studied. (Figure 4.1 in chapter 4 gives a graphical representation of the distribution of lengths in the logarithmic scale. As seen in the figure, the modal length is close to 200 (actually 250). The mutual preference of residues in various pairs were calculated as described earlier. Pairs having strongest preference values and the distance from each other at which the strongest preference is seen (at some most likely unique distance from each other within the sequences) as calculated from spectral analysis of the serial correlations and verified by checking in the sequences of the data-base, are presented in Tables I and II. Table I contains preferences and the distances (periodicities) corresponding to the preferences. Since several strong preferences are seen in case of homo-pairs up to 5 strong preferences have been mentioned along with respective preferred distances between them.

Table II contains the preferences in case of hetero-pairs. Only the strongest peak and the periodicity corresponding periodicity has been mentioned. There are 44 hetero-pairs with relatively strong preferences. The remaining hetero-pairs have not been considered for calculation of the weights of the sequences as they carry relatively small weight. Only pairs having more than 5% preferences have been listed (and considered) as compared to a random sequence. The random and the natural sequence were compared by plotting the graphs of sequence weights *vs* log(length) and weight per residue *(i.e.,* sequence weight /sequence length) vs log(length)

sequence weight vs log(length)

In the case of random sequences as clear from figure 5.1 (a), the weight is found to be a function of length, as it increased with length. This is expected because as the length increases, more number of residues are present and the probability of occurrence of a pair which has not yet occurred in the sequence, increases. Since this fact holds true for the natural sequences also, one does see in figure 5.1(b) a general increase in the weight with increase in length. But one finds fluctuations in this pattern in figure 5.1(b) in the natural sequences). Any additional weight than expected (as in random case) at any length is due to strong intrinsic preferences among the residues in the sequence. The weight of a natural sequence of length 1000 is 60 units, whereas for a random sequence it is about 50 (the simulated random sequences has the same amino acid composition as the natural one). This deviation from the graph of random sequences is a measure of non-randomness of the natural sequences, which though small do exist and is probably sufficient to impart biological activity. It may be necessary to mention that non-randomness need not always impart biological functionality but biological functionality mostly implies non-randomness.

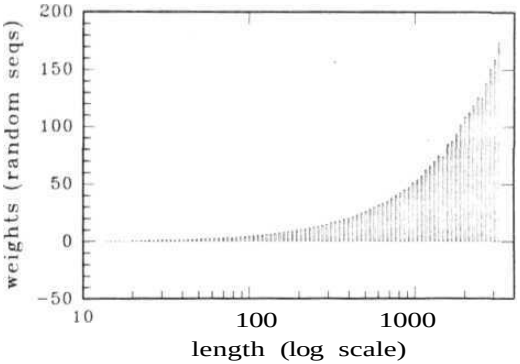

Fig 5.1a: The lengths of simulated random sequences are plotted on the X-axis against their respective weights on the Y-axis. Weight of a sequence is defined as the sum of the weights of the constituent pairs present at the given periodicities as required as in Table I and Table II

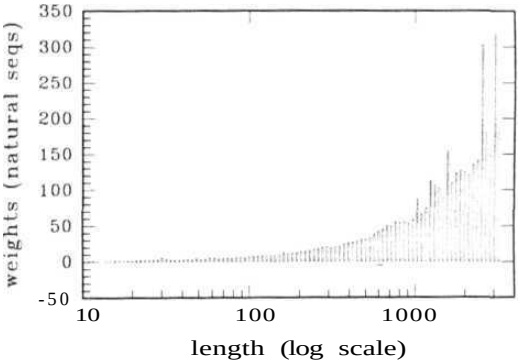

Fig 5.1b: The lengths of natural sequences as found in the whole of the data-base as plotted on the X-axis against their weights on the Y-axis. In this plot all sequences have been considered (not only the ones with length greater than or equal to 200 residues and all residues even beyond 200 are included).

Sequence weight / sequence length vs log(length)

When we calculate the weight/residue in the case of a random sequence it is found to be 0.046 whereas in the case of a natural sequence it is about 0.063. The additional weights are due to non-randomness in the primary sequences of the natural sequences.

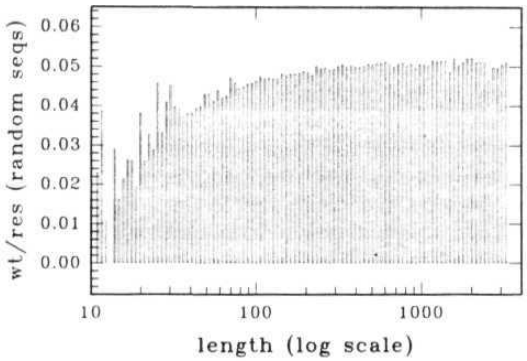

Fig 5.2a: The weight per residue ratio of sequences in case of simulated random sequences. The weight per residue is obtained by dividing the sequence weight by its length. The ratio appears to converge to a constant value from the sequence length 100 onwards. This is because correlations beyond 100 have not been taken into account

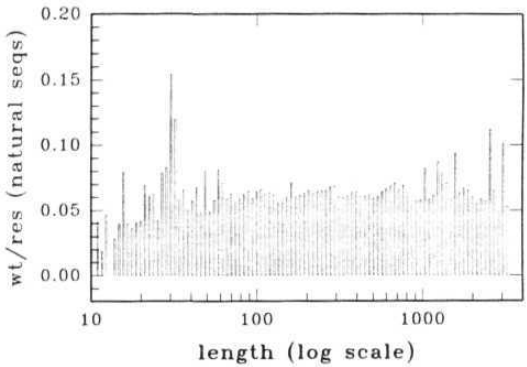

Fig 5.2b: The weights per residue in case of natural sequences. The ratios have been obtained by dividing the sequence weight by its length, as above. The ratio in this case is higher than those seen for the random sequences

In the case of random sequences as well as natural sequences we find that the ratio of sequence weight to sequence length converges to a constant value around the length of 100. But below the length of 100 one finds fluctuations and not a constant value. This is because we have not considered the preferences which correspond to periodicity greater than 100 (It may be recalled that it was done because the length of the series being only 200, periodicity values greater than 100 cannot be accurately detected in the series and must be ignored; however this could be overcome by taking considering larger lengths but that could reduce the number of sequences being considered for the calculations). Since correlations beyond 100 have been ignored we obtain a constant value of weight (and weight per residue) beyond length of 100. Using this technique we have a rough way of distinguishing a random sequence from a natural one based on these weights alone. This method can be made more accurate by introducing more number of pairs, even though they may carry less weight.

We calculated the weights of a number of sequences of the following proteins- histones, calmodulin, heat shock proteins, myoglobins and cytochrome c. In

all these cases, it was seen that natural sequences had more weight than random sequences. This confirms the correctness of our inferences made earlier regarding the weights of natural and random sequences as deduced from figures 5.1(a&b) and 5.2(a&b) In table III, we have summarised the comparison of weights of sequences of the five selected families. It can be seen that all of them have weights greater than random sequences. The difference of the weights from that of a random sequence gives a measure of the non-randomness of the sequence of a protein. In that sense myoglobin sequences have maximum non-randomness and calmodulin sequences have least non-randomness among the 5 sets of the protein families. However, less amount of non-randomness do not imply that the sequences do not have a conserved structure. It only suggests that in the process of evolution from a random sequence to a meaningful sequence it acquired its meaning *(i.e.,* biological functionality) much earlier compared to other sequences.

Ptitsyn (1984) has remarked that proteins are largely random polypeptides which have been edited in the course of evolution to impart biological functionality. We have found that weights of simulated random polypeptides are not markedly lower from lower from those of the natural ones but they are *always* lower than them. The difference in their weights is then a measure of the non-randomness. For smaller lengths it is not expected to be very significant but for larger lengths one finds considerable differences. If one assumes that the natural proteins have emerged by edition of random polypeptides then those sequences which are older in the evolutionary scale may have smaller weight differences than more recent ones. Of the same 5 families of proteins selected, we calculated their weights and based on the difference of the weights of the corresponding simulated random polypeptides, from them we have calculated the evolutionary status of these proteins in the increasing order of evolution as follows:

Calmodulin < heat shock proteins < cytochrome c < histones < myoglobin.

These results may be applied in protein designing. Simon (1985) has regarded a protein sequence as a sequence of short overlapping sequence of oligopeptides having significantly stable conformations which are responsible for the folding . In such a case calculating the preferred distances between two pairs, within the sequences by time-series techniques is very useful. The results are useful in designing and engineering protein structures.

t is important to add in conclusion that real data used in the computations of these preference values may have "trends" and "seasonalities" which we have not made any explicit effort to eliminate. These may have some influence on our results, but as the finally used values are experimental ones obtained from the data-base itself, their effects are expected to be minimal (Beran, 1992).

We would also like to mention that several small compromises have been made in the calculation of the periodicity and construction of the knowledge-base of periodicities. They are:

i) We have considered the positional distributions only upto 200. The basic reason for this was that more than half the sequences (5034) could be used.

ii) We did not consider all the periodicities but only those which had high intensities.

However we believe that this does not qualitatively affect the nature of our results. Improvement in the method is possible for instance, by using more sequences (from the latest data-base, now having several tens of thousands of sequences) and by including more number of periodicities. That is expected to make the results more marked than they are now.

Chapter 6

Entropy of protein sequences

6.1 Living organisms and their entropy

The second law of thermodynamics states that in all processes the entropy of the system plus the surroundings, *i.e.,* the entropy of the universe always increases until equilibrium condition is attained, at which point the entropy is the maximum possible under prevailing conditions of temperature and pressure. In other words, the ultimate driving force for all chemical and physical processes is the tendency for the entropy of the universe to be maximised.

The molecular complexity and the orderliness of structure of living organisms, in contrast to the randomness of the surrounding inanimate matter, also obey the laws of thermodynamics though for some time it was not believed so. It is explained as follows:

Living organisms are open systems, *i.e.,* they exchange both matter and energy with their surroundings. They are not in equilibrium with the surroundings but appear to be so because they are in a steady-state condition. An open system is said to be in a steady-state when the rate at which matter and energy are absorbed is practically balanced by the rate at which they leave the system. Living organisms absorb a useful form of energy (free energy) from their surroundings and exchange it with less useful forms like heat, *etc.,* which quickly randomise in the surroundings and increase its entropy. In short living organisms are non-equilibrium open systems which maintain their orderliness at the expense of the orderliness of their surroundings.

6.2 Entropy: phenomenological and statistical basis

Imagine a vessel with a barrier in the middle and gas on one side and vacuum on the other side of the barrier. When the barrier is lifted the gas will gradually distribute itself homogeneously among all the volume elements of the system. On the other hand if the vessel is filled with gas and the barrier inserted at intervals it is very unlikely that one will find all the gas molecules on one side of the barrier.

The example of heat transfer relates to the classical description of entropy, while the distribution of gas among volume elements as described above, relates to the statistical description. The connection between heat flow and the tendency of a system to proceed towards a random distribution of maximum number of states (spontaneous process) was used by Boltzmann to formulate the statistical laws of entropy. The statistical interpretation has not

only helped to illustrate thermodynamic behaviour with molecular models, it has also led to the use of entropy as a characteristic of non-molecular systems where statistical considerations govern molecular behaviour.

6.3 Entropy in terms of molecular statistics

Boltzmann suggested a relation between 2 types of spontaneous unidirectional processes; the flow of heat from a hotter body to a colder body and the tendency of a system towards greater disorder. The order of a system is associated with the number of possible states. Obviously if a system exists in a single state the probability of finding it in that state is one. Although when Boltzmann and Planck said that entropy is a function of W (probability), they used this symbol to define the number of microscopic configurations or states of the system that are compatible with its macroscopic state. This is the reciprocal of the probability of finding the system in one of the states. Therefore W goes from 1 to large numbers, while probability goes from 0 to 1. Since S is a function of probability we write

$$S = f(W) \tag{1}$$

Entropy is an extensive property, therefore the *combined entropy* of two systems A and B is given by their sum as

$$\begin{aligned} S_{AB} &= S_A + S_B \\ &= f(W_A) + f(W_B) \end{aligned} \tag{2}$$

However *combined probabilities* result in a product and the number of configurations is

$$W_{AB} = W_A \bullet W_B \tag{3}$$

where W_A and W_B are the entropies of A and B respectively

From this reasoning Boltzmann and Planck formulated the logarithmic relation

$f(W) = k \bullet \ln(W)$ Where k is called Boltzmann constant

(4)

6.4 Link Between the classical and statistical thermodynamics

a) Classical derivation

$\frac{dq}{T} = dS$, Where dq is change in heat and dS is change in entropy

(5)

In an isothermal reversible (very small) volume change AV we can obtain entropy changes of expansion of gas from volume V1 to V2

$$\Delta H = p\Delta V$$
$$= RT \cdot \frac{\Delta V}{V}$$

(6)

$$\frac{\Delta H}{T} = R \cdot \frac{\Delta V}{V}$$

$$\Delta S = \int_{V1}^{V2} \frac{\Delta H}{T} = R \cdot \ln\left(\frac{v2}{v1}\right)$$

(7)

b) Statistical derivation

If a gas expands isothermally the change in number of configurations depends only on the increase in volume elements available to the same number of gas molecules. The number of spatial configurations available to a single molecule is proportional to the volume V and in case of N number of molecules it is proportional to V^N

$$\Delta S = S_2 - S_1$$
$$= k \cdot \ln V_2^N - k \cdot \ln V_1^N \quad \text{therefore,}$$
$$\Delta S = k \cdot \ln\left(\frac{V2}{V1}\right)^N$$
$$= R \cdot \ln\left(\frac{V2}{V1}\right) \text{ where } R = N \cdot k$$

(8)

hence equations (7) in the classical derivation and equation (8) in the statistical derivation are equivalent.

Shannon and Weaver (1949) and also Szilard (1925) have proposed a relation between entropy and information. To be informative, a message must a pattern of order. If the message is garbled so that the meaning is partly or wholly lost, then information disappears and order is replaced by disorder. (For example when ice melts the orderly crystal structure is replaced by more random motion of liquid molecules.) This has led Edstall and Gutfreund to propose that the production of information is in effect the lowering of entropy (increase in order).

Proteins and nucleic acids are informational molecules (activity of protein is decided by its 3-D structure which depends on its precise sequence) hence

we presume that they should have lower entropy compared to random sequences (as protein sequences contain information). The information for the activity of the protein is inherent in its sequence. Hence we feel that it is appropriate to calculate the entropy of protein sequences based on the information present in it, i.e., its primary sequence. The information in the form of pair-preferences and pair-repulsions is seen as a result of short, medium as well as long range correlations (Meeta Rani and Mitra, 1994b). However at present, we restrict ourselves to calculation of the entropy based on short and medium range correlations only, i.e., only upto 10 neighbours.

6.5 Entropy as a measure of order in protein sequences

Statistical analysis clearly shows that in spite of apparent randomness there exists a set of well-defined preferences in the primary structures of proteins (Black *et al,* 1976; Saroff *et al,* 1984; Vonderviszt *et al,* 1986). These preferences are exerted over a considerable range, e.g., upto the tenth neighbour (Cserzo and Simon, 1989) and beyond (Meeta Rani and Mitra, 1994b). The pair preferences of residues were used to predict replaceability of amino acids in site-directed mutagenesis (Tudos *et al,* 1990) and in predicting domain boundaries in multi-domain proteins (Vonderviszt and Simon, 1986). The non-randomness of the pair distribution of the residues is related to neighbourhood structure determination properties of proteins (Simon, 1986). which has been used to calculate 3-D structures of globular proteins (Simon *et al* 1991).

Here we suggest another application, namely a simple procedure to distinguish a naturally occurring protein sequence from a random polypeptide on the basis of lower entropy due to pair preferences. This is expected to help in locating the protein coding sequences *(I.e.,* exons) in DNA. It can also help us design new proteins.

6.5.1 Methodology

The frequencies of occurrence of the various amino acid pairs (20X20, *i.e.,* 400 pairs) were obtained from the Swiss prot protein sequence data-bank (release 26, 1994) containing 31808 sequences and 10,875,091 residues. For reasons of convenience, sequences that are shorter than 20 residues were not considered (802 sequences). No selective filtering of the data-base was attempted. We feel that screening of the data-base may introduce a fresh and additional bias (rather than to remove any bias present). The sequences present in the data-base is obviously not a representative random sample; there exists a strong bias introduced due to several factors: (i) biological importance (ii) simplicity and ease of structure and sequence determination (iii) personal interests/ preferences in the class of molecules and last, (iv) the current trends and thrusts in the area. Apart from eliminating all sequences smaller than twenty residues, no other conditioning of the data-base was done.

6.5.2 Calculation of frequencies of the 400 pairs in the data-base

The frequencies were obtained for the pairs separated by 0..9 residues (1st to 10th neighbours) by direct counting. If we consider a pair of residues A and B that are separated by any residue (denoted as X), *i.e.,* a sequence of $A.X_n.B$, where n assumes a value between 0 and 9, then the frequencies of occurrences of such sequences are determined. The observed frequencies were normalised by the average frequencies of the respective pairs. This was calculated using the formula for the conditional probability for residue A and B, assuming independence. In other words, the average probability for the pair $A.X_n.B$ was taken as p(A).p(B), where p(A) is the probability of finding the residue A in the whole data-base. This was done for all the 20X20, *i.e.,* 400 elements of the 10 matrices (corresponding to the 1st to 10th neighbours). A value greater than 1 suggests that the particular pair $A.X_n.B$ is preferred more than average. Similarly, a value less than unity suggests that the corresponding pair is less preferred. These values have been reported by Cserzo and Simon (1989) and Meeta Rani (1990).

6.5.3 Calculation of mixing entropies of the pairs

According to Boltzmann distribution,

$$N_a = N_0 \cdot e^{-(\Delta F / RT)} \tag{9}$$

where AF is the difference of free energy between the state "a" and the ground state "0" and N_a and N_0 denote the populations of the two states respectively.

$$\Delta F = \Delta H - T\Delta S \quad \text{where AH is change in enthalpy and AS is change in entropy} \tag{10}$$

We do not expect that the enthalpy will be significantly different between the two states. Hence we can write as an approximation,

$$N_a = N_0 \cdot e^{(\Delta S / R)} \tag{11}$$

and can determine AS from the data-base using the formula

$$\Delta S = R \cdot \ln(N_a / N_0) \tag{12}$$

As mentioned in the previous section the pair preference values are considered as conditional probabilities. These values are equivalent to equation 12, where N_a is the observed frequency of a given pair and N_0 is the expected frequency in the data-base assuming complete independence This ratio gives the pair preference or conditional probability.

The combined entropy or "mixing entropy" for each of these pairs was

obtained by taking the natural logarithm of their conditional probabilities multiplied by the gas constant R. Ten such 20X20 matrices (each corresponding to n value of 0 to 9) were obtained corresponding to the 1st to 10th neighbours, and were referred to as the mixing entropy matrices. The total mixing entropy for each chain of the data-base was computed using the mixing entropy matrices. To calculate the weight of a given sequence, we add (since probabilities are multiplicative, their logarithms are additive) the entropy values for all the pairs present in the sequence using the table of the mixing entropy. For comparison, for every natural protein sequence, a random chain of the same length was simulated using a Monte Carlo technique, taking into consideration the natural abundances of various amino acids. The mixing entropies of these random sequences were also calculated using the same procedure.

In an earlier work, using fourier transformation of the correlations, we have shown that long range interaction does exist in significant amount in protein sequences. Therefore we cannot say that the mixing entropy calculated by this technique is unique, because neighbours beyond the tenth have been ignored. Nevertheless, we have a consistent and uniform procedure to compare various sequences.

All the computations have been performed on an IBM compatible PC-AT/486 and all the programs were written in Turbo Pascal 6.0.

6.6 Results

To find out the distribution of sequence lengths in the data-base, we plotted a histogram of distribution as shown in figure 6.1. For a clear picture, a logarithmic scale was chosen for the x-axis and the sequences were divided into 100 classes (equal interval in the logarithmic scale). The number of sequences in each interval was counted and the histogram was finally plotted. As mentioned, sequences shorter than twenty residues were not considered. The most probable (the most common) sequence length appears to be close to 400. This graph is important because the frequencies of various classes can be directly obtained from this graph. We also computed the amino acid composition for the whole data-base; the values are in excellent agreement with our earlier results in which a much smaller data-base was used. These values (frequencies of occurrences of various residues) were used in the simulation of random sequences by Monte Carlo technique.

To calculate the entropy of a given sequence, the mixing entropy matrix is required. As explained earlier, the matrix was obtained by direct enumeration of the data-base followed by normalisation and taking logarithms. Since neighbours upto 10th position were considered, this is a 20X20X10 matrix, *i.e.,* a matrix with 4000 elements. This matrix has been sorted and the pair preferences (attractions and repulsions) have been reported in the tables I and II.

Table I: Preference values for the various aa residue pairs In the decreasing order (attractions)

No.	log(P)	R1	R2	F	No.	log(P)	R1	R2	F	No.	log(P)	R1	R2	F
1	1.0562	Cys	Cys	3	2	0.81894	Cys	Cys	7	3	0.78686	Cys	Cys	5
4	0.75369	Cys	Cys	6	5	0.72849	Cys	Cys	4	6	0.72542	Cys	Cys	10
7	0.6826	Cys	Cys	9	8	0.67796	Cys	Cys	8	9	0.56974	Cys	Cys	2
10	0.49977	His	His	1	11	0.49912	His	His	4	12	0.49121	Cys	Cys	1
13	0.48353	Gin	Gln	1	14	0.43523	Pro	Pro	3	15	0.41914	Trp	Trp	7
16	0.41029	His	His	3	17	0.40854	His	His	2	18	0.40831	Gln	Gln	
19	0.4044	Pro	Pro	4	20	0.39975	His	His	5	21	0.39198	Pro	Pro	2
22	0.3893	Gln	Gln	2	23	0.385	Pro	Pro	6	24	0.38432	Trp	Trp	
25	0.38171	Gln	Gln		26	0.37395	Trp	Trp	8	27	0.36899	Trp	Trp	4
28	0.36883	Gly	Gly	3	29	0.36544	Gln	Gin	7	30	0.34392	Glu	Glu	1
31	0.3421	Gly	Gly	6	32	0.33378	Pro	Pro	5	33	0.33198	Trp	Trp	6
34	0.33066	Gln	Gln	6	35	0.3221	Ala	Ala	4	36	0.31353	Trp	Trp	9
37	0.31341	Gly	Gly	9	38	0.31274	Glu	Glu	7	39	0.31109	His	His	6
40	0.30968	Trp	Trp	10	41	0.30843	Arg	Arg	1	42	0.3053	Ala	Ala	1
43	0.30341	Trp	Trp	2	44	0.30226	Pro	Pro	7	45	0.3011	Gln	Gln	8
46	0.29136	Trp	Cys	7	47	0.29126	Trp	Trp	1	48	0.2865	Pro	Pro	1
49	0.28538	Gln	Gln	10	50	0.28369	Glu	Glu	3	51	0.28243	Arg	Arg	2
52	0.28152	Glu	Lys	3	53	0.2806	Gln	Gin	9	54	0.27785	Pro	Pro	9
55	0.27695	His	His	9	56	0.27565	Lys	Lys	1	57	0.27328	Gln	Gin	5
58	0.26749	Pro	Pro	8	59	0.26719	His	His	10	60	0.26235	Arg	Arg	3
61	0.26186	Lys	Lys	3	62	0.26116	Pro	Pro	10	63	0.2606	Glu	Glu	4
64	0.25864	Ala	Ala	3	65	0.25765	Ala	Ala	2	66	0.25367	Ser	Ser	1
67	0.25192	Arg	Arg	4	68	0.2496	Lys	Lys	4	69	0.24875	Glu	Lys	4
70	0.24584	Lys	Glu	4	71	0.24391	Lys	Lys	2	72	0.2424	Lys	Lys	5
73	0.24146	Glu	Glu	8	74	0.23953	Lys	Lys	8	75	0.23614	Asn	Asn	2
76	0.23422	Arg	Arg	7	77	0.23287	Asn	Asn	1	78	0.23133	Lys	Lys	7
79	0.23021	Cys	His	4	80	0.22976	Glu	Glu	2	81	0.22864	His	Cys	8
82	0.22789	His	His	7	83	0.22578	Tyr	Cys	2	84	0.22542	Ser	Ser	4
85	0.22423	His	His	8	86	0.22391	Tyr	Tyr	5	87	0.22382	Cys	His	1
88	0.22356	Ser	Ser	2	89	0.22203	Asn	Asn	4	90	0.21979	Gly	Gly	2
91	0.21951	Lys	Lys	6	92	0.21886	His	Cys	1	93	0.21884	Glu	Arg	3
94	0.21865	Thr	Thr	2	95	0.21841	Tyr	Tyr	1	96	0.21471	Cys	Gly	4
97	0.21229	Ser	Ser	3	98	0.21217	Asn	Asn	8	99	0.2111	Gly	Gly	4
100	0.21063	Lys	Lys	9	101	0.20919	Gly	Gly	8	102	0.20826	Phe	Phe	4

In these two tables, positive values (of log (P)) represent attractions and negative values represent repulsions. F refers to the relative separation of the two residues R1 and R2. It is to be noted that the interaction is not symmetric, *i.e.,* preference of R1 for R2 is not the same as the preference of R2 for R1. This table lists preferences only upto 10th neighbours and hence is not exhaustive. To conserve space, only the first 100 preferences (both attractions and repulsions) have been tabulated.

Table II: Preference values for the various aa residue pairs in the decreasing order (repulsions)

No.	log(P)	R1	R2	F	No.	log(P)	R1	R2	F	No.	log(P)	R1	R2	F
1	-0.28743	Trp	Pro	1	2	-0.27652	Pro	Met	1	3	-0.27069	Glu	Pro	1
4	-0.24067	Cys	Met	6	5	-0.23983	Cys	Met	1	6	-0.22593	Glu	Ser	1
7	-0.22484	Cys	Met	9	8	-0.22439	Cys	Glu	3	9	-0.21927	Cys	Met	2
10	-0.21806	Glu	Met	3	11	-0.21535	Glu	Met	7	12	-0.21498	His	Asp	1
13	-0.21317	Cys	Glu	1	14	-0.21133	Gly	Glu	3	15	-0.21129	Cys	Glu	7
16	-0.20999	Tyr	Ala	1	17	-0.20649	Pro	Met	5	18	-0.2037	Gln	Asp	2
19	-0.19875	Cys	Met	5	20	-0.19623	His	Glu	1	21	-0.19581	Pro	Met	4
22	-0.19446	Pro	Met	8	23	-0.19266	Gly	Lys	5	24	-0.19242	Lys	Met	4
25	-0.19133	Tyr	Met	5	26	-0.19055	Glu	Gly	4	27	-0.18874	Cys	Met	8
28	-0.18617	Cys	Glu	4	29	-0.18581	Pro	Met	3	30	-0.18493	Cys	Glu	8
31	-0.18444	Cys	Gln	3	32	-0.18404	His	Lys	1	33	-0.18403	Gly	Met	4
34	-0.18392	Gly	Lys	3	35	-0.18312	Glu	Ser	2	36	-0.18176	Asn	Ala	2
37	-0.18121	Ser	Met	3	38	-0.1807	Glu	Cys	4	39	-0.18054	Lys	Met	3
40	-0.18015	Gly	Glu	2	41	-0.17845	Asp	Gln	1	42	-0.17781	Ser	Met	1
43	-0.17535	Lys	Pro	2	44	-0.17369	Pro	He	1	45	-0.17351	Ala	Asn	4
46	-0.17315	Cys	Ala	10	47	-0.17243	Pro	Met	7	48	-0.17232	Cys	Ala	I
49	-0.1722	Leu	Met	2	50	-0.17141	Gly	Lys	4	51	-0.1712	Tyr	Ala	4
52	-0.17051	Ala	Cys	7	53	-0.17019	Pro	Lys	2	54	-0.17015	Gly	Asn	3
55	-0.1682	Asn	Gly	2	56	-0.16764	Asp	Gly	2	57	-0.16731	Phe	Met	I
58	-0.16701	Thr	Met	3	59	-0.16654	Glu	Trp	3	60	-0.16644	Asn	Gly	5
61	-0.16545	Lys	Met	7	62	-0.16545	His	Met	6	63	-0.16506	Trp	Met	8
64	-0.16467	Trp	Met	10	65	-0.16464	Ala	Cys	9	66	-0.16418	Tyr	Ala	3
67	-0.16293	Glu	Pro	2	68	-0.16237	Cys	Ala	8	69	-0.1613	Asp	Met	4
70	-0.16114	Cys	Met	4	71	-0.1605	Ser	Met	7	72	-0.16041	Glu	Gly	3
73	-0.1604	Ala	Asn	1	74	-0.15936	Tyr	Met	2	75	-0.15904	Arg	Met	9
76	-0.15882	Phe	Ala	I	77	-0.15876	Tyr	Ala	7	78	-0.15829	Gly	Gln	3
79	-0.15823	Cys	Ala	4	80	-0.1577	Trp	Lys	3	81	-0.15619	Pro	Asn	5
82	-0.15559	Asn	Gly	4	83	-0.15529	Glu	Cys	7	84	-0.15528	Gln	Met	6
85	-0.15526	Glu	Met	4	86	-0.15484	Gln	Met	7	87	-0.15483	Thr	Arg	1
88	-0.15403	Val	Met	5	89	-0.15343	Gly	Met	10	90	-0.15256	Pro	Asn	3
91	-0.15185	Tyr	Met	9	92	-0.15167	Ser	Met	4	93	-0.15131	Glu	Pro	5
94	-0.15089	Pro	Lys	6	95	-0.15082	Thr	Met	9	96	-0.15043	Asp	Met	10
97	-0.15024	Tyr	Ala	10	98	-0.15002	Val	Met	6	99	-0.14968	Cys	He	6
100	-0.14956	Arg	Met	4	101	-0.14893	Ala	Trp	3	102	-0.14864	Met	Trp	1

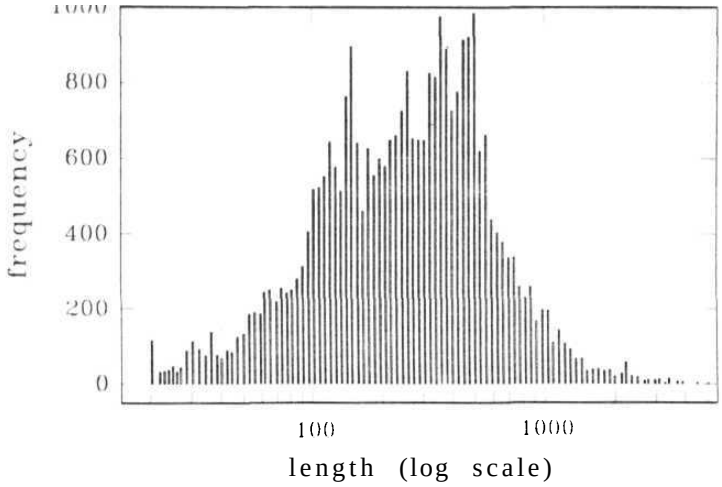

Fig 6.1: A histogram showing the distribution of the lengths of the sequences present in the data-base used. For reasons of convenience, lengths are plotted on a logarithmic scale on the X-axis. The modal sequence length is around 400 residues and the maximum lengths around 4000 and sequences shorter than 20 have been ignored.

In Figure 6.2(a), we present the total weight of a sequence in the data-base plotted against its length. As in the earlier case, the 31000 weights were divided into 100 classes and the mean ± std dev for each class were plotted as a function of the logarithm of mean length of the class. As before, all classes have equal width in the logarithmic scale used. Figure 6.2(b) presents the same information for the randomised sequences. The difference between the two graphs is significant. This is somewhat expected, as the natural sequences follow positive preferences (attractions) and the total weight increases with the chain length. Both the graphs follow an exponential pattern (since the x-axis is logarithmic), *i.e.,* the weight is linearly related to the sequence length. However, the slopes are quite different, as expected. Also, the random (simulated) sequences show far less scatter of weights.

The linear dependence of the weights can be seen more clearly if the weight per residue is plotted against the sequence length. This can be obtained simply by dividing the calculated weight by its length. For natural sequences, this is shown in Figure 63(a) and the corresponding graph for random sequences is shown in figure 6.3(b). As expected, the weight /residue for random sequences is a constant for all practical purposes. For natural sequences, the "noises" are considerable. The reasons for this are not very clear at this moment.

For a more clear picture, we have plotted the distribution of weight/ residue for both natural and random sequences in figure 6.4. The distributions overlap

considerably, but a clear pattern is apparent.

These weights obtained can be correlated with the mixing entropy after multiplication with R (the gas constant). The entropy values so obtained may be considered as mixing or preference entropies based on a standard state which is given by the data-base used. It may therefore be observed that the most probable value of the average entropy (figure 6.4) is close to zero for the natural sequences. On the other hand, the random chains have approximately 0.1 entropy unit higher entropy per residue because

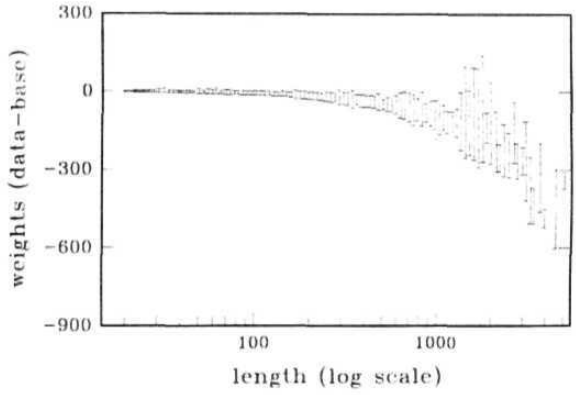

Figure 6.2(a): The weight due to entropy of the sequences in the data-base are plotted against their lengths. The sequences have been grouped into 100 classes on a logarithmic scale based on their lengths. The vertical bars represent standard deviations of the samples in their respective classes.

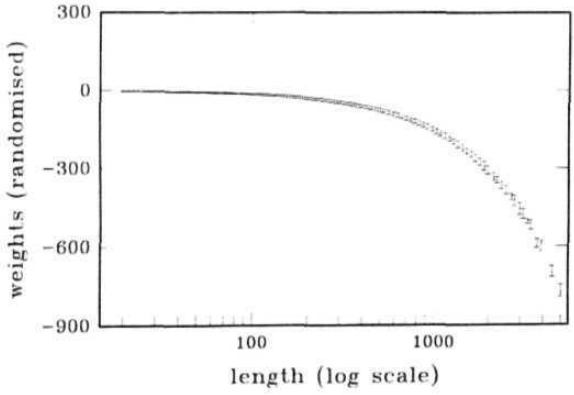

Figure 6.2(b): The weights due to entropy of the random sequences generated using Monte Carlo procedure having an identical length distribution as the sample data-base. For ease of comparison the graph has been plotted in an identical manner as fig. 6.2(a).

of the lack of positional preferences. This can be an appreciable amount for chains of larger lengths (please see Figure 6.2(a)). If a standard state based on completely random sequences were used, the simulated chains would have shown entropy per residue values close to zero (most probable value). However the natural sequences would have shown negative entropies, *i.e.,*

the x-axis in figure 6.4 has been shifted to by 0.1 units to the right.

The overlapping part of the two curves is considerably large. It is because the mixing entropy of short natural fragments are close to the random ones. However, the confidence of distinction between real and random sequences increases with the increasing sequence length. The natural and random-like amino acid sequences are distinguishable if the chain is longer than 70 residues. Note that this length reminds the lower limit of length of protein structural domains. The procedure described above can be a test for identification of protein-coding open reading frames.

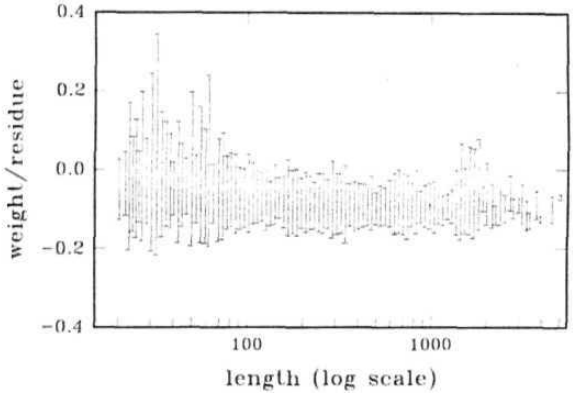

Fig 6.3(a):The weight per residue plotted as a function of the sequence length for natural sequences. The abscissa is taken in the same way as in figures 6.2(a and b). Note the fluctuations in this graph compared to the one below.

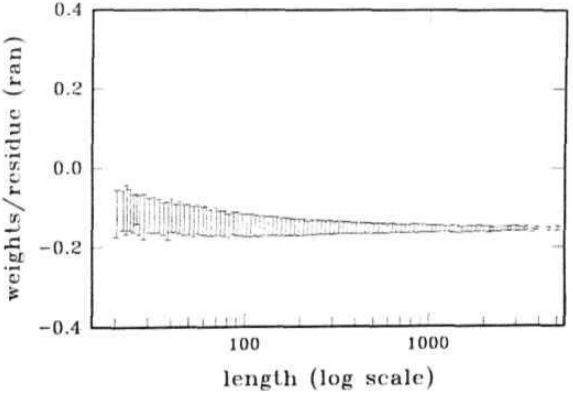

Fig. 6.3(b): The weight per residue is plotted as a function of sequence length in case of random sequences. We note that considerably less fluctuations are seen in the graph.

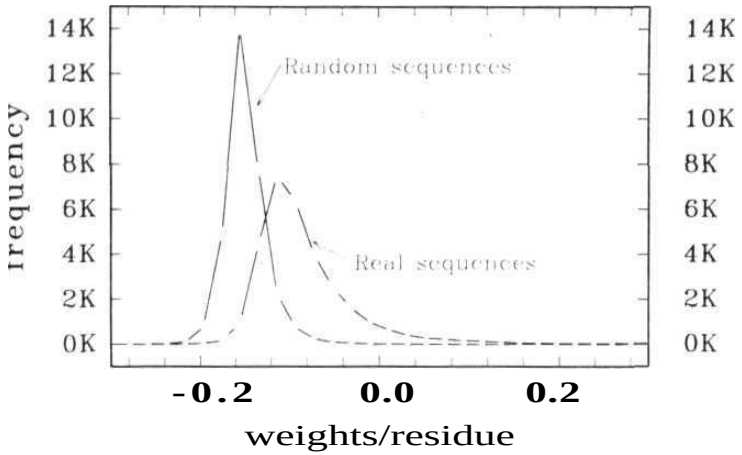

Fig. 6.4: A plot of the frequency distribution for the natural and simulated sequences. The 31,006 weights are classified into 100 classes and their frequencies were obtained. The two distributions are clearly different. Only the important region of the graph is presented and the tails extend beyond the graph shown.

6.7 Discussions

From the results mentioned above, it is clear that the primary sequences of proteins is significantly governed by the pair preferences; this is expected since the folding of the primary sequence into a 3-D compact structure requires a considerable degree of co-operativity amongst the residues. This interaction is mostly of van der Waal's kind; a very strong interaction may preclude the diversity of protein sequences. In this light, the smallness of the "mixing entropy" values is not surprising. It is to be noted that the results obtained are dependent on the data-base used: there is no reason to believe that the data-base used in our computation is a "random representative sample". However, it may not be desirable to doctor the data-base by selective elimination of sequences- this may introduce more bias than we intend to remove. On the other hand, since the data-base is sufficiently large in size, bias is likely to be relatively of lesser importance. Another point to be noted that the computations used do not take into account pair preferences beyond the 10th neighbour. In an earlier work, we have demonstrated that such preferences do exist to a significant extent (Meeta Rani and Mitra, 1994) and may not be ignored. However, long range preferences were not considered in this work because the actual counts may be rather small and based on the present calculations would not be very significant.

Summary: A large data-base with over 31,000 sequences and 10 million residues has been analysed in order to be able to distinguish natural protein sequences from random polypeptides. A weight corresponding to "mixing entropy" has been developed. The entropy values of natural sequences are lower than their random counterparts of same length and similar amino acid composition. Finally it can be argued that natural sequences are a special set of polypeptides with additional qualification of biological functionality that can be quantified using the entropy concept that has been developed.

References

Allinger N L (1976), Calculation of molecular structure and energy by force-field methods, *Adv. Phys. Org. Chem.* 13, 1

Anfinsen C B (1973), Principles that govern the folding of protein chains; *Science* **181,** 223-230

Bailey J L (1976), *Techniques in protein chemistry,* 2nd edition (New York : Elsevier)

Beran J (1992), Statistical methods for data with long range dependence; *Stat. Sci.* 7, 404-427

Black J A, Harkins R.N., and Stenzel P. (1976), Non-random relationships among amino acids in protein sequences, *Int. J. peptide. protein Res.* 8, 125-130.

Blout E R, de Loze C, Bloom S M and Fasman G D, (1960), The dependence of conformations of synthetic polypetides on amino acid conformation, *J Am. Chem. Soc.* 82, 3787-3789.

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

Branden C and Tooze J, (1991), *Introduction to protein structure* (New York & London: Garland Publishing Inc.)

Brouwer L E J (1975), *Collected works,* (eds) A Heyting and H Freudenthal, (New York: Elsevier North Holland);

Chou P Y and Fasman G D (1974a), Conformational parameters for amino acids in helical, β-sheets and random coil regions calculated from proteins, *Biochemistry,* 13,211-221.

Chou P Y and Fasman G D (1974a), Prediction of protein conformation, *Biochemistry,* 13, 222-244

Chou P Y and Fasman G D (1974a), P-turns in proteins, *J Mol. Biol.*, **115, 135-**175

Chou J J and Zhang C T (1993) A joint prediction of the folding of 1490 proteins from their genetic codons, *J theor. Biol.*, **161,** 251-262.

Cserzo M and Simon I (1989), Regularities in primary structures of proteins, *Int. J. peptide. protein Res.* 34, 184-195

Daniell P J (1946), Discussion on "Symposium on auto-correlation in time-series", *Suppl. J R Statist. Soc.* 8, 88

Devaney R L (1988), Fractal patterns arising in chaotic dynamical systems in *The Science of Fractal Images;* (eds) H O Peitgen and D Saupe, (New York: Springer Verlag).

Dewdney A K (1989), A Pandora's box of minds, machines and metaphysics, *Scientific American* **261,** 92

Dose K (1976), Ordering Process; in *Protein structure and Evolution* (eds) J L Fox, Zdenek Deyl and Anton Blazej, (New York: Marcel Dekker), 165-166

Dunker A K and Reuckert R R (1969), Observation of molecular weight determination on poly acry lamide gel, *J Biol. Chem.* **244**, 5074.

Fitch W M and Margoliash E (1967), Construction of phylogenetic trees, *Science,* **155,** 279-284.

Haschemeyer R H and Haschemeyer D E V (1973), *Proteins: a guide to study by physical and chemical methods,* (New York : Wiley).

Hurewicz and Wallman (1941), *Dimension theory,* Princeton University Press

Kendall M (1976), *Time-series,* 2nd edition (London: Charles Griffin)

Kendall M Stuart A and Ord J K (1983), *The advanced theory of statistics,* 4th edition (London, High Wycombe: Charles Griffin)

Kendrews J C (1961), The three dimensional structure of a protein molecule, *Scientific American.* **205,** 96-110

Kolaskar A S and Ramabrahamam V (1983), Conformational properties of pairs of amino acids; *Int. J. Peptide Protein Res.,* **22,** 83-91

Kotelchuck D and Scheraga H A (1968), *Proc. Natl. Acad. Sci. US,* 61, 1163-1170. (not consulted)

Kotelchuck D and Scheraga H A (1969a), *Proc. Natl. Acad. Sci. US,* **62,** 14-21 (not consulted)

Kotelchuck D, Dygert M and Scheraga H A (1968b), *Proc. Natl. Acad. Sci. US,* 63, 615-622. (not consulted)

Mao B, Chou K C and Zhang C T (1994), Protein folding classes: a geometric interpretation of the amino acid composition of globular proteins, *Protein engineering,* 7, 319-330.

Maxfield F R and Scheraga H A (1975), *Macromolecules*, 8, 491-493. (not consulted)

Tanaka S and Scheraga H A (1975), *Macromolecules, 9,* 142- 158 (not consulted)

Lehninger A L (1986), *Biochemisry* (New York: Worth Publishers Inc.)

Levitt M and Chothia C, (1976) Structural patterns in globular proteins, *Nature (London),* **261,** 552-558.

Lifson S and Sander C (1979), Antiparallel and parallel P-strands differ in amino acid residue preference, *Nature (London),* **282,** 109-111

Mandelbrot B B (1983), *The fractal geometry of nature* (New York: W H Freeman &Co.)

Menzer (1943), What is dimension?, *American Mathematical Monthly* **50,** 2-7

Meeta Rani (1990), "Fractal dimensions of protein sequences", *M.Phil. Dissertation,* University of Hyderabad, India

Mitra C K and Meeta Rani (1993), Protein sequences as random fractals, *J. Biosciences,* **18,** 213-220

Meeta Rani and Mitra C K (1994), Periodicities in protein sequences, *J Biosciences,* 19, 255-266

Meeta Rani and Mitra C K (1994), Correlation analysis of the distribution amino acid residues in protein sequences, *J Biosciences,* 19, 1994, 101-108.

Mitra C K , Cserzo M, Simon 1 and Meeta Rani , (1994), Proteins as a special subset of polypeptides, communicated.

Monod J, Symmetry and function of biological systems at the macromolecular level, in *Nobel Symposium-II,* (eds) Arne Engstrom and Bror Strandberg, 15-27.

Nandy A (1994), Recent investigations into global characteristics of long DNA sequences, *Ind. J Biochem. Biophys.,* **31,** 149-155.

Oppenheimer P E, (1986), Real time design and animation of fractal plants and trees, *Computer Graphics,* 20, 4.

Orbach R (1986), Dynamics of fractal networks, *Science,* **231,** 814-819

Peng C K, Buldyrev S V, Goldberger A L, Havlin S, Sciortino F, Simons M and Stanley H E (1992), Long range correlations in nucleotide sequences, *Nature (London),* **258,** 168-170.

Ptitsyn O B (1984), Protein as an edited copolymer, *Mol. Biol. USSR,* **18,** 574-590

Richardson J (1981) The anatomy and taxonomy of protein structures, *Adv. prot. Chem.* 34, 167-339

Richardson J (1985a), Describing patterns of protein tertiary structures, *Methods in Enzymol.* **115,** 349-358.

Richardson J (1985b), Schematic drawings of protein structures, *Methods in Enzymol.* **115,** 359-380.

Rose G D (1978), Prediction of chain turns in globular proteins on a hydrophobic basis *Nature (London),* **272,** 586-590

Sanger F and Tuppy H (1961), The amino acid sequence in the phenylalanyl chain of insulin, *Biochem. J.,* 49, 463-490.

Sanger F and Thompson E O P (1963), The amino acid sequence in the glycyl chain of insulin, *Biochem. J.,* **53,** 353-374

Saroff H.A. (1984), *Bull. Math. Biol.* 46, 661-672 (not consulted)

Shannon C E and Weaver W (1949), *The Mathematical theory of communication,* University of Illinois, Urbana, Illinois.

Shapiro A L, Vinuela E and Maizel J V (1967), Molecular weight estimation of polypeptide chains by electrophoresis in SDS-polyacrylamide gel, *Biochem. Biophys. Res. Commun,* **28,** 815-820.

Szilard L (1925), *Zeit Phys.* 53, 753. (not consulted).

Simon I (1986), Proteins as general crystals. *J theor. Biol.* **123,** 121-124.

Simon I, Glasser L and Scheraga H A (1991), Calculation of protein conformation as an assembly of stable overlapping segments: application to bovine pancreatic trypsin inhibitor, *Proc. Natl. Acad. Sci. US.* 88, 3661-3665.

Sorm F and Knichal V (1962) On proteins: Mathematical approach to the evaluation of similarities in protein structures, *Collection Czechoslav. Chem. Commun.* 27, 1988

Solovyev V V, Korolev S V and Lim H A (1993), A new approach for the classification of functional regions of DNA sequences based on fractal representation, *Int. J Genome Res.* 1 109-128.

Stapleton H J, Allen J P, Flynn C P, Stinson D G and Kurtz S R (1980), Fractal form of proteins, *Phys. Rev. Lett.* 45, pp 1456.

Tudos E, Cserzo M and Simon I (1990), *Int. J. peptide protein Res.* 36, 236-239.

Rogers C A (1970) *Hausdorff measures*, Cambridge University Press

Vonderviszt F, Matrai Gy and Simon I (1986), Characteristic residue environment of amino acids in proteins, *Int. J. peptide protein Res* 27, 483-492

Vonderviszt F and Simon I (1986), A possible way for prediction of domain boundaries in globular proteins from amino acid sequences, *Biochim. Biophys. Res. Comm.* **139,** 11-17.

Voss R F and Clarke J (1975), 1/f noise in music and speech, *Nature (London)* 258, 317-318.

Voss R F (1988), Fractals in Nature, in *The Science of Fractal Images* (eds: H O Peitgen and D Saupe), Springer-Verlag, New York, p21-70

Voss R F (1992), Evolution of long range fractal correlations and 1/f noise in DNA base sequences; *Phys. Rev. Lett.* 68, 3805-3808

Weber K and Osborn M (1969), The reliability of molecular weight determinations by dodecyl sulphate-polyacrylamide gel electrophoresis, *J Biol. Chem.* **244,** 4406

Williams J, Clegg J B and Mutch M U (1961), Coincidence and protein structure, *J Mol Biol.* 3, 533.

Wilmanns M and Eisenberg D (1993), Three-dimensional profiles from residue-pair preferences: identification of sequences with p/a barrel fold, *Proc. Natl. Acad. Sci. US,* 90, 1379-1383

Zehfus M (1986), Continuous compact protein domains, *Proteins, 2,* 90-110.

Zehfus M (1994), Binary discontinuous compact protein domains, *Protein engineering,* 7, 335-340.

Zhang C T and Chou K C (1992), An optimization approach to predicting protein structural class from amino acid composition, *Protein Science,* 1, 401-408

BIO-DATA

Name	Meeta Rani
Father's Name	C L Patel
Date of Birth	July 2nd, 1966
Place of Birth	Bilaspur, MP (India)
Secondary School	Central School, Kanchanbagh, Hyderabad (English, Hindi, Social Sciences, General Sciences and Mathematics)
Higher Secondary School	Central School, Kanchanbagh, Hyderabad (Biology, Maths, Physics & Chemistry)
Graduation Degree	Andhra Mahila Sabha, Hyderabad (Botany, Zoology and Chemistry)
Master's Degree	University of Hyderabad (Biochemistry)
Master of Philosophy	University of Hyderabad Dissertation: Fractal dimensions of protein sequences

List of Publications:

1. Fractal Dimensions of Protein sequences, M.Phil. Dissertation, University of Hyderabad, Hyderabad, India, 1990.
2. Protein Sequences as Random Fractals, *J Biosciences,* 18, 213-220, 1993.
3. Periodicities in Protein Sequences, *J Biosciences,* 19, 255-266, 1994.
4. Correlation Analysis of Distribution of Amino Acids in Protein Sequences, *J Biosciences,* 19, 101-108, 1994.
5. Proteins as a special subset of polypeptides,..*communicated*

Address for correspondence
Meeta Rani, Department of Biochemistry
School of Life Sciences
University of Hyderabad
Hyderabad 500 134
India

Address (Residential)
Meeta Rani, C/o Shri C L Patel
18-8-254/6/A
Rakshapuram Colony
Post: Saidabad, 500 659
Hyderabad, India