PROTEIN SEQUENCE ANALYSIS OF SWISS PROT DATABASE BY USING MARKOV MODEL

A dissertation submitted for the degree of
DOCTOR OF PHILOSOPHY

by

SURYA PAVAN YENAMANDRA

Department of Biochemistry
School of Life Sciences
University of Hyderabad
Hyderabad, India

June 2005
Enrollment No: 01LBPH08

DEDICATED TO MY PARENTS

UNIVERSITY OF HYDERABAD
Department of Biochemistry
School of Life Sciences

Certificate

*This is to certify that the thesis entitled "**PROTEIN SEQUENCE ANALYSIS OF SWISS PROT DATABASE BY USING MARKOV MODEL** " **is** based **on** the research work carried out by Mr. Surya Pavan Yenamandra in fulfillment for the Degree of Doctor of Philosophy under my guidance. This work has not been submitted for any degree or diploma to any other university.*

13/6/05

Chanchal K Mitra **Supervisor**	**Head** **Dept. of Biochemistry**	**Dean** **School of Life Sciences**

Dean, School of Life Sciences
University of Hyderabad.
Hyderabad-500 134. (India)

UNIVERSITY OF HYDERABAD
Department of Biochemistry
School of Life Sciences

DECLARATION

I here by declare that the work presented in this thesis entitled **"PROTEIN SEQUENCE ANALYSIS OF SWISS PROT DATABASE BY USING MARKOV MODEL"** has been carried out by me under the supervision of Prof. Chanchal K Mitra and that this work has not been submitted for a degree or diploma at any other university.

Y. Surya Pavan
Surya PaVan Y
(Enrollment No. 01LBPH08)

13/6

Date:

Prof. Chanchal K Mitra
(Supervisor)

ACKNOWLEDGEMENTS

It gives me immense pleasure to express my profound gratitude and deep respect to my teacher and guide, Prof. Chanchal K Mitra for teaching me research methodology and introducing me to this fascinating field of research. He has been inspiring with thought provoking discussions and invaluable lectures.

My indebtedness and sincere respects are due to Dr.S.M. Bendre for her helpful guidance and encouragement. I have been fortunate to have her obliging discussions.

It is my pleasure to thank Prof. Nikolay A Kolchanov and Prof. Yuri Matushkin, Institute of Cytology and Genetics, Novosibirsk, Russia, for their immense hospitality shown during my stay at Novosibirsk, Russia. I thank all the collaborators for making the ILTP (DST and RAS) project successful.

I thank the Dean, School of Life Sciences and all the faculty members of the School for all their help as and when needed.

I thank the Head, Department of Biochemistry for his cooperation during research.

I am thankful to all my M.Sc Animal science faculty members at University of Hyderabad for encouraging me to take up higher studies.

I also thank Mr. Pattabhi sir, and all the non-teaching staff of the School of Life Sciences, for their assistance on various occasions.

I also thank staff members of CMSD for their assistance and support.

I wish to record my thanks to CSIR, New Delhi for financial assistance.

I wish to thank my M.Sc batchmates for their encouragement.

I wish to thank my be loved seniors Ravi Prasad Aduri, D. Raghuvar Gopal, Puli Ramesh, Y.Sailu, Hassina, Jyothsna, Suneeth, Shanthi for their encouragement.

A special note of thanks to Vyomkesh, Ajoy, Raju, Sharmila, Rajni kanth, Chandu, Sudar, Jaddu, Suhitha, Mahipal, Mallik, Sivaram, Sridhar for their support and pleasant company during my research tenure.

I wish to thank my friendly and cooperative labmates Salomi, Ashok, Shasi Rekha, Goutham, Ramesh and my previous labmates Shailly Varma, Anitha, Venu and Sridevi for creating a cheerful work atmosphere. My stay on this campus has been pleasant with the association of all the scholars of School of Life Sciences. I am thankful to all research scholars of Life Sciences and research friends of School of Chemistry for cheerful company.

I cannot forget my close friends Murali, Narendra and Jaya Ram for their constant encouragement thought out my life. I hope and wish for their prosperity.

The blessings and best wishes of my parents, sister, brother-in-law, brother and sister-in-law have made me what I am and I owe everything to them. Last, but certainly not the least, all the children in my family (Chittu, Riddu and Tinkku) deserve a word of thanks for their smiles.

TABLE OF CONTENTS

INTRODUCTION
TO
PROTEINS

1 INTRODUCTION

1.1 MOTIVATION

Protein secondary structure prediction and protein folding are the problems of the past and the present. Few of the methods were found partially successful regarding secondary structure prediction. However, understanding the protein folding pattern remains a major problem till date. We have tried to study the information content present in the protein primary sequences and look for the existence of any short-range interactions of amino acids in the protein sequences. In the present study, we have applied fractal geometry and information theory to understand the behavior of amino acids in the protein sequences and analyze any role in protein folding.

1.2 INTRODUCTION TO PROTEINS

The term "protein" (prote-first and eidos-like) is derived from the Greek word, which may be translated as "of first rank or position". This term was suggested by Berzelius to Mulder in 1838 to apply to a complex class of nitrogenous organic compounds seen in living organisms. During the late nineteenth century and early twentieth century investigators were concerned with the separation of proteins from their complex environment and determination of their constituent amino acids. The most elegant and significant work in this field, however, came from the laboratory of the German chemist Emil Fisher. His studies have revolutionized research concerning the structures of carbohydrates, fats and polypeptides. Of particular note is the work of Mulder, Liebig, Schutzenberger and others on the isolation of amino acids from protein hydrolyzates and that of Fisher, who deduced how these amino acids are linked in the intact protein (Fisher, 1907).

The first protein to be crystallized was hemoglobin, in 1840, by evaporation of the blood of earthworm; second was globulin from Brazil nut in 1877, and third, ovalbumin in 1889. The first enzyme to be crystallized was urease (Sumner, 1926). Abel (Abel, 1926) crystallized the first protein hormone insulin, in 1926 and Stanley (Stanley, 1935), the first virus, the tobacco mosaic virus.

Functionally, proteins are the most diverse of all biological macromolecules (Lehninger et al, 1993). All proteins, whether from the most ancient lines of bacteria or from the most complex highly developed forms of life, are constructed from the same ubiquitous set of 20 amino acid.

1.3 CHEMICAL NATURE OF PROTEINS:

Proteins are nitrogenous complex organic compounds found in all living cells with molecular weight ranging from a few hundred to several million Daltons. All proteins on hydrolysis produce a group of smaller compounds that can be identified as amino acids. Careful studies show that proteins are polymers made up of amino acids joined head to tail. Proteins with length of upto 20 amino acids are called peptides, for example, oxytocin, vasopressin etc. Proteins with about 100 residues are considered small proteins, e.g., insulin, and those with more than 300 residues are considered to be large. But some proteins can be very huge like viral coat proteins.

Basically proteins are polymers of amino acids but may contain many non-protein elements also. The average composition of proteins is carbon 49-51%; hydrogen 7-8%; oxygen 23-25%; nitrogen 16-17%; sulphur 0-3%. Elements such as iron, iodine, copper, manganese, phosphorus, zinc, heme, cofactors etc. are found in specific proteins as prosthetic groups.

1.4 PROTEIN CLASSIFICATION AND DIVERSITY

Structurally and functionally proteins are the most diverse of all biological molecules. They vary in shapes and sizes, physical and chemical properties, biological function etc. on the basis of which they are often classified.

CLASSIFICATION OF PROTEINS:

1.4.1 CLASSIFICATION BASED ON FUNCTION

On the basis of their functions, proteins may be classified as follows:

i) Enzymes:- These are the most specialized type of proteins with catalytic activity e.g. Trypsin, RNA/ DNA Polymerase etc. All the reactions of organic biomolecules in the cell are catalyzed by enzymes. A large number of enzymes have been isolated which catalyses different types of specific reactions.

ii) Storage proteins:- Gliadin (wheat), zein (corn) and hordein (barley) are the storage proteins of plant origin and are responsible for the supply of nutrients required for the growth of the embryonic plants. Ovalbumin and casein are the storage proteins found in egg white and milk respectively.

iii) Transport proteins:- Many small ions and molecules are transported in the blood and within cells by binding to carrier proteins. The most important carrier protein is hemoglobin, present in red blood cells, and transports oxygen to blood. Cation transporting membrane ATPases are also included in this category.

iv) Regulatory proteins:- These proteins help in the regulation of cellular and physiological activity both in plants and animals. The most important are the hormones which are responsible for the regulation of metabolic activities e.g. insulin, regulates sugar metabolism, the deficiency of which causes diabetes. Many growth hormones are also included in this category.

v) Contractile proteins:- They are found in muscles and organelles helping in movement, e.g., myosin, actin, dynein (found in cilia and flagella) etc.

vi) Protective proteins in vertebrate blood:- Antibodies complement together, help in defense mechanism against foreign antigens. Fibrin and thrombin help in blood clotting thus preventing blood loss from injuries.

vii) Toxins:- They are toxic to living organisms thus helping organisms producing them in offensive and defensive purposes, e.g., clostridium toxin, snake venom, ricin etc.

viii) Hormones:- They act as modulators of biochemical functions, e.g., epinephrine helps in contraction of most smooth muscles and coherin regulate peristaltic rhythmicity of kidney tubules and intestinal musculature.

ix) Structural proteins:- They form structural components of organelles and cells. E.g., collagen connective tissue and elastin in tendons.

1.4.2 CLASSIFICATION BASED ON COMPOSITION

The proteins can be classified on the basis of chemical composition in to two broad groups.

1) Simple or holoproteins - consists only of amino acids.

2) Conjugated or heteroproteins - consists of one or several polypeptide chains and a non-protein part, it may be metal ion, inorganic group, a low molecular weight or a high molecular weight organic compound (sugar, polysaccharide, lipid and nucleic acid). In all these cases, the non-protein component may be bound tightly or only associated loosely. The tightly bound non-protein component is called prosthetic group.

The important proteins of this group are:

i) Nucleoproteins:- Protamines and histones, due to their basic character, combine with deoxyribonucleic acid in the nuclei to form deoxyribonucleoproteins. Ribosomes are ribonucleoproteins, which are combination of ribonucleic acids and proteins.

ii) Phosphoproteins:- These proteins contain phosphoric acid, which esterifies the alcohol group of serine and threonine. Casein from milk, vitellin and vitellenin of egg yolk are well known in this class.

iii) Mucoproteins:- They are proteins combined with more than 4% carbohydrates, e.g., salivary amylase.

iv) Glycoproteins:- These proteins contain less than 4% carbohydrates, e.g., proteins found in snake venom.

v) Lipoproteins:- These are proteins containing lipids like Phosphatidyl glycerides etc.

1.4.3 CLASSIFICATION BASED ON MOLECULAR SHAPE

Based on the shape, two large categories of proteins are

a) Fibrous proteins.

b) Globular proteins.

a) Fibrous proteins: They are elongated, tightly linked polypeptide chains made of several coils. They are insoluble, of animal origin and generally resistant to proteolytic digestion. They are components of structural organs like wool, silk, hair, skin, horn, nails, hoofs, quills, connective tissue and bones. On the basis of solubility, they can further be sub-divided into

i) Collagens:- They are found in the connective tissue of animals. They are insoluble in water and converted to gelatins by boiling in dilute acids or alkalis.

ii) Elastins:- They are seen in elastic tissues like tendons and arteries. They cannot be converted to gelatins like collagens.

iii) Keratins:- They are proteins of hair, wool, quill, hoofs, nails etc. and contain large amount of cysteine.

b) Globular proteins:- These are spheroids or ellipsoids and are generally soluble in water, dilute salt solutions or dilute acids and bases.

Based on other solubility they are classified as

i) Albumins:- They are readily soluble in water and are coagulable, e.g., ovalbumin and serum albumin.

ii) Globulins:- They are insoluble or sparingly soluble in water. Their solubility is increased by addition of small amounts of salts and they are coagulable by heat, e.g., immunoglobulins.

iii) Histones:- They are basic proteins and produce relatively large amounts of arginine and lysine on hydrolysis.

iv) Protamines:- They are strongly basic, low molecular weight proteins, associated usually with nucleic acids and have high nitrogen content.

1.5 GENERAL STRUCTURE OF PROTEINS

1.5.1 AMINO ACIDS

Amino acids are the chemical constituents of proteins, and are characterized by a central alpha carbon atom. The alpha carbon atom indicates the primary position from which the numbering follows for all subordinate groups. Four substituents are connected to this Ca: one substituent is the alpha proton -H, another is the side chain -R that gives rise to the chemical variety of the amino acids, the third is the carboxylic acid functional group (-COOH), and the fourth is the amino functional group ($-NH_2$). The alpha carbon is the asymmetric center of the molecule for all 20 amino acids except glycine, which has only a proton as its side chain (R group). The configuration about the alpha carbon center must be the L-isomer for proteins synthesized on the ribosome. This is probably an accident of chemical evolution where the L-isomer happens to be the one chosen for early prebiotic systems and fixed into evolutionary history. In general, the amino acids almost never occur in equal amounts in proteins. Although the basic components of all proteins are the same 20 amino acids, they differ in the order of sequence from one protein to another.

1.5.2 R-GROUPS OF AMINO ACIDS

There are 20 amino acid residues based on the kind of side chain (or R-group) in them (Tablel). The R-group can be classified on several basis

1.5.2.1 Aliphatic or aromatic

Aliphatic groups are found in Alanine, Valine, Isoleucine, Leucine, Aspartate, Glutamate, Lysine etc. Aromatic groups are found in Phenylalanine, Methionine, Tryptophan and Tyrosine.

1.5.2.2 Polarity and charge

On the basis of polarity and charge they can be divide into

i) Non-polar side chains: They are seen in Alanine, Valine, Leucine, Isoleucine, Proline, Phenylalanine, Methionine and Tryptophan. These non-polar side chains in proteins tend to cluster together on to the inner side of the molecule (for soluble proteins).

ii) Basic side chains: Lysine, Arginine and Histidine contain basic side chains.

iii) Acidic side chains: Aspartic acid and Glutamic acid have this property.

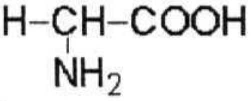
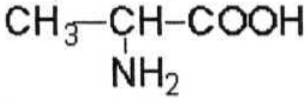
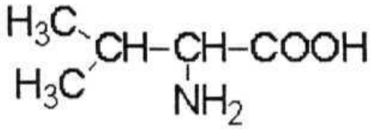
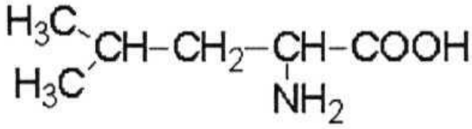
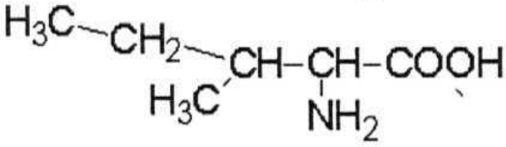
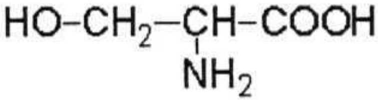
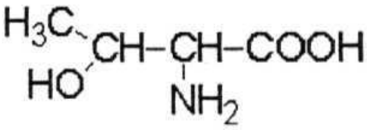
iv) Uncharged polar side chain: They are seen **in** Glycine, Asparagines, Glutamine, Cysteine, Serine, Threonine and Tyrosine.

Table 1 List of all the 20 common amino acids and the pK values of their ionizing groups at 25°C (Albert L. Lehninger, 1988)

Amino Acid	Symbol	pK_1 (COOH)	pK_2 (NH_2)	pKR Group
Amino Acids with Aliphatic R-Groups				
Glycine	Gly - G	2.34	9.60	
Alanine	Ala - A	2.34	9.69	
Valine	Val - V	2.24	9.70	
Leucine	Leu - L	2.36	9.60	
Isoleucine	Ile - I	2.36	9.80	
Non-Aromatic Amino Acids with Hydroxyl R-Groups				
Serine	Ser-S	2.21	9.15	
Threonine	Thr-T	2.63	10.43	
Amino Acids with Sulfur-Containing R-Groups				
Cysteine	Cys-C	1.71	10.78	8.33
Methionine	Met-M	2.12	9.30	
Acidic Amino Acids and their Amides				
Aspartic Acid	Asp - D	2.09	9.82	3.86
Asparagine	Asn-N	2.10	8.80	
Glutamic Acid	Glu - E	2.19	9.67	4.25
Glutamine	Gln - Q	2.17	9.13	
Basic Amino Acids				
Arginine	Arg-R	2.17	9.04	12.48
Lysine	Lys - K	2.18	8.95	10.53
Histidine	His-H	1.82	**9.17**	6.00
Amino Acids with Aromatic Rings				

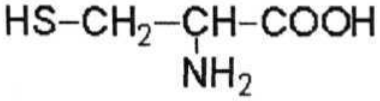
Phenylalanine	Phe - F	2.20	9.20	
Tyrosine	Tyr - Y	2.20	9.11	10.07
Tryptophan	Trp-W	2.40	9.40	
		Imino Acids		
Proline	Pro – P	2.00	10.60	

Table 2: Linear structure of R groups of 20 common amino acids

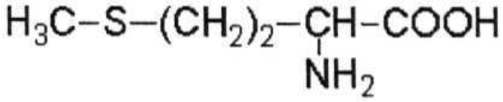
Amino Acid	**Structure**
Amino Acids with Aliphatic R-Groups	
Glycine	
Alanine	
Valine	
Leucine	
Isoleucine	
Non-Aromatic Amino Acids with Hydroxyl R-Groups	
Serine	
Threonine	

Amino Acids with Sulfur-Containing R-Groups

Cysteine

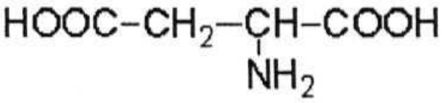


Methionine

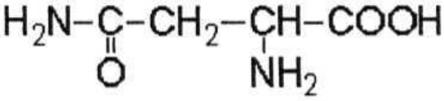


Acidic Amino Acids and their Amides

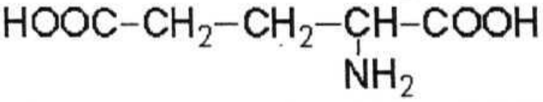
Aspartic Acid



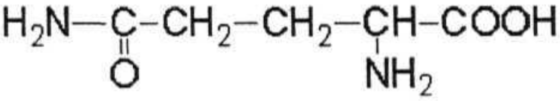
Asparagine



Glutamic Acid

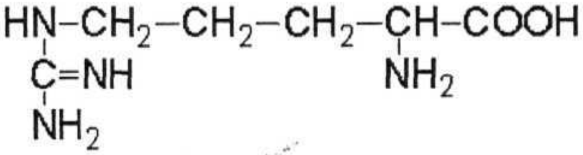


Glutamine

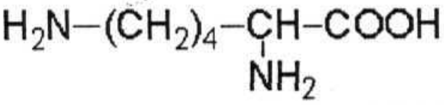


Basic Amino Acids

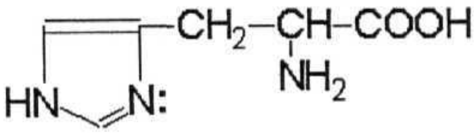
Arginine



Lysine

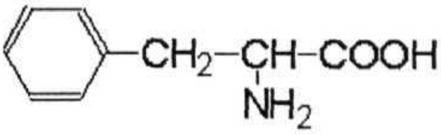


Histidine

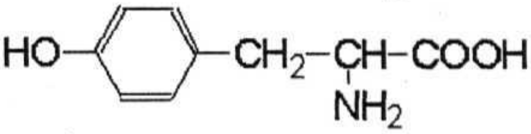


Amino Acids with Aromatic Rings

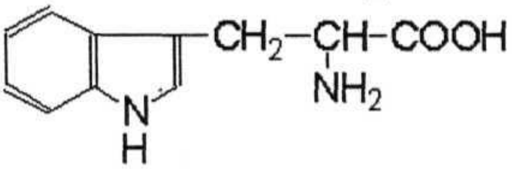
Phenylalanine



Tyrosine

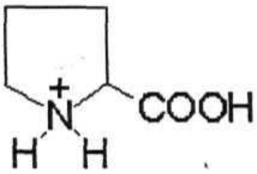


Tryptophan



Imino Acids

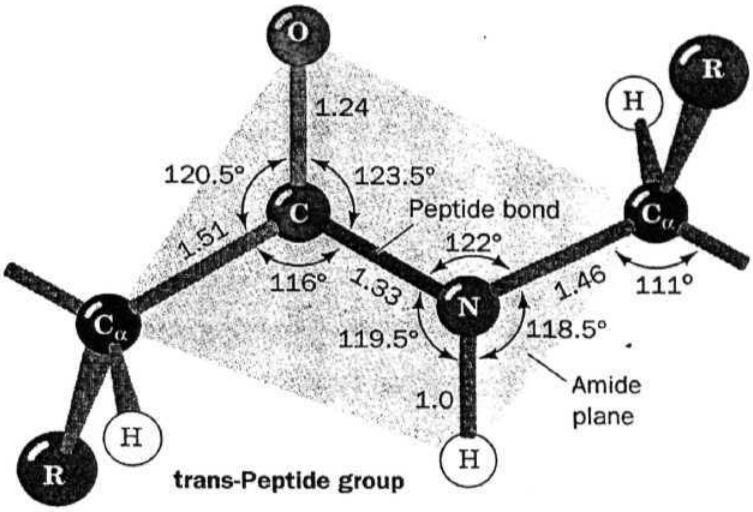
Proline



1.5.3 PEPTIDE BOND

Amino acids are joined by a peptide bond between the carboxyl end of one amino acid and the amino end of another. This bond is formed by loss of a water molecule. The peptide linkage (-CO-NH-) has partial double bond character due to resonance and this is responsible for the planarity of the peptide group (Fig1). Hence, all the atoms of the peptide bond together with C_α atoms lie in a common plane with the carbonyl oxygen and amine hydrogen in trans configuration.

Figure 1 represents typical trans-peptide bond with –CO-NH- linkage (Albert L. Lehninger, 1988).



Consequently the only adjustable geometric features of the dipeptide backbone involve rotations about the single covalent bonds that connect each residue C_α to the adjacent planar peptide groups. Rotations about the C_α-N bond are labeled with a Greek letter Φ (phi) and rotations about the C_α –Carbonyl carbon are labeled Ψ (psi). Due to free rotations about the C_α –C and C_α-N bonds, Φ and Ψ can in principle, take all possible values from -180° to +180°, but in reality the range of permissible values for Φ and Ψ are quite restricted. This restriction arises mainly due to two factors:

i) Steric effects with nearby groups (the R-group of the amino acid residue and/or other atoms in the backbone chain).

ii) Long range interactions with other groups or residues, e.g., hydrogen formation, steric hindrance from distant residues etc. Studies on the conformation of dipeptide groups give us valuable information on the interaction with nearby atoms or with its own side chain (R-group). Such interactions, when taken together with long-range interactions, certainly play an important role in the folding of the protein chains.

1.5.4 POLYPEPTIDE CHAIN

In order to form the amino acid monomers into a polymeric chain, amino acids are condensed with one another through dehydration synthesis. This reaction occurs when water is lost between the carboxylic functional group of one amino acid and the amino functional group of the next to form a C-N bond. These polymerization reactions are not spontaneous; however they can be arranged to occur through the energy-driven action of the ribosome. Ribosomes are complexes of proteins and RNA that translate a gene sequence in the form of mRNA into a protein sequence. The 20 amino acids listed above (Table 1) are encoded by the genes and are incorporated by the ribosomal machinery during protein synthesis. Other minor amino acids are incorporated by ribosomes, but are derived by post-translation modifications.

The reverse reaction, involving hydrolysis of the peptide bond, is thermodynamically spontaneous but kinetically very slow. It can be accomplished chemically, but only under very vigorous conditions. For example, treatment with strong acid (1 molar HCl) and boiling at 100°C can hydrolyze the peptide bonds. So, the reverse hydrolysis reaction actually happens very slowly under normal conditions. Thus, proteins are chemically and biologically stable unless they are deliberately depolymerized. The decomposition of a polypeptide chain into individual amino acids can also be facilitated by hydrolytic enzymes. All proteins are heteropolymeric (i.e., they contain most or all the different amino acids). Only rarely do regions of proteins consist of sequences composed of just a few amino acids. Any region of a typical protein will therefore have a chemically heterogeneous environment. This heterogeneity is further amplified by the higher levels of protein structure, as we will see.

1.5.5 ALLOWED CONFORMATIONS OF POLYPEPTIDES

Many geometric features of a polypeptide chain are fixed owing to bonded interactions between adjacent atoms. These geometric features include all bond lengths and bond angles and vary only very slightly irrespective of the sort of protein structure they occur in. The backbone of the peptide bond has substantial double bond character due to electron delocalization over a Π-orbital system involving the carbonyl oxygen, carbonyl carbon and the amide nitrogen atoms of backbone peptide lying in a common plane with the carbonyl oxygen and amide nitrogen in trans configuration. Consequently the only adjustable geometric features of the polypeptide chain backbone involve rotations about the single covalent bonds that connect C_α atom of each residue to the adjacent planar peptide groups. Rotations about the C_α-N bond are labeled with a Greek letter Φ (phi) and rotations about C_α-Carbonyl carbon are labeled Ψ (psi). In principle, both Φ and Ψ can have any value from -180° to +180° so that the conformations of the polypeptide chain can be described in terms of their Φ-Ψ conformational angle, which automatically takes account of the fixed geometric features of the polypeptide backbone. As a result any polypeptide conformation can be represented as a point on a plot of Φ versus Ψ, where Φ and Ψ have values that range from –180° to +180°.

If all possible Φ and Ψ combinations are investigated, it is found that owing to various sorts of unfavorable steric interactions only a few restricted regions of conformations are possible. The conformationally allowed regions of the Φ and Ψ in the Ramachandran plot (Ramachandran and Sasisekharan, 1968) shows how the accessible regions of Φ and HP space are limited by steric interactions among the polypeptide backbone and side chain groups assuming that atomic groups behave as rigid spheres having appropriate van der Waals radii.

1.6 SECONDARY STRUCTURES

To minimize to the extent of hydrophobic group exposure to solvents and to preserve the favorable energy contributions of the hydrogen bonded interactions formed between the polypeptide backbone and surrounded water in the protein's unfolded state, are the driving forces for secondary structure formations. The secondary structure is assigned based on hydrogen bonding patterns as those initially proposed by Pauling, before any protein structure had ever been experimentally determined (Pauling and Corey, 1950), (Pauling et. al., 1951).

1.6.1 HELICAL STRUCTURES

In an α-helix the polypeptide backbone follows the path of a right-handed helical spring to form an arrangement in which each residue's carbonyl group forms a hydrogen bond with the amide -NH group of the residue four amino acids further along the peptide chain.

All residues in an alpha-helix have nearly identical conformations in which Ψ = -45° to -50° and Φ = -60° so they lead to a regular structure in which each 360° of helical turn incorporates approximately 3.6 residues **and** rises 5.6 A along the helix axis is **1.5** A (Fig 2). The alpha helix is **more stable** than other helical structures due to both good hydrogen bonds and tight packing.

In secondary structure database element 'H' represents α-helix (i, i+4), 'G' represents 3_{10} helix (i, i+3), T represents pi-helix (i, i+5). T is very rare.

Figure 2 shows different ways of diagrammatic representation of regular α-helix (i, i+4) (Albert L. Lehninger, 1988).

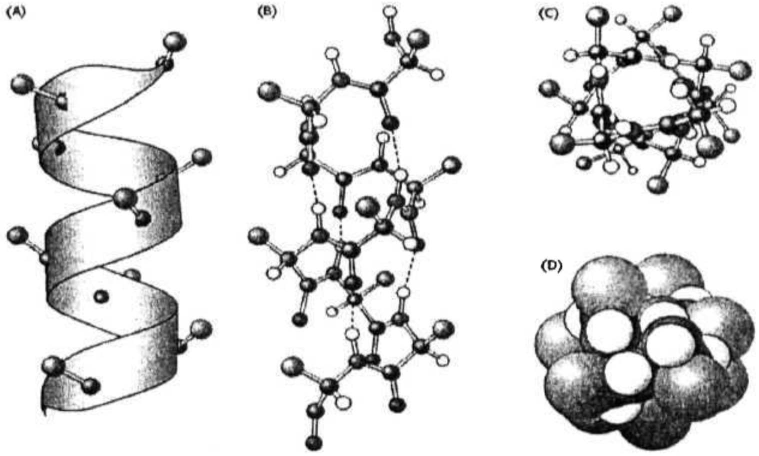


Fig 2: A (side view of a-helix), B (Ball and stick view shows how the residues are getting into helical shape by forming H-bonds), C (Top view of Ball and stick model of regular a-helix), D (Space filled view of a-helix (i, i+4) in protein sequences)

1.6.2 BETA STRUCTURES

p-sheets are formed when regular hydrogen bonds are formed between **the amide -NH and** carbonyl –O groups of the peptide backbones of adjacent chains of almost fully extended polypeptide chains. Β-sheets **can occur** in 2 different arrangements parallel (Fig 3a) and antiparallel.

In parallel P-sheets the chains are arranged with the same N to C polypeptide sense whereas in antiparallel p-sheets (Fig 3b) the chains are arranged with opposite N-to-C sense.

Figure 3a show parallel β sheet representation

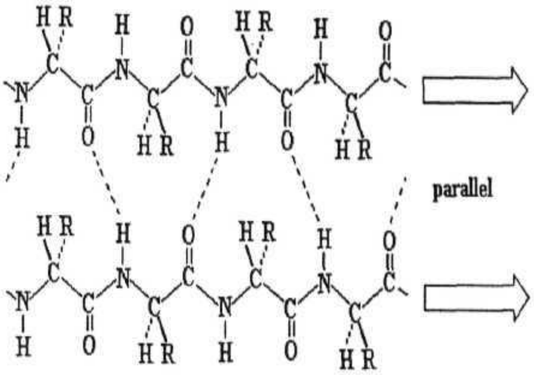


Fig 3a: This parallel (3 sheet representation is obtained when regular hydrogen bonds are formed between the amide -NH and carbonyl -O groups of adjacent chains of the peptide backbones. If both the chains are in either N-C direction or in C-N direction then they are parallel β sheets.

Figure 3b show anti parallel β sheet representation

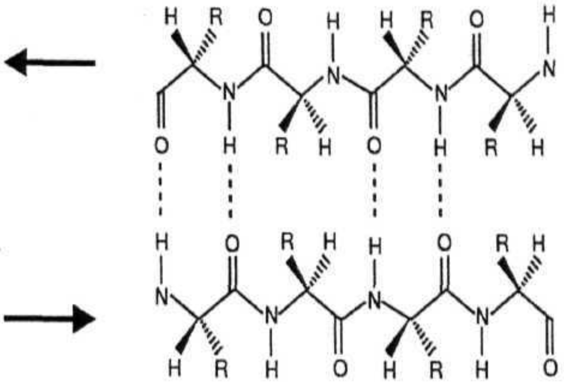


Fig 3b: If the adjacent chains interacting have opposite direction to each other then they give anti-parallel (3 sheet representation.

1.6.3 NONREPETITIVE STRUCTURES

Folding a polypeptide chain to a compact globular form requires a way to change the direction of the polypeptide chain, for example, to connect the adjacent ends of the polypeptide chains in an antiparallel P-sheet. An efficient and commonly observed way to do this is by formation of tight loop in which a residue's carbonyl group forms a hydrogen bond with the amide -NH group of the residue three positions farther along the polypeptide chain. The resulting structure is called a P-bend and it reverses the direction of the polypeptide chain. Glycine is very commonly found in P-bend structure. Random coil (C), Turn (T), Beta bridge (B) and Bend (S) (Fig 4) are common secondary non-repetitive elements (Fig 5) in the secondary structure database.

Figures 4 and 5 are typical non-repetitive bend structures

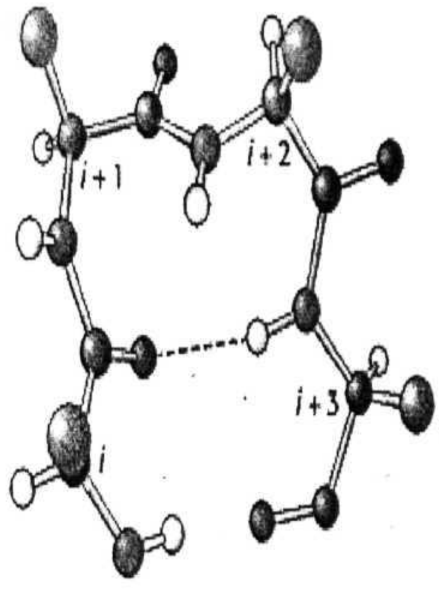


Fig 4: This figure shows the representation of bend structure. The pattern is very similar to 3_{10} helix but it is non repetitive in nature giving rise to bend shape to the region.

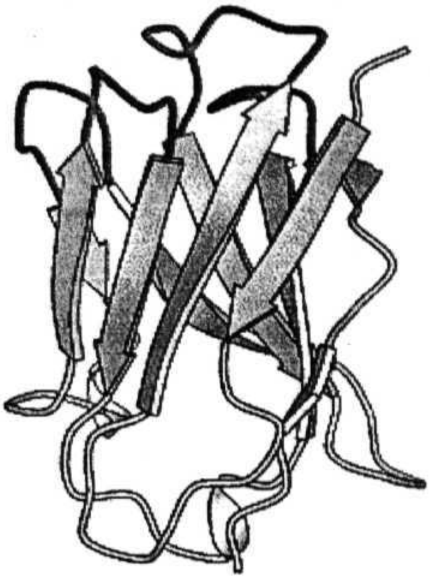
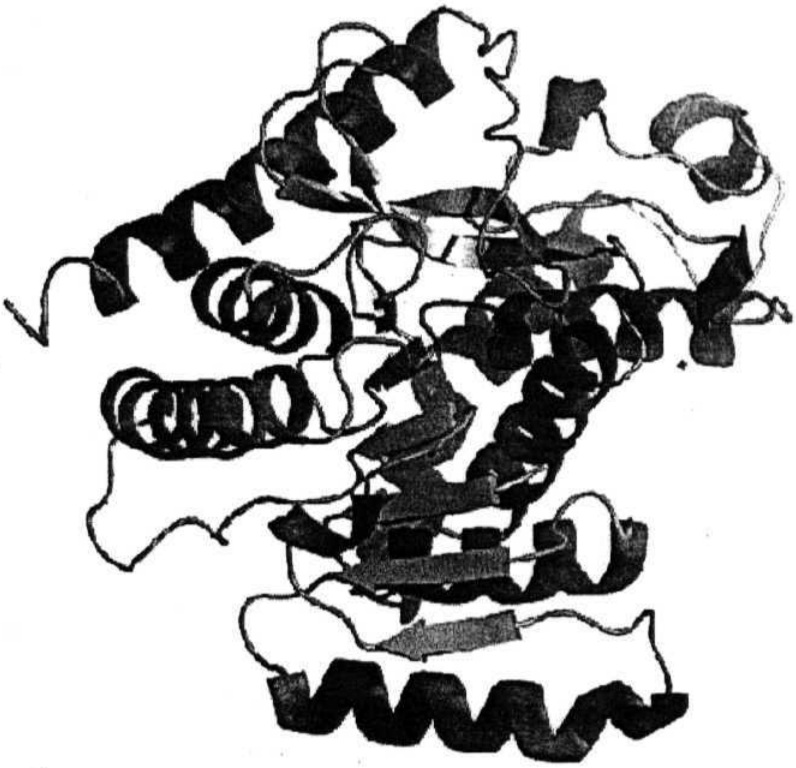


Fig 5: This figure represents the tertiary structure of illustrated protein. The highlighted dark region represents the non-repetitive structure of the protein

1.7 TERTIARY STRUCTURES

The secondary structural elements are packed to form the tertiary structures. The tertiary structures (Fig 6) may be composed of what are known as 'Structural domains'. Domains are sub-regions of the polypeptide chains (encoded by the sub-regions of a structural gene) that are autonomous in the sense that they possess all the characteristics of a complete globular protein. One of the commonly observed arrangements is a special case of the right-handed crossover connections packing a layer of helices on one side of the sheet, then the polypeptide chain returns **to the** middle of the sheet and winds out to the opposite side. This pattern is often known as "nucleotide binding domain" since most of these proteins bind a mononucleotide or dinucleotide cofactor in the middle of the C-terminal end of the p-sheet.

Figure 6 below shows various secondary structure elements packed into a tertiary level of protein (Lactate Dehydrogenase).



Domains serve as modular bricks to aid in efficient assembly. Existence of separate domains is important in simplifying the protein folding process in small separable steps. Hence, tertiary structures of all globular proteins are made up of one or more domains giving each protein its unique 3D structure.

1.8 PROTEIN STABILITY

Thennodynamic measurements indicate that native proteins are only marginally stable entities under physiological conditions. The free energy required to denature them is ~0.4kJ/mol of amino acid residues so that 100-residue proteins are typically stable by only around 40kJ/mol. In contrast, the energy required to break a typical hydrogen bond is ~20 kJ/mol. Various stable configurations of polypeptide chain were demonstrated earlier by Pauling (Pauling and Corey, 1953). Electrostatic interactions, hydrogen bonding and hydrophobic forces are the non-covalent influences to which proteins are subject, each have energetic magnitudes that may total thousands of kilojoules per mole over an entire protein molecule. The non-covalent forces that play a major role in stabilizing the proteins are as follows.

1.8.1 ELECTROSTATIC FORCES

They include charge-charge interactions, charge-dipole interactions and dipole-dipole interactions. Each interaction shows different functional dependence on the distance between the interacting groups. For instance, in case of charge-charge interaction if the charges in question are buried within the protein then their interaction energy is increased as dielectric constant in regions inaccessible to water is much lower than the dielectric constant of water (because the interaction energy is inversely proportional to the dielectric constant of medium). However this interaction may be more complex than the one described above.

Amino acids like Lysine, Arginine, Aspartate, Glutamate and sometimes Histidine are charged at physiological pH and hence, contribute to charge-charge interactions.

1.8.2 VAN DER WAAL FORCES

The van der Waals attractive forces arise due to favorable interactions among the induced instantaneous dipole moments that arise from fluctuations in the electron charge densities of neighbouring non-bonded atoms. Such forces are generally very small yielding energies of 0.1 to 0.2 kcal/mole but can be add up as the number of such interactions are large and hence, can contribute significantly to the total conformational energy.

1.8.3 HYDROGEN BONDING

The strength of the hydrogen bond is due to the unshielded nature of the single proton that makes up the hydrogen nucleus. An attractive interaction exists between the lone pair of electrons of either nitrogen or oxygen atoms and the hydrogen atoms. The attraction is usually directed along the lone pair orbital axis of hydrogen bond acceptor group. The distance between the non-bonded atoms is reduced when a hydrogen bond is present.

The polypeptide chain contains a number of both hydrogen bond donor and acceptor groups in their backbone structure as well as in the side chains. Water also has donor hydroxyl groups and acceptor oxygen groups. Formation of the maximum number of hydrogen bonds (mainly with water molecules) would require the complete unfolding of the polypeptide chain. Since water itself is highly hydrogen bonded the bonds within water molecules must be broken for forming hydrogen bonds with the protein atoms.

The strategy followed by most proteins is to maximize the number of intramolecular bonds of the peptide but to keep most of the potential hydrogen-bond forming side chains on the protein surface for interaction with water molecules.

In transmembrane proteins most of the amino acid side chains must be non-polar *(e.g.* Ala, Val, Leu, Ile, Phe). The very polar CONH groups (peptide bonds) of the polypeptide backbone of transmembrane segments must participate in hydrogen bonds (H-bonds) in order to lower the cost of transferring them into the hydrocarbon interior. This H-bonding is most easily accomplished with alpha helices for which all peptide bonds are H-bonded internally. It can also be accomplished with beta-sheets provided that the beta-strands form closed structures such as the beta-barrel. All membrane proteins of known three-dimensional structure adhere to these principles.

1.8.4 HYDROPHOBIC FORCES

Hydrophobic forces relate to en tropic factors and concern the solvent much more than the solute. Solvent exposure of hydrophobic groups results in the introduction of some extent of ordering in the surrounding water which is energetically unfavorable owing to which water tends to withdraw in the region of non polar hydrophobic molecules forming a rigid hydrogen-bonded network with itself and restricting the number of possible orientations of water molecules forming the water-hydrophobic group interface.

Proteins contain a number of amino acids with predominantly non-polar side chains (Alanine, Valine, Isoleucine, Leucine, and Phenylalanine). Due to hydrophobic forces the amino acids are generally buried in the interior of the folded structure of the protein.

1.8.5 DISULFIDE BONDS

Disulfide bonds form as a protein folds to its native conformation, they function to stabilize its three dimensional structure. The relative reducing chemical character of the cytoplasm, however, greatly diminishes the stability of intracellular disulfide bonds. In fact, almost all proteins with disulfide bonds are secreted to more oxidized extracellular destinations where their disulfide bonds are effective in stabilizing protein structures.

1.9 QUATERNARY STRUCTURE

Quaternary structure is that level of form in which units of tertiary structure (separate polypeptide chains) aggregate to form homo- or hetero- multimers. The subunits are held together by non-covalent interactions.

1.9.1 HETERO-MULTIMERS

In this case we see different tertiary domains aggregating together to form a unit. Sometimes, we find that several domains are found in a single enzyme complex, either in a single polypeptide chain, or as an association of separate chains. A good example is the F-1 ATPase (Fig 7).

Figure 7 shows a heteromultimer protein (Fl ATPase) at quaternary level.

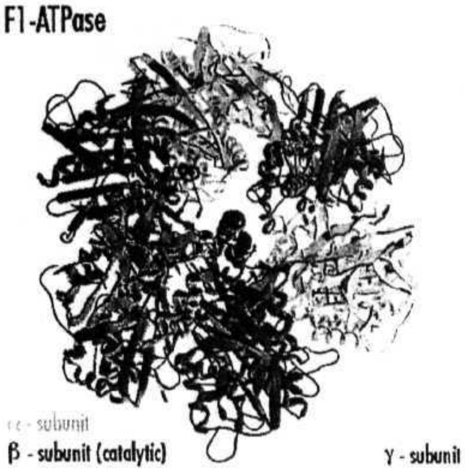


Fig 7: Illustration of Fl ATPase showing different polypeptide chains (different patterns) with highlighted a, β and γ subunits interacting at quaternary level of heteromultimer. The individual tertiary level polypeptides interact with each other by non-bonded interactions giving rise to quaternary level structure.

1.9.2 HOMO-MULTIMERS

It is more common to find copies of the same polypeptide chain associating non-covalently. Such complexes are usually, though not always symmetrical. Because proteins are inherently asymmetrical objects, the multimers almost always exhibit rotational symmetry about one or more axes. The majority of the enzymes of the metabolic pathways seem to aggregate in this way, forming dimers, trimers, tetramers, pentamers, hexamers, octamers, decamers, dodecamers, or even tetradecamers in the case of the chaperonin GroEL (Fig 8).

Figure 8 shows the Homomultimer (GroEL) at quaternary level.

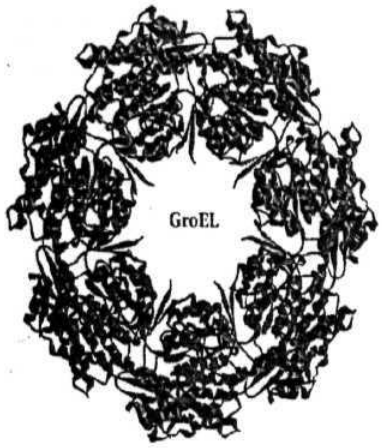


Figure 8: The GroEl chaperonin is a typical example for Homomultimer at quaternary level. The majority of the enzymes of the metabolic pathways seem to aggregate in this way, a tetradecamers in the case of the chaperonin GroEL.

The reason for this is the allosteric cooperativity that results in increased catalytic efficiency, effectively a `sharing' of the small conformational changes that accompany substrate binding and catalytic activity. A good, well-studied example is the `breathing motion observed in the haemoglobin tetramer.

1.10 PROTEIN FOLDING

In a polypeptide chain the amide planes are free to rotate about single bonds to the connecting alpha-carbon atoms. These rotations and twistings give the main backbone chain of the protein its 3D conformation. In addition to this the different side-chains present on the protein make the number of conformations virtually infinite. The protein adopts the conformation of lowest energy from these and hence, a unique folding pattern predominates at physiological conditions. The spontaneity of the renaturation process of a denatured protein to its native conformation indicates that it is energetically favourable.

Causes of folding: A protein in solvent folds only if the AG (i.e., change in the free energy) of the process of going from unfolded to folded state is negative.

AG = AH - T AS

Where AH is the change in the enthalpy of the system (enthalpy is the heat content of a system at constant pressure) and AS is the change in the entropy of the system (entropy is a measure of the disorder of the system). Despite the fact that the covalent bond energies range in the value from 30 to 230 kcal/mole and intramolecular non-bonded interaction energies range from 0.1 to 0.6 kcal/mole, protein folding is mainly directed by the latter, because usually no new bonds are formed or old bonds broken in the folding process. Major intramolecular forces that are involved in the protein folding are mainly the electrostatic forces, van der Waals forces, hydrogen bonds, hydrophobic forces and disulfide bonds.

Moreover, the folding of a polypeptide chain depends on the sequence of the residues as well as the nature of the residues but the different properties of the R-groups does not give any idea about that. For instance, one may find polar R-group in the interior and a non-polar R-group in the exterior regions of a protein but the folding of the protein is such, that most of the polar groups are on the exterior and non-polar groups are in the interior of the protein.

Each R-group is so specific that it cannot be easily replaced by another one. The properties of some side chains, which play known significant roles in the folding, are mentioned below.

Glycine has only a H as a side chain. It being very small and having no bulky side chain hindrance can adopt unusual dihedral angles giving rise to kinks in the main chain and capability to feed the main chain through tight places in the protein molecules. Hence glycine is well preserved during evolution but its frequency is restricted as too great an amount of glycine can make the chain too flexible. Side chains like Valine, Isoleucine and Leucine are branched. Branching limits internal flexibility by stiffening the main chain and side chains. Stiff chains are easier to fix in a certain position. The decrease in the entropy on chain folding is not large and hence, chain folding is facilitated.

Polar and neutral side chains form H-bonds. For instance, serine and threonine have hydroxyl groups and acid amides, asparagines and glutamine act as hydrogen donors to carboxyl groups functioning as acceptors hence, all these can form hydrogen bonds. At high pH the -OH group of tyrosine can form strong hydrogen bonds. Histidine has pK value of 6 in the physiological pH range and may remain charged or uncharged after taking up a H atom from the solution. Charged residues of Aspartate, Glutamate, lysine and arginine are found on the protein surfaces and increase the solubility of the protein globule.

CURRENT APPROACH

Protein secondary structure prediction and protein folding are the problems of the past and the present. Few of the methods were found partially successful regarding secondary structure prediction (Naderi-

Manesh et. al., 2001) (Gibrat et. al., 1987). However, understanding the protein-folding pattern is a major problem till date.

A very brief account of studies done so far and their limitations is given as follows:

According to the work done by (Blout *et al*, 1960) seven types of residues were assigned α-helix forming and breaking properties on the basis of experiments with their synthetic homopolymers. Helix forming residues were those that assumed the a-helix conformation, and those, which did not, were helix-breaking residues. Since only seven residues were considered one doesn't get a complete picture of the roles of all residues in protein folding.

Dirkx (Dirkx *et al* 1972, 1975) determined the frequency of each of the 20 residues in the a-helix, P-sheet and reverse turn regions of the database. This was not a purely probabilistic method, as one required prior information regarding which regions had a-helix and P-sheet conformations.

Based on the x-ray analysis (Chou and Fasman, 1974a,b) gave the conformation of a protein as well as suggested a method for the mechanism of protein folding but it wasn't a probabilistic method.

(Periti *et al* 1974) used doublets for prediction of a-helices and P-sheets with a purely probabilistic method.

Triplets of three adjacent residues were used by Kabat and Wu to predict secondary structure as well as complete chain folding (Kabat and Wu, 1972).

The limitation of most of these methods is that they are not based on the primary sequence data information.

Since the primary sequence has lot of information for a protein to fold (Anfinsen, 1950). We tried to analyze the information present in the primary sequence that may help us in predicting the active structure of the protein. In the present study, various attempts were made to understand the problem by analyzing the protein primary sequence as follows.

a. We tried to study positional frequencies of all the amino acids in the primary sequences to understand the distribution behavior of amino acids.

b. We tried to implement the fractal studies on the frequency distributions of amino acids in the protein sequences. This study helps in identifying the amino acids that are more significant in the proteins with respect to positions.

c. We implemented information theory, to check whether the adjacent amino acids show any short-range interaction during protein folding.

d. We implemented Shannon's information theory on the secondary structure elements to check whether there are any long-range interactions in the protein sequences.

References:

Abel, J. J. (1926) Crystalline insulin, Proc. Natl. Acad. Sci. USA. 12:132-135.

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

Beghin F., Dirkx J. (1975) Proceedings: A simple statistical method to predict protein conformations, Arch Int Physiol Biochim. 83 (1): 167-8

Blout, E. R. De lozé, C. Bloom, S. M. and Fasman G. D. (1960) The dependence of the conformations of synthetic polypeptides on amino acid composition, pp 3787 - 3789

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

Chou, P. Y. and Fasman, G. D. 1974b Prediction of protein conformation, Biochemistry. 13 222-244

Crowfoot, D. (1935) X-ray single crystal photographs of insulin, Nature. 125: 591-592.

Fischer, E. (1907) Syntheses of polypeptides, XVII. Ber. Dtsch. Chem. Ges. 40: 1754-1767

Gibrat, J. F. Garnier, J. Robson B. (1987) Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs, J Mol Biol. 198(3): 425-43.

Harvengt, B. Dirkx J. (1972) Correlations in the primary structure of proteins and evidence for relationships among the proteins, Arch Int Physiol Biochim. 80(1): 191-2.

Kabat, E. A. and Wu, T. T. (1972) Construction of a Three-Dimensional Model of the Polypeptide Backbone of the Variable Region of Kappa Immunoglobulin Light Chains, Proc. Nat. Acad. Sci. U.S 69. 960-964.

Lehninger, A. L. Nelson, D. L. (1993) Principles of Biochemistry, Second edition, Worth Publishers, New York.

Naderi-Manesh, H. Sadeghi, M. Arab, S. Moosavi A.A. Movahedi (2001) Prediction of protein surface accessibility with information theory, Proteins. 42(4): 452-9.

Pauling, L. and Corey, R.B. (1950) Two Hydrogen-bonded spiral configurations of the polypeptide chain, J. Am. Chem. Soc. 72: 5349.

Pauling, L. and Corey, R.B. (1953) Stable configurations of polypeptide chains, Proc. Roy. Soc. Lond. B 141: 21-33

Pauling, L. Corey, R.B. and Branson, H.R. (1951). The structure of proteins, two hydrogen-bonded helical configurations of the polypeptide chain, Proc. Natl. Acad.Sci. USA. 37: 205-511.

Periti, P. (1974) A Bayesian approach to the recognition of discrete patterns with an application to a problem of protein molecular structure, Boll Chim Farm. 113(4): 187-218

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

Stanley, W. M. (1935) Isolation of a crystalline protein possessing the properties of tobacco mosaic virus, Science 81: 644-645

Sumner, J. B. (1926) The isolation and crystallization of the enzyme urease, J.Biol. Chem. 69: 435-441

NATURAL SEQUENCES ARE DIFFERENT FROM RANDOM SEQUENCES

2.1 INTRODUCTION

Protein primary structure is defined as the linear sequence of amino acid residues in a polypeptide chain. The primary structure is responsible for the final three dimensional (3D) folded structure of the protein (*Anfinsen 1993),* which is the biologically active form. There were several attempts to relate primary structure to protein folding. A protein sequence differs from another sequence only in the number, kind of amino acid residues and the sequence of their arrangement. In principle, sufficient number of changes, including additions, deletions or substitutions, in the sequences of any protein can change it into another protein. The central question of protein evolution is, how the mutational changes in amino acid sequence leads to change in the structure and stability and thereby change the protein function.

The primary structure, fold to form 3D, a roughly spherical structure, but with a definite pattern that is called its tertiary structure. The median length of a protein sequence based on the Swiss-Prot database is ~350. However, the distribution of sequence length is quite broad and sequences smaller than 50-100 residues are quite common and on the larger side sequences can be as long as 400-500 residues in length. This again suggests that the final structure is more important than the primary sequence. Just as the 26 letters of the English alphabet can make thousands of meaningful words, these 20 amino acids can make 20^{350} (~10^{455}) different kinds of naturally occurring proteins each with a unique sequence (assuming a typical length of ~350 residues). In reality, we find less than 10^5 sequences in the living system (e.g., human genome is reported to have only ~30,000 genes and many genes code for only one protein). This again suggests that most of the theoretically possible sequences are not biologically meaningful, as they do not meet the essential requirement of a well-defined three-dimensional structure with an useful activity. The sequences that are biologically significant (naturally occurring) are not random sequences but are quasi-random; they represent a minute fraction of the total number of theoretically possible sequences. The typical length of a protein sequence ranges from 10-5000 and the average (modal) length is about ~350. Since any given random sequences of polypeptide does not necessarily form a biologically meaningful active protein (though it may have a unique 3D structure), the secret of the biological activity of protein lies in their specific primary sequences, i.e. the specific distribution of amino acid residues in their sequences.

A series of random numbers has no correlations between successive terms. Since we believe that protein sequences are not random, we have tried to study the positional distribution and correlation between successive amino acid residues in the primary sequences of the proteins. This would help us distinguishing in general between sequences that do not exist in nature (i.e., random polypeptides that are not biologically meaningful) and sequences that are common in nature (i.e., natural polypeptides). In other words, we would be able to characterize and distinguish, at least in principle, the allowed primary sequences in natural proteins.

In the present section, major emphasis was to analyze the positional distribution patterns of amino acids in the natural sequences. This current study may give a clearer idea on how amino acids behave in the protein primary sequences and how different they are from random sequences. Present studies are also related to check whether the distributions of the real sequences follow a stationary distribution.

2.2 METHODOLOGY

For most of the position distribution analysis, we used Swiss-Prot protein sequence databank, Release 37, 1999. It contains over 77,976 protein sequences and about 28,262,817 residues. Lengths of protein sequence vary from a few residues to thousands.

A basic question at this point is whether the data bank used can be considered as a random sample (of all the known and unknown sequences) without replacement. This task has been attempted earlier by Doolittle (Doolittle, *1981).* When a particular protein is sequenced, it is selected based on several considerations: (i) general interest in the protein, (ii) biological significance, (iii) ease of purification and host of other factors. General interest in a protein gives rise to certain homologies that are present in the databases. However, their total number is relatively small and we do not expect that this will bias the database 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 database. The major causes of bias in the database we have used come from fragments and short sequences.

Keeping in mind the above condition, we have ignored fragments and short sequences (less than 256 residues) present in the database. We have used this to avoid one-sided bias in the results. After trimming the database (i.e., removing fragments and short sequences) we found that there were 41,408 protein sequences and a total of 22,408,660 amino acid residues in the Swiss-Prot databank. This working database was used for our further studies (referred to as the working database in the following sections unless explicitly mentioned otherwise)

Reasons for ignoring short sequences and fragments:

1. Since specific information can be available at the beginning of protein sequences, we have removed fragments and short sequences, which may interference with characteristics of larger sequences while analyzing.
2. Based on the sequence length distribution of Swiss-Prot database, one can say that short sequences are more in number and may interfere and reflect the properties of larger proteins and their beginning regions.
3. The selection for trimming protein sequences below 256 residues is not specific, and the number selected doesn't have any significant reason and the same can be done by using any other number like 128, 512 etc.
4. Selecting residues above 256 residues is one of the convenient ways for better computations.

The bias present in the Swiss-Prot protein sequence databank can be attributed to several factors, but it is certain that the database cannot be considered a random sample. The proteins that have been selected for sequencing are not chosen at random. Therefore a complete protein database derived from the genomic data of a given organism may be a better representative sample. However, no such database is available at present. The ASTRAL SCOP (http://astral.berkeley.edu) (Brenner *et al,* 2000) database suggests a novel idea in selecting a set of non-redundant, and *possibly* a random representative protein sequence database. We have chosen two sets of protein sequences, one with less than 40% and other with less than 95% sequence similarity (Brenner *et al,* 2000) for our reference. However, removing similar sequences does not assure that the database is unbiased. Some of the computations have been done even with this non-redundant database for comparison.

An algorithm has been developed for simulating random protein sequences. For every natural protein sequence, a random chain of the same length was simulated using Monte Carlo technique, taking into consideration the natural abundances of various amino acids of Swiss-Prot database. We simulated 41,408 random sequences and further analysis in the same way as of Swiss-Prot database for comparison..

2.2.1 POSITIONAL DISTRIBUTION STUDIES

Frequencies of 20 common amino acids were calculated using C++ programming language on Linux operating system. The individual positional frequencies of amino acids were calculated using the working database. The frequencies were constructed by actual counting the number of amino acids present of type X (one of the 20 common amino acids) at the first position in the working database that corresponds to the frequency of X at position 1. In a similar manner, the frequencies of all the 20 amino acids were computed at each position upto 256. For convenience the frequencies were converted into percentage distributions for all common amino acids. Similar approach was carried out for the 20 common amino acids using the random database. The obtained positional distributions were used for further analysis.

CORRELATION STUDIES: Correlation studies were carried out to check whether there are any differences among the distributions of amino acid of natural and simulated (random) sequences. The initial 50 positional frequencies of various amino acids were considered and their distributions are compared between the natural (Swiss Prot working database) and random sequences. The first two positions were skipped for obvious reasons (many of the sequences in the database has been obtained from the corresponding cDNA sequences and the first amino acid is often Methionine which is coded by AUG that also stands for the start codon).

2.2.2 STEADY STATE DISTRIBUTION OF PROTEIN SEQUENCES

We have implemented statistical Analysis of Variance (ANOVA) (Sokal *et al)* test to check whether the distributions of amino acids follow a steady state behavior. The Swiss-Prot working database was divided into 8 equal mini databases for comparison. Each mini database consists of 5176 protein sequences. Positional ranges of 10 to 20, 50 to 60, 100 to 110, 150 to 160 and 200 to 210 were selected for calculating the frequencies of 20 amino acids from the 8 mi ni-databases. The eight frequency values obtained for each of the 20 amino acids, corresponding to eight different mini databases, each at different selected positions, are compared.

The obtained frequencies from all the eight mini databases were normalized (converted to per million) and used for comparing the distributions among the selected positions. The frequency counts were converted to per million by using the formulae:

$$\frac{\text{Actual count of aminoacid residue at selected region * 1.0E+ 6}}{\text{Total counts of all aminoacids in the minidatabase up to 256th position}} \qquad 0)$$

We compared the normalized frequencies to test (ANOVA) for differences.

Similar computations were carried out on the ASTRAL SCOP non-redundant sets. We have chosen two database sets of protein sequences with less than 40% and 95% sequence similarity.

NON-REDUNDANT PROTEIN DATABASE

We have used ASTRAL SCOP non-redundant databases **with less than 40%** and less than 95% similar protein sequences in it.

We have selected both the databases and checked for the distribution behavior of amino acids in these sequences at various positional ranges as described above.

Non-redundant set with less than 40% sequence identity

Details of the database: There are total of 5176 protein sequences in this set. The database has 1312 protein sequences after removing the fragments and short sequences less than 256 residues. We have divided the protein database into 4 mini databases. Each mini database has 328 protein sequences in an equal ratio.

Frequency counts were calculated for all the selected regions of 10 to 20, 50 to 60, 100 to 110, 150 to 160 and 200 to 210 in the four mini databases created.

The 4 frequency values for a given amino acid, corresponding to four mini databases, each at different selected positions, are now compared.

The obtained frequencies from all the four mini databases were normalized per million as described in the Eq 1.

ANOVA test was conducted to compare the frequency distribution at various selected positional ranges.

Non-redundant set with less than 95% sequence identity

Details of database: There are total of 5176 protein sequences. After removing the fragments and short sequences, the database contains 1995 protein sequences. We have divided the protein database into 5 mini databases. Each mini database consists of 399 protein sequences.

Similar approach was followed as above to check the distribution of various amino acids at the selected regions. ANOVA test is conducted to compare distributions of various amino acids at the selected regions.

To further substantiate that the selected regions in the working database follow a stationary distribution we used the following formulae (Cover and Thomas, 1991).

$$\mu_i \mathrm{P}_{ij} = \mu_j \qquad (2)$$

wherein μ_i is the probability value of the ith amino acid at the selected region (for illustration we have selected 50 - 60th position).

Transition matrix: After obtaining the pair frequency count (20x20 matrix) from the selected region in the working database, each individual row of the matrix was normalized by taking the sum for each row and dividing each element of that row by the corresponding row sum. This matrix is now called the transition matrix.

$$\mathrm{p}(\mathrm{j} \mid \mathrm{i}) = \frac{\text{Frequency of } i\text{th with adjacent } j\text{th amino acid}}{\text{Rowsum of } i\text{th row.}}$$

p (j|i) = Conditional probability of *j*th given the probability of *i*th residue in succession.

P_{ij} is the transition matrix of the adjacent *i*th, *j*th residues at the selected region.

If the product of μ_i and P_{ij} is equal to μ_j then the distributions are said to follow a steady stationary distribution.

2.2.3 COMPARISON OF TWO DIFFERENT DISTRIBUTIONS BY LOG ODDS METHOD

We have implemented log likelihood ratio method for comparing two different distributions of amino acids at the selected regions (Durbin *et al*)

$$S(x) = \log \frac{P(x \mid \text{model} +)}{P(x \mid \text{model} -)} = \sum_{i=1}^{L} \log \frac{a^{+}_{x_{i-1}x_i}}{a^{-}_{x_{i-1}x_i}} = \sum_{i=1}^{L} \beta_{x_{i-1}x_i} \qquad (3)$$

where x is the sequence and $\beta_{x_{i-1}x_i}$ are the log likelihood ratios of corresponding transition probabilities.

Model + = proposed non-stationary chain (selected 10-20 region).

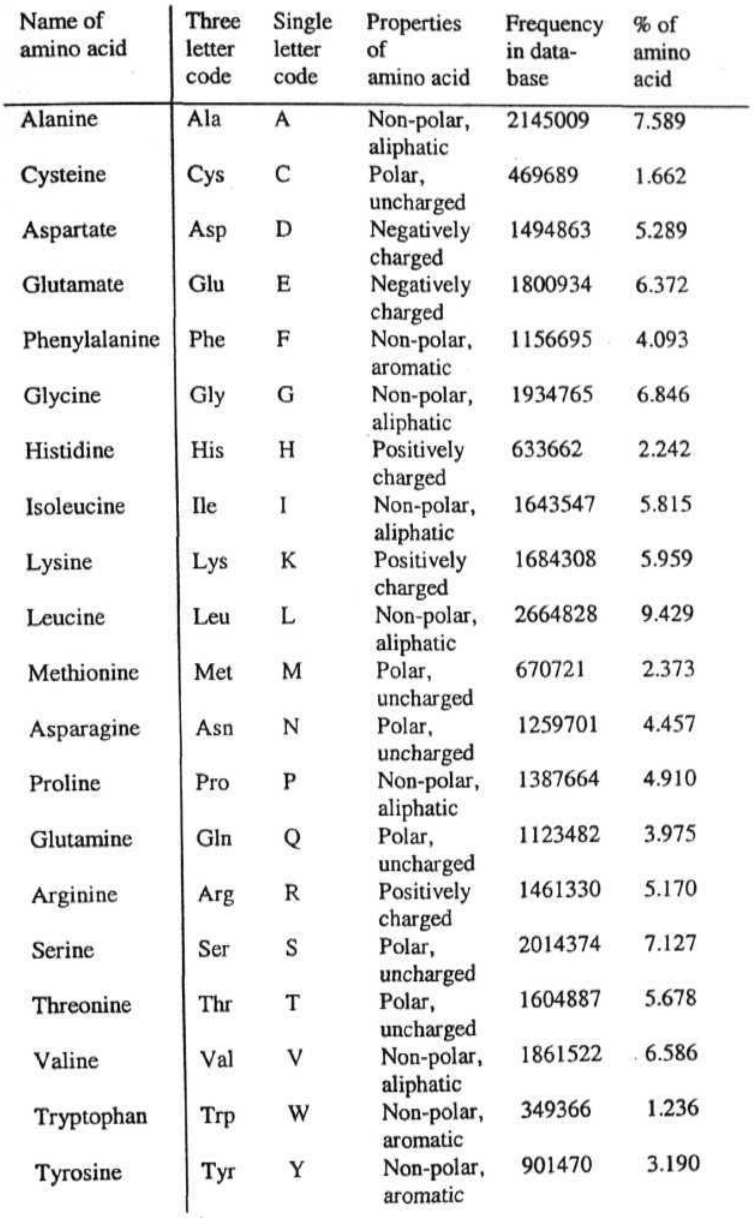
Model - = proposed stationary chain (selected 50-60 region).

All the computations have been carried out on an IBM compatible PC using C++ under Linux (gcc).

2.3 RESULTS AND DISCUSSIONS

The computational studies have revealed that there are total 77,976 sequences with 28,262,817 residues in the Swiss Prot protein databank (Release 37). Of the 20 common amino acid residues, Leucine is the most abundant and Tryptophan is the least abundant residue in the databank. A clear description of amino acids with their respective properties and abundance values in the Protein Databank is given in Table1.

Table 1: Amino acids abundance in the database.



Name of amino acid	Three letter code	Single letter code	Properties of amino acid	Frequency in data-base	% of amino acid
Alanine	Ala	A	Non-polar, aliphatic	2145009	7.589
Cysteine	Cys	C	Polar, uncharged	469689	1.662
Aspartate	Asp	D	Negatively charged	1494863	5.289
Glutamate	Glu	E	Negatively charged	1800934	6.372
Phenylalanine	Phe	F	Non-polar, aromatic	1156695	4.093
Glycine	Gly	G	Non-polar, aliphatic	1934765	6.846
Histidine	His	H	Positively charged	633662	2.242
Isoleucine	Ile	I	Non-polar, aliphatic	1643547	5.815
Lysine	Lys	K	Positively charged	1684308	5.959
Leucine	Leu	L	Non-polar, aliphatic	2664828	9.429
Methionine	Met	M	Polar, uncharged	670721	2.373
Asparagine	Asn	N	Polar, uncharged	1259701	4.457
Proline	Pro	P	Non-polar, aliphatic	1387664	4.910
Glutamine	Gln	Q	Polar, uncharged	1123482	3.975
Arginine	Arg	R	Positively charged	1461330	5.170
Serine	Ser	S	Polar, uncharged	2014374	7.127
Threonine	Thr	T	Polar, uncharged	1604887	5.678
Valine	Val	V	Non-polar, aliphatic	1861522	6.586
Tryptophan	Trp	W	Non-polar, aromatic	349366	1.236
Tyrosine	Tyr	Y	Non-polar, aromatic	901470	3.190

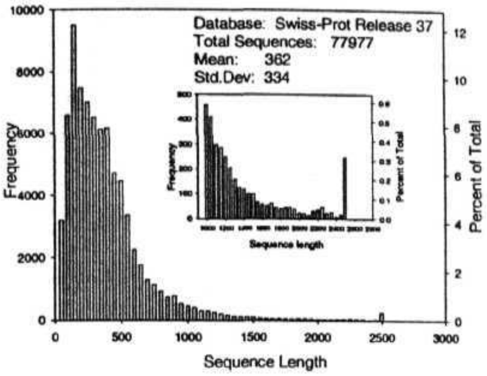


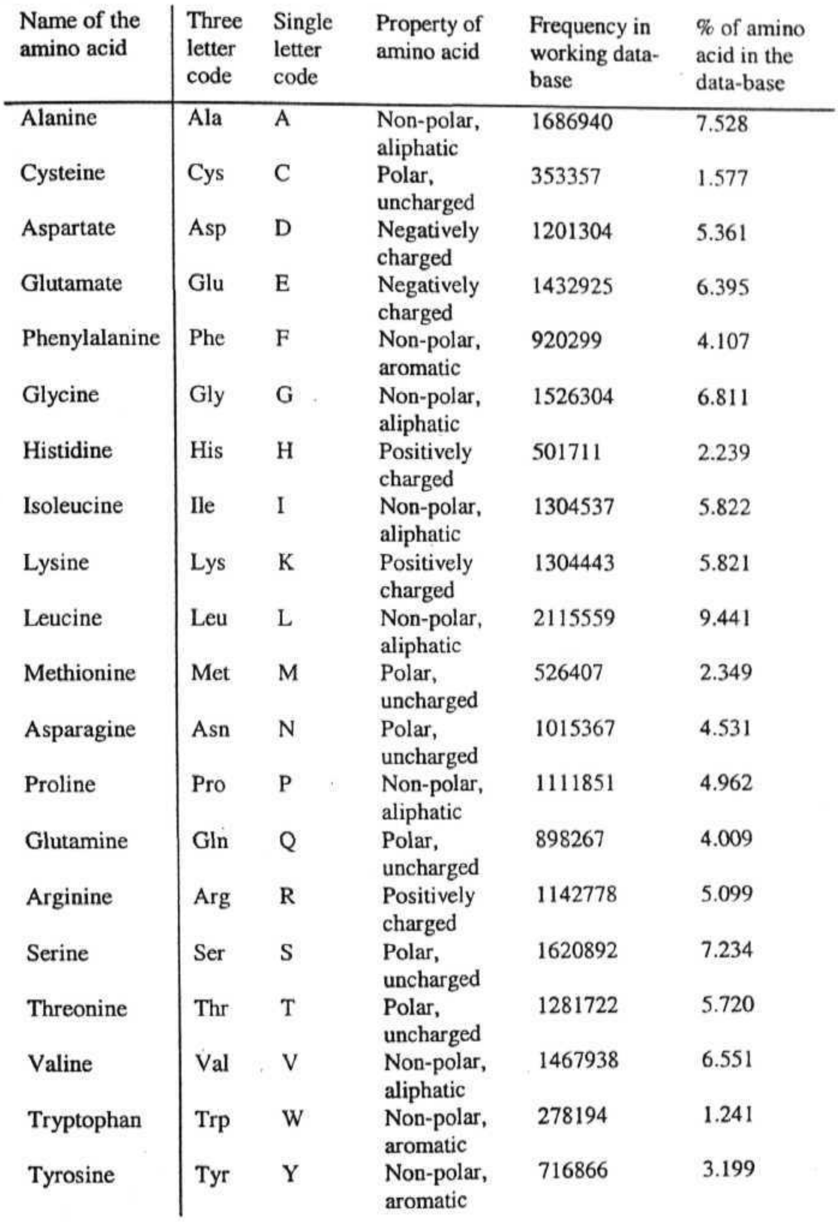
Fig 1: This figure gives a clear description of sequence length distribution of proteins in the complete database. The tail region (distributing from 1000 to 3000 sequence length Vs frequency) is highlighted.

The Fig 1 gives clear description of sequence length distribution of proteins in the complete Swiss-Prot database. The distribution of sequence length is quite broad and sequences smaller than 50-200 residues are quite common and on the larger side sequences are as long as 400-500 residues in length. In the above figure, the tail region of sequence distribution is highlighted. Moreover, the mean of sequence length distribution is 362.

For our studies, we have ignored fragments and short sequences (less than 256 residues) present in the database and this was done to avoid one-sided bias in the results.

After trimming the database (i.e., removing fragments and short sequences) we found that there are 41,408 protein sequences and a total of 22,408,660 amino acid residues in the SWISS PROT databank. This working database was used for our further studies (referred to as the working database in the following sections unless explicitly mentioned otherwise). The amino acid composition in the working database is given in Table 2.

Table-2: Amino acids in the working database.



Name of the amino acid	Three letter code	Single letter code	Property of amino acid	Frequency in working data-base	% of amino acid in the data-base
Alanine	Ala	A	Non-polar, aliphatic	1686940	7.528
Cysteine	Cys	C	Polar, uncharged	353357	1.577
Aspartate	Asp	D	Negatively charged	1201304	5.361
Glutamate	Glu	E	Negatively charged	1432925	6.395
Phenylalanine	Phe	F	Non-polar, aromatic	920299	4.107
Glycine	Gly	G	Non-polar, aliphatic	1526304	6.811
Histidine	His	H	Positively charged	501711	2.239
Isoleucine	Ile	I	Non-polar, aliphatic	1304537	5.822
Lysine	Lys	K	Positively charged	1304443	5.821
Leucine	Leu	L	Non-polar, aliphatic	2115559	9.441
Methionine	Met	M	Polar, uncharged	526407	2.349
Asparagine	Asn	N	Polar, uncharged	1015367	4.531
Proline	Pro	P	Non-polar, aliphatic	1111851	4.962
Glutamine	Gln	Q	Polar, uncharged	898267	4.009
Arginine	Arg	R	Positively charged	1142778	5.099
Serine	Ser	S	Polar, uncharged	1620892	7.234
Threonine	Thr	T	Polar, uncharged	1281722	5.720
Valine	Val	V	Non-polar, aliphatic	1467938	6.551
Tryptophan	Trp	W	Non-polar, aromatic	278194	1.241
Tyrosine	Tyr	Y	Non-polar, aromatic	716866	3.199

To understand the behavior of common amino acids in the protein sequences we have analyzed the positional distribution of each amino acid. This approach may give a clear idea on the general behavior of all the common amino acids in the protein sequences.

2.3.1 POSITION DISTRIBUTION OF COMMON AMINO ACIDS IN THE SWISS PROT DATABASE

To understand the distribution pattern of all the common 20 amino acids in the protein database we have calculated the frequencies of amino acids from the N-terminal of sequence at a block interval of 50 upto 500th position. These frequency values were converted into percent distribution and graphically represented upto 500th position.

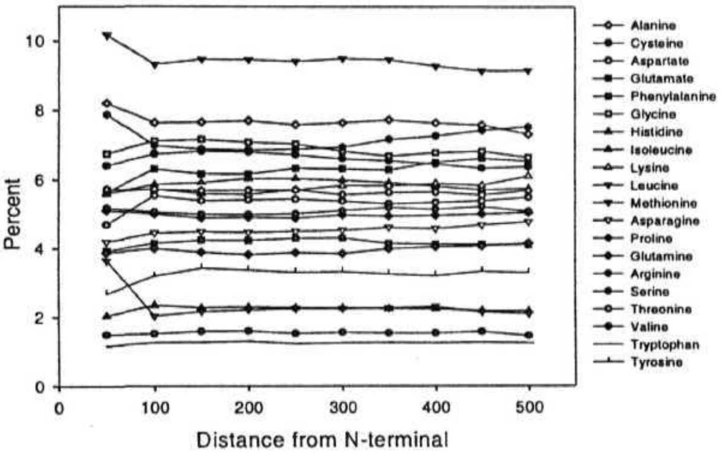


Figure 2 shows the distribution of 20 common amino acids in the protein sequences upto 256th position. The percentage distribution of amino acids from the N-terminal of sequence at every 50th position (assuming as one interval) up to 500th position has been plotted. Except at the beginning positions, the distributions followed a steady state in most of the amino acids.

In Fig 2 the amino acids Alanine, Leucine, Methionine, Lysine, Serine showed high frequency distribution at beginning regions (upto 30th position) and attained a steady distribution at later positions. The amino acids Glutamate, Aspartate, Glycine and Tyrosine showed low frequency distribution at beginning regions and achieved a steady state behavior at later positions (after 30 to 40 positions). Overall, the amino acids showed a steady positional distribution, except at few beginning regions. In the present section, the observed positional behavior of amino acids was confirmed by using various statistical methods.

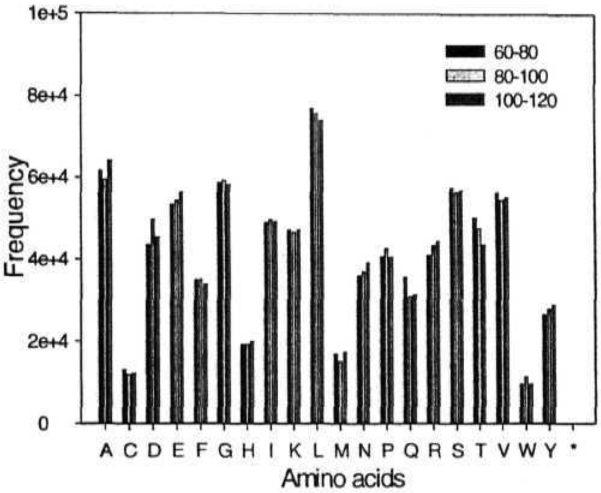


Fig 3 shows how amino acids are distributed in the selected regions of protein sequences. Frequencies of various amino acids are calculated at position intervals of 60-80, 80-100 and 100-120. No significant change was observed in the distribution of amino acids.

* indicates the non-standard amino acids.

Figure 3 gives the frequency distribution of amino acids in the selected regions (60-80, 80-100 and 100-120) of protein sequences. It is evident from the figure that no significant change in the frequencies of amino acids was observed in the selected regions of protein sequences of the working database.

Individual amino acid positional distributions are graphically represented for further analysis. Figure 4 shows the distribution behavior of Alanine and Cysteine. Alanine showed a steady state distribution (~7.5%) all over except at the beginning 30-40 positions. Moreover, Cysteine showed steady state distribution (-1.5) all over in the observed regions of the database.

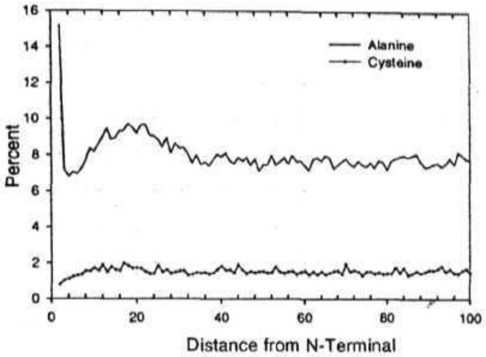


Figure 4: This figure shows the distribution of Alanine and Cysteine from N-terminal region. Alanine distribution is high at the beginning positions and is maintained at a steady state after 30-40 positions. Where as Cysteine showed a steady behavior all over.

Note: We have skipped the first position for amino acid distribution studies, as it is a well known that Methionine by default dominates the first position in most of the protein sequences.

Figure 5 shows the positional distribution of Leucine and Lysine in the database. The Leucine and Lysine showed steady state distribution of ~9.4 and ~5.8 respectively. In the figure, the positional frequencies at beginning regions (upto 30th position) of both amino acids is high compared to the steady state distribution observed in the later positions.

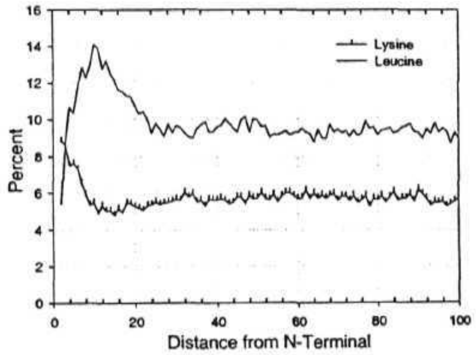


Figure 5: This figure shows frequency distributions of Leucine and Lysine. The Leucine distribution is more at initial positions and achieved a steady distribution after 30-40 positions. The Lysine distribution is at steady state allover, except at few beginning regions.

The figure 6 shows the positional distribution of Tryptophan and Tyrosine. The distribution of these amino acids has been maintained in a steady state (-1.2 (Trp) and ~2.2 (Tyr) respectively) all over the positions considered.

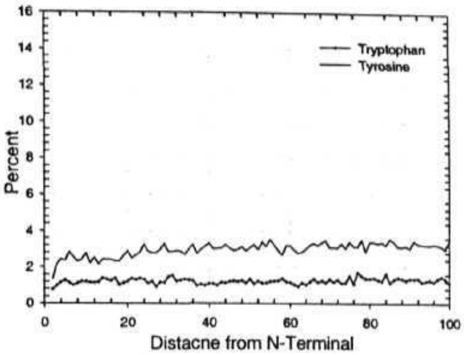


Figure 6: In this figure there is no significant change in the distribution of both Tryptophan and Tyrosine in the protein sequences of the database up to observed 100 positions from the N-terminal region.

2.3.2 DISTRIBUTION STUDIES FROM C-TERMINAL REGION

The N-terminal distribution studies showed that the common amino acids follow a steady state distribution all over in the protein sequences except at few beginning regions. These observations made us curious to check the distribution behavior of amino acids from the C-terminal region. The distribution studies at C-terminal end of protein sequences reveal that the common amino acids follow a steady state behavior all over.

Figure 7 and 8 shows the positional distribution behavior of Alanine, Cysteine, Glutamate, Valine, Tryptophan and Tyrosine from C-terminal region. As it is evident from the figures the distributions of these amino acids are maintained at

~7.5%, ~1.5%, ~6.3%, ~6.5%, ~1.2% and -3.1% respectively. There is a clear steady state distribution observed in the amino acids considered.

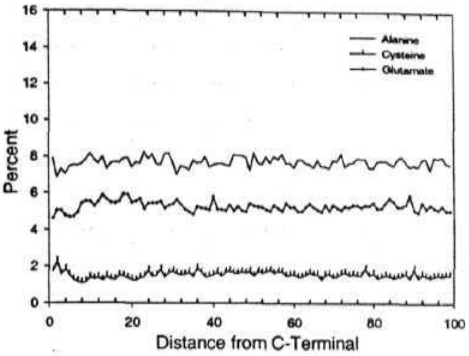


Figure 7: This figure shows the distribution of amino acids from the C-terminal region. There is no significant change in the distribution of amino acids from the C-Terminal region.

Note: The X axis shows position distribution from the beginning of C-terminal and it is indicated as starting point (zero).

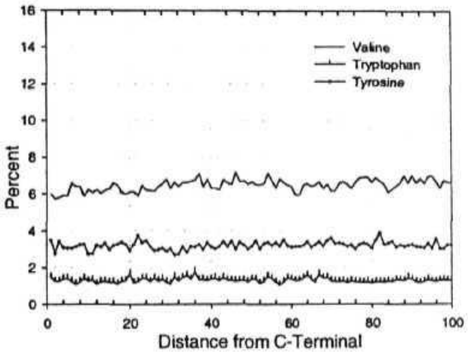


Figure 8: This figure shows the distribution of amino acids, Valine, Tryptophan and Tyrosine from the C-terminal region. In fact there is no significant change in the distribution of amino acids from the C-terminal region.

Note: The X axis shows position distribution from the beginning of C-terminal and it is indicated as starting point (zero).

2.3.3 DISTRIBUTION STUDIES OF COMMON AMINO ACIDS OF RANDOM (SIMULATED) SEQUENCES

The percent distributions of amino acids of random sequences are shown in the figure 9. We have considered the amino acids Alanine, Cysteine, Aspartate, Glutamate and Phenylalanine for distribution studies. As expected, figure 9 shows a uniform distribution for the amino acids selected upto 50th position. Moreover, their positional distributions are maintained as of the amino acid abundances of the natural sequences (Table 2).

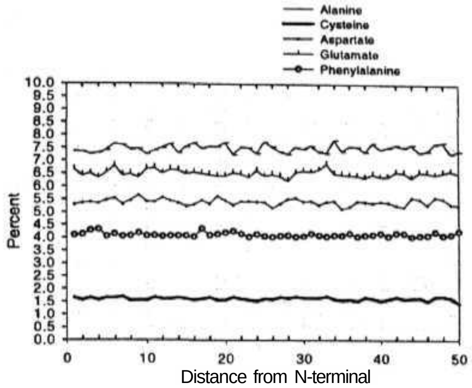


Figure 9: This figure reveals the individual amino acid distribution in the random sequences. The distributions of amino acids illustrated show a uniform random behavior.

2.3.4 TO SHOW NATURAL SEQUENCES ARE DIFFERENT FROM SIMULATED (RANDOM) SEQUENCES.

CORRELATION STUDIES

Correlation studies were carried out by comparing amino acid positional distributions of real and random databases. The correlation analysis was done by obtaining the frequency value of amino acid of type 'X' (any common amino acid) at *ith* position *(i* goes from 1 to 50 positions in our studies) from Swiss Prot database and comparing with frequency value of amino acid of type 'X' at *i*th position of random database. The frequency values of the first two positions were skipped for obvious reasons (many of the sequences in the database has been obtained from the corresponding cDNA sequences and the first amino acid is often Methionine which is coded by AUG that also stands for the start codon). After skipping the first two positions, the remaining 48 frequency data points obtained for the amino acid of type 'X' from Swiss Prot and random sequences were analysed. We graphically represented the frequency values obtained from Swiss Prot database on X-axis and frequency values obtained from random database on Y-axis. The frequency values were calculated as described in the section 2.2.1. If the values compared at each position are similar, then the distribution of amino acid of type X is significantly correlated at that position. If the frequency values plotted for comparison follow a straight line passing through origin (with coefficient of correlation equal to 1), then we expect that the distributions are strongly correlated. If the coefficient of correlation is close to zero then we expect that the distributions compared show different pattern of behavior with no correlation.

In Fig 10, the positional distributions of the Alanine (one of the representative amino acids chosen for illustration) were considered for correlation analysis. In Fig 10, X-axis shows the positional frequency values (converted in to percent) of Alanine from Swiss Prot database upto 50th position. The Y-axis shows the positional frequency values of Alanine taken from random sequences upto 50th position. The frequency values obtained from these two databases were now correlated to check whether there are any similarities in the distribution.

The coefficient of correlation of Alanine (Fig 10) obtained for the compared distributions is 0.017. This shows that the frequency distributions of Alanine of Swiss Prot are different from random sequences.

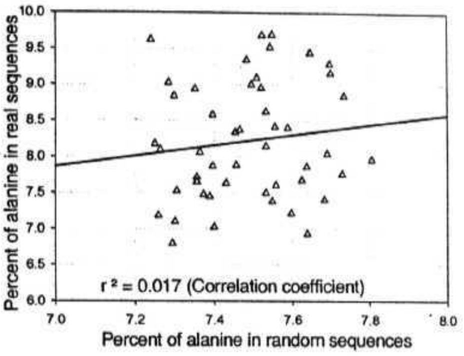


Figure 10: This figure gives correlations of the Alanine distribution in both real and random protein sequences. There is no significant correlation in between the Alanine distribution in the real and random sequences. X-axis shows the distribution of real sequence and Y-axis the distribution of random sequence database.
Comment: no correlation between the two variables.

Correlation studies of Cysteine and Tryptophan were illustrated in a similar manner. Figure 11 shows the correlation studies of Cysteine residue distribution in real and random sequences. The distribution values of Cysteine obtained from natural sequences up to 50th position, represented at X-axis are compared with Y-axis values of Cysteine obtained from random sequences. The comparison is done in a similar manner as illustrated above (Fig 10). The correlation coefficient of Cysteine for the compared distributions is 0.006. This shows that the positional distributions of Cysteine compared between the databases show no significant correlation.

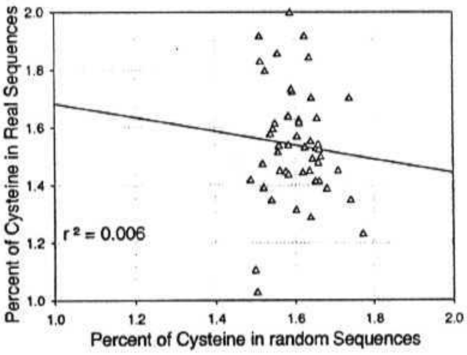


Figure 11: This figure gives correlations of the Cysteine distribution in both real and random protein sequences. There is no significant correlation in between the Cysteine distribution in the real and random sequences. X-axis shows the distribution of real sequence and Y-axis the distribution of random sequence database.
Comments: no correlation is seen.

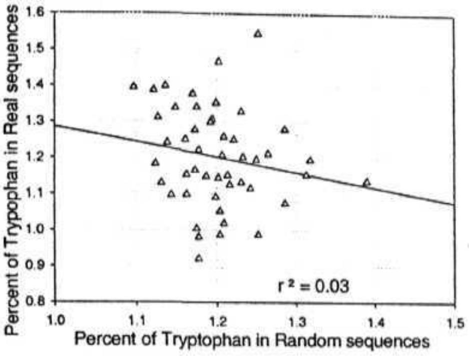


Figure 12: This figure gives regression correlations of the Tryptophan distribution in both real and random protein sequences. There is no significant correlation in between the Tryptophan distribution in the real and random sequences. X-axis shows the distribution of real sequence and Y-axis the distribution of random sequence database.
Comments: no correlation is seen

We have obtained a correlation coefficient of 0.03 (Fig 12) for the distributions of Tryptophan (upto 50th position) compared between the Swiss-Prot and random databases. Each individual distribution value of Tryptophan obtained from Swiss Prot and random database was compared to check for distribution similarities. The coefficient of correlation of compared distributions is 0.03, which shows that they follow different pattern of positional distributions.

It is evident from above correlation studies that the amino acids of the natural sequences follow different pattern of positional distribution in comparison with the positional distributions of the random sequences. This suggests that the natural sequences are different from the random sequences with respect to positional distribution.

2.3.5 AMINO ACIDS IN THE PROTEIN SEQUENCES FOLLOW A STEADY STATE DISTRIBUTION

The primary aim of our study is to show that the amino acids follow a steady state distribution. The method was clearly illustrated by taking the Alanine distribution from Swiss-Prot mini databases. The Table 3 below shows the frequencies of Alanine converted per million based on Eq 1 from the selected regions.

Table 3: Frequency values of Alanine converted per million in the selected positional ranges.

Mini- databases	10-20 position	50-60, position	100-110 position	150-160 position	200-210 position
1	4047.376	2926.669	3080.624	3193.072	3048.173
2	3740.219	3076.096	3091.945	3128.924	2865.539
3	3854.931	3033.079	3218.732	3313.067	3189.299
4	3541.737	2905.537	3024.777	2924.404	2897.990
5	3719.088	3109.302	2933.460	3279.106	3033.834
6	3179.488	3185.525	2737.997	2660.265	2632.341
7	3289.672	2762.147	2778.750	2890.444	2921.385
8	3047.418	2810.447	2848.181	2784.788	2747.808

Using the frequency counts of Alanine at various positions, we compared the distributions among the selected positions.

ANOVA test; This statistical test is implemented to check any differences among the distributions of amino acids at various selected positions.

Note: The samples compared should be similar in size.

Example: We have calculated frequencies of Alanine at 50 to 60 and 100 to 110 positions from 8 mini databases and checked for any differences among the distributions by using ANOVA test.

a = number of samples (2) i.e. 50-60, 100-110.

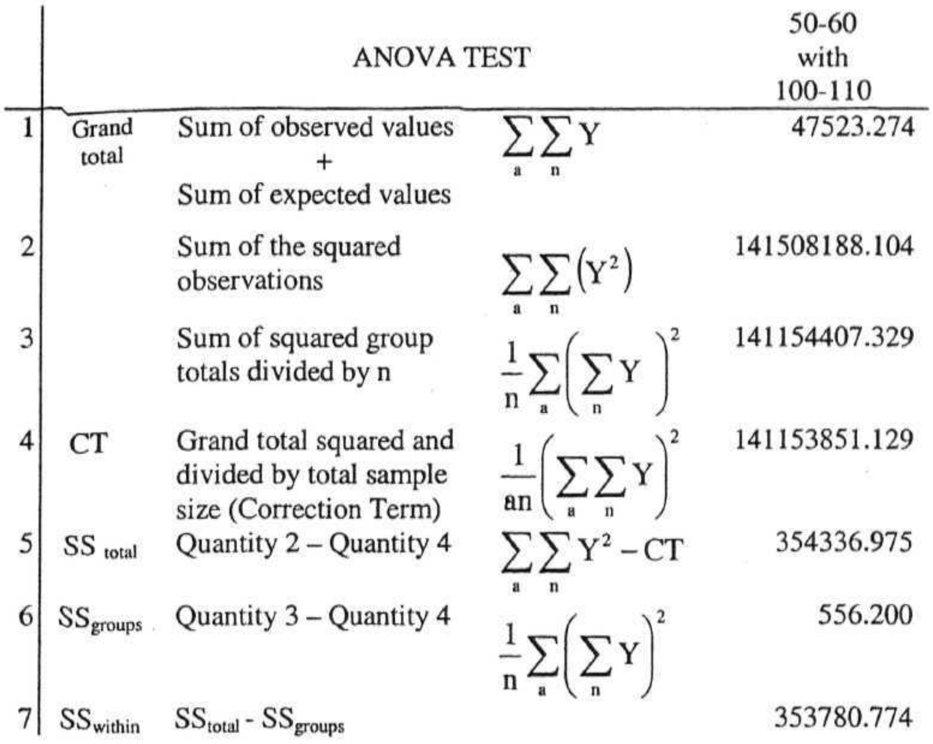
n = sample size (8) i.e. mini databases

Critical value: This value is obtained through statistical table by comparing the total number of samples (2) with the degrees of freedom (14)

Degrees of freedom = a (n-1) = 2 (8-1) = 14.

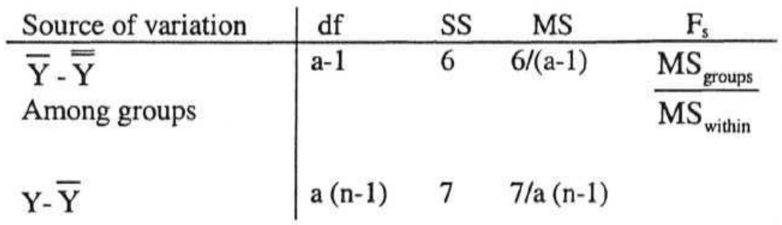
The obtained critical value is 4.07. If the F value obtained by ANOVA test is more than the critical value then the samples compared follow non-stationary distribution and vice versa.

Table 4: Comparison of Alanine distribution of the selected positional ranges (50-60 with 100-110) using ANOVA (Sokal *et al*).



	ANOVA TEST			50-60 with 100-110
1	Grand total	Sum of observed values + Sum of expected values	$\sum_a \sum_n Y$	47523.274
2		Sum of the squared observations	$\sum_a \sum_n \left(Y^2\right)$	141508188.104
3		Sum of squared group totals divided by n	$\frac{1}{n}\sum_a \left(\sum_n Y\right)^2$	141154407.329
4	CT	Grand total squared and divided by total sample size (Correction Term)	$\frac{1}{an}\left(\sum_a \sum_n Y\right)^2$	141153851.129
5	SS_{total}	Quantity 2 – Quantity 4	$\sum_a \sum_n Y^2 - CT$	354336.975
6	SS_{groups}	Quantity 3 – Quantity 4	$\frac{1}{n}\sum_a \left(\sum_n Y\right)^2$	556.200
7	SS_{within}	SS_{total} - SS_{groups}		353780.774

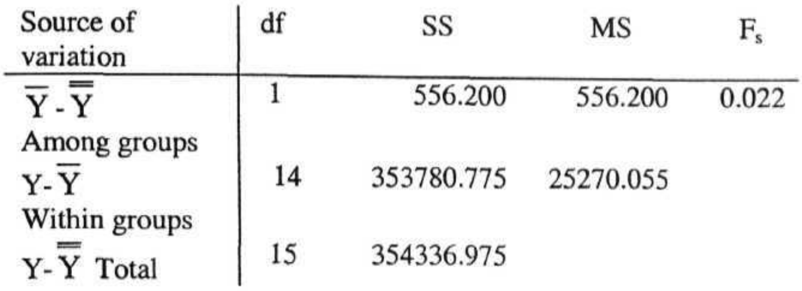
ANOVA table is constructed as follows:



Source of variation	df	SS	MS	F_s
$\overline{Y} - \overline{\overline{Y}}$ Among groups	a-1	6	6/(a-1)	$\frac{MS_{groups}}{MS_{within}}$
$Y - \overline{Y}$	a (n-1)	7	7/a (n-1)	

Within groups $Y-\bar{\bar{Y}}$ Total	an-1	5

ANOVA table for 50-60 with 100-110 positional range

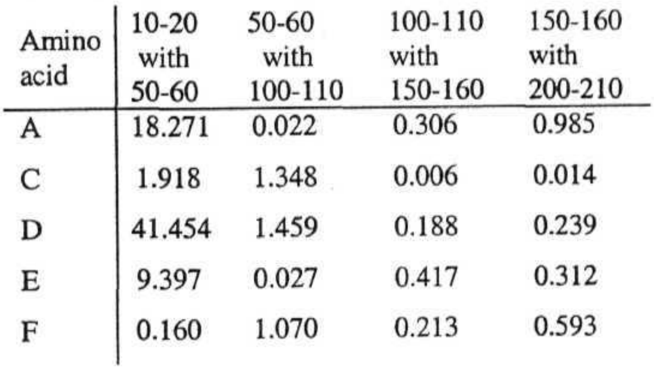


Source of variation	df	SS	MS	F_s
$\bar{Y}-\bar{\bar{Y}}$ Among groups	1	556.200	556.200	0.022
$Y-\bar{Y}$ Within groups	14	353780.775	25270.055	
$Y-\bar{\bar{Y}}$ Total	15	354336.975		

Conclusions: The obtained F value 0.022 was tested for significance by checking with the statistical F table. The value obtained is below the critical value (4.07) proving that the samples compared follow a steady state distribution. This shows that the distributions of Alanine at 50-60 and 100-110 positional regions show a steady state behavior.

To understand the steady state behavior of amino acids in the database, we followed the above ANOVA statistical method and calculated the F values (Table 5) for all common amino acids comparing the distributions of selected positional ranges of the 8 mini databases of Swiss Prot.

Table 5: The table shows the F values of all the common amino acids after comparing the selected regions by using ANOVA method.



Amino acid	10-20 with 50-60	50-60 with 100-110	100-110 with 150-160	150-160 with 200-210
A	18.271	0.022	0.306	0.985
C	1.918	1.348	0.006	0.014
D	41.454	1.459	0.188	0.239
E	9.397	0.027	0.417	0.312
F	0.160	1.070	0.213	0.593

G	0.066	0.065	0.112	0.051
H	16.072	0.682	0.439	1.334
I	0.675	0.022	2.372	0.023
K	6.020	0.386	0.149	0.328
L	54.506	0.524	0.001	0.002
M	5.676	4.314	0.155	0.013
N	5.699	1.526	1.846	0.739
P	1.805	0.024	0.049	0.053
Q	1.011	3.684	0.015	0.005
R	0.363	0.989	0.318	0.081
S	15.943	2.927	0.874	0.982
T	1.537	4.178	0.443	0.300
V	0.422	1.300	0.026	0.379
W	0.894	2.911	0.948	2.999
Y	18.215	3.902	0.392	0.005

In the Table 5 we have compared the differences among the distributions of two positional ranges by implementing ANOVA test. If the obtained F value is below the critical value (4.07) then the distributions follow a steady state behavior and vice versa. It is evident from the Table 5 that the amino acids (Alanine, Aspartate, Glutamate, Histidine, Lysine, Leucine, Methionine, Asparagine, Serine and Tyrosine) at the positional range 10-20 compared with 50-60 show F value above the critical value. This shows that the distributions compared within these amino acids at those positional ranges (10-20 with 50-60) follow non-stationary distribution. This non-stationary behavior at beginning regions is clearly evident from the figure 4 and 5 shown earlier in this section. However, the F values obtained while comparing other positional ranges show

below critical value suggesting a steady state distribution at those regions. Overall, the result shows that the amino acids follow a steady state distribution in the database except at few beginning regions.

ANOVA method was implemented on non-redundant sets to check the positional distribution behavior of amino acids in those datasets.

Non-redundant set with less than 40% sequence identity

We compared the distributions (non-redundant sets) of all the amino acids of the selected positions by following the illustration shown in Table 4. ANOVA test was implemented on non-redundant sets and the F values calculated are shown in the Table 6 below.

For this set the Critical value is calculated as shown below

a = number of samples (2) (Two selected regions are compared)

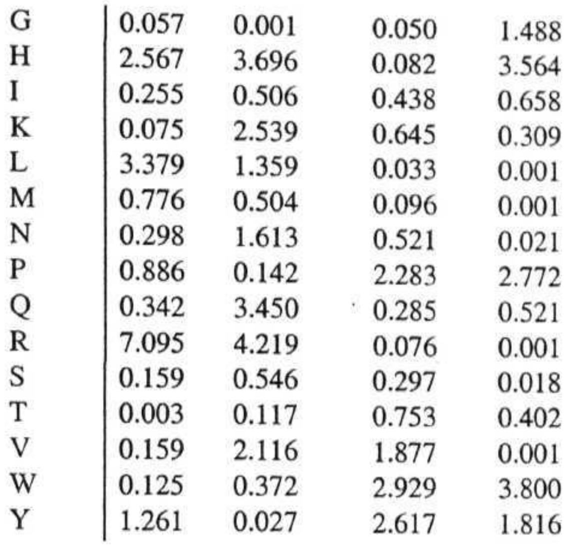
n = sample size (4) (Four mini databases)

Critical value: This value is obtained through statistical table by comparing the total number of samples (2) with the degrees of freedom (6). The critical value is 11.1 for this set.

Degrees of freedom = a (n-1) = 2 (4-1) = 6.

Table 6: This table shows **the F values calculated for ASTRAL non-redundant sets with less than 40% sequence identity.**

Amino acids	10-20 with 50-60	50-60 with 100-110	100-110 with 150-160	150-160 with 200-210
A	3.307	2.565	0.039	0.635
C	9.000	0.062	0.863	0.019
D	0.456	0.043	0.598	0.202
E	1.264	0.045	1.459	0.112
F	0.704	1.676	0.628	2.559



G	0.057	0.001	0.050	1.488
H	2.567	3.696	0.082	3.564
I	0.255	0.506	0.438	0.658
K	0.075	2.539	0.645	0.309
L	3.379	1.359	0.033	0.001
M	0.776	0.504	0.096	0.001
N	0.298	1.613	0.521	0.021
P	0.886	0.142	2.283	2.772
Q	0.342	3.450	0.285	0.521
R	7.095	4.219	0.076	0.001
S	0.159	0.546	0.297	0.018
T	0.003	0.117	0.753	0.402
V	0.159	2.116	1.877	0.001
W	0.125	0.372	2.929	3.800
Y	1.261	0.027	2.617	1.816

Non-redundant set with less than 95% sequence identity

The obtained F values of amino acids of non-redundant sets with less than 95% sequence identity are shown in the Table 7 below.

Critical values is calculated as follows

a = number of samples (2)

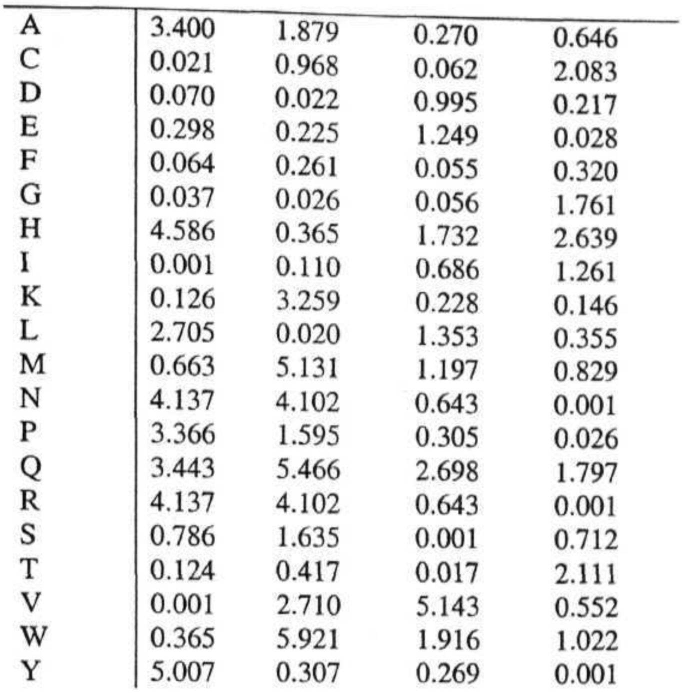
n = sample size (5)

Critical value: This value is obtained through statistical table by comparing the total number of samples (2) with the degrees of freedom (8). The critical value is 7.50 for this set.

Degrees of freedom = a (n-1) = 2 (5-1) = 8.

Table 7: This table shows the F values calculated for ASTRAL non-redundant sets with less than 95% sequence identity.

Amino acid	10-20 with 50-60	50-60 with 100-110	100-110 with 150-160	150-160 with 200-210



A	3.400	1.879	0.270	0.646
C	0.021	0.968	0.062	2.083
D	0.070	0.022	0.995	0.217
E	0.298	0.225	1.249	0.028
F	0.064	0.261	0.055	0.320
G	0.037	0.026	0.056	1.761
H	4.586	0.365	1.732	2.639
I	0.001	0.110	0.686	1.261
K	0.126	3.259	0.228	0.146
L	2.705	0.020	1.353	0.355
M	0.663	5.131	1.197	0.829
N	4.137	4.102	0.643	0.001
P	3.366	1.595	0.305	0.026
Q	3.443	5.466	2.698	1.797
R	4.137	4.102	0.643	0.001
S	0.786	1.635	0.001	0.712
T	0.124	0.417	0.017	2.111
V	0.001	2.710	5.143	0.552
W	0.365	5.921	1.916	1.022
Y	5.007	0.307	0.269	0.001

The F values of non-redundant data sets of less than 40% and 95 % sequence identity are shown in the Table 6 and 7. The obtained F values of both the sets in the compared positions are found below the critical values (11.1 and 7.5 respectively). This shows that the positional distribution of amino acid of both non-redundant sets follow steady state behavior all over from beginning regions. This is quite exception with Swiss-Prot database wherein the distribution attains steady state after few beginning positions.

To further substantiate that the selected 50 to 60 region in Swiss-Prot database follow a stationary distribution, we have used the following equation Eq 2.

$$\mu_i P_{ij} = \mu_j$$

where μ_i is the probability value of the *i*th amino acids at the selected region (50 – 60). P_{ij} is the transition matrix of *i*th and *j*th amino acids at the selected region.

We expect that the product of μ_i and P_{ij} should be equal to μ_j suggesting that the distribution follows a stationary behavior.

Table 8: This table shows the P_{ij} transition values for the selected 50 to 60 region.

	A	C	D	E	F	G	H	I	K	L
A	0.107	0.012	0.046	0.062	0.041	0.074	0.026	0.053	0.054	0.093
C	0.069	0.021	0.052	0.069	0.041	0.087	0.025	0.051	0.051	0.092
D	0.080	0.012	0.059	0.073	0.041	0.078	0.018	0.065	0.058	0.097
E	0.082	0.013	0.059	0.092	0.030	0.062	0.020	0.061	0.068	0.087
F	0.064	0.018	0.056	0.056	0.048	0.070	0.024	0.056	0.053	0.095
G	0.077	0.017	0.054	0.059	0.040	0.081	0.031	0.060	0.059	0.086
H	0.070	0.021	0.039	0.061	0.040	0.080	0.030	0.054	0.052	0.098
I	0.069	0.022	0.056	0.062	0.040	0.064	0.021	0.060	0.061	0.091
K	0.069	0.013	0.061	0.073	0.031	0.059	0.021	0.064	0.075	0.091
L	0.081	0.014	0.052	0.061	0.040	0.073	0.023	0.051	0.067	0.097
M	0.071	0.016	0.059	0.058	0.036	0.073	0.040	0.054	0.050	0.103
N	0.065	0.015	0.049	0.057	0.040	0.072	0.021	0.065	0.059	0.089
P	0.074	0.012	0.056	0.071	0.035	0.076	0.023	0.048	0.048	0.083
Q	0.074	0.014	0.043	0.064	0.037	0.062	0.023	0.053	0.055	0.094
R	0.073	0.016	0.053	0.071	0.048	0.074	0.023	0.054	0.053	0.097
S	0.067	0.013	0.056	0.050	0.043	0.073	0.025	0.051	0.058	0.090
T	0.081	0.015	0.047	0.054	0.041	0.079	0.028	0.058	0.051	0.098
V	0.082	0.017	0.055	0.062	0.040	0.061	0.018	0.059	0.056	0.100
W	0.060	0.024	0.051	0.052	0.037	0.065	0.025	0.052	0.064	0.089
Y	0.063	0.017	0.059	0.057	0.042	0.078	0.026	0.053	0.056	0.084

	M	N	P	0	R	S	T	V	W	Y
A	0.021	0.035	0.046	0.042	0.048	0.071	0.057	0.071	0.012	0.027
C	0.016	0.047	0.046	0.048	0.051	0.071	0.053	0.063	0.012	0.031
D	0.017	0.036	0.052	0.032	0.044	0.074	0.049	0.066	0.013	0.033
E	0.021	0.051	0.042	0.041	0.056	0.056	0.053	0.064	0.01	0.028
F	0.019	0.044	0.050	0.038	0.043	0.091	0.057	0.075	0.01	0.030

G	0.020	0.043	0.045	0.037	0.051	0.075	0.052	0.065	0.012	0.033
H	0.019	0.037	0.055	0.045	0.054	0.063	0.049	0.076	0.014	0.040
I	0.016	0.053	0.052	0.042	0.047	0.070	0.066	0.065	0.009	0.033
K	0.019	0.050	0.042	0.040	0.058	0.063	0.055	0.065	0.01	0.040
L	0.020	0.042	0.048	0.043	0.052	0.073	0.063	0.061	0.01	0.027
M	0.022	0.048	0.046	0.038	0.050	0.066	0.054	0.070	0.013	0.027
N	0.016	0.051	0.059	0.040	0.048	0.083	0.055	0.069	0.013	0.033
P	0.016	0.040	0.070	0.041	0.061	0.076	0.060	0.066	0.013	0.027
Q	0.021	0.040	0.057	0.070	0.060	0.068	0.054	0.065	0.011	0.030
R	0.026	0.041	0.044	0.040	0.063	0.065	0.050	0.058	0.012	0.035
S	0.020	0.044	0.051	0.038	0.045	0.097	0.060	0.070	0.013	0.032
T	0.019	0.044	0.054	0.038	0.042	0.071	0.067	0.069	0.017	0.025
V	0.022	0.039	0.053	0.035	0.050	0.070	0.065	0.070	0.015	0.026
W	0.026	0.057	0.039	0.042	0.064	0.060	0.063	0.074	0.024	0.031
Y	0.019	0.053	0.048	0.047	0.048	0.068	0.060	0.075	0.010	0.034

Table 9: This table shows the individual probability values of *i*th amino acids at 50 to 60 region (μ_i). The μ_j is obtained by multiplying the P_{ij} transition matrix with μ_i. The values of μ_i and μ_j are approximately equal proving that the selected region 50 to 60 follows a stationary distribution.

Amino acids	μ_i	μ_j
A	0.076	0.077
C	0.015	0.015
D	0.055	0.054
E	0.063	0.064
F	0.041	0.040
G	0.072	0.072
H	0.023	0.024
I	0.058	0.057
K	0.058	0.059
L	0.093	0.093
M	0.020	0.020
N	0.044	0.044
P	0.049	0.050
Q	0.041	0.041
R	0.051	0.051
S	0.072	0.073

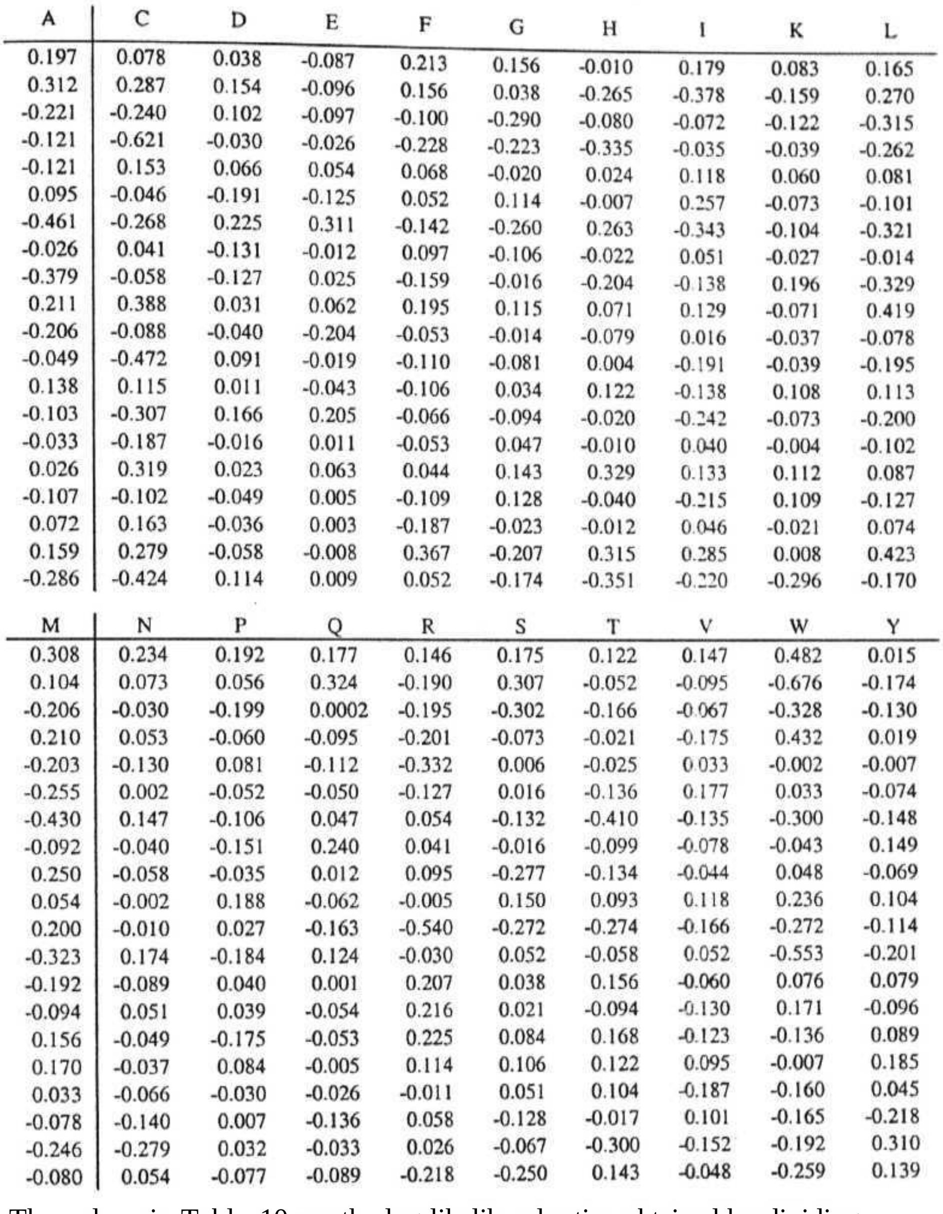
T	0.058	0.058
V	0.068	0.067
W	0.012	0.012
Y	0.032	0.031

The primary use of Eq 2 is to show that the selected 50 to 60 positional regions follow a steady state distribution. We estimated μ_j by multiplying the P_{ij} transition matrix (Table 8) with μ_i individual probability values (Table 9) of /th amino acid at the selected region. The observed values of μ_j are approximately similar to μ_i individual probability values of the selected region. This shows that the selected positional range of distribution 50 to 60 show steady state behavior. This result supports the Figure 4 and 5 in which the amino acid distribution follows steady state behavior after few beginning positions.

2.3.6 COMPARISON OF DISTRIBUTIONS BETWEEN THE BEGINNING (10 TO 20) AND MIDDLE REGIONS (50 TO 60) OF PROTEIN SEQUENCES

We have used log-odds method to differentiate the distributions of 10 to 20 with 50 to 60 regions. We calculated the p_{ij} transitions (20 X 20) matrix of *i*th and *j*th amino acids for both the selected regions. By using log-odds method of Eq 3, we could show that the compared 10-20 with 50-60 regions shows differences in their positional distributions.

TABLE 10: Log odds ratio values of compared regions (10-20 and 50-60 regions). The negative and positive values show no hint of similarity between the regions compared.



A	C	D	E	F	G	H	I	K	L
0.197	0.078	0.038	-0.087	0.213	0.156	-0.010	0.179	0.083	0.165
0.312	0.287	0.154	-0.096	0.156	0.038	-0.265	-0.378	-0.159	0.270
-0.221	-0.240	0.102	-0.097	-0.100	-0.290	-0.080	-0.072	-0.122	-0.315
-0.121	-0.621	-0.030	-0.026	-0.228	-0.223	-0.335	-0.035	-0.039	-0.262
-0.121	0.153	0.066	0.054	0.068	-0.020	0.024	0.118	0.060	0.081
0.095	-0.046	-0.191	-0.125	0.052	0.114	-0.007	0.257	-0.073	-0.101
-0.461	-0.268	0.225	0.311	-0.142	-0.260	0.263	-0.343	-0.104	-0.321
-0.026	0.041	-0.131	-0.012	0.097	-0.106	-0.022	0.051	-0.027	-0.014
-0.379	-0.058	-0.127	0.025	-0.159	-0.016	-0.204	-0.138	0.196	-0.329
0.211	0.388	0.031	0.062	0.195	0.115	0.071	0.129	-0.071	0.419
-0.206	-0.088	-0.040	-0.204	-0.053	-0.014	-0.079	0.016	-0.037	-0.078
-0.049	-0.472	0.091	-0.019	-0.110	-0.081	0.004	-0.191	-0.039	-0.195
0.138	0.115	0.011	-0.043	-0.106	0.034	0.122	-0.138	0.108	0.113
-0.103	-0.307	0.166	0.205	-0.066	-0.094	-0.020	-0.242	-0.073	-0.200
-0.033	-0.187	-0.016	0.011	-0.053	0.047	-0.010	0.040	-0.004	-0.102
0.026	0.319	0.023	0.063	0.044	0.143	0.329	0.133	0.112	0.087
-0.107	-0.102	-0.049	0.005	-0.109	0.128	-0.040	-0.215	0.109	-0.127
0.072	0.163	-0.036	0.003	-0.187	-0.023	-0.012	0.046	-0.021	0.074
0.159	0.279	-0.058	-0.008	0.367	-0.207	0.315	0.285	0.008	0.423
-0.286	-0.424	0.114	0.009	0.052	-0.174	-0.351	-0.220	-0.296	-0.170

M	N	P	Q	R	S	T	V	W	Y
0.308	0.234	0.192	0.177	0.146	0.175	0.122	0.147	0.482	0.015
0.104	0.073	0.056	0.324	-0.190	0.307	-0.052	-0.095	-0.676	-0.174
-0.206	-0.030	-0.199	0.0002	-0.195	-0.302	-0.166	-0.067	-0.328	-0.130
0.210	0.053	-0.060	-0.095	-0.201	-0.073	-0.021	-0.175	0.432	0.019
-0.203	-0.130	0.081	-0.112	-0.332	0.006	-0.025	0.033	-0.002	-0.007
-0.255	0.002	-0.052	-0.050	-0.127	0.016	-0.136	0.177	0.033	-0.074
-0.430	0.147	-0.106	0.047	0.054	-0.132	-0.410	-0.135	-0.300	-0.148
-0.092	-0.040	-0.151	0.240	0.041	-0.016	-0.099	-0.078	-0.043	0.149
0.250	-0.058	-0.035	0.012	0.095	-0.277	-0.134	-0.044	0.048	-0.069
0.054	-0.002	0.188	-0.062	-0.005	0.150	0.093	0.118	0.236	0.104
0.200	-0.010	0.027	-0.163	-0.540	-0.272	-0.274	-0.166	-0.272	-0.114
-0.323	0.174	-0.184	0.124	-0.030	0.052	-0.058	0.052	-0.553	-0.201
-0.192	-0.089	0.040	0.001	0.207	0.038	0.156	-0.060	0.076	0.079
-0.094	0.051	0.039	-0.054	0.216	0.021	-0.094	-0.130	0.171	-0.096
0.156	-0.049	-0.175	-0.053	0.225	0.084	0.168	-0.123	-0.136	0.089
0.170	-0.037	0.084	-0.005	0.114	0.106	0.122	0.095	-0.007	0.185
0.033	-0.066	-0.030	-0.026	-0.011	0.051	0.104	-0.187	-0.160	0.045
-0.078	-0.140	0.007	-0.136	0.058	-0.128	-0.017	0.101	-0.165	-0.218
-0.246	-0.279	0.032	-0.033	0.026	-0.067	-0.300	-0.152	-0.192	0.310
-0.080	0.054	-0.077	-0.089	-0.218	-0.250	0.143	-0.048	-0.259	0.139

The values in Table 10 are the log likelihood ratios obtained by dividing the distribution of 50-60 with 10-20 transitions. For convenience the 20 X 20 matrix was divided in to 10 X 20 matrix. The Table 10 shows both positive and negative values, the negative values are obtained when the 10-20 region transition values are more than the values of 50-60 region and vice versa. This shows that the distributions are different among the selected regions.

CONCLUSION

With respect to positional frequencies, there is a significant difference in the distribution behavior of amino acids compared between natural and random sequences. Correlation studies clearly indicate that the Swiss Prot protein sequences follow a different pattern of distribution with respect to random or simulated sequences. The positional distribution studies from N-terminal region in Swiss Prot database showed a high or low frequency of amino acids at the beginning regions and attained steady state distribution after 30 to 40 positions. This behavior was not observed in the studies made from the C-terminal end of the protein sequences, wherein the distribution is steady all over the positions considered. As expected the distribution behavior of amino acids in the random database showed uniform or random distribution.

We supported the view of steady state distribution of amino acids in Swiss Prot database by using ANOVA (Analysis of Variance) method. The test clearly pointed out the difference of distribution of amino acids at the beginning positions compared to rest of the positions selected. Moreover, the non-redundant sets (40% and 95% similarity) analysed for distribution showed a steady behavior of amino acids all over the positions considered. By using other statistical methods (log odds method), we further substantiated the steady state behavior of amino acids in the Swiss Prot database.

Based on the analysis of positional distributions of amino acids in the primary sequences, we conclude that Swiss Prot (natural) sequences follow a steady state distribution and are different from random sequences.

References:

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

Brenner, S.E. Koehl, P. and Levitt, M. (2000) The ASTRAL compendium for protein structure and sequence analysis, Nucleic Acids res. 28, 254-256.

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

Chou, P. Y. and Fasman, G. D. 1974b Prediction of protein conformation, Biochemistry. 13 222-244

Cover, T.M. and Thomas, J.A (1991) Elements of Information Theory. Wiley.

Doolittle, R. F. (1981). Similar amino-acid sequences - chance or common ancestry? Science. 214, 149-159.

Durbin, R. Eddy S. Krogh A. and Mitchison, G. (1998) Biological Sequences analyses, Cambridge University press. pp 51.

Feller, W. An Introduction to Probability Theory and its Applications (John Wiley and Sons, Inc., New York, 1950).

Lehninger, A. L. Nelson, D. L. (1993) Principles of Biochemistry, Second edition, Worth Publishers, New York.

Meeta, R. Mitra, C. K. Cserzo, M. and Simon I. (1995) Proteins as special subsets of polypeptides, J. Bioscience. 20 579-590.

Mitra, C. K. and Sen, A. (2001) Towards A Dynamical Systems Approach to Protein sequence structure, Calcutta Statistical Association Bulletin. 51, 203-204.

Sokal R. and Rohlf James, F. Introduction to Biostatistics, pp 164.

FRACTAL STUDIES
ON
PROTEINS

3.1 INTRODUCTION

Mandelbrot (Mandelbrot, 1983) introduced the term 'fractal' (from the latin *fractus*, meaning 'broken') to characterize spatial or temporal phenomena that are continuous but not differentiable.

According to Mandelbrot a Fractal object is defined as a rough or fragmented geometric shape that can be subdivided in parts, each of which is (at least approximately) a reduced/size copy of the whole.

In other words, fractals are non-compact sets. It is useful to understand the characteristics of a fractal with simple examples. The length of arteries in human body is very large and depends on how accurately it is measured. Similarly, the length of the branches of a tree shows similar behavior pattern. The coastlines of various countries are very large, and are similarly dependent on the degree of precision of measurement. Such systems are often classified by considering a parameter called dimension. Therefore, fractals are characterized based on their dimensions.

3.1.1 WHAT ARE FRACTALS

Mandelbrot (Mandelbrot, 1983) also called "Father of fractal geometry" defines fractals as sets for which the Hausdorff-Besicovitch dimension strictly exceeds the topological dimension. For a discussion on the Hausdroff-Besicovitch dimension, please refer to Hurewicz and Wallman (Hurewicz, *et al)* and Rogers (Rogers, *et al).* Topological dimension (designated as D_T) is rather simple in concept and corresponds to Euclidean dimension of space into which the fractal set is embedded.

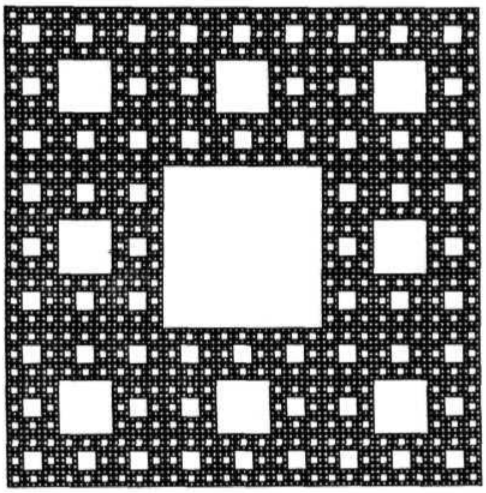
According to Barnsley (Barnsley, 1988) a fractal may be described as a "geometrically complicated subset of a geometrically simple set of points in space". As a set of points constitute a line, surface or space or any other object, fractals too are set of points arranged according to some given rule or equation. Hence, they are also geometrical entities and their shape depends upon the arrangement of points as characterized by the given rule or equation. Naturally, different rules or equations give rise to different fractals.

3.1.2 HOW FRACTALS ARE CREATED

Most geometrical fractals can be obtained by repeated application of some rule to simple set, e.g. a line or a closed figure. After an indefinite number of applications of the same rule the set or figure is a fractal. For e.g., the Cantor dust (Mandelbrot, 1983) or the Mandelbrot set (Peitgen *et al,* 1986) are created this way. In nature, fractals occur in the form of trees, rivers, braches of arteries, anatomical structure of organs and organelles of animal and plants and these shapes are produced naturally as a result of what is known as non-linear dynamics. Systems showing non-linear dynamics are those that respond disproportionately to stimuli. For instance, the nonlinear chaotic processes shape environment like seashores, atmospheres and geologic faults the results are fractals like coastlines, clouds and rocks respectively.

Unlike more familiar Euclidean constructs, every attempt to split a fractal into smaller pieces results in the resolution of more structure. Fractal objects and processes are therefore said to display 'self-invariant' (self-similar or self-affine) properties (Hastings and Sugihara, 1993). Self-similar objects are isotropic upon rescaling, whereas rescaling of self-affine objects is direction-dependent (anisotropic). Thus the trace of particulate Brownian motion in two-dimensional space is self-similar, whereas a plot of the x-coordinate of the particle as a function of time is self-affine (Brown, 1995).

Figure 1. A Sierpinski Square Fractal



The Sierpinski square fractal is illustrated in Figure 1. This pattern is constructed starting with a solid (filled) square, divides it in 9 smaller congruent squares, and removes the interior of the center square (i.e. the boundary is not removed). This sequence of steps is then applied again to each one of the 8 remaining Solid squares which will be divided in 9 smaller congruent squares, and so on. The construction can be repeated infinitely.

3.1.3 PROPERTIES OF FRACTALS

Fractal properties include scale independence, self-similarity, complexity, and infinite length or detail. Fractal structures do not have a single length scale, while fractal processes (time series) cannot be characterized by a single time scale (West and Goldberger 1987). Nonetheless, the necessary and sufficient conditions for an object (or process) to possess fractal properties have not been formally defined. Indeed, fractal geometry has been described as "a collection of examples, linked by a common point of view, not an organized theory" (Lorimer et al. 1994).

Fractal theory offers methods for describing the inherent irregularity of natural objects. In fractal analysis, the Euclidean concept of 'length' is viewed as a process. This process is characterized by a constant parameter D known as the fractal (or fractional) dimension. The fractal dimension can be viewed as a relative measure of complexity, or as an index of the scale-dependency of a pattern. Excellent summaries of basic concepts of fractal geometry can be found in Mandelbrot (1983), Frontier (1987), Schroeder (1991), Turcotte (1992), Hastings and Sugihara (1993), Lam and De Cola (1993).

The fractal dimension is a summary statistic measuring overall 'complexity'. Like many summary statistics (e.g. mean), it is obtained by 'averaging' variation in data structure (Normant and Tricot 1993). In doing so, information is necessarily lost. The estimated fractal dimension of a lakeshore, for example, tells us nothing about the actual size or overall shape of the lake, nor can we reproduce a map of the lake from D alone. However, the fractal dimension does tell us a great deal about the relative complexity of the lakeshore, and as such is an important descriptor when used in conjunction with other measures.

3.1.4 FRACTAL DIMENSIONS

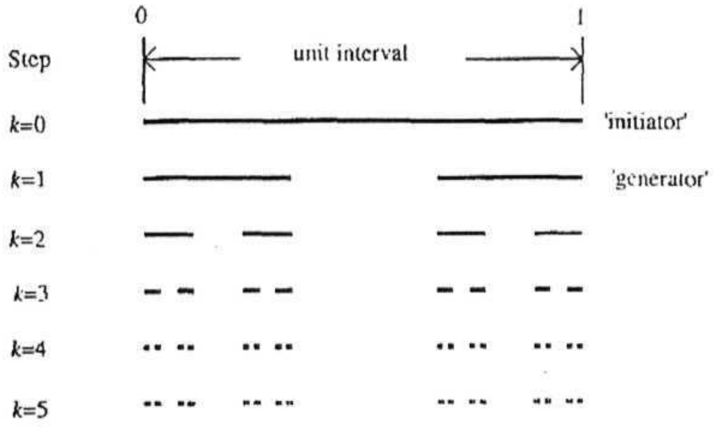
Fractal dimensions are numbers associated with fractals, which gives an idea about how densely a fractal occupies the metric space to which it belongs. Dimension in classical geometry always has an integral value.

For instance, a point, line, plane and space have dimensions of 0, 1, 2 and 3 respectively and are called their topological dimensions. Classical geometry cannot attach any meaning to a figure with a fractional value of dimension but fractal geometry can, and it calls such figures fractals.

Fractal dimensions have fractional values. Those sets in which the topological dimension is equal to 0, the value of fractal dimension is greater than 0 but less than 1, i.e., D lies between 0 and 1 e.g., Cantor dust, the original Cantor set (Mandelbrot, 1983).

The older but simplest example is the Cantor Set. The Cantor set is a famous construction in mathematics, which is much older than the relatively recent interest in fractals and fractal geometry. The Cantor set is a very simple set to describe (although it is not quite so easy to study), and it is perhaps the simplest example of a fractal. See the figure 2 below

Figure 2 illustrates the geometric construction of cantor set.



The Cantor set can be constructed by starting with a line segment, and then removing the middle third of the segment. This leaves two sub-segments. The geometric construction is iterated by removing the middle thirds of these, and so on. In the case of the original cantor set D = (log 2)/(log 3) = 0.6309 > 0 but topological dimension of the cantor dust is 0. Therefore, we see that D has a fractional value and is greater than topological dimension.

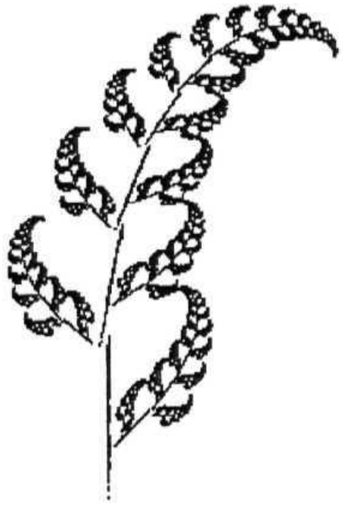
Fractal sets in which topological dimension is 1 and fractal dimension is greater than 1 (usually between 1 and 2) and hence, a fractional, are fractal curves and fractal lines. The fractal dimension is always greater than or smaller than the topological dimension, but in most cases it is greater. However, a fractal may have an integral value of D so long as D > D_T i.e. so called "Szpilrajn inequality" is obeyed. For instance, in case of path followed by a particle in Brownian motion, D_T = 1 but fractal dimension is 2 i.e., D is an integer in this case. Therefore fractal dimension need not be fractional value all the time.

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 England (its values is 1.2). Fractal dimensions can be attached to clouds, trees, coastlines, feathers, network of neurons in the body, dust in the air at any instant, the clothes we are wearing, the distribution of frequencies of light reflected by a flower, the color emitted by the sun and wrinkled surface of the sea during a storm. These numbers allow us to compare sets in the real world with the laboratory fractals such as the attractors of the IFS (Iterated function systems).

3.1.5 FRACTALS IN THE BIOLOGICAL SCIENCES

Nature, and biology in particular, abounds in fractals. Fractals are seen in the shape of clouds, rivers, mountains and in the anatomical structures of animals and plants. For instance, the branching pattern of trees, blood vessels, intestinal villi, neuronal networks, corals; the arrangement of the elements of the vascular bundles; the convolutions in the brain of mammals etc. Various intricate shapes of micro-organisms, leaf (Fig 3), which cannot be easily described in classical geometry, can be generated by iterated function systems (IFS) on a computer.

Figure 3 illustrates the generated leaf from fractal IFS approach (Mingtian, Ni and Xin Li, 2001).



3.1.6 IMPORTANCE OF FRACTAL STRUCTURES 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 available for absorption (as in the intestine), distribution and collection (by the blood vessels, bile duct 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.

Fractal like structures plays a vital role in the healthy mechanical and electrical dynamics of the heart. First, for example, a fractal-like network of coronary arteries and veins conveys blood to and from the heart muscles. Hans Van Beek and James.B.Bassingthwaighte of the university of Washington recently used fractal geometry to explain anomalies in the blood flow patterns to the healthy heart (Goldberger *et al,* 1990). Interruptions of this arterial flow may cause a myocardial infraction (heart attack).

A fractal like canopy of connective tissue fibers within the heart, the chordae tendinae, tethers the mitral and tricuspid valves to the underlying muscles. If the tissues break, there can be severe regurgitation of blood from ventricles to atria, followed by congestive heart failure.

Fractal architecture is also evident in the branching pattern of certain cardiac muscles as well as the His-Purkinje system, which conducts electrical signals from the atria to the cardiac muscles of the ventricles.

The one way to characterize fractal is to calculate their dimensions. The fractal dimensions of some organelles in the human body have been calculated (Mandelbrot, 1983)

Membrane dimensions over a wide range = 2.17.

The outer mitochondrial membrane has a fractal dimension of 2.09. The inner mitochondrial membrane has a dimension of 2.53.

Endoplasmic reticulum has a fractal dimension of 1.72.

Fractal structures in the living systems arise from the low dynamics of embryonic development and evolution. These processes, like others that produce fractal structures, exhibit deterministic chaos.

3.1.7 FRACTAL DYNAMICS IN BIOLOGY

Biologists have traditionally modelled nature using Euclidean representations (1, 2, 3 topological dimensions) of natural objects or series. Examples include the representation of heart rates as sine waves, conifer trees as cones, animal habitats as simple areas, and cell membranes as curves or simple surfaces. However, scientists have come to recognize that many natural constructs are better characterized using fractal geometry. Biological systems and processes are typically characterized by many levels of substructure, with the same general pattern being repeated in an ever-decreasing cascade. The relevance of fractal theory to biological problems is dependent on objectives. Estimating stand board-feet in forest, a Euclidean representation of a tree trunk (as a cylinder or elongated cone) may be quite adequate. However, for an ecologist interested in modelling habitat availability on tree trunks (say, for small epiphytes or invertebrates), fractal geometry is more appropriate. Using a fractal approach, the complex surface of tree bark is readily quantified. A forester's diameter tape ignores the surface roughness of the bark, giving a crude estimate of the circumference of the trunk. For an insect 10 mm in length, the 'distance' that it must travel to circumnavigate the trunk is much greater than the measured diameter value. For an insect of length 1 mm, the distance travelled is greater still. This has consequences on the way that the tree trunk is perceived by organisms of different sizes. If the bark has a fractal dimension of D = 1.4, an insect an order of magnitude smaller than another perceives a length increase of $10^{D-1} = 10^{0.4} = 2.51$, or a habitat surface area increase of $2.51^2 = 6.31$. By contrast, for a smooth Euclidean surface, D = 1 and both insects perceive the same 'amount' of habitat. The higher the fractal dimension D, the greater the perceived rate of increase in length (or surface) with decreasing scale.

Relationships that depend on scale have profound implications in human physiology (West and Goldberger 1987), ecology (Loehle 1983; Wiens 1989), and many other sub-disciplines of biology. The importance of fractal scaling has been recognized at virtually every level of biological organization. Fractal geometry may prove to be a unifying theme in biology (Kenkel and Walker 1993), since it permits generalization of the fundamental concepts of dimension and length measurement. Most biological processes and structures are decidedly non-Euclidean, displaying discontinuities, jaggedness, and fragmentation. Classical measurement and scaling methods such as Euclidean geometry, calculus and the Fourier transform assume continuity and smoothness. However, it is important to recognize that while Euclidean geometry is not realized in nature, neither is strict mathematical fractal geometry. Specifically, there is a lower limit to self-similarity in most biological systems, and nature adds an element of randomness to its fractal structures. Nonetheless, fractal geometry is closer to nature than is Euclidean geometry (Deering and West 1992).

3.1.7.1 FRACTAL DYNAMICS IN PROTEIN SEQUENCES

The native structure of a globular protein is difficult to describe in classical geometry and no universal algorithm exists today that can reliably predict the folding pattern of a globular protein from its primary sequence. The native conformation of a protein depends on the primary sequence, i.e., the sequence of the amino acid residues -just as the shape of a fractal depends on the arrangement of points on it. Hence in fractal geometry it should not be difficult to describe the shape of a globular protein. The fractal description of protein structures and shapes haven't been studied thoroughly till date. However, several workers have reported dimensions of several iron containing proteins using Moussbauer spectroscopy (Stapleton *et al,* 1980).

The correlations between various amino acid residues are important parameters in the overall structure of the protein. These correlations are often described in terms of coefficients, but a more realistic description uses the concept of fractals. Geometrically, fractals are most useful in describing an irregular shape in a rational way. Protein sequences can be considered as a multi-fractal, with the different amino acid residues, showing different fractal dimensions (Meeta and Mitra, 1993).

The fractal dimensions are related to the scaling parameters and correlation lengths of the corresponding residues. As different residues show different fractal dimensions, we conclude that different residues are associated with different scaling parameters and correlation lengths. Dewey et al (1997). have demonstrated interesting relations introducing generalized thennodynamic parameters with these quantities (scaling parameters that are in turn related to fractal dimensions). Proteins show an intrinsic self-similarity in the compactness (folding) and packing of their structure. This is a simple form of fractal behavior (at the level of the secondary structures), but it has important consequences for the morphology of the protein and for the dynamics of protein folding. Protein folding concept has been a confusing phenomenon till date. Based on the positions of alpha carbon backbone several studies have been done on individual proteins to check the fractal properties (Wang and Huang (1990), Isvoran et. al (2001), Isvoran et. al (2001)).

We have studied the behavior of both primary and secondary structure elements (based on their positional distribution in the secondary structure database) and their possible preferences in the protein sequences by using fractal approach. These studies may provide us valuable insight into the general principles that govern the protein folding, rather than predicting the folding of a given protein. In the present study, we have used secondary structure database, which is a version of the PDBFINDER (Hooft et al, 1996) obtained based on DSSP program (Kabsch & Sander 1983). No selective filtering of the database was attempted.

3.2 METHODOLOGY

3.2.1 FRACTAL DIMENSION OF AMINO ACIDS OF PRIMARY SEQUENCES

We have followed the box counting method to determine the fractal dimensions of the relative normalized frequency curves for the 20 amino acid residues of Swiss-Prot working database (as illustrated in the section 2.3.1, Fig 4 and Fig 5 of chapter 2). The positional frequency values of the amino acids were calculated upto 256 position. The graph was constructed by actual counting the number of amino acids of type X present at the first position in, the whole database that corresponds to the frequency of amino acid X at position 1. In a similar manner, the frequencies of all the 20 amino acids were computed upto 256 position. These set of frequency points (i.e. the relative normalized frequencies at different positions, a total of 256 points) are distributed in a plane. Divide the plane into a number of square grids (taken for computational convenience as 4^1, 4^2 , 4^3, 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 (with 256 frequency points), sub divisions beyond 32 X 32 = 1024 boxes are not physically meaningful and hence were not carried out. Mathematically fractal dimension is calculated by using

$$D_A = \lim_{n\to\infty} \frac{\ln N_n(A)}{n \ln 2} \qquad (1)$$

$N_n(A)$ = Number of boxes counted at the *nth* level.

2" = The number of subdivisions at the *nth* level. Maximum value of n was 5 in this study. The details of this procedure has been described elsewhere (Meeta *et al,* 1993). Larger values of n cannot be used, as the sequence length considered is only 256.

3.2.2 FRACTAL DIMENSION OF SECONDARY STRUCTURE ELEMENTS

For our studies, we have considered polypeptide chains of length greater than or equal to 256. This number is a compromise between the typical sequence length and the need of having longer sequences to study dependencies. After trimming (i.e., removing all sequences smaller than 256), the working database (secondary structure) has 19,752 polypeptide chains. The total number of symbols present in this working database are approximately eight millions. The DSSP program (Kabsch *et al,* 1983) assigns each residue's secondary structure symbol to one of eight classes: α-helix (H), 3_{10} helix (G), β-strand (E), β-bridge (B), coil (C), Turn (T) and Bend (S). The symbol I (π-helix) was ignored, because they are very few in number in the database. The most abundant elements in the database are H, C and E. The least abundant symbols are G, S, B and T. The fractal dimensions of secondary structure elements were calculated by using Hausdroff-Besicovitch (box-counting principle) dimension (Meeta *et al,* 1993). The dimensions were calculated up to the fifth level and the fifth level values were considered for analysis. Based on the above equation Eq 1 the fractal dimension was calculated for the secondary elements.

3.2.3 TEST OF INDEPENDENCE OF AMINO ACIDS IN THE PROTEIN SEQEUNCES.

The positional behavior of amino acids was analysed by using fractal geometric approach (section 3.2.1). Moreover, study of adjacent pair dependencies of amino acids in the protein sequences is useful for understanding the folding problem. In this current section, we emphasized to check whether there are any adjacent dependencies of amino acids in the protein sequences? How to test the existence of dependencies of amino acids in the protein sequences? We followed Basawa and Prakasa Rao (Basawa and Prakasa Rao, 1980) statistical test to check pair dependencies of amino acids in the individual protein sequences. We applied Equation 2 on each individual protein sequences for testing the adjacent amino acid dependencies.

$$Z = \sum_{i=1}^{20}\sum_{j=1}^{20}\frac{\left(N_{ij} - N_i N_j\right)^2}{\dfrac{N_i N_j}{N}} \qquad (2)$$

Nij = Pair frequency count of *i*th and *j*th adjacent amino acids.

Ni = Frequency count of *i*th amino acids, *i* goes from 1 to 20.

Nj = Frequency count of *j*th amino acids, *i* goes from 1 to 20.

N = Total frequency of residues in the protein sequence.

Since there are 20 x 20 combinations (400) for the pair frequencies, the degrees of freedom (19 x 19) will be very large.

df = (number of row elements – 1) x (number of column elements – 1)

df = (20 - 1) x (20 - 1) = 19 x 19 = 381.

Z is distributed, as a $\chi 2$ variate with 19 x 19 degrees of freedom, under the hypothesis of independence (i.e. no dependency). As the degrees of freedom is rather large, Z value is approximated as a normal variate by the transformation (using Eq 3)

$$Y = \sqrt{(2Z)} - \sqrt{2\times 19^2 - 1} \qquad (3)$$

Y is expected to have an asymptotic normal distribution with mean zero and variance one under hypothesis of independence. The test procedure is to reject the hypothesis of independence if the observed value |Y| is significantly large.

3.3 RESULTS AND DISCUSSIONS

After trimming the Swiss-prot database (i.e., removing fragments and short sequences) we found that there are 41,408 protein sequences and a total of 22,408,660 amino acid residues. This working database (referred to as the working database in the following sections unless explicitly mentioned otherwise) was used for estimating fractal dimension values of amino acids of primary sequence.

3.3.1 FRACTAL DIMENSION OF AMINO ACID OF PRIMARY SEQUENCE

The positional distributions of all the amino acids were calculated as described in earlier section 2.3.1. The box counting method was implemented on all the 20 amino acid distributions. Based on the above Eq 1 we have calculated dimensions upto fifth level and the fifth level dimension values were taken for analysis.

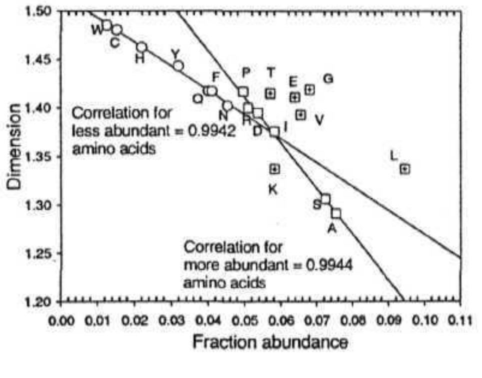


Figure 4: The fractal dimension values of the common amino acids showed interesting observations. There is a negative correlation between the fractal dimensions of amino acids with respect to abundance. We further sub-divided the graph into three different groups.

The plot showed very interesting results. The amino acids in the database showed a negative correlation between the fractal dimensions and their abundance values. Based on the dimensional behavior of amino acids, we could group them into three categories.

Group I: This group of amino acids are less abundant in the database with obtained high fractal dimensional values. The following amino acids W (1.24%, 1.48), C (1.52%, 1.48), H (2.2%, 1.46), Y (3.19%, 1.44), Q (4%, 1.41), F (4.1%, 1.41), N (4.53%, 1.4) with their respective percent abundance and fractal dimension values, are grouped in this category. Linear regression analysis of the plot (Fig 4) shows that these amino acids are negatively correlated with respect to their abundance and fractal dimensional values. The abundance of these amino acids is <5% (5% is the expected equiprobable abundance of amino acids) in the database. This group of amino acids are positionally more preferred (with high fractal dimension values) in the database comparing to other group of amino acids.

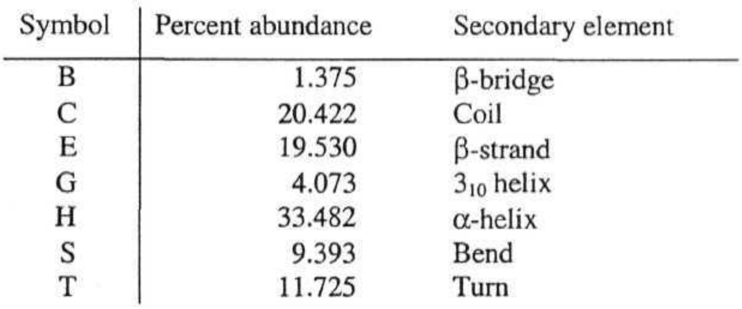
Group II: This group of amino acids are more abundant (>5%) in the database. The following amino acids P (5%, 1.41), R (5.1%, 1.4), D (5.36%, 1.4), I (5.82%, 1.38), S (7.23%, 1.30), A (7.52%, 1.30) with their respective percent abundance and fractal dimension values, are grouped in this category (Fig 4). These amino acids show a negative correlation between the fractal dimension and their abundance values. This group of amino acids are positionally less preferred in the database (with less fractal dimension values) comparing to other group of amino acids.

Group III: This group of amino acids are more abundant in the database. The following amino acids T (5.71%, 1.41), K (5.82%, 1.33), E (6.4%, 1.41), V (6.55, 1.39), G (6.8%, 1.42), L (9.44%, 1.33) with their respective percent abundance and fractal dimension values, are grouped in this category. These amino acids show no negative correlation with respect to abundance and dimensional values. This group is observed to be an exception to other groups.

3.3.2 PERCENT ABUNDANCE OF SECONDARY STRUCTURE ELEMENTS

The database used to study the fractal properties of secondary structure elements consists of 19,752 polypeptide chains. The composition of the seven secondary elements in the database was calculated. The highest abundant secondary element was found to be α-helix (~33.4%) and least abundant element is β-bridge (-1.37%). The various secondary structure elements and their composition are given in the in the Table 1.

Table 1 Percent abundance of the secondary elements in the database



Symbol	Percent abundance	Secondary element
B	1.375	β-bridge
C	20.422	Coil
E	19.530	β-strand
G	4.073	3_{10} helix
H	33.482	α-helix
S	9.393	Bend
T	11.725	Turn

3.3.3 POSITIONAL DISTRIBUTION OF THE SECONDARY STRUCTURE ELEMENTS

The positional distribution values of the secondary structure elements were calculated upto 256 position. Fig 5A (H and C) and 5B (E and T) shows the distributions of four selected elements. The graph was constructed by actual counting the number of H present at the first position in the whole database that corresponds to the frequency of H at position 1. In a similar manner, the frequencies of all the seven elements were computed upto 256 position. If the distributions were random, uniform distribution would be expected and the graph would have been a horizontal line. It is important to note here that the sample size is sufficiently large and statistical fluctuations, commonly associated with small samples can be safely ignored. Further, one notes that the distribution patterns are characteristic, i.e., each structural element has its own distinct positional distribution pattern. Of the seven secondary elements, the element C (coil) is significantly predominant in the first two positions (1-2) in the database. Similarly, a-helix is significant in its absence in the first few positions. This is not really unexpected, as the regular structure needs few previous residues to build up some order.

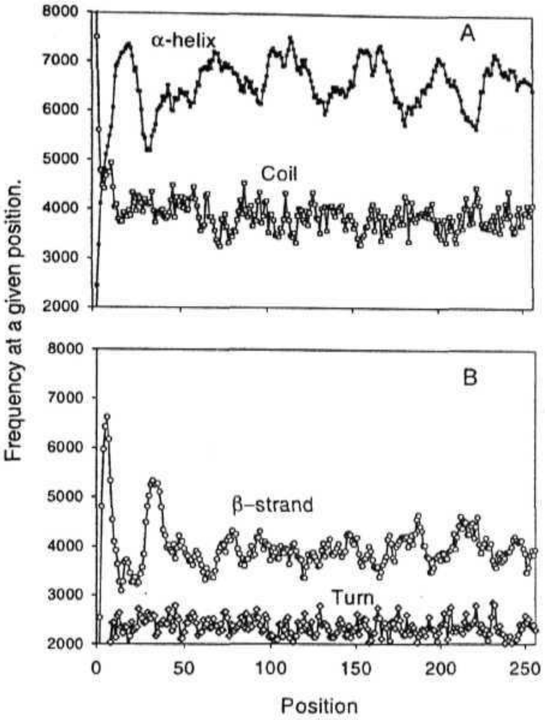


Figure 5: Positional distributions of secondary structure elements in database [The distributions of α-helix, β-strand, coil and turn elements were calculated and can be visualized in panels A and B of the figure. Similarly, other three secondary elements positional distributions were also calculated using the same method]

3.3.4 FRACTAL DIMENSION OF SECONDARY STRUCTURE ELEMENTS

The fractal dimensions calculated for the secondary structure elements showed interesting results. Based on the percent abundance and respective fractal dimensions of the elements, we categorized the seven structural elements into two different groups.

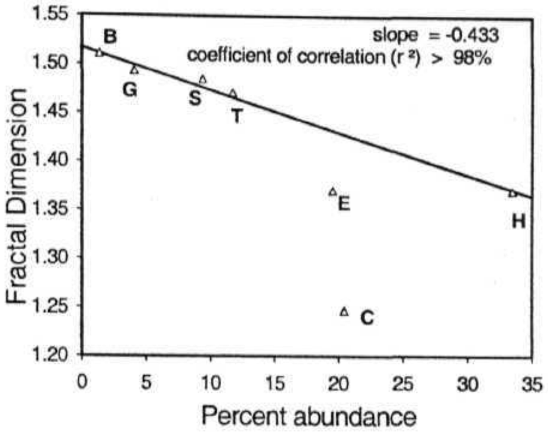


Figure 6: Fractal dimension values of seven representative secondary elements based on box counting algorithm [Fractal dimensions were calculated up to the fifth level and the fifth level values were considered. The Y-axis gave the fractal dimension values, comparing to the abundance of those respective secondary elements in the X-axis. Based on the abundance and fractal dimension values, the elements were categorized into two groups]

Group I The percent abundance of secondary structure elements B, G, S, T and H in the database are 1.38, 4.1, 9.4, 11.73 and 33.50, respectively. The fractal dimensions are 1.51, 1.492, 1.48, 1.47 and 1.37 respectively. There is an overall negative correlation (-0.433 and the correlation coefficient is greater than 98%) with respect to percent abundance and fractal dimensions of the elements in this group (Fig 6). The least abundant element in this category is the P-bridge (B). The fractal dimension of this element is the highest in this group, suggesting that the distribution of this structural element is less random or more ordered. This shows that the least abundant element will be preferred more comparing with the more abundant elements. The percent abundance of α-helix (H) in the database is the highest (33.48%), compared to other secondary structure elements. However, the fractal dimension is the least of all the elements in this category, which shows that the preferences are less. In other words, the distribution of this element is more uniform (more randomly distributed) in the database comparing with other elements.

Group II The secondary structure elements C (coil) and E (β-strand) fall in to this category. The percent abundance of C and E is 20.42 and 19.53 and the dimensional values are 1.25 and 1.37, respectively. This shows that even though the abundance of C is more, this element is preferred less, showing a random distribution in the database. The fractal dimension of element E is more, but the relative abundance of E, comparing with C is approximately same. This shows that the element E is relatively more preferred (more ordered) in the database, comparing to the element C.

3.3.5 TEST OF INDEPENDENCE OF AMINO ACIDS IN THE PROTEIN SEQEUNCES.

In this section, we clearly demonstrated the existence of amino acid pair dependencies in the individual protein sequences by following Baswa and Prakasa Rao statistical test of independence. This test was conducted on each individual protein sequence of Swiss-Prot working database. We calculated the pair frequency counts of amino acids (20x20 combinations) in the individual sequences and implemented Eq 2 to test dependence hypothesis. Under the hypothesis of independence (i.e. no dependency), the obtained Z value (by applying Eq 2) for each sequence is distributed as a $\chi 2$ variate with 19×19degrees of freedom.

As the degrees of freedom is rather large, Z value is transformed by using Eq 3. The test procedure is to reject the hypothesis of independence (in favor of dependency) if the observed value |Y| is significantly large.

We have plotted the values of the test statistic Y for all the sequences in the working database as a histogram in Figure 7. Under hypothesis of independence the Y is expected to have asymptotic normal distribution with mean zero and cumulative probability plot to be straight line.

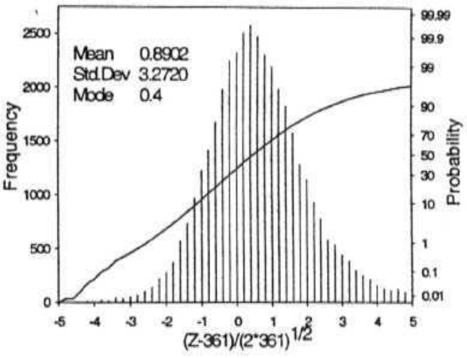


Figure 7: The figure describes **the** values of transformed variateY in the form of histogram. The shape of the distribution appears close to normal distribution with mean of 0.890 and mode of 0.4. The cumulative probability plot of the frequency distribution is not a straight line.

Even though the graph shows that most of the sequences lie inside the acceptance region (±2.58 at 99% level of significance), the mean of the distribution is significantly different from zero (~0.9) with mode of 0.4. The cumulative probability plot is not a straight line.

The result of the test clearly indicates a hope of dependency among the adjacent amino acids. This conclusion provided us a hope to apply Markov dependencies on adjacent amino acids to check the existence of short-range interactions if any in the protein sequences.

CONCLUSION

The fractal dimensions of amino acids of the primary sequence were calculated based on the positional distributions. The amino acids could be grouped into three distinct classes, based on their fractal dimensions. Amino acids in group I (with less than 5% abundance in the database) and group II (with more than 5% abundance) showed a negative correlation between the dimensional values and their abundances. The amino acids in group HI showed an exceptional behavior from the other two groups of amino acids.

The present study of secondary structure elements shows a negative correlation of fractal dimension with respect to abundance. The secondary structure elements that are more abundant are positionally less preferred with respect to less abundant elements.

Moreover, the positive result of statistical test for pair dependencies of adjacent amino acids in the protein sequences gave a hope to apply Markov dependencies for checking the existence of short-range interactions if any in the protein sequences.

References

Barnsley, M. (1988) in Fractals everywhere, Academic Press, Inc

Basawa, I.V. and Prakasa Rao, B.L.S. (1980): Statistical inference for stochastic processes. Academic press, Inc. London-New York.

Brown, S.R. (1995) Measuring the dimension of self-affine fractals: examples of rough surfaces. In: C.C.Barton and P.R. La Pointe (eds.), Fractals in the earth sciences. Plenum Press, New York pp. 77-87.

Deering, W. and West, B J. (1992) Fractal physiology, IEEE Engin. Med. Biol. 11:40-46

Frontier, S. (1987). Applications of fractal theory to ecology. In; Legendre, P. and L. Legendre (eds), Developments in numerical ecology. pp. 335-378. Springer, Berlin.

Goldberger, Rigney and West, Scientific American 262, 34 (1990)

Hastings H.M., and Sugihara, G. (1993) Fractals - A User's Guide For The Natural Sciences, Oxford University Press, Oxford, United Kingdom. 235 pp.

Hooft, R. W. W., Sander, C, and Vriend, G. (1996) The PDBFINDER database: A summary of PDB, DSSP and HSSP information with added value. CABIOS, 12, 525-529.

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

Isvoran, A., Licz, A. and Morariu, V. V. (2001) Determination of the fractal dimension of Ascarys Lumbricoides Trypsin Inhibitor, Revue Roumaine Chimie 48 (8), 51-53

Isvoran, A., Licz, A., Unipan, L. and Morariu, V. V. (2001) Determination of the fractal dimension of the lysozyme backbone for three different organisms, Chaos Solitons Fractals. 12, 757-760

James, S. B. and Dewey, T. G. (1997) Multifractal analysis of solvent accessibilities in proteins, Phys Rev E. 52, 880-887

Kabsch, W. and Sander, C. (1983) A dictionary of protein secondary structure Biopolymers, 22, 2577-2637

Kenkel, N.C. and Walker, D.J. (1993) Fractals and ecology, Abst. Bot. 17: 53-70.

Lam, N. S. and L. De Cola (eds) (1993) Fractals in geography, Prentice Hall, Englewood Cliffs, NJ, U.S.A.

Loehle, C. (1983) The fractal dimension and ecology, Specul. Sci. Tech. 6: 131-142.

Lorimer, N.D. Haight, R.G. and Leary, R.A. (1994) The fractal forest: fractal geometry and applications in forest science. U.S. Department of Agriculture, Forest Service. North Central Forest Experimental Station, General Technical Report, NC-170. 43 pages.

Mandelbrot, B.B. (1983) The Fractal Geometry of Nature, W.H.Freeman & Company, New York,

Meeta, R. and Mitra, C. K. (1993) Protein sequence as random fractals, J Bios. 18, 213-220

Meeta, R. and Mitra, C. K. (1996) Pair-preferences: a Quantitative Measure of regularities in protein sequences, J Biomol Struct Dynamics. 13,935-944

Mingtian, Ni. and Xin, Li. (2001) Lecture: "Procedure-based Simple Plant Simulation", Department of Computer Science and Engineering University of Nebraska-Lincoln.

Normant, F. and Tricot, C. (1993) Fractal simplification of lines using convex hulls, Geogr. Anal. 25: 118-129.

Peitgen and Richter, (1986) The Beauty of Fractals, Springer Verlag.

Rogers, (1970) Hausdroff Measures, Cambridge University Press

Schroeder, M. (1991) Fractals, chaos, power laws, Minutes from an infinite paradise. Freeman, New York.

Stapleton, Allen, Flynn, Stinson and Kurtz, (1980) Phys. Rev. Lett 45, 1456

Turcotte, D.L. (1986) Fractals and fragmentation, J. Geophys. Res. 91: 1921-1926.

Wang, C. X. and Huang, F. H. (1990) Fractal study of tertiary structure of proteins, Phys Rev A41, 7043-7048

West, B.J. and Goldberger, A.L. (1987) Physiology in fractal dimensions, Am. Sci. 75: 354-365.

Wiens, J.A. (1989) Spatial scaling in ecology, Funct. Ecol. 3: 385-397.

Wiens, J.A. and Milne, B.T. (1989) Scaling of 'landscapes' in landscape ecology, or landscape ecology from a beetle's perspective, Land. Ecol. 3: 87-96.

INFORMATION CONTENT OF PROTEIN SEQUENCES

4.1 INFORMATION THEORY

Information theory is concerned with the mathematical laws governing the transmission and reception of information. More specifically, information theory deals with the measurement of information, the representation of information (such as encoding), and the capacity of communication systems to transmit and receive information.

Encoding can refer to the transformation of information from one form to another during transmission or reception.

Information theory was first developed in 1948 by the American electrical engineer Claude E. Shannon in his article, A Mathematical Theory of Communication (Shannon, 1948). The need for a theoretical basis for communication arose from the increase in the complexity and the crowding of communication channels, such as telephone, teletype networks and radio communication systems. Information theory also encompasses all the other existing forms of information transmission and storage, including television and the electrical impulses transmitted in computers and in magnetic and optical data recording. The term information refers to the transmitted messages: voice or music transmitted by telephone or radio, images transmitted by television systems, digital data in computer systems and networks, and even nerve impulses in living organisms. More generally, information theory has found application in such varied fields of inquiry as cybernetics, cryptography, linguistics, psychology, and statistics.

The most extensively studied type of communication system consists of several components. An information source (such as a person speaking) produces information, or a message, that is to be transmitted. A transmitter, such as a telephone and amplifier, or a microphone and radio transmitter, converts the message into electronic or electromagnetic signals. These signals are transmitted through the channel, or medium, such as a wire or the atmosphere. The channel, in particular, is susceptible to interference issuing from many sources, which distorts and degrades the signals. (Examples of interference, known as noise, include the static in radio and telephone reception and the “snow” in television picture reception.) The receiver, such as a radio receiver, reconstructs the signal back into the original message. The final component is the reception, such as a person listening to the message.

4.1.1 INFORMATION CONTENT

A fundamental concept in information theory is that the amount of information in a message, called information content, is a measurable mathematical quantity. The term content does not refer to the meaning of the transmitted message, but to the probability that a given message will be received from a set of possible messages (Feller, 1950).

The highest value for the information content is assigned to the message that is the least probable. If a message is expected with certainty its information content is 0. If a coin is tossed, for example, the combined message "heads or tails", describing the result, has no information content. The two separate messages "heads" or "tails", on the other hand, are equally probable and have probabilities of one-half. In order to relate information content (I) to probability, Shannon introduced a simple formula

$$I = - \log_2 p$$

in which p is the probability of a message being transmitted and $\log_2$ is the logarithm of p to a base 2. Using this formula, it is found that the messages "heads" or "tails" have information content equal to $\log_2 2 = 1$.

The information content of a message can be understood in terms of the number of possible symbols that represent a message. In the example above, if "tails" is represented by a 0, and "heads" by a 1, there is only one choice to represent the message: 0 or 1. The 0 and the 1 are the digits of the binary system (Number Systems), and the choice between those two symbols corresponds to the so-called binary information unit, or bit. If a coin is tossed three times in a row, the eight equally possible results (or messages) can be represented as 000, 001, 010, 011, 100, 101, 110, or 111. These messages correspond to the numbers 0, 1, ... 7 written in binary notation. The probability of each message is one-eighth, and its information content is $\log_2 8 = 3$, which is the number of bits needed to represent each message.

4.1.2 ENTROPY

In most practical applications, one must choose among messages that have different probabilities of being sent. The term entropy has been borrowed from thermodynamics to denote the information content of these messages (Cramer, 1946) (Schrodinger, 1948). Entropy can be understood intuitively as the amount of "disorder" in a system. In information theory the entropy of a message equals its average information content. If in a set of messages, the probabilities are equal, the formula for the total entropy can be given as H = -Z pi log p_i, where p_i is the probability of /th message (Cover and Thomas, 1991).

4.1.3 ENCODING AND REDUNDANCY

If messages are transmitted consisting of random combinations of the 26 letters of the English alphabet, the space, and five punctuation marks, and if it is assumed that the probability of each message is the same, the average entropy in bits to encode each character is $H = -\sum 1/32 \log_2 1/32 = 5$ where probability of each message p_i is 1/32 for 32 characters considered. This means that five bits are needed to encode each character, or message: 00000, 00001, 00010 ... 11111. Efficient transmission and storage of information require the reduction of the number of bits used for encoding. This is possible when processing English texts because letters are far from being completely random. The probability is extremely high, for example, that the letter following the sequence of letters "INFORMATIO" is an "N".

It has been shown that the entropy of ordinary written English is about one bit per letter. This indicates that the English language (like every other language) has a large amount of redundancy incorporated in it, which is called natural redundancy. This redundancy enables a person, for example, to understand messages in which vowels are missing, or to decipher unclear handwriting.

4.1.4 STUDY OF INFORMATION ON PROTEIN SEQUENCES

Functionally, proteins are the most diverse of all biological macromolecules. All proteins, whether from the most ancient lines of bacteria or from the most complex highly developed forms of life, are constructed from the same ubiquitous set of 20 amino acid (Lehninger and Nelson, 1993). The three dimensional structure of a protein is uniquely encoded in the primary sequence and in principle (cf. chaperones) it is possible to predict the most probable three dimensional structure of a protein based solely on the primary sequence (Anfinsen, 1973). The function of a protein is closely related to its three-dimensional structure but the prediction of the function from a given structure remains difficult. In reality this has remained an unsolved problem even today although several empirical treatments of the problem are available in the literature (Chou and Fasman, 1974a, b) to predict the structure from the primary sequence.

The median length of a protein sequence based on the Swiss-Prot database is ~350. However, the distribution of sequence length is quite broad and sequences smaller than 50-100 residues are quite common and on the larger side sequences can be as long as 400-500 residues in length. This again suggests that the final structure is more important than the primary sequence. As an example, there are 20^{350} ($\sim 10^{455}$) possible sequences (assuming a typical length 350 for a protein) whereas we find less than 10^{5} sequences in the living system (e.g., human genome is reported to have only ~30,000 genes and most genes code for only one protein). This again suggests that most of the theoretically possible sequences are not biologically meaningful, as they do not meet the essential requirement of a well-defined three-dimensional structure. Therefore we can conclude that the native sequences are a very special subset (Meeta Rani et al., 1995) (that have some function) of the full set of all possible sequences (most of them will be random without any function). This is comparable to the common English language, in which case we have 26 letters (ignoring the case) and all possible combinations are not meaningful- only a very special subset constitutes words. Again, all possible combinations of words are not meaningful, only a very special subset constitutes sentences. A common English text will have a sequence of highly correlated sentences. We see a similar hierarchical level of structures in protein sequences. Many have been identified qualitatively by careful analysis and observation.

In this work, we tried to show that there are some approaches by which we can analyze the functional significance of pattern of alignment of amino acids. We used the principles of information theory, in particular the concept of entropy, to study the sequences.

The entropy of a random variable X with a probability distribution of p(X) will be defined as, following Shannon Equation 1 (Cover and Thomas, 1991),

$$H(X)= - E\,(\log p(X))= - \Sigma\, p(X) \log p(X) \qquad (1)$$

which is functionally equivalent to Boltzmann H- theorem (without the negative sign). Boltzmann showed that in a spontaneous process (collisions) H never increases and can be identified as the entropy of the system, when p (X) is taken as the distribution function of the velocities in an ideal gas (Eq 2).

$$S =- kH = k \ln W \qquad (2)$$

where S is the entropy (of a given system) and W is the fraction of all possible arrangements where the given configuration is realized (k is the Boltzmann constant). This statistical mechanical derivation of entropy as a measure of order (or randomness) is formally equivalent to the classical thermodynamics. A rigorous derivation including quantum statistics must include an additional factor or statistical weight while calculating the probability W. The probability function need not correspond to the equilibrium state of the system but the equilibrium state can be identified with the most probable probability distribution corresponding to W. This gives us a very powerful but simple technique to measure the degree of order in a given system. The equilibrium state has the most probable distribution and highest entropy. The most ordered system has the lowest probability and the lowest entropy. As a system moves towards equilibrium, the randomness increases and the entropy also increases. The Shannon definition of entropy only lacks the -k factor required in the classical thermodynamics. To obtain the functional significance of pattern of amino acids we calculated Markov approximations and used them to estimate the entropies of those approximations.

4.1.4.1 MARKOV CHAIN

A Markov sequence is one in which each term has a direct dependence on the immediate preceding term. This gives rise to a sequence that is a simple or first order Markov sequence. We also have higher order Markov sequences in which each term depends directly on the two, three or more number of previous terms. In this way there is always a strong correlation between successive elements of the sequence. However, this is an example of short-range interactions or correlations, as we do not expect that two terms separated by a significant distance to be correlated. We expect short-range interactions to be present in protein sequences and therefore we also expect Markov dependence. However, it is well known that long-range interactions play a very important role in the overall folding in a protein sequence and therefore both short range and long-range order are expected.

The primary sequence of a protein is believed to be directly responsible for the secondary structure. However, the secondary **structure** is directly responsible for the folded conformation of the molecule and therefore the primary sequence is only indirectly responsible for the overall three dimensional structure of the protein. Therefore it is imperative that analysis may be carried out in a similar fashion, i.e., the primary sequence may be used to predict the secondary structure and only the •secondary structure may be used to predict the final folded three dimensional conformation of the molecule. It is difficult, if not impossible, to predict the tertiary structure starting directly with the primary sequence. On the other hand, various attempts, mostly empirical or semi-empirical, to predict the secondary structure from the primary sequence has been quite successful. Dewey *et al.* have demonstrated interesting relations by implementing Shannon information theory to check the redundancy of hydrophilicity and solvent accessibility of Concanavalin A protein (Bonnie and Dewey, 1995). In our present studies, the information content upto 3rd order Markov approximations was calculated on a global perspective in estimating the local interactions if any in the protein sequences of Swiss-Prot database. Similar approach was followed on other non-redundant ASTRAL (http://astral.berkeley.edu) (Brenner et.al., 2000) protein databases and on the randomly simulated protein sequences.

It is not possible to extend this process to very high order, as the size of the database is a limiting factor. Alternative methods for estimating the entropy of natural protein sequences are therefore required. Further, we have studied entropy on the secondary structure models for getting better information. In analogy with the English language, we can compare words with the secondary structure present in the sequence. Various secondary structures are already identified in the literature and they are reasonably few in numbers. We could compute the information content of such sequences using a modified form of the Shannon formula.

4.2 METHODOLOGY

To obtain the functional significance of pattern of amino acids we calculated Markov approximations and used them to estimate the entropies of those approximations. A Markov sequence is one in which each term has a direct dependence on the immediate preceding term. This gives rise to a sequence that is a simple or first order Markov sequence. We can also have higher order Markov sequences in which each term depends directly on the two, three or more number of previous terms. In this way there is always a strong correlation between successive elements of the sequence. However, this is an example of short-range interactions or correlations, as we do not expect that two terms separated by a significant distance to be correlated. We expect short-range interactions to be present in protein sequences and therefore we also expect Markov dependence. However, it is well known that long-range interactions play a very important role in the overall folding in a protein sequence and therefore both short range and long-range order are expected. Aggregation of residues into a meaningful way has been reported earlier in a different way by using fractal studies (Bonnie and Dewey, 1995).

The frequencies of occurrence of the various amino acid pairs (20 x 20, i.e. 400 entries), triplet (20 x 20 x 20, i.e. 8000 entries) and quadruplets (20 x 20 x 20 x 20, i.e. 160,000 entries) were obtained from Swiss Prot

Protein Sequence Databank (Release 26, 1994). For reasons of convenience, all sequences that are shorter than 256 residues were excluded. No selective filtering of the database was attempted. We feel that screening of the database may introduce a fresh and additional bias (rather than to remove any bias already present). It is difficult to remove any existing bias in the database when the sources of the bias are not well known or completely characterize. The swiss-prot database contains a number of fragments (incomplete sequences; approximately 6000) that should be excluded, as the initial starting point is not known. A number of sequences are also very small and very small peptides do not possess a definite three dimensional structure. Sequences larger than ~100 residues are known to have a well defined structure useful for binding a substrate. However, we must mention that the choice of 256 residue cut-off is purely arbitrary but our results are not strongly dependent on this (cut-off value). We have also shown that the initial part of the sequences follow a different distribution (overall composition) of amino acids and this stabilizes only after ~50 residues [Mitra and Sen, 2001). Thereafter the distribution becomes stationary. If the chosen sequences are only 50 residues, we will be ignoring the stationary distribution. If the chosen sequences are -100 residues long, we shall be observing both the initial and stationary distributions. If the chosen sequences are very long, we shall be seeing only the stationary distributions (but the number of very long sequences are rather less). Very short sequences contribute more towards noise than information.

The bias present in the Swiss-Prot protein sequence databank can be attributed to several factors, but it is almost certain that the database cannot be considered a random sample. The proteins to be selected for sequencing are not chosen at random. Therefore a complete protein database derived from the genomic data of a given organism may be a better representative sample. However, no such database is available at present. The ASTRAL SCOP (http://astral .berkeley.edu) database suggests a novel idea in selecting a set of non-redundant, and *possibly* a random representative, protein sequence database. We have chosen a set of protein sequences with less than 40% and 95% sequence similarity [Brenner et al., 2000] for our reference. However, removing similar sequences does not assure that the database is unbiased. Therefore, the same computations have been repeated with this non-redundant database also.

To compare our results we simulated, 41,408 random sequences using a Monte Carlo method, (same as the number being analyzed in Swiss-prot) having the same amino acids composition as in the Swiss-Prot database. These random sequences were analyzed in the same way as real sequences.

4.2.1 MARKOV APPROXIMATIONS ON VARIOUS PROTEIN DATABASES

Case I: The symbols are independent and equiprobable. Since there are 20 common amino acids the case I approximation will have 20 identical entries to evaluate. This result is independent of any database.

Case II: The symbols are independent. Frequency counts of the various databases were calculated and evaluated. The total entries are 20, one for each amino acid residue.

First order Markov dependence: The frequencies of the pairs of amino acid counts were taken in to consideration and followed up for evaluation. The total entries are 400 (20 x 20). This corresponds to a first order Markov sequence.

Second order Markov dependence: The frequencies of triplets of amino acids are counted and the total entries are 8000 (20 x 20 x 20). This corresponds to a second order Markov process.

Third order Markov dependence: The frequencies of quadruplets of amino acids are counted and the total entries are 160,000 (20 x 20 x 20 x 20). As the number of entries is large, the actual frequencies are small and *for a small database can give rise to significant errors.* For this same reason, higher order approximations cannot be reliably performed on the available databases.

Apart from the above calculations we have done studies to check any preferences among amino acids in the database. This was calculated by leaving one amino acid in between. All the Markov approximations are calculated as above by leaving one amino acid in between. The pattern that we followed for calculating the counts for the first second and third order approximations are A_iXA_j, $A_iA_jXA_k$ and $A_iA_jA_kXA_l$ respectively, where A;, Aj, A_k and A_l are any common residue (labeled by i, j, k and 1) and X is any intervening amino acid.

4.2.2 ENTROPIES FOR THE MARKOV APPROXIMATIONS

We followed Shannon's limit *H* for calculating the entropies of Markov approximation (Cover and Thomas, 1991) Eq. (1).

Case I: The amino acids are equiprobable and independent, i.e. $\log_2 20$. This is the maximum achievable entropy (4.322 bits per residue).

Case II: The amino acids are independent and the p_i value is the fraction or proportion of individual amino acids.

$$I = -\sum_{i=1}^{20} p_i \log_2 p_i \tag{3}$$

Entropy for first order approximation (I): The pair frequency counts of amino acids were evaluated for entropy calculation.

$$I = -\sum_{i=1}^{20} p_i \sum_{j=1}^{20} p(j \mid i) \log_2 p(j \mid i) \tag{4}$$

Transition matrix: After obtaining the pair frequency count (20x20 matrix) from the database, each individual row was normalized by taking the sum for each row and dividing each element by the corresponding row sum. This matrix is now called the transition matrix.

$$p(j \mid i) = \frac{\text{Frequency of ith with adjacent jth amino acid}}{\text{Rowsum of ith row.}}$$

and is the conditional probability of j given the probability of ith residue in succession.

Entropy for second order approximation (I): The triplet frequency counts of amino acids were evaluated for entropy.

$$I = -\sum_{i=1}^{20}\sum_{j=1}^{20} p(i,j) \sum_{k=1}^{20} p(k \mid i,j) \log_2 p(k \mid i,j) \tag{5}$$

$$p(k \mid i,j) = \frac{\text{Frequency of triplets of ith jth with adjacent kth amino acid}}{\text{Rowsum of i, jth row.}}$$

p(k|i,j) = Conditional probability of kth residue given the adjacent probabilities of *i*th, *j*th residues and

p(i,j) = Joint probability of *i*th and *j*th amino acids in succession.

$$p(i,j) = \frac{Aij}{(\text{Total number of residues} - \text{Total number of sequences})}$$ where

$Aij = $ Actual count for Pair of i and j adjacent amino acids .

Entropy for third order approximation (I): The quadruplet frequency counts of amino acids were evaluated for entropy.

$$I = -\sum_{i=1}^{20}\sum_{j=1}^{20}\sum_{k=1}^{20} p(i,j,k) \sum_{l=1}^{20} p(l \mid i,j,k) \log_2 p(l \mid i,j,k) \tag{6}$$

$$p(l \mid i,j,k) = \frac{\text{Frequency of triplets of ith jth kth with adjacent lth amino acid}}{\text{Rowsum of } i, j, k\text{th row.}}$$

p(l|i,j,k) = Conditional probability of lth residue given the adjacent probabilities of i,j,k residue.

p(i, j, k) = Joint probability of ith, jth and kth amino acids in succession.

$$p(i,j,k) = \frac{Aijk}{(\text{Total number of residues} - 2 * \text{Total number of sequences})}$$

$Aijk = $ Actual count for triplet of i , j and k adjacent amino acids

This approach was followed for all the databases considered.

4.2.3 CALCULATION OF MARKOV APPROXIMATIONS ON SECONDARY STRUCTURE ELEMENTS

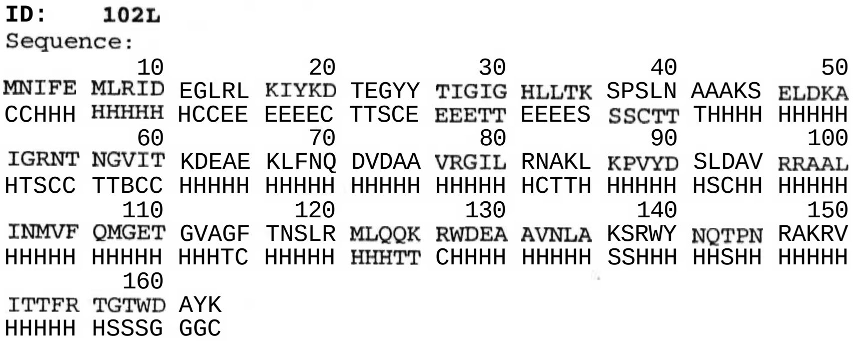
To study the entropy of secondary structure elements we have used secondary structure database, which is a version of the PDBFINDER (ftp.cmbi.kun.nl/pub/ molbio/data/pdbfinder2) (Hooft et.al., 1996) with the secondary structure elements assigned based on DSSP program (Kabsch and sander, 1983). No selective filtering of the database was attempted.

The DSSP code is frequently used to describe the protein secondary structures with a single letter code. DSSP is an acronym for "Dictionary of Protein Secondary Structure", which was the title of the original article actually listing the secondary structure of the proteins with known 3D structure (Kabsch and Sander 1983). The secondary structure is assigned based on hydrogen bonding patterns as those initially proposed by Pauling et al. in 1951 (before any protein structure had ever been experimentally determined).

- G = 3-turn helix (3_{10} helix). Min length 3 residues.
- H = 4-turn helix (alpha helix). Min length 4 residues.
- I = 5-turn helix (pi helix). Min length 5 residues.
- E = beeta sheet in parallel and/or anti-parallel sheet conformation (extended strand). Min length 2 residues.
- B = residue in isolated beta-bridge (single pair beta-sheet hydrogen bond formation)
- S = bend (the only non-hydrogen-bond based assignment)
- C = Random coil
- T = hydrogen bonded turn (3, 4 or 5 turn)

The helices (G,H and I) and sheet conformations are all required to have a reasonable length. This means that 2 adjacent residues in the primary structure must form the same hydrogen bonding pattern. If the helix or sheet hydrogen bonding pattern is too short they are designated as T or B, respectively.

The secondary structure elements obtained from a DSSP program from a protein sequence was described not in terms of the amino acid sequence but by a sequence like:



```
ID:    102L
Sequence:
         10         20         30         40         50
MNIFE MLRID EGLRL KIYKD TEGYY TIGIG HLLTK SPSLN AAAKS ELDKA
CCHHH HHHHH HCCEE EEEEC TTSCE EEETT EEEES SSCTT THHHH HHHHH
         60         70         80         90        100
IGRNT NGVIT KDEAE KLFNQ DVDAA VRGIL RNAKL KPVYD SLDAV RRAAL
HTSCC TTBCC HHHHH HHHHH HHHHH HHHHH HCTTH HHHHH HSCHH HHHHH
        110        120        130        140        150
INMVF QMGET GVAGF TNSLR MLQQK RWDEA AVNLA KSRWY NQTPN RAKRV
HHHHH HHHHH HHHTC HHHHH HHHTT CHHHH HHHHH SSHHH HHSHH HHHHH
        160
ITTFR TGTWD AYK
HHHHH HSSSG GGC
```

The above illustration shows the primary sequence of Hydrolase (O-Glycosyl) with ID 102L and the line below to the primary sequence represents their respective secondary structure elements.

For our studies, we have considered polypeptide chains of length greater than or equal to 256. This number is a compromise between the typical sequence length and the need of having longer sequences to study dependencies. After trimming (i.e., removing all sequences smaller than 256), the working database of secondary structure contains 19,752 polypeptide chains. The total number of symbols present in this working database are approximately eight millions. The DSSP program assign each residue's secondary structure symbol to one of eight classes: α-helix (H), 3_{10} helix (G), β-strand (E), (3-bridge (B), coil (C), Turn (T) and Bend (S). The symbol I (π-helix) was ignored because of very few in number ~30 in the database. We could able to compute the information content of such sequences using a modified form of the Shannon formula. Most of the calculations were been done on Linux Operating System by using C++ language (gcc).

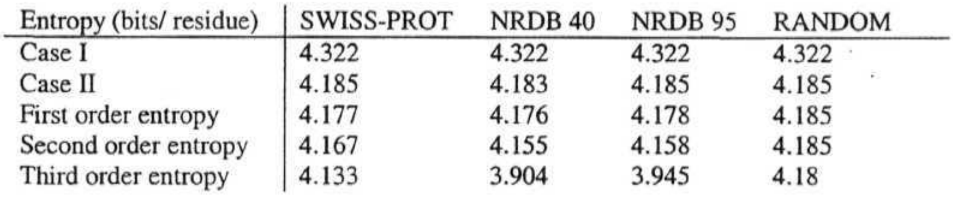
4.3 RESULTS AND DISCUSSIONS

After trimming the Swiss-prot database (i.e., removing fragments and short sequences) we found that there are 41,408 protein sequences and a total of 22,408,660 amino acid residues. This working database was used for our further Markov approximation studies (referred to as the working database in the following sections unless explicitly mentioned otherwise).

4.3.1 ENTROPY OF PROTEIN PRIMARY SEQUENCES ON VARIOUS DATABASES

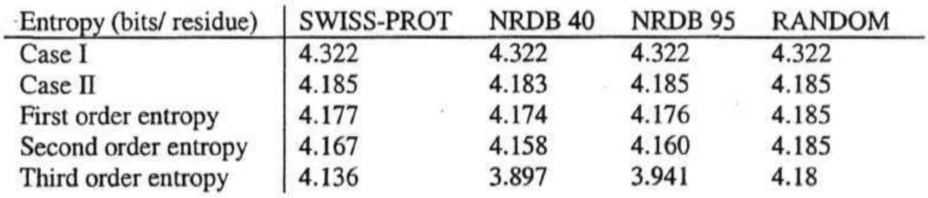
The following are the entropy (bits per symbol) values for various databases studied. The Table 1 below gives the values of entropies from case I approximation until third Markov approximation.

Table 1: Entropy (bits per residue) calculated upto the 3rd order in different databases



Entropy (bits/ residue)	SWISS-PROT	NRDB 40	NRDB 95	RANDOM
Case I	4.322	4.322	4.322	4.322
Case II	4.185	4.183	4.185	4.185
First order entropy	4.177	4.176	4.178	4.185
Second order entropy	4.167	4.155	4.158	4.185
Third order entropy	4.133	3.904	3.945	4.18

Table 2: Entropy (per bit) calculated upto the 3rd order in different databases with one gap between the elements of the pairs.



Entropy (bits/ residue)	SWISS-PROT	NRDB 40	NRDB 95	RANDOM
Case I	4.322	4.322	4.322	4.322
Case II	4.185	4.183	4.185	4.185
First order entropy	4.177	4.174	4.176	4.185
Second order entropy	4.167	4.158	4.160	4.185
Third order entropy	4.136	3.897	3.941	4.18

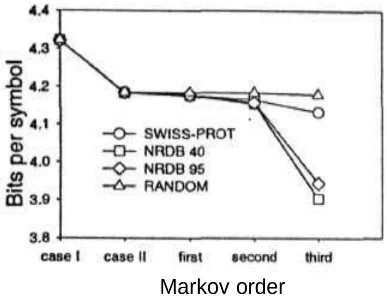


Figure 1: This figure shows the entropy values (bits per symbol) calculated for case I to third order Markov approximations (without gap). The values show that there is a gradual decrease of entropy with more complexity in Markov order approximation.

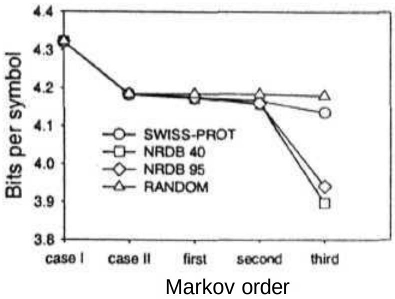


Figure 2: This figure shows the entropy values (bits per symbol) calculated for case I to third order Markov approximations (with one gap). The values show that there is a gradual decrease of entropy with more complexity in Markov order approximation.

The above table (Table 1) clearly shows that the entropy of proteins, expressed in bits per residue, for the equiprobable distribution is log 20 = 4.322. As expected, this value is same for all cases. However, if we introduce the actual probabilities of the 20 different residues as observed in a given database, we see a significant decrease in the entropy. In case II, we find the entropy to be 4.185 for the swiss-prot database. The values are very similar for other databases as the abundances of the 20 different amino acids are very similar. As we increase the complexity of the model, we capture more significant patterns of protein sequences and the conditional uncertainty of the next residue is reduced. The first order model Eq. (5) has given an estimate of 4.177 bits per symbol (Swiss-prot) while the second order model Eq. (6) gives an estimate of 4.167 bits per symbol (Swiss-prot). For the third order model Eq. (7), there is an entropy decrease of 0.186 bits per symbol comparing with the case I. Although the entropy changes are relatively small, they are nevertheless significant in **our** opinion. The entropies reported are in bits/residue and in a reasonably large protein may well add up to a significant contribution. The higher order entropies could not be calculated as the databases are not large enough (we may estimate the population size of the non-redundant sequence database to be 1-2 $x10^5$ sequences). However, the present NRDB databases are too small for an accurate evaluation of the third order entropies.

If we compare the NRDB 40 and NRDB 95 databases, the entropy values obtained seems to be similar with minor differences with respect to Swiss-prot database upto 2nd order model (figure 1). This is somewhat expected as the non-redundant databases are made based on lack of large-scale homology present. Therefore the preferences that are used for the computations are not altered significantly. Differences were observed in entropy values between Swiss-prot and non-redundant databases (NRDB 95 and NRDB 40) at the 3rd order model. This difference might be because of insufficient data in non-redundant databases with respect to Swiss-Prot Databank. So the values obtained in the non-redundant databases seem to be unreliable to check the 3rd order models, mainly because of their much smaller size.

The preference studies for amino acid pairs separated with a single residues (i.e., AjXAj where X is any residue) were computed and it was found that there is no significant differences between the elements of the transition matrix (as compared to the ungapped pairs). Information content computed from these data (Table 2) is therefore very similar to the ungapped estimates (Table 1).

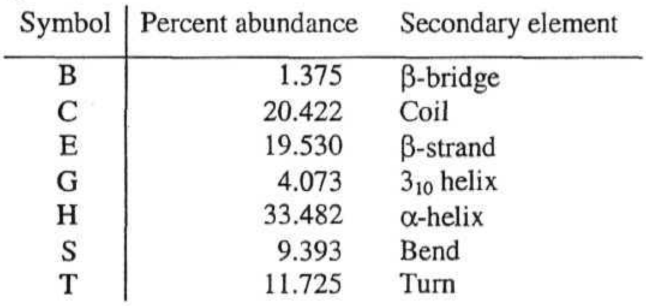
The random simulated database generated using a Monte Carlo technique in which the proportions of the amino acid residues are conserved as to Swiss prot database composition. The information content of these simulated sequences (Table 1) shows that there is no significant change in the information with increase in the complexity of the model. This behavior is as expected, and supports the view that the amino acid residues in the natural sequences show Markov dependency contrary to random simulated sequences.

It is not possible to extend this process to very high order, as the size of the database is a limiting factor. Alternative methods for estimating the entropy of natural protein sequences are therefore required.

4.3.2 ENTROPY OF SECONDARY STRUCTURE SEQUENCE

The obtained secondary structure database was simplified by DSSP program to a sequence of helixes (H), extended sheets (E), beta sheets (S), and other four symbols (T, C, B and G) for unstructured loops. The database used to study the information content of secondary structure elements consists of 19,752 polypeptide chains. The composition of the seven secondary elements in the database was calculated. The highest abundant secondary element was found to be α-helix (~33.4%) and least abundant element is β-bridge (~1.37%). The various secondary structure elements and their composition are given in the in the Table 3.

Table 3: **The** percent abundance of secondary elements in **the** database

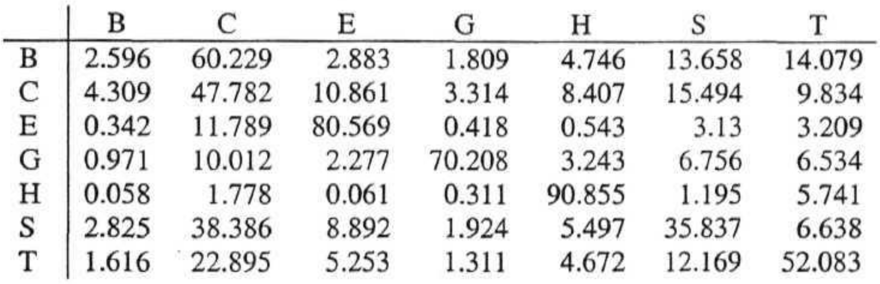


Symbol	Percent abundance	Secondary element
B	1.375	β-bridge
C	20.422	Coil
E	19.530	β-strand
G	4.073	3_{10} helix
H	33.482	α-helix
S	9.393	Bend
T	11.725	Turn

4.3.3 TRANSITION OF SECONDARY STRUCTURE ELEMENTS

The abundances for various structural elements do not specify how the elements are clustered. A sequence of H is more likely to be found rather than a random distribution. The highest transition probabilities are observed in the transition of H to H (~90.8%), i.e., the persistence of transition of 'H to H' (alpha helix) in the database is very high. Similarly, the occurrence of other high abundant transition elements are E to E, G to G, T to T and C to C, which shows that the continuous persistence of these elements is very common in the observed secondary database. In contrast, the less persistent elements are the transition of B to B, i.e., that essentially means that one B is very unlikely to be followed by another (B). These characteristics are summarized in the transition matrix that has been shown in Table 4.

Table 4: Transition probability matrix of the secondary structure (in percent) elements



	B	C	E	G	H	S	T
B	2.596	60.229	2.883	1.809	4.746	13.658	14.079
C	4.309	47.782	10.861	3.314	8.407	15.494	9.834
E	0.342	11.789	80.569	0.418	0.543	3.13	3.209
G	0.971	10.012	2.277	70.208	3.243	6.756	6.534
H	0.058	1.778	0.061	0.311	90.855	1.195	5.741
S	2.825	38.386	8.892	1.924	5.497	35.837	6.638
T	1.616	22.895	5.253	1.311	4.672	12.169	52.083

The maximum entropy calculated for 7 states (Case I) is $\log_2 7 = 2.807$ bits. The Case II entropy of the secondary structure is 2.413 bits (Table 5). However, the first order (conditioned) entropy is 1.375 bits. This shows that neighboring secondary structure elements are strongly correlated with a difference of 1.038 bits with respect to case II.

Table 5: Entropy (bit per symbol) calculations for secondary sequences.

Entropy orders	Case I entropy	Case II entropy	First order entropy
Bits per symbol	2.807	2.413	1.375

We note the presence of a relatively stronger correlation for the secondary structures. This is not unexpected, as the original amino acid residues have now been categorized into a more meaningful set of secondary structures. However, long range correlations may still be present and need to be investigated in more details.

CONCLUSION

The present work shows that Markov dependencies are clearly evident in the protein primary sequences of various databases studied. The higher order Markov approximations and their entropy calculations showed that short-range interactions are evident between the neighboring amino acids in the protein primary sequences. Insertion of a gap between the adjacent amino acids showed relatively no significant decrease in the entropy with ungapped estimates.

In addition, based on Markov studies, a significant correlation is observed between adjacent secondary structure elements. We believe considerable work still needs to be done, but an abstract description of the protein structure may be obtained from the primary sequence and the secondary structure elements. Once it is achieved, the process can be further refined to include evolutionary processes.

REFERENCES

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

Bonnie, J. S. and Dewey, G. T. (1995) Multifractal and decoded walks: Applications to protein sequence correlations, Phys. Rev. E. 52, 6588-6592.

Brenner, S.E., Koehl, P.and Levitt, M. (2000) The ASTRAL compendium for protein structure and sequence analysis, Nucleic Acids res. 28, 254-256.

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. 1974b Prediction of protein conformation, Biochemistry. 13 222-244

Cover, T.M. and Thomas, J.A (1991) Elements of Information Theory. Wiley.

Cramer, H. Mathematical Methods of Statistics (Princeton University Press, Princeton, 1946).

Feller, W. An Introduction to Probability Theory and its Applications (John Wiley and Sons, Inc., New York, 1950).

Hooft, R. W. W, Sander C. and Vriend, G. (1996) The PDBFINDER database: A summary of PDB, DSSP and HSSP information with added value. CABIOS, 12, 525-529.

James, S. B., and Dewey, T. G. (1997) Multifractal analysis of solvent accessibilities in proteins, Phys Rev E. 52, 880-887

Kabsch, W. and Sander, C. (1983) A dictionary of protein secondary structure, Biopolymers. 22, 2577-2637

Lehninger, A. L. and Nelson, D. L. (1993) Principles of Biochemistry, Second edition, Worth Publishers, New York.

Meeta, R., Mitra, C. K. Cserzo, M. and Simon I. (1995) Proteins as special subsets of polypeptides, J. Bioscience. 20 579-590.

Mitra, C. K. and Sen, A. (2001) Towards A Dynamical Systems Approach to Protein sequence structure, Calcutta Statistical Association Bulletin. 51,203-204.

Schrödinger, E. Statistical Thermodynamics (Cambridge University Press, Cambridge, 1948).

Shannon, C. E. (1948); Bell System Tech. J. 27, 379, 623 C. E. Shannon and W. Weaver, The Mathematical Theory of Communication (University of Illinois Press, Urbana, 1949).

CONCLUSIONS

SUMMARY

The brief summary of the work is given below

Natural sequences are different from random sequences

Based on the positional distribution behavior of amino acids, we conclude that the natural sequences show different pattern of distribution compared with the amino acids of random sequences.

- The frequency distributions of amino acids analysed from N-terminal region of Swiss-Prot database showed a steady state distribution except at the beginning regions. However, the distribution studies from C-terminal region showed a steady state distribution allover the positions considered.
- As expected the positional distribution of amino acids of simulated sequences showed a uniform distribution.
- The positional distribution of amino acids of natural sequences was compared with the amino acids of random (simulated) sequences.
- The correlation of amino acid distributions of Swiss-Prot and random sequences show that the natural sequences follow different pattern of distribution compared with random sequences.
- The steady state behavior of amino acids observed during the distribution studies was further confirmed by implementing ANOVA statistical method. The observed difference of distribution of amino acids at the beginning regions compared with steady state behavior at remaining regions of Swiss-Prot

sequences was statistically confirmed by using ANOVA test. Similar analysis was carried out on the non-redundant databases (sets with less than 40% and 95% sequence similarity) and the studies revealed that the amino acids follow a steady state distribution with similar pattern of distribution behavior all over the positions considered.

- Various other statistical methods (log-odds) were implemented and substantiated the steady state behavior of amino acids in the Swiss-Prot sequences.

Fractal studies on protein sequences

- To understand the behavior of amino acids, we implemented fractals studies on primary and secondary sequences of proteins.

- The fractal dimensions of amino acids of the primary sequence were calculated strictly based on the positional distributions. Based on the dimensional values, the amino acids were further grouped in to three categories. The group I (with less than *5%* abundance in the database) and group II (with more than *5%* abundance) amino acids showed a negative correlation between the dimensional values and their abundances. The amino acids in group III showed an exceptional behavior from the other group of amino acids.

- The secondary structure elements also showed a negative correlation of fractal dimension with respect to abundance. The secondary structure elements are grouped into two categories based on their behavior.

- The result of statistical test (Baswa and Prakasa Rao) for pair dependencies among the adjacent amino acids in the protein sequences gave a hope to apply Markov dependencies for checking the existence of short-range interactions if any in the protein sequences.

Information content of protein sequences

- Based on Shannon's information studies, we observed that there are clear evidences of Markov dependencies in the protein primary sequences of various databases studied.
- The higher order Markov approximations and their entropy calculations showed that short-range interactions are evident between the neighboring amino acids in the protein primary sequences.
- Insertion of a gap between the adjacent amino acids showed relatively no significant decrease in the entropy compared with ungapped estimates.
- Based on Markov studies, a strong correlation was observed between adjacent secondary structure elements. We believe considerable work still needs to be done, but an abstract description of the protein structure may be obtained from the primary sequence and the secondary structure elements. The process needs further refinement to understand the folding problem.

List of Publications and conferences attended

1. A MARKOV MODEL FOR PROTEIN SEQUENCES.
Y. Surya Pa van[1], C. K. Mitra[1]* and S. M. Bendre[2]

BGRS 2004, Kluwer Academic publishers
Accepted for publication on May 18th 2005.

2. FRACTAL STUDIES ON THE PROTEIN SECONDARY STRUCTURE ELEMENTS.
Y Surya Pavan and C K Mitra*

Indian Journal of Biophysics and Biochemistry
Accepted for publication on May 25th 2005.

3. COMPUTATIONAL STUDIES OF COVALENTLY MODIFIED GLUCOSE OXIDASE
B.S.B Salomi, **Y.Surya Pavan** and C K Mitra

Submitted to Bioinformatics, India.

4. Presented a poster at International conference on **"Bioinformatics Genome Regulation and Structure"** 12-17 July 2002 at Novosibirsk, Russia.