CHAPTER ONE: INTRODUCTION

1.1 Intrinsically Disordered Proteins

Intrinsically disordered proteins (IDPs) are defined as the regions of protein or of an entire protein that does not have a definite rigid three-dimensional (3D) structure (Romero et al. 2001). The IDPs participate in different cellular functions such as transcription, post translational modification and signaling. Intrinsically disordered regions can generally be identified in experiments such as in X-ray diffraction, where missing electron density indicates disorder; in protease digestion, where sites of hypersensitivity identify the disordered region; and in circular dichroism, where low intensity in the 210nm-240nm region and lack of aromatic-related peaks in the 280nm-310nm region indicates lack of structure (disordered regions). According to Dunker et al. (2005), these regions are flexible and often tend to combine with partners to attain structural stability, some of the examples being combining with proteins, ribonucleic acid (RNA), deoxyribonucleic acid (DNA), metals and other binding partners.

Native protein structure tends to change to new confirmations with changes in the environment. These changed protein confirmations will have properties intermediate between the native and the unfolded conformations (Anfinsen et al. 1961). Biologically active proteins in general appear in three different forms or conformations: the solid-like form that is the stable ordered form; the molten globule form, where the protein is partially folded, globule like, but with a dynamic core (Ptitsyn, 1995); and the random coil form where the protein is completely irregular, (Romero et al. 2004). According to the protein trinity hypothesis proposed by Dunker et al. (2001), a protein can be present in any of the above mentioned

forms, and their form of appearance depends on the sequence and the nature of the surroundings.

The central dogma of molecular biology states that DNA codes for mRNA, which in turn codes for the amino acid sequence, which then fold into 3D structure in order to carry out function. This protein folding step is sometimes called the second half of the genetic code. The ability of the proteins to change to stable folded conformations is believed to be the central property of their amino acid sequences. Hence, it is considered that amino acid sequence of a protein contains all the information required for a protein to fold into a native structure, which is biologically active (Anfinsen et al. 1961). The amino acid sequence of a protein determines its 3D structure, which in turn governs the biological activity. This suggests that amino acid sequence codes for the 3D structure of a protein.

Since the amino acid sequence codes for the 3D structure, Romero et al. (1997) conjectured that the amino acid sequence could also code for the absence of structure or ID. To test this hypothesis, attempts were made to develop computer programs that use amino acid sequence as the input and then predict structure or disorder as the output. One set of predictors using the amino acid sequences as source of information is Prediction of Naturally Disordered Regions (PONDR) ®.

Although these intrinsically disordered regions or proteins have a functional diversity, they do have some common features such as a low hydropathy, a high net charge, and a high flexibility. Because of these characteristics, amino acid compositions have been used in developing sequence-based predictors of structure and disorder. Analysis shows that disordered proteins are enriched in glycine, serine, aspartic acid and glutamic acid (G, S, D & E) whereas ordered regions contain higher amounts of cystine, phenylalanine, tryptophan,

and tyrosine (C, F, W & Y). Some functions performed by natively disordered proteins are: cleavage by activation, molecular recognition and molecular assembly. IDPs can participate in different molecular functions while in the disordered state or when they are transforming from disorder to order and vice versa.

It is interesting to see how these IDPs are involved in different cellular functions, mainly in signaling, and it is evident that most of the higher organisms, especially those with nucleus, have a higher number of disordered proteins. The reason for this could be that these proteins have to take part in signaling, so they require a flexible structure rather than a rigid structure. Due to the role these IDPs play in these functions they have become a subject of great importance during the last decade, especially in the field of structural biology.

1.2 Evolution

It is believed that the earth was formed 4.5 billion years ago, and that early life dates back to 3.8 billion years ago. There have been many theories about what could have happened during this period. Some say while the earth was cooling, the atmosphere responsible for the formation of life was developed (Alexander Oparin, 1938). Later this was studied by Stanley Miller (1953) who conducted experiments, in which he synthesized organic compounds using the inorganic compounds thought to be present in the early atmosphere. Starting with the postulated early atmosphere of methane gas (CH_4), hydrogen (H_2), water (H_2O) and ammonia (NH_3), Stanley's experiments produced several of the amino acids such as glycine, alanine and aspartic acid.

Evolution is the process through which organisms develop different traits and pass both the novel and old traits. The study of this evolutionary process can highlight not only the biological diversity present in nature, but can also yield information about common,

highly conserved feature. For example, almost all organisms use basically the same four nucleotides and the same twenty amino acids yet they have different physiological appearances and biological functions. Since it is believed that all biological life forms have a common ancestor, the study of life from evolutionary perspective will help us understand and analyze the differences and similarities between different biological life forms.

In the modern cell, genetic information is carried by DNA, and the control of various chemical reactions by catalysis is carried out by protein structures called enzymes. In some viruses, RNA rather than DNA carries genetic information. Also, RNA is crucial for protein synthesis, providing both messenger RNA, which carries the genetic information, and the ribosomal RNA, which is responsible for linking the amino acids together via a catalytic mechanism. Other examples of RNA catalysts, called ribozymes, have been discovered. The findings that RNA can carry the genetic information and can carry out chemical catalysis led to the proposal of RNA World Hypothesis, which views early life as depending entirely on RNA. The evolution from the RNA World to the modern world must have involved a number of steps that are not entirely clear, but the widespread use of RNA in present day life are thought by many to be simply remnants from the RNA World.

Based on the hypothesis of the RNA World and the other theories of how life started, it seems likely that not all amino acids were present at the time the life started. In 2000 Edward Trifonov developed an analysis of many factors to propose the temporal order of occurrences of the amino acids in nature. The temporal order from oldest to newest amino acids given by his analysis is the following: G/A, V/D, P, S, E/L, T, R, N, K, Q, I, C, H, F, M, Y and W. In this analysis the amino acids that were synthesized in the Miller's experiment came early, and the amino acids related to the codon capture came later in

evolution. Codon capture in evolution is defined as the process in which a codon may be missing along with its anticodon from the coding sequence, and this codon may become used later in evolution along with its same anticodon or a different anticodon. In this way the codon is captured (Osawas et al. 1989).

In his analysis Trifonov analyzed 40 different criteria to estimate the order of appearance of the amino acids through early evolution. These 40 criteria were separated into single-factor based and multiple-factor based criteria, and then an average ranking was used to suggest get to the new temporal order of occurrences of the amino acids in evolution. To date the above order of occurrences of amino acids serves as the model of order for the age of the different known amino acids.

1.3 Last Universal Ancestor (LUA)

There have been a number of theories on what the common ancestor of all the living beings, on earth would have been like. What was it made of? What functions did it carry out? What genetic material did it contain? Was it a simple organism or a complex one? Was it thermophilic or mesophilic in nature? The Last Universal Ancestor (LUA), which is also called the Last Universal Common Ancestor, is defined as the form of life that diverged into the three primary lineages, namely the archea, the bacteria and the eukaryotes.

Based on the features of the present day organisms, it is proposed that some of the features in common to all the living beings today came from this last common ancestor. The obvious features in common include the ability to code for amino acids in eukaryotes and bacteria, the ability to assemble amino acids into proteins by translation of mRNA, the ability to use glucose as the energy source via glycolysis, and the ability to capture the energy in the

form of ATP. According to some researchers, there is no single organism that can be called a "universal" ancestor.

However, it should be kept in mind that the Last Universal Ancestor is not the very first living organism on earth. Another hypothesis is that the archea was first to be split from the Last Universal Ancestor. Since the archea can be stable to extreme conditions of the temperature and the salinity, it is considered that the LUA might have lived in the deep oceans. Another observation concludes that the archea and eukaryote share a common ancestor from the bacteria. This conclusion is based on the result that the nucleus of the eukaryote is as old as the archeal line is. Some studies suggest that the Last Universal Ancestor was a complex life form with a body of the mesophilic eukaryotes before they got accustomed to the oxygen environment. Some of the analysis done to study LUA suggests that the LUA could have already had very complex genetic machinery and metabolic activities.

Contrary to the previous observations that the LUA could have already used genetic heredity, some studies say it could be the physical nature of the LUA may have been inherited rather than the genetic machinery. This LUA might have had many cells which evolved into a single biological unit. However, it is mostly known that the LUA has diverged into the three primary lineages where the archea and the bacteria have a similar set of metabolic functions and genes that are not a part of the eukaryotes. How this simple organism evolved into much complex eukaryotes is not known, but it is mostly believed that the LUA could have originated in a thermophilic environment. Since various proposals regarding the Last Universal Ancestor, lack sufficient evidence and since the various

proposals are mostly mutually inconsistent, it can be said that the nature of this hypothetical organism is still debatable and an open question.

1.4 Estimation Of Ancestral Proteins

Many of the studies related to the LUA have been done to better understand its nature. Some studies use the gene sequence analysis, but it is mostly understood that protein sequences can be a more informative source to reveal the nature of the LUA. Knowledge of amino acid compositions of the early proteins can help us understand the process of protein evolution, including the known amino acids that were introduced into the genetic code.

When we examine the amino acid compositions of the Last Universal Ancestor, we find that some of the amino acids are over-represented and some are under-represented relative to their modern counterparts. Based on this, it can be suggested that those amino acids which are over-represented could have been present much earlier than the ones that are under-represented due to additions in the course of evolution. Most importantly these kinds of studies have the potential to help us understand the nature of the LUA. The amino acid compositions or the proteomic feature will not only help us understand the nature of the LUA but also various other characters of the early proteins such as optimal PH and temperature which would have been aided in the growth of the life.

Analyzing the proteomes of organisms from the three primary lineages, i.e., archea, bacteria and eukaryotes, will help us find the nature of the proteins in the LUA. Since these ancient proteins would contain some traces of the early biological processes, genetic information and biological functions, it is important to investigate each of these lines.

In the present era of informational sciences it has become very easy to store and retrieve data, and due to the increase in available biological data it has become possible to

analyze these kinds of proteomic approaches. Based on the current amino acid compositions many researchers have come up with different approaches to build and infer the ancient amino acid compositions. Some of the methods are based on the G+C content, Maximum Parsimony, Maximum Likelihood and Expectation Maximization. In these methods the conserved residues, that is the residues that do not change over the course of evolution, would help us find the nature of the primitive proteins and their evolution in particular.

CHAPTER TWO: LITERATURE REVIEW

2.1 Disordered Proteins

Just as it is important to understand protein evolution, protein structure becomes an interesting subject of study, since the structure of the early proteins can help us understand how they might have worked in early biological processes. In this Literature Review section I present the various theories and studies in the lines of disordered nature of proteins and protein evolution. Proteins that don't have a fixed 3D structure have been called intrinsically disordered proteins (IDPs) and such regions have been called intrinsically disordered regions (IDRs). Based on some features such as charge/hydropathy, amino acid residues and net charge, a predictor was built (Romero et al. 1997). This study was supported by various other researchers working with biophysical methods to induce structures. Weinreb et al. (1996) studied the protein "non-A beta component of Alzheimer's disease amyloid plaque" (NAC). Based on their study the team confirmed that this protein was natively unfolded and further stated that many other biologically significant proteins or their domains are totally or partly unstructured

Based on the above study Baskakov et al. (1999) used the naturally occurring solute, Trimethylamine N-oxide (TMAO) to force two unstructured proteins to fold. This group found that TMAO can be used to help unstructured regions of a protein to change to a unique structure. Campbell et al. (1999) used the term intrinsic structure disorder for the protein, Human C-Fos which is biologically active and also intrinsically disordered at the C-terminus. Romero et al. (2001) conducted a study on the sequence complexity of the disordered proteins. In this study they concluded that the low complexity proteins have higher content of the residues arginine, lysine, glutamic acid, proline and serine and lower content of cystine,

tryptophan, tyrosine, isoluecine and valine. It was also noticed that during the crystallization processes used for X-ray diffraction studies the missing data in the peaks were mostly disordered regions rather than structured regions that moved as rigid body domains. The studies mentioned above were conducted in the early stages of understanding the disordered nature of proteins. Based on these studies and their results, many experiments have been carried out emphasizing the importance of disorder proteins and their roles in biological processes that will be discussed in the following paragraphs.

One of the remarkable findings by Uversky et al. (2000) on natively unfolded (unstructured or disordered) proteins suggests that these IDPs fall in the regions with large net charge and low hydrophobicity when the net charge is plotted on the vertical axis and mean hydrophobicity on the horizontal axis. The study was done on two different sets of proteins: 1.folded proteins that are not in complexes with any other partners; and 2.natively unfolded proteins that lack compact structures. Hence, low hydrophobicity and high net charge is considered to be a characteristic of many natively unfolded proteins. Romero et al. (1998) in their study on the proteins from Swiss-Prot databases concluded that as many as 15,000 or more proteins are either completely disordered or have long disordered regions. Williams et al. (2001) came up with a study on different databases to check the nature of the intrinsically disorder proteins. The team concluded that disordered proteins were depleted in the residues tryptophan, cystine, phenylalanine, isoleucine, tyrosine, valine, leucine and asparagines, and that the disordered proteins are enriched in the residues alanine, arginine, glycine, glutamine, serine, proline, glutamic acid and lysine, hinting that the former set of residues are order promoting and the latter set are disorder promoting.

Dunker et al. (2001) concluded that disordered proteins or disordered regions can participate in number of molecular functions such as molecular recognition and activation by proteolytic cleavage. They stated that the disordered nature of a protein is an encoded property, and that there are larger amounts of intrinsically disordered-proteins in nature than in the proteins in the Protein Data Bank (PDB).

Flaugh and Lumb (2001) proposed and tested their hypothesis that molecular crowding can change the structure of the protein from disordered to ordered. They concluded that molecular crowding does not always stimulate ordered structure in intrinsically disordered- proteins. Bienkiewicz et al. (2002) worked on a protein called protein p27 (Kip 1) and concluded that some proteins derive the kinetic advantage from the intrinsic structural disorder to hinder themselves from forming a complexes. Dunker et al. (1998) found that IDPs or ID regions upon binding to a partner played an important role in molecular recognition. This was later supported by Dyson and Wright (2002) who conducted a study on the proteins in eukaryotes, concluding that these proteins coupled with their partners can be helpful or involve in some of the important biological processes of molecular recognition. This was also suggested by Romero et al. (2004).

Oldfield et al. (2005) proposed a method to predict the molecular recognition elements (MoREs). The importance of the study on MoREs, also called as molecular recognition feature (MoRFs), was established by several other researchers such as Mohan et al. (2006) and Vacic et al. (2007). Uversky et al. (2005) suggested that disordered proteins that become involved in complex formation or in various interactions have a valid identification that helps the other partners to identify them. The identifications are present mainly in the disordered regions of the protein. This study considered several signaling

proteins such as BRCA1, p53, RNase E and many others that have carefully studied disordered region.

Brown et al. (2002) proposed an idea that disordered-proteins evolve more rapidly than the ordered proteins due to the anecdotal accounts of higher rates of evolution in disordered proteins compared to the rates in ordered proteins. Based on these findings, Radivojac et al. (2002) came up with an idea to build a scoring matrix based on the disordered proteins, so that the multiple sequence alignments can use the features of the disordered proteins in calculating the similarity and the phylogenetic relationships.

Iakoucheva (2002) used the PONDR® VL-XT predictor the proteins involved in the signaling and cancer. It was found that intrinsically disordered proteins play a key role in cancer. Verkivkher (2003) studied the protein p27 where he found a transition of the disorder/order and it was agreed that the transition takes place depending on the functional requirements to form intermolecular interface. Vucetic (2003) proposed the flavor-function hypothesis for disordered-protein function. The hypothesis is that different types of disordered proteins perform different functions.

Uversky (2003) in his study concluded that proteins take one of the three routes: the folding route, the misfolding route, or the non-folding route. The results and deductions out of the above studies helped build predictors to predict the disordered regions of proteins that are presented in the next paragraph.

Iakoucheva et al. (2004) developed DISorder-enhanced PHOSphorylation (DISPHOS) predictor to predict protein phosphorylation sites, based on amino acid frequencies and disorder information. The team concluded that protein phosphorylation is predominantly in intrinsically disordered protein regions. Bracken et al. (2004) proposed that

the combination of computation, prediction and experimentation such as spectrometry can help improve the accuracy of the prediction of disorder regions or proteins. Obradovic et al. (2005) proposed and developed the VSL1 predictor for predicting the disordered-proteins using two different levels based on the sequence length and meta-predictor for assigning the weights for both shorter and longer length sequence predictors. Peng et al. (2005) combined various predictors for disordered regions to obtain accuracy higher than the existing predictors; predictors such as VL3H and VL3P were combined to produce PONDR® VL3.

In the past few years there has been a great amount of research in the field of intrinsically disordered-proteins, especially in different spheres such as their roles in the molecular functions of biology and biological activities. Uversky et al. (2006) performed a bioinformatics analysis of the proteomes of the high risk and low risk human papilomaviruses (HPV) key proteins, which are considered to be intrinsically disordered. The finding of Skarabana et al. (2006) revealed that intrinsically disordered proteins are likely to play a key role in neurodegenerative process. This study suggested the use of monoclonal antibodies as structural probes to study the IDPs in the development of neurodegenerative process. Uversky et al. (2007) conducted a study on different protein databases and disorder predictors. This study emphasized on the use of the knowledge gained from the disordered nature of proteins to obtain structural and functional information related to a protein, protein families and protein domains. This method could help in finding different proteins with same structural proteins and phylogenies. The above studies are based on bioinformatics approach, whereas experimental emphasis on intrinsic disorder in proteins can be the focus of experiments such as those carried out by Baskakov et al. (1999) and by Daughdrill et al. (1997). Other experimental evidence of intrinsic disorder in proteins is provided by the

missing coordinates in x-ray crystal structure methods. These missing coordinates or residues are sometimes found to be involved in complex formation in different experiments.

Currently there has been a great increase in the study of intrinsically disordered regions and intrinsically disordered proteins especially in relation to their involvement in disease. Uversky et al. (2008) used the interrelations of intrinsically disordered proteins, cell signaling, and human diseases, to develop a list of proteins involved in the human diseases, for example, p53, tau-protein, α-synuclein and BRCA1 and introduced the concept of D^2, that is "disorder in disorders."

To sum up, the importance of IDPs has increased over the time. The research in this field has been able to answer some of the key issues related to the structure and function of proteins. The roles and functions of IDPs in various diseases are being investigated by many researchers currently.

2.2 Protein Evolution

Life's origin on earth is one area of study that has been most debatable so far. Many theories about the origin of life on the earth exist. None of these studies have proven accurate or established the exact environment or the manner in which the process took place. In this section of the study, various theories and the research that have been carried out in the lines of protein evolution from early life are discussed. First it is believed, for life to begin, it needed congenial atmosphere in which to grow. This atmosphere comprised of chemical components, precisely the inorganic compound such as H_2, H_2O, N_2, CO_2, methane and ammonia. Given energy in the form of electrical sparks, which were intended to mimic lighting storms, several amino acids were produced (Miller, 1953). Not only Miller but also many other researchers, for example Hutchinson (1964), Mayr (1964), Meinschien (1965),

and Cairns-Smith (1966), have all agreed that inorganic carbon compounds were produced naturally in the environment before life could evolve on earth.

Next was the period described now as the RNA World Hypothesis, which argues that early genetic material could have been RNA-like molecules which were self reproducing and able to do most of the function of DNA in the present world. Orgel (1994) conducted an extensive study on the role of RNA in the primitive world and concluded that the catalytic role of RNA was indeed the early step, though its emergence and birth are not known. Orgel (2000) suggested that the genetic material in the primitive world could have been the simple nucleic acid-like molecules; RNA analogous to RNA called (L)-a-threofuranosyl oligonucleotides or TNAs. Though they have threose rather than ribose, they generally function like RNA.

Nelson at al., (2000) stated that the first or early genetic material on earth could have been polypeptide nucleic acid (PNA), a precursor of RNA which emerged from the polymerization of the ***N***-(2-aminoethyl) glycine (AEG); since the PNA can be easily polymerized, it could have been the first genetic material. Most of the reviews concluded that much before the emergence of DNA and protein enzymes, there was a much simpler form of life based on RNA. Vlassov (2005) in his review concluded that the prebiotic conditions were linked with freezing rather than warm and wet conditions, which was a key factor in the RNA World Hypothesis. His study responded to the confusion related to the RNA World Hypothesis. Even though the RNA World Hypothesis is well accepted by many there are still a number of key issues to be resolved. Vlassov focused on the missing links between the RNA World and the miniribozymes and their role in molecular evolution. The hypothesis says that catalytic RNA could cleave and ligate the phosphodiester bonds. In addition, recent

observations and studies on ribosomal proteins point out that RNA was indeed the first genetic material. This RNA in the primitive world might have been subjected to rapid chemical degradation. Hence it is possible that the present day RNA's chemical material might differ from that of the RNA in the primitive world.

Many conflicting theories about the existence of the LUA, or the last known ancestor, have been proposed, but the idea of LUA from which the three primary lineages have emerged is still accepted widely. Woese (1998) stated that the LUA is not a discrete entity. Rather it is a group of cells that lived and emerged as a biological entity. Harris et al. (2003) carried out a molecular analysis of the conserved regions of ribosomal RNA in the modern day genome, and these conserved regions tracked back to the three sets of species that converged into a single unit, the LUA. Giulio (2007) in his work argued that the Last Universal Common Ancestor (LUCA) existed and that it was an anaerobic organism not an aerobic organism. Wais (2005) believed that archaebacteria, which has the features of both eubacteria and eukaryotes, could lead to the identification of the LUA. Giulio (2003) postulated that the LUCA was a hyperthermophile and the direct ancestor of the bacteria and archea, whereas the ancestor of the eukaryote was a mesophile. Giulio (2001) proposed that the LUCA could have been a thermophile or a mesophile.

In order to understand the history and the nature of the LUA, it is important to have proteomic knowledge, since this can provide some insight into the very nature of molecular evolution. Trifonov (2000) came up with the temporal chronological order of the occurrences of the amino acids in the genetic code. According to him, not all amino acids existed at the same time but the ones that were produced in the Miller's experiment came first and those related to codon capture came late. In a study related to the proteome of the LUA, Brooks et

al. (2001) implemented an algorithm to calculate the amino acid compositions of the LUA. From this study they concluded that the amino acid residues could be more or less frequent in primitive proteins than in modern ones, since amino acids were slowly added into the genetic code in the course of evolution. The under-represented amino acids could have been late into the genetic code and the over represented could have been early amino acids. Brooks et al. (2002) used a set of 65 proteins Clusters of Orthologous Groups of proteins (COGs) which were estimated to be present in the LUA. They calculated the compositions of the amino acids in the LUA. Based on the compositions of the amino acids with respect to the modern proteome they concluded that some amino acids came late into existence. Later, Brooks et al. (2004) developed an improved algorithm using Maximum Likelihood and Expectation-Maximization approach to calculate the amino acid composition of same set of 65 COGs believed to be present in the LUA.

2.3 Hypothesis

It is very important to know how primitive proteins have changed in the course of evolution, and how each amino acid could influence the structure of the proteins in general. It is generally believed that early life had the RNA molecules that were self-reproducing and were able to take part in the catalysis of various biological activities. Slowly the amino acids were added into the genetic code during the course of evolution. That means most of the proteins in the LUA could have had the high amounts of the ancient or the early amino acids and low amounts of the modern or the late amino acids. This was primarily proved by Brooks et al. in 2002. Based on the differences in the frequencies of the amino acids of the LUA and the modern day proteins, it was concluded that some amino acids were over-represented and were under-represented. It is a well known paradigm that the structure of the proteins is

dependent on the amino acid sequence and that structure is a prerequisite for the function of the proteins. Hence, it would be interesting to see how these primitive proteins might have looked with regard to their structures and functions. Some of the information related to the structure of the proteins is acquired by the amino acid compositions.

According to Romero et al. (1997), the structure of proteins – especially the disordered – proteins is related to the compositions of the amino acid residues glycine, aspartic acid, proline, glutamic acid and serine. According to Trifonov's proposal chronological order of the amino acids, the above-mentioned amino acids entered into the genetic code first. In other words, they are early or ancient amino acids and related to the disordered nature of IDPs. The ancient or the primitive proteins likely had disordered or unstable structures, and probably would not have been able to carry catalytic activity unless they associated with other partners such as RNA molecules or metal cofactors. Therefore, we hypothesize that the primitive proteins in the LUA had less stable structures. These primitive proteins likely had unstable structures, because the entry of order-promoting amino acids such as cystine, phenylalanine, tryptophan and tyrosine was late into the genetic code, much after the disorder-promoting residue emerged. Thus, the content of these modern amino acids would be less in the primitive proteins when compared to the modern ones. We also propose that the enzymes along with their counterpart did not have a rigid 3D structure, and that the non-enzymes had unstable structure to a larger extent than the enzymes. Both the sets of proteins were likely disordered in the LUA, with non-enzymes being the more disordered than enzymes.

Dividing the set of input data into enzymes and non-enzymes would help to reveal if the enzymes, which are supposed to be structured in order to act as catalysts, were ordered in

the primitive world or not. That is, if they had higher contents of order-promoting residues cystine, phenylalanine, tryptophan and tyrosine. Accordingly, their counterparts would have lesser amounts of these residues and higher contents of disorder-promoting residues, that is, glycine, aspartic acid, proline, glutamic acid and serine.

Studies on proteins from various organisms from the three kingdoms of life that is from archea, bacteria and eukaryotes show that proteins are intrinsically disordered. These intrinsically disordered proteins mainly function in signaling, molecular recognition, transcription and post-translational modifications (Dunker et al. 2000; Ward et al. 2004). Since in evolution some traits are passed on to the next generations, and since these particular proteins from these three kingdoms of life are intrinsically disordered which diverged from the LUA, it can be assumed that ancestral proteins in the LUA were likely to have been disordered.

Further, it would be interesting to see how these amino acid compositions have changed over time relative to the current proteins set. To test this hypothesis the proteins in the Protein Data Bank (PDB) were checked for the absences of a particular residue and their probabilities of not having that residue in a given sequence could reveal the importance of that particular residue in the structure of the proteins. The probabilities of not having the residue for the 65 COGs divided into enzymes and non-enzymes compared with the present structured protein sets can help us offer some insight into proteomic knowledge of the primitive world.

CHAPTER THREE: METHODS AND MATERIALS

3.1 Estimation of amino acid composition of the Last Universal Ancestor

The algorithm used by Brooks et al. (2004), to estimate the amino acid composition of the LUA was used to test the hypothesis proposed in section 2.3. The algorithm uses the Maximum Likelihood (ML) and Expectation-Maximization (EM) steps to infer the amino acid compositions of the ancestor sequences. There are two different approaches to reconstruct the ancestral sequences: Maximum Parsimony (MP) and Maximum Likelihood (ML). Both of these methods have certain disadvantages that do not allow them to yield accurate results of the ancestral sequences and their compositions. The MP method works on a statistical approach to infer the phylogenetic trees by cutting short the steps in evolution. The ML method works on the principle of maximizing the probabilities of the observed data and inferring the sequence compositions. The MP method is less accurate, whereas the ML method needs a prior assumption of the compositions of the amino acid sequences that will be reconstructed which is believed to be same as the extant descendants (Yang et al. 1995). In the present method of inferring the ancestral amino acid compositions Brooks et al. (2004) have used a combination of ML and EM methods. Both of these methods are explained in the following paragraphs.

In a traditional ML method, sets of sequences are used that are aligned along with the phylogenetic tree related to these sequences. The probability distribution of every amino acid over the root of the phylogenetic tree is calculated, and then this probability distribution is maximized by the likelihood of the observed residues at the external node, that is the modern day sequences (Yang et al. 1995). As stated earlier, the prior assumptions of the amino acid compositions are made at the root of the phylogenetic tree. In the ML approach, the

empirically calculated substitution matrix by Jones et al. (1992) is used so as to generate the probability distribution of the amino acid compositions of the ancestral sequences. The following is an example of the ML approach used in this algorithm for reconstructing the ancestral sequences.

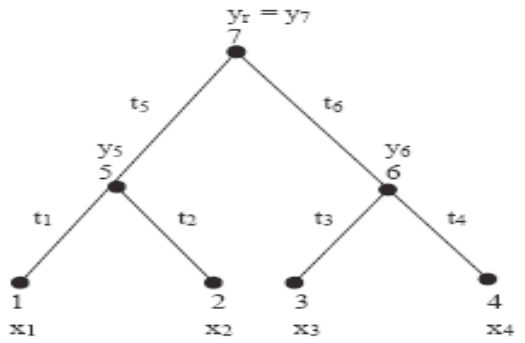


Fig 3.1: Phylogenetic tree used to illustrate ML method (Brooks et al. 2004).

Here in Fig 3.1 the example tree has four external nodes, 1-4, four aligned extant protein sequences, internal nodes 5-6, and a root node 7. The branch lengths are denoted by t_1-t_6, the inferred ancestral sequences are represented by y_5-y_7, and the modern day sequences are represented by x_1-x_4. The data at a particular site, say j in this case, is represented as $x_i^{(j)}$. That is the amino acid in the i^t extant sequence at site j and similarly the $y_i^{(j)}$ is the amino acid at the i^{th} in the inferred ancestral sequence (Brooks et al. 2004). So the probability of x at site j is given by the equation:

$$f(x^{(j)};\theta) = \sum_{y_r^{(j)}} p(y_r^{(j)})\, f(x^{(j)} \mid y_r^{(j)};\theta)$$

Eq 3.1: Probability of an amino acid at site j (Brooks et al. 2004).

Here θ signifies to the branch lengths of the phylogenetic tree, p(i), which is the preassumed probability of the amino acid at the root, and the function is the conditional

probability of the root node at the site *j* (Brooks et al. 2004). Upon further simplification for the values of *y*, that is ancestral sequences, we get the equation of the log-liklihood score:

$$l = \sum_{j} \log \left[\sum_{y_r^{(j)}} p(y_r^{(j)}) f(x^{(j)} \mid y_r^{(j)}; \theta) \right]$$

Eq 3.2: Log likelihood score for an amino acid (Brooks et al. 2004).

EM is used to calculate the estimates of parameters in the probabilistic models just like the ones explained above. This method contains two different steps. That is the expectation step, or the E step and the maximization step, or the M step (Brooks et al. 2004). In the E step for the given observed values, the expectation likelihood values are calculated for the missing data from the likelihood score, which are then used by the M step to maximize the likelihood, and then the maximization values, are used to start the next E step (Brooks et al. 2004). This process is repeated until the values congregate. In the algorithm, the EM method re-estimates the values and validates the log-likelihood of the observed data. Using the methods ML and EM, the aim is to calculate the amino acid frequencies of the ancestral sequences and the prior probability p*(i)* shown in equation 3.1. EM can be used in three steps to get the amino acid frequencies of ancestral sequences (Brooks et al. 2004).

The First step is to make an assumption of the amino acid compositions p*(i)* of the ancient sequences. The second step is the E step where the probability distribution is calculated for the residues at the root of the phylogenetic tree. Then the number of occurrences of every amino acid in the ancestral sequence is calculated. In the final step, M step, the ML of the amino acid compositions are calculated dividing the number of occurrences of amino acids from the E step by the total number of amino acids in the

ancestral sequences (Brooks et al. 2004). This value is again used for the next E step (Brooks et al. 2004). These steps are repeated until the values of the two consecutive iterations congregate. That is the difference in the log-likelihoods of two consecutive iterations should be less than 0.0001 (Brooks et al. 2004). The final estimate is the frequency of the amino acids in the ancestral sequences and the estimate probability p*(i)* (Brooks et al. 2004).

3.2 Phylogenetic Tree

To obtain the phylogenetic tree for the data set used in these experiments Brooks et al used the program 'protdist' from the software Phylogeny Inference Package (PHYLIP), which was developed by Felsenstein in the year 1993. A phylogenetic tree, for example the one in Fig.3.1, is a tree that links one species to its descendant or to its most common ancestor. The nodes represent the species; the edges represent the time. One main limitation of the phylogenetic tree is that the final outcome of the analysis of the evolutionary tree is the phylogeny of the character but not that of the species. There are different kinds of trees, for example rooted, unrooted, cladogram, phylogram, dendogram and chronogram. In the present experiment the phylogenetic tree generated by PHYLIP, which is publically available, has many different executable programs such as the PROTRAPS, PROTDIST, PROMLK, SEQBOOT, PROML and CONSENSE only for the phylogenetic analysis of the protein sequences and has different programs for DNA sequences, restriction sites, gene frequencies and other biological aspects. Here Brooks et al. used PROTDIST, which computes the distance for the protein sequences based on the maximum likelihood and the Dayhoff PAM matrix, JTT matrix, and the PBM model. These distances can also be corrected for gamma-invariant-sites-distributed rates of change in different sites. The distances depend on the hidden markov models or the rates of evolution. These distances

obtained can be used in the distance matrix program. The PHYLIP program is used to obtain the tree and to create the distance matrix for the set of ancestral sequences. This distance matrix and an un-rooted tree using the PHYLIP neighbor-joining program are used to estimate the amino acid frequency in the ancestral sequences.

3.3 ClustalW

ClustalW is multiple sequence alignment software, which is freely available and used to align nucleotide and protein sequences. Multiple sequences can be aligned globally, that is the whole length of the protein sequences, or locally, that is a small region of the sequences, to infer the relation between the organism and their evolutionary background. Thompson et al. (1994) developed ClustalW. This program was an extension to the existing ClustalV, in which new features where included. These new features added improvements to the multiple alignment of protein sequences. ClustalW uses dynamic program, merges the small sub-alignments obtained by the method of progressive alignments. To optimize the alignment parameters, the Gonnet series of residue comparison matrix is used by default for the protein sequences. ClustalW is more advanced than its other previous Clustal series of programs in many ways for example, generating different output file formats. File formats, such as PHYLIP, GDE and NEXUS to name a few, can be used. Apart from this, the ClustalW software is built with some complicated gap penalty values such as the sequence weighting calculated directly from the guide tree; initial gap penalties such as the gap opening penalty (GOP), which is the penalty for opening a gap of any length; the gap extension penalty (GEP), which is the penalty for extending the gap; and the position specific gap penalties. In addition, there is flexibility in the weight matrices such as BLOSUM and PAM. Here Brooks et al. used the Gonnet series for proteins. The important parameters that the ClustalW needs

for multiple alignment of sequences are the distance matrix/pairwise alignment, the guide tree and progressive alignment. Firstly, all the sequences are aligned pairwise or separately to obtain the distance matrix to get to the divergent sequences of the pairs. Secondly, using the distance matrix, the guide tree is calculated, and finally the progressive alignment is obtained using the guide tree.

3.4 Protein Data of interest

The idea of Brooks et al. was to look into the proteomes of the LUA to check how the amino acid frequencies have changed in the course of evolution. Here we are not looking into the nature of the LUA, but the nature of the proteins in the LUA. The LUA is believed to be the single-celled ancestor from whom the three primary lineages, the eubacteria, archea and eukaryotes, have diverged. On the basis of different studies and analysis, the LUA is believed to have 325 proteins from these three primary lineages. But only 65 proteins related to these three primary lineages were used for the experiment as explained below. The organisms from these three primary lineages include two archaea, *Archaeoglobus fulgidus* and *Methanobacterium thermoautotrophicum*; one eukaryote, *Saccharomyces cerevisiae*; and five eubacteria, *Aquifex aeolicus*, *Thermotoga maritima*, *Synechocystis* PCC6803, *Escherichia coli*, and *Bacillus subtilis* (Brooks et al. 2002). The Clusters of Orthologous Groups (COG) database was used to select the orthologous proteins (that is, these are homologs that are the result of direct speciation) which are common to these eight species, since it was proposed that the orthologous proteins are more likely to be present in the LUA. Around 215 out of the 315 COG families are known to be present in the eight species mentioned above. The reason to select these eight species was to increase the accuracy and validate the reconstruction of ancestral sequences. Since the phylogenetic tree of these eight

species would meet the above mentioned requirement, hence, they were selected (Brooks et al. 2002). Paralogs (that is, these are homologs that are the result duplication), which are formed before the speciation, would overlap and nullify the phylogenetic tree. So, to overcome these problems, only orthologos, which were either having one protein in one species for all the eight species in common, were selected other than in yeast, and those paralogs which emerged after the speciation, were removed from the reconstruction of ancestral proteins (Brooks et al. 2002). After this filtering, the protein set was reduced to 105 from 215. Along with this Brooks et al. removed all the proteins that were present likely due to the lateral transfer, since this would also nullify the reconstruction of sequences. This problem was overcome by removing all the proteins which were not consistent with the small subunit rRNA tree, since it is known that the species phylogenetic tree has to be consistent with the small subunit rRNA tree. Finally, 65 proteins were left for the analyses, as shown in Table 3.1. These proteins are mostly involved in translation, replication and transcription processes (Brooks et al. 2002).

Table 3.1: The list of all the 65 COG protein used for the experiment is given Brooks et al. (2002)

COG ID	NAME OF THE PROTEIN (COG)
COG0012	Predicted GTPase
COG0013	Alanyl-tRNA Synthetase
COG0016	Phenylalanyl-tRNA synthetase alpha subunit
COG0030	Dimethyladenosine transferase
COG0048	Ribosomal protein S12
COG0049	Ribosomal protein S17
COG0051	Ribosomal protein S10
COG0052	Ribosomal protein S2
COG0060	Isoleucyl-tRNA synthetase
COG0063	Predicted sugar kinase
COG0072	Phenylalanyl-tRNA synthetase beta subunit
COG0080	Ribosomal protein L11
COG0081	Ribosomal protein L1

COG0085	DNA-directed RNA polymerase beta subunit/140 kD subunit
COG0087	Ribosomal protein L3
COG0088	Ribosomal protein L4
COG0090	Ribosomal protein L23
COG0091	Ribosomal protein L2
COG0092	Ribosomal protein L22
COG0092	Ribosomal protein S3
COG0093	Ribosomal protein L14
COG0094	Ribosomal protein L5
COG0096	Ribosomal protein S8
COG0097	Ribosomal protein L6
COG0098	Ribosomal protein S5
COG0099	Ribosomal protein S13
COG0100	Ribosomal protein S11
COG0102	Ribosomal protein L13
COG0103	Ribosomal protein S9
COG0112	Glycine hydroxymethyltransferase
COG125	Thymidylate Kinase
COG0126	3-Phosphoglycerate Kinase
COG0143	Methionyl-tRNA synthetase
COG0164	Ribonuclease HII
COG0172	Seryl-tRNA synthetase
COG0177	Predicted EndoIII-related endonuclease
COG0180	Tryptophanyl-tRNA synthetase
COG0184	Ribosomal protein S15p/S13E
COG0185	Ribosomal protein S19
COG0186	Ribosomal protein S17
COG0197	Ribosomal protein L16/L10E
COG0200	Ribosomal protein L15
COG0201	Preprotein translocase subunit SecY
COG0202	DNA-directed RNA polymerase beta subunit/40 kD subunit
COG237	Dephospho-CoA kinase
COG0244	Ribosomal protein L10
COG0250	Transcription antiterminator
COG0256	Ribosomal protein L18
COG0258	5’-3’ exonuclease (including-terminal domain of Po1I)
COG0441	Threonyl-tRNA synthetase
COG0442	Prolyl-tRNA synthetase
COG0452	Phosphopantothenoylcystine synthetase/decarboxylase
COG0459	Chaperonin groEL (HSP60 family)
COG0468	RecA/RadA recombinase
COG0495	Leucyl-tRNA synthetase
COG0522	Ribosomal protein S4 and related proteins
COG0525	Valyl-tRNA synthetase
COG0532	Translation initiation factor 2 (GTPase)

COG0533	Metal-dependent proteases with possible chaperone activity
COG0541	Signal recognition particle GTPase
COG0550	Topoisomerase IA
COG0552	Signal recognition particle GTPase
COG0575	CDP-diglyceride synthetase
COG0592	DNA polymerase III beta subunit
COG1758	DNA-directed RNA polymerase subunit K/omega

3.5 Enzymes & Non-Enzymes

The 65 COGs were further divided into enzymes and non-enzymes. The idea was to test software developed by Brooks et al. on enzymes and non-enzymes separately to compare the ancestral enzymes with ancestral non-enzymes in terms of their compositions. The proteins which end with 'ase' were termed as enzymes and the rest were non-enzymes; there were 34 enzymes and 31 non-enzymes from the 65 COGs listed above. The reason for making this separation is as follows. If the RNA World Hypothesis is true, then the enzymes likely arose after the transition to the current biological world and are therefore likely to be more recently evolved proteins. Furthermore, the non-enzymes are almost all RNA-associated proteins and so are possibly proteins that may have arisen while the RNA World still existed, in which case these would be the more ancient proteins.

3.6 Application of the software on Data Set

The algorithm developed to reconstruct the ancestral sequence and the frequencies of the amino acids in the ancient proteins by Brooks et al is publically available at http://compbio.cs.princeton.edu/ancestralaa. The 65 COGs protein sequences for each species mentioned in the data set section from the above table were collected from COG database http://www.ncbi.nlm.nih.gov/COG. These sequences for each organism were collected and made into a single protein sequence. Then all the eight long sequences were aligned using

publically available ClustalW with default parameters. The gaps in the aligned sequences were removed to keep the most conserved regions of the proteins. Then the Brooks et al software was used to estimate the amino acid compositions of the LUA with the help of this aligned sequences and the tree provided by Brooks et al. [(((1 : 25,2 : 25) : 25,(3 : 25,4 : 25) : 25) : 25, ((5 : 25,6 : 25) : 25,(7 : 25,8 : 25) : 25) : 25)], it was midpoint rooted. These procedures were used on the enzymes set and non-enzymes set with the same phylogenetic tree, and their ancient amino acid compositions were obtained.

Since the hypothesis is to check the possible disordered nature of the early proteins, an amino acid substitution matrix based on disordered proteins was used as the distance matrix in the process of multiple alignments of the above-mentioned eight proteome sequences with ClustalW. This matrix was developed by Radivojac et al. (2002) to optimize the sensitivity and specificity when aligning the disordered regions. The scoring matrices such as BLOSUM, PAM and others have been built on different criteria such as evolutionary properties, minimum number of base changes per codon, amino acid compositions and structures. Radivojac et al. (2002) argue that, since these matrices have idiotypic advantages, they are widely used. But all these scoring matrices concentrate on the ordered proteins. What about intrinsically disordered proteins, which are very common? These intrinsically disordered proteins have higher numbers of insertions and deletions in them as compared to the ordered proteins, justified building a new scoring matrix based on the above stated criteria. The matrix based on ordered proteins would not work as well on a disordered proteins alignment, since there has to be a greater penalty for these proteins. The matrix developed by Radivojac et al. is shown in Table 3.2.

C	10																			
S	0	3																		
T	1	1	4																	
P	-2	0	-1	6																
A	-1	1	0	-1	3															
G	-3	0	-2	-1	0	5														
N	-1	1	0	-1	-1	0	4													
D	-3	0	-1	-2	-1	-1	1	4												
E	-4	-1	-1	-1	-1	-2	0	2	4											
Q	-3	0	0	-1	-1	-2	1	0	0	5										
H	-1	-1	0	-2	-2	-1	2	-1	-1	1	8									
R	-1	-1	-1	-2	-2	-2	0	-2	-1	1	0	5								
K	-3	-1	0	-1	-1	-2	0	-1	0	0	-1	2	4							
M	0	-2	-1	-2	-1	-4	-2	-4	-3	-1	-2	-1	-2	7						
I	0	-2	-1	-2	-1	-5	-3	-4	-3	-2	-2	-2	-2	1	4					
L	-1	-2	-2	-1	-1	-4	-3	-4	-3	-2	-2	-2	-2	2	2	4				
V	1	-2	0	-1	0	-4	-3	-4	-2	-2	-2	-2	-2	1	3	1	4			
F	-1	-2	-2	-3	-2	-4	-2	-4	-4	-2	0	-3	-3	1	1	1	0	7		
Y	0	-2	-1	-3	-2	-3	-1	-4	-3	0	2	-2	-2	-1	0	0	-1	4	8	
W	-5	-3	-5	-1	-5	-4	-3	-4	-4	-1	-2	0	-3	-1	-2	-2	-4	-1	3	13
	C	**S**	**T**	**P**	**A**	**G**	**N**	**D**	**E**	**Q**	**H**	**R**	**K**	**M**	**I**	**L**	**V**	**F**	**Y**	**W**

Table 3.2. Scoring Matrix based on Disordered-Proteins Radivojac et al (2002)

The matrix based on the disordered proteins was used to obtain the ClustalW alignment of the eight species and the gaps were removed. Then the software developed by Brooks et al. was used on the two sets, namely the enzymes and non-enzymes, and their LUA was inferred.

3.7 Frequency of 'C' 'F' 'W' 'Y'

After the results were obtained from the above analysis (see table 4.1 & 4.2), the next step was to check for the frequency of the modern amino acids, that is, Cystine, Phenylalanine, Tryptophan & Tyrosine in ordered protein sequences. From Protein Data Bank (Berman et al. 2000) the sequences for all the structured proteins were selected by downloading the file 'pdbseqres'. These sequences were checked for redundancies and all the repeating sequences were deleted that is 100% non-redundant sequences and those sequences, which had lengths greater than 50 residues, were selected for the analysis. The

possible combinations of 4 amino acids from the total 20 amino acids were generated, which produced 4845 possible combinations. The aim was to check in a particular sequence whether a specific set of 4 amino acids was absent from the sequence. This experiment was performed by implementing two different algorithms. The first algorithm would take the combinations from the combinations file and then check for their absence in the sequences contained in the 100% non-redundant 'pdbseqres' file. The numbers of sequences for each combination were recorded by the algorithm.

In the second algorithm, it would first check for the absence of a single residue and store the protein _id in an array, and then depending on the combination set it would take the common protein_id. This procedure was applied for the set to 2 amino acids combination from the 20 amino acids (190 different combinations). The set of 3 amino acids from 20 amino acids yielded 1140 possible combinations. These combinations were divided in different groups; if the combination contains a cystine or an aromatic amino acid then it is classified as group1. The set, which had 2, that is a cystine and an aromatic or any two Aromatic amino acids, is included in group2. Sets with these 3 amino acids, that is one cystine and 2 aromatic amino acids or 3 aromatic amino acids correspond to group3. The set, which contains cystine and all aromatic amino acids (i.e., CFWY) is termed group4. Finally, if there are no cystine and aromatic amino acids then it is classified as group0. So for the possible combinations of 4 we have 1826 in group0, 2240 in group1, 714 in group2, 64 in group3 and 1 in group4. Similarly for possible combinations of 3 we have 560 in group0, 480 in group1, 96 in group2 and 4 in group3. For the possible combinations of 2 the group0 had 120, 64 in group1 and 6 in group2. Then it was also checked to see the absence of single residue in entire input file to calculate the probability of not having that residue. The entire

procedure was applied on the PDB sequences, which were 100% non-redundant, without any duplicates and the sequence length 50 and above. All the short sequences less than 50 length residues were removed because it is clear that shorter chains would have high chances of not having the residue and hence this would invalid the results. There were 28048 chains after the filtering from the 'pdbseqres' data file from PDB. The two algorithms gave the same results, these validating the results for PDB sequences.

The same methods were applied to a set of another input sequences taken from PDB, but monomers only. Monomers are single chain sequences, which are not part of a complex. There were a total of 3727, 100% non-redundant sequences used for this analysis. These sequences had no hetro atoms and a length greater than 49 residues. This set was checked for the different combinations of amino acids, that is combinations of 4, 3, 2 and 1 from the 20 amino acids. The probabilities of not having the residue in the monomer set of proteins were calculated. Both of the algorithms gave the same results.

Swiss-Prot, a database developed by Swiss Institute of Bioinformatics, contains the highly annotated proteins information and has more than 250,000 sequences in their database (Bairoch and Apweiler, 1997). These sequences are publically available. The sequences were checked for redundancy and for lengths greater than 50 residues. Sequences that met these criteria were selected for the analysis. After the filtering there were 190,604 sequences left to be analyzed. The same procedure of checking for the absence of the possible combinations of amino acids of combinations 4, 3, 2 and 1 from the 20 amino acids was adapted. The probabilities of not having the single residue for a sequence in the Swiss-Prot set of proteins were calculated using the same two algorithms mentioned above.

After applying the above two methods on all the input data sets, the frequency for each group was calculated by dividing the number for sequences for that group by the total number of combinations in that particular group. Then the histograms were drawn based on the results. The probabilities of not having the residues were plotted against the age of the amino acids to see how their trend appeared over time.

3.8 Length Bias

Although the monomer set were a subset of the PDB, there was a great extent of length bias. The longest sequence in the monomer set was of 1267 residues, whereas in PDB the longest sequence was 2060 and there weren't many sequences in the length range 1267 to 2060 in the PDB file. It was essential to check for the length ranges and the number of sequences for each length range in order to remove this length bias and validate the results. For this purpose the length ranges of 50 to 200, 201 to 400 and so on were selected. For each length range the number of sequences was obtained using a simple Perl script for both the PDB and monomer set, after which a histogram for the length distribution and the number of sequences was drawn. The length bias would definitely affect the outcome of the comparison of PDB and monomer. In order to avoid this, Bootstrapping was applied.

3.9 Bootstrapping

Based on the length distribution from the above method, bootstrapping was applied for the length range 50 to 400 residues. Bootstrapping is a statistical approach used for resampling the present data to estimate the parameters such as standard variation, average, median and mode. In a sense this is a way to approximate the distribution of the data. The resampling is done by generating random data from the observed data of equal length, equal size and equal number. By randomly selecting the data set with replacement, from the original data set a

larger data set is generated. It has great advantages, such as estimating the complex identities such as standard error and confidence intervals. There are many types of bootstrapping for example, smooth, parametric, case, wild, resampling residual and statistical-pivotal. Generally it is good to run the resampling 10,000 times but 1000 times would give an idea of how the final output would look. In this case the bootstrapping is done by taking the length distribution in consideration. First the number of sequences for each length (50-400) was obtained using a simple Perl script for both monomer and PDB datasets. The average of both for each single length is calculated. This average will be the value or the number of times the sequences has to be randomly selected with replacement from the original data set, thus making the same number of sequences in both PDB and monomer datasets. This process is done once for the data generated. The probability of not finding a residue in the sequences of both the data set was calculated using Perl scripts. The input file consisted of 13818 sequences for both PDB and monomer datasets after the resampling using Bootstrapping approach. The graph for the age of the amino acids versus the probability of not finding the residue is plotted. The above method was developed into and algorithm to perform the resampling and calculating the probabilities for 1000 times using Perl scripting Language. The same method is followed for the construction of both the samples of PDB and monomer datasets. Then the average values for each amino acid is calculated, the lower bound that is the lower value of each residue from the 1000 samples was subtracted from the average and finally the upper bound that is highest value for each residue from the 1000 samples was subtracted from the average. These two values are used as the upper and lower bound for their respective residues. The age and the probability of not having a residue in the data sets

were plotted on a graph. There were totally 13,818,000 sequences in both monomer dataset and PDB after resampling.

CHAPTER FOUR: RESULTS

The results of the present experiments are detailed here in two sections. The first section contains the analysis carried out using the Brooks et.al, Software on the 65 COGs, enzymes and non-enzymes. The second section contains the results of the analysis done on the sequences from PDB, Monomer dataset and Swiss-Prot databases. The summary of the analysis and the results are discussed in the last section Summary of Findings.

4.1 LUA Comparisons

The hypothesis was to see if the primordial proteins were intrinsically disordered based on the amino acid compositions of the proteins the LUA. First, the LUA of the 65 COG proteins was estimated using Brooks and team's software. The data set selected was different from the one that Brooks and team used for their analysis, since there have been considerable changes in the sequences of the COGs database. Hence, it was possible that some of our data might not be exactly the same as results reported earlier. Indeed, there was a 2% of change on average for the amino acid compositions obtained here when compared with the compositions obtained by Brooks et al. The compositions of modern day proteins structured proteins from PDBselect 25 (25% redundant data) were collected.

Table 4.1 LUA values of the 65 COGs and amino acid compositions from PDBselect.

Amino Acids	Compositions from PDBselect	LUA of 65 COGs
G	0.0716	0.0762
A	0.0770	0.0782
D	0.0583	0.0467
V	0.0672	0.0924
P	0.0457	0.0435
S	0.0619	0.0223
E	0.0665	0.1080
T	0.0563	0.0424
L	0.0868	0.0846

R	0.0493	0.0693
N	0.0458	0.0290
I	0.0561	0.0817
Q	0.0395	0.0210
H	0.0241	0.0200
K	0.0637	0.0896
C	0.0174	0.0039
F	0.0398	0.0322
Y	0.0350	0.0266
M	0.0222	0.0240
W	0.0144	0.0081

It can be inferred from the table that the frequencies if the amino acids such as Cystine, Phenylalanine, Tryptophan and Tyrosine have increased in the structured proteins over the LUA. The reason for this increase could be the emergence of the proteins mainly enzymes to help in biosynthesis, since these set of proteins (enzymes) require rigid structure to help in catalysis.

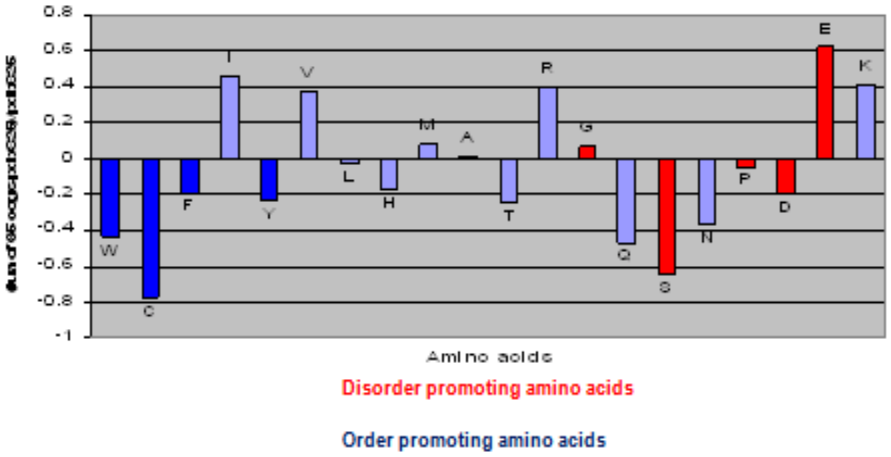


Figure 4.1 Comparison of LUA and the pdbselect25.

As shown in Table 4.1 and Figure 4.1, when the amino acid composition of the LUA proteins were compared with the compositions of the sequences in PDBselect, 25 % identity, it revealed that there was considerable increase in the compositions of modern, order-

promoting, amino acids residues and a decrease in the early, disorder-promoting, amino acids.

Then the datasets were split into enzymes and non-enzymes and the result of the analysis is detailed in the Table below.

Table 4.2 LUA values of Enzymes, Non-Enzymes, compositions of pdbselect and disport.

Amino Acids	LUA of Enzymes (Gonnet series)	LUA of Non-Enzymes (Gonnet series)	Compositions from PDBselect	Compositions from Disprot
G	0.0694	0.0904	0.0716	0.0741
A	0.0758	0.0863	0.0770	0.0810
D	0.0510	0.0378	0.0583	0.0580
V	0.0900	0.1051	0.0672	0.0541
P	0.0411	0.0468	0.0457	0.0811
S	0.0259	0.0142	0.0619	0.0865
E	0.1187	0.0782	0.0665	0.0989
T	0.0408	0.0497	0.0563	0.0556
L	0.0937	0.0628	0.0868	0.0622
R	0.0618	0.0956	0.0493	0.0482
N	0.0284	0.0293	0.0458	0.0382
I	0.0821	0.0761	0.0561	0.0324
Q	0.0226	0.0196	0.0395	0.0527
H	0.0209	0.0190	0.0241	0.0193
K	0.0765	0.1143	0.0637	0.0785
C	0.0053	0.0022	0.0174	0.0080
F	0.0357	0.0201	0.0398	0.0244
Y	0.0297	0.0202	0.0350	0.0213
M	0.0221	0.0273	0.0222	0.0187
W	0.0087	0.0049	0.0144	0.0067

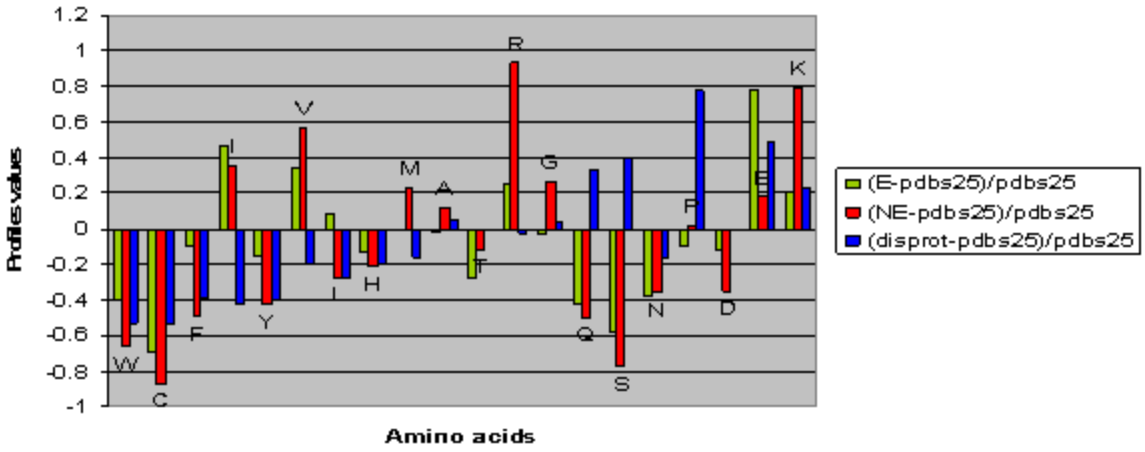


Figure 4.2 Comparison of Enzymes, Non-Enzymes and disprot over pdbselect25.

It is known that PDB has the ordered set of proteins and hence when compared with the compositions of typical disordered proteins from disport database, the early, disorder-promoting, residues are enriched in the disport and the modern, order-promoting, residues are enriched in the pdbselect25 except for the amino acids D and S. The results seem to be the same, when the LUA compositions of the non-enzymes were compared over the compositions of the PDBselect. However, this does not seem to be the case when the LUA compositions of enzymes were compared over the PDBselect compositions. The disorder-promoting residues Proline, Glycine, serine and Aspartic Acid seem to be depleted in the enzymes of the LUA.

Table 4.3 LUA values of Enzymes and Non-Enzymes based on different scoring matrices.

Amino Acids	LUA of Enzymes (Gonnet series)	LUA of Non-Enzymes (Gonnet series)	LUA of Enzymes (Disorder Matrix)	LUA of Non-Enzymes (Disorder Matrix)
G	0.0694	0.0904	0.0678	0.0897
A	0.0758	0.0863	0.0794	0.0903
D	0.0510	0.0378	0.0528	0.0374
V	0.0900	0.1051	0.0965	0.1097
P	0.0411	0.0468	0.0408	0.0447
S	0.0259	0.0142	0.0274	0.0145

E	0.1187	0.0782	0.1143	0.0774
T	0.0408	0.0497	0.0418	0.0514
L	0.0937	0.0628	0.0903	0.0612
R	0.0618	0.0956	0.0606	0.0997
N	0.0284	0.0293	0.0287	0.0269
I	0.0821	0.0761	0.0759	0.0717
Q	0.0226	0.0196	0.0241	0.0174
H	0.0209	0.0190	0.0213	0.0200
K	0.0765	0.1143	0.0772	0.1150
C	0.0053	0.0022	0.0050	0.0019
F	0.0357	0.0201	0.0340	0.0199
Y	0.0297	0.0202	0.0291	0.0199
M	0.0221	0.0273	0.0245	0.0258
W	0.0087	0.0049	0.0086	0.0053

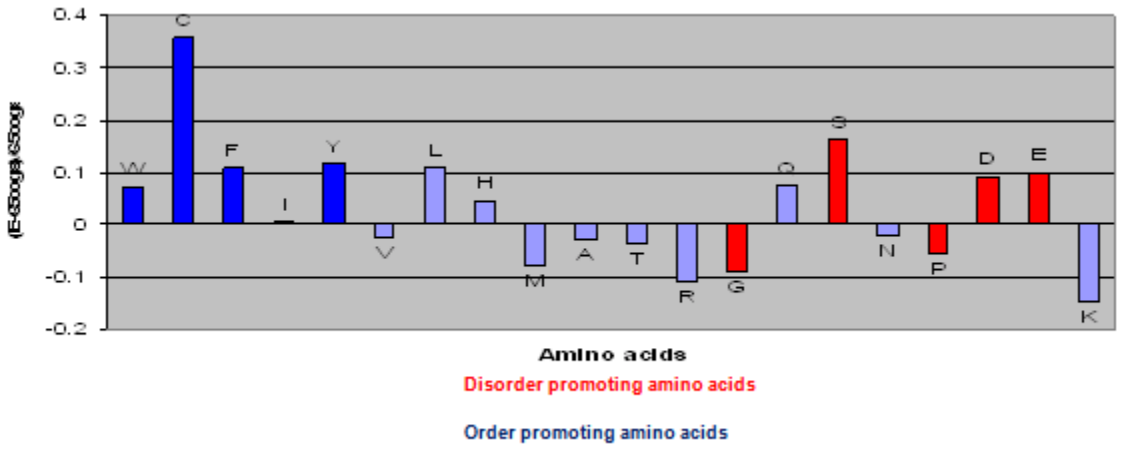


Figure 4.3 Comparison of compositions of LUA enzymes with 65 COGs.

The comparisons of the values obtained for LUA enzymes with those of 65 COGs, both calculated based on the Gonett series of scoring matrix in the phylogenetic analysis, are shown in the Figure 4.3. It can be observed that the modern order-promoting amino acid residues are depleted in the ancient enzymes and the early disorder-promoting residues are generally enriched in the 65 COGs, with the exception for S, D and E. Therefore, there is enrichment in the modern order promoting residues in the enzymes and the depletion in the disorder promoting residues.

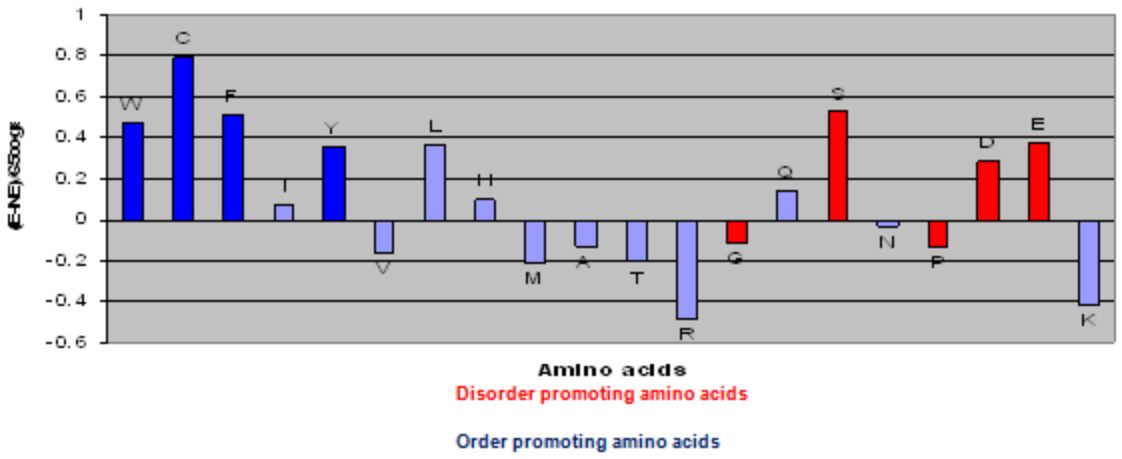


Figure 4.4 Comparison of LUA of compositions of enzymes with non-enzymes.

The comparison of the values for the LUA enzymes with non-enzymes based on the Gonett series of scoring matrix in the phylogenetic analysis is shown in Figure 4.4. It can be inferred from the graph that there is an increase in the modern order-promoting residue content in enzymes and depletion in the ancient amino acids with exception in S, D and E. The opposite is seen in the case of non-enzymes that is the enrichment in the disorder-promoting residues.

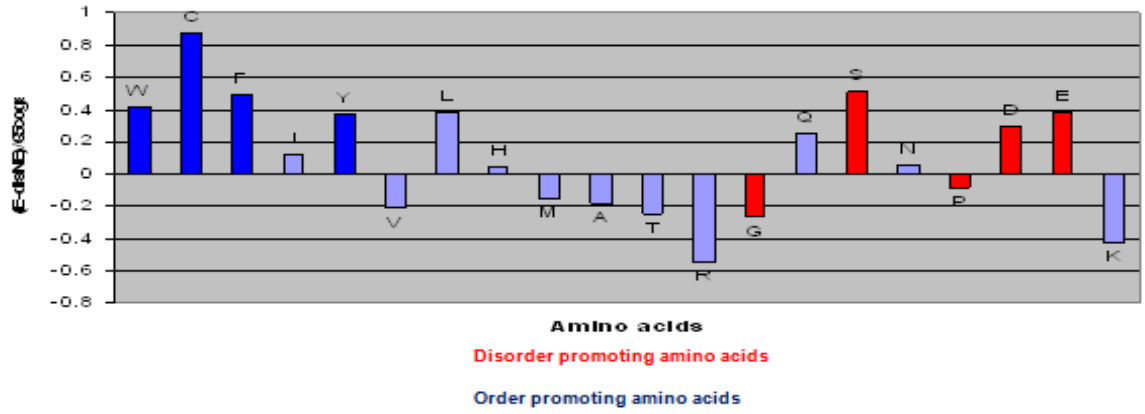


Figure 4.5 Comparison of LUA of values of enzymes with non-enzymes (disorder matrix)

The comparison of the LUA enzymes based on the Gonnet series of scoring matrix with the LUA non-enzymes based on the matrix of the disorder proteins are shown in Figure 4.5. It can be observed that there is an enrichment of the order-promoting residues in enzymes. The non-enzymes are enriched in the disorder promoting residues over the 65 COGs.

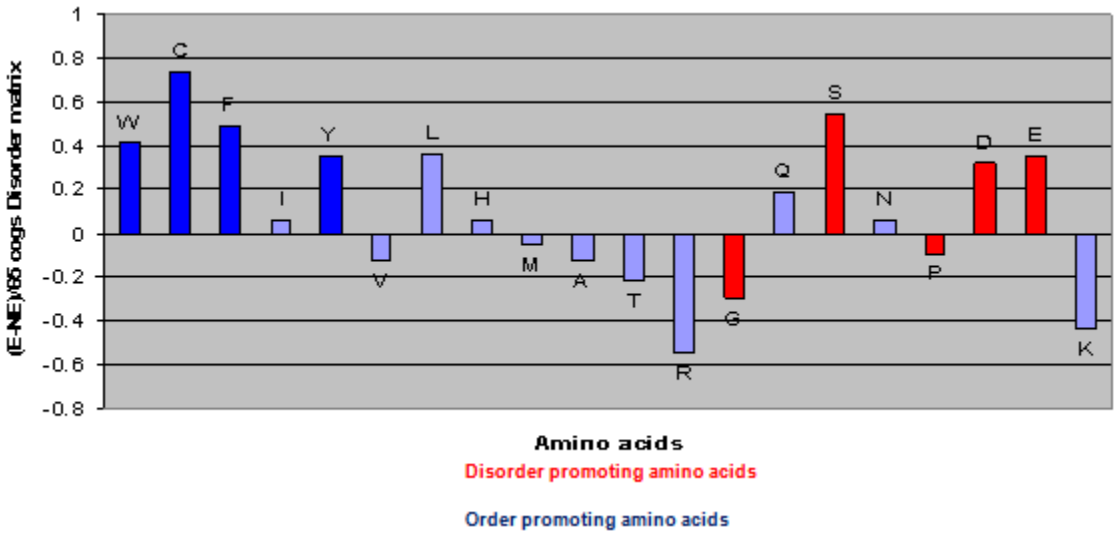


Figure 4.6 Comparison of LUA of compositions of enzymes (disorder matrix) with non-enzymes (disorder matrix)

The comparsion of the LUA enzymes with the LUA non-enzymes using the Disorder matrix in the phylogenetic analysis can be seen in Figure 4.6. It can be observed from the graph that there is an increase in the modern order-promoting residues and depletion in the ancient disorder promoting residues, with the exception for S and D residues.

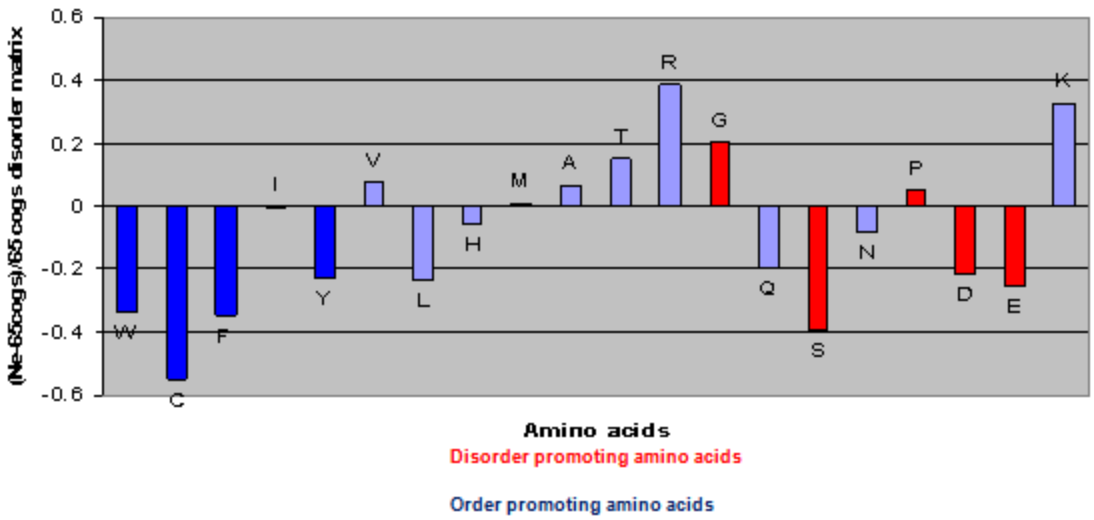


Figure 4.7 Comparison of LUA of compositions of non-enzymes (disorder matrix) with 65 COGs (disorder matrix)

The comparison of the LUA non-enzymes with entire 65 COGs calculated based on the Disorder matrix in the phylogenetic analysis can be seen Figure 4.7. It can be inferred from the graph that the non-enzymes are depleted in the modern amino order-promoting residues and are enriched in disorder-promoting residues over the 65 COGs.

4.2 Results of the analysis on protein sequences

This section deals with the results of the analysis on different data sets: 1. PDB, 2. Swiss-Prot, and 3. Monomers, as well as the enzymes and non-enzymes subgroups from the COGs dataset. When the frequency of different groups (group0, group1, group2 etc, from the methods section) were checked it was observed that there was a high frequency for the combination of CFWY to be absent than any other combinations of the amino acids in PDB dataset. It was also observed that there were highest probabilities for combinations with Histidine and Methonine to be missing in PDB. Monomer dataset also showed the same

trend. However the combination CHWY showed the highest frequency of being absent in the sequences from Swiss-Prot dataset.

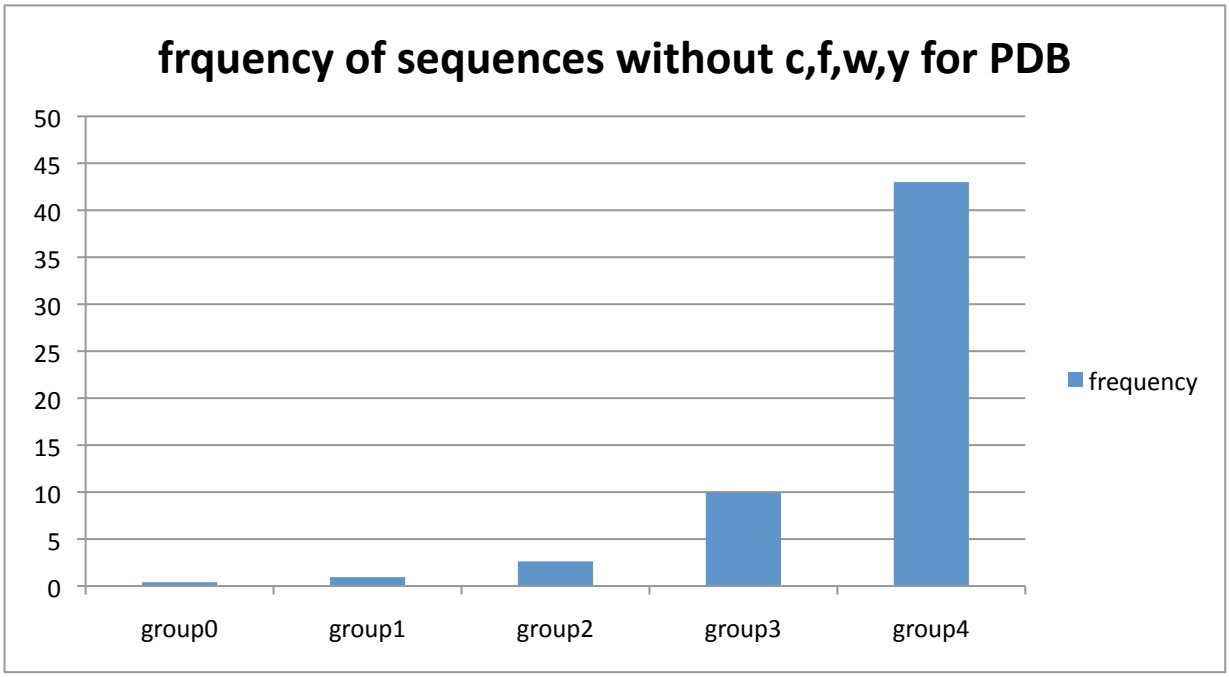


Figure 4.8 Average frequencies for each group (PDB).

The total number of sequences for each group obtained from the analysis was divided by the number of combinations in each group. This defines the per combination or the average frequency for each group. The frequency of the number of sequences per group is seen, it is observed that group4 which contains all the order promoting residue CFWY shows the highest frequency for being absent in PDB data set in Figure 4.8.

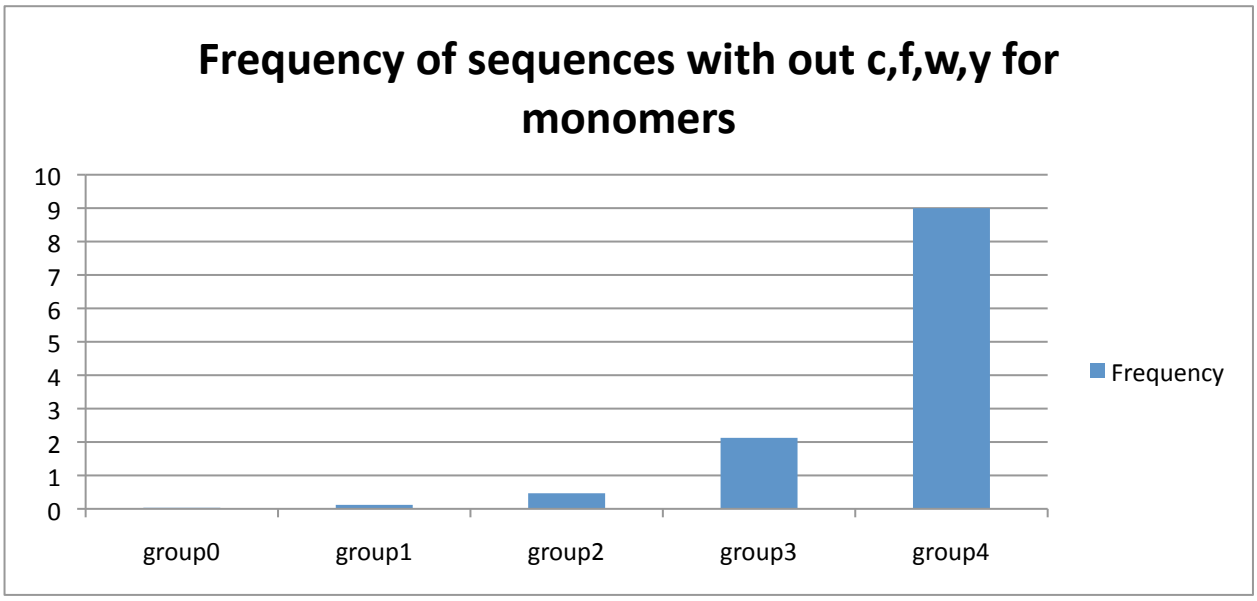


Figure 4.9 Average frequencies for each group (Monomers).

The total number of sequences for each group obtained from the analysis was divided by the number of combinations in each group. This is per combination or the average frequency for each group. The frequency of number of sequences per group of four residues in the Monomers dataset is plotted. It is inferred that the group4 shows the highest frequency of not being in the dataset as seen in Figure 4.9.

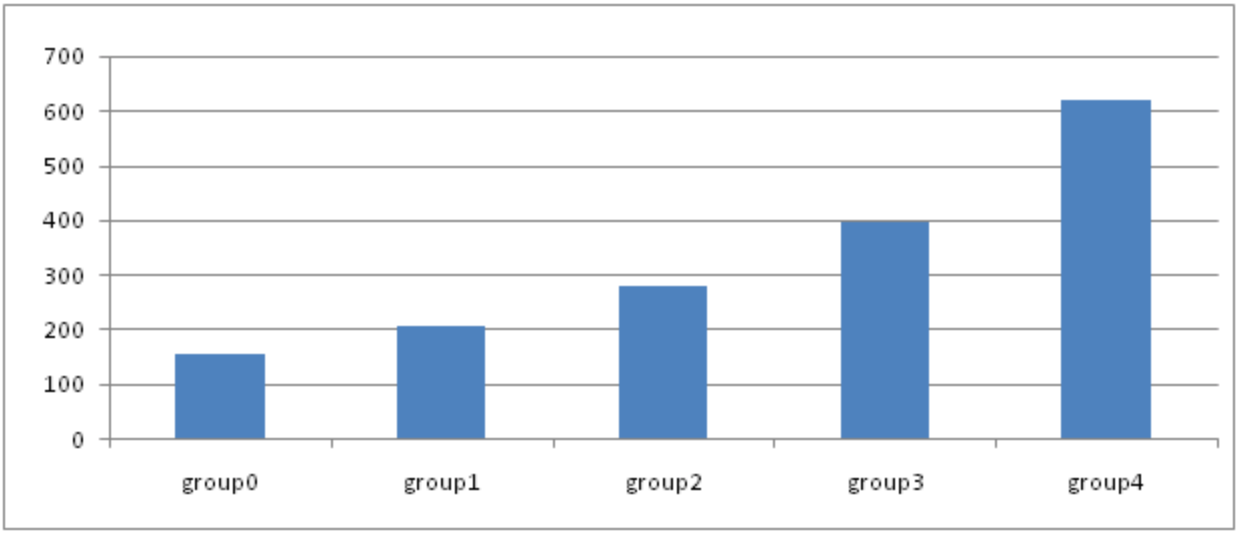


Figure 4.10 Average frequencies for each group (Swiss-Prot).

The total number of sequences for each group obtained from the analysis was divided by the number of combinations in each group. This is per combination or the average frequency for each group. The frequency of number of sequences for a group of four amino acids for the Swiss-Prot dataset is plotted and the highest frequency is observed in the order promoting group that is CFWY, group4 as seen in Figure 4.10.

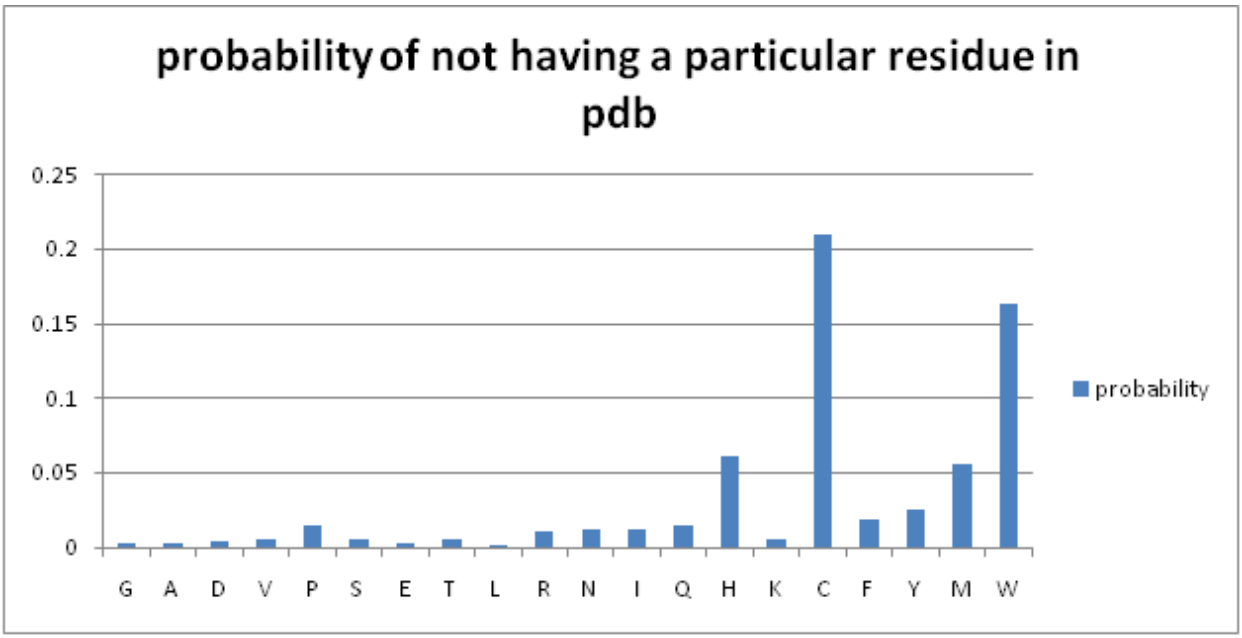


Figure 4.11 Probability of not having a single residue (PDB).

The probability of not having the particular residue in the dataset (100% non-redundant sequences) from PDB is plotted in Figure 4.11. Residues on X-axis are ordered by decreasing age, whereas the probability of not having a residue is plotted on Y-axis. Figure 4.11 clearly shows that Cystine has the highest probability for not being found in the sequence from the PDB dataset. Notice that in the Figure 4.11, the overall probability of not having modern amino acids (bars to the right) is high.

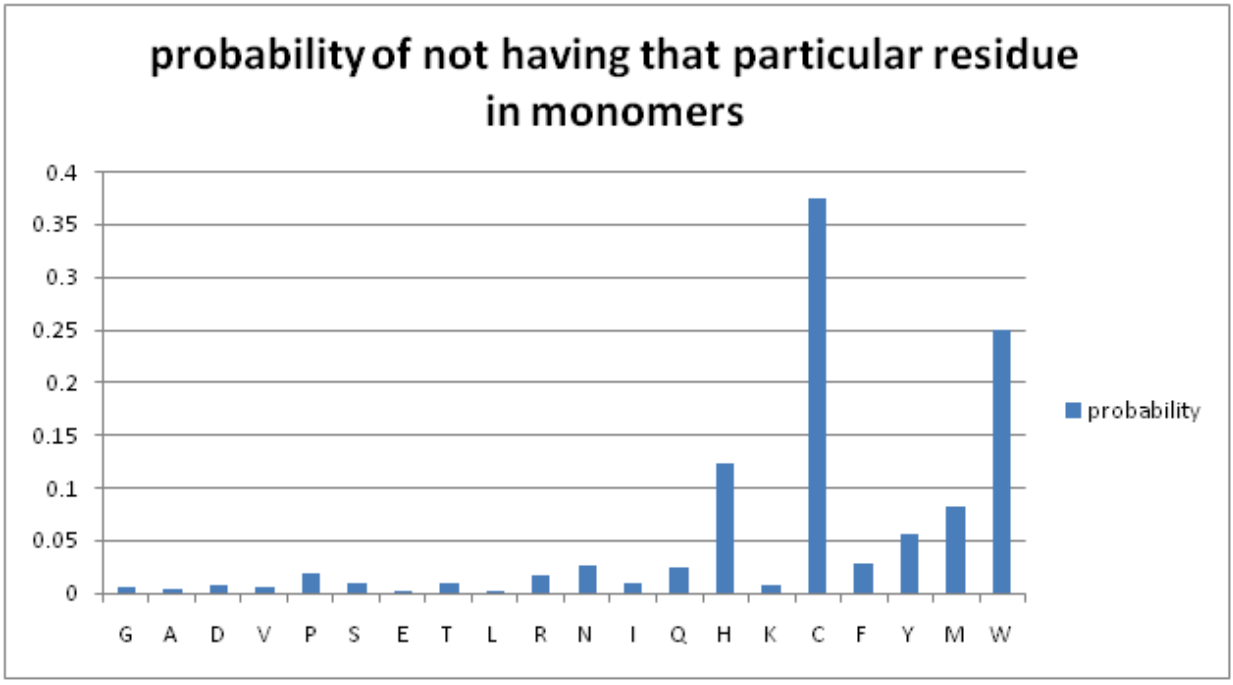


Figure 4.12 Probability of not having a single residue (Monomer).

The probability of not having the particular residue in the dataset (100% non-redundant sequences) from Monomer dataset is plotted. Residues on X-axis are ordered by decreasing age and the probability of not having a residue is plotted on Y-axis. Cystine has the highest probability for not being found in the sequence from the Monomer dataset too. The overall probability of not having modern amino acids (bars to the right in Figure 4.12) is high.

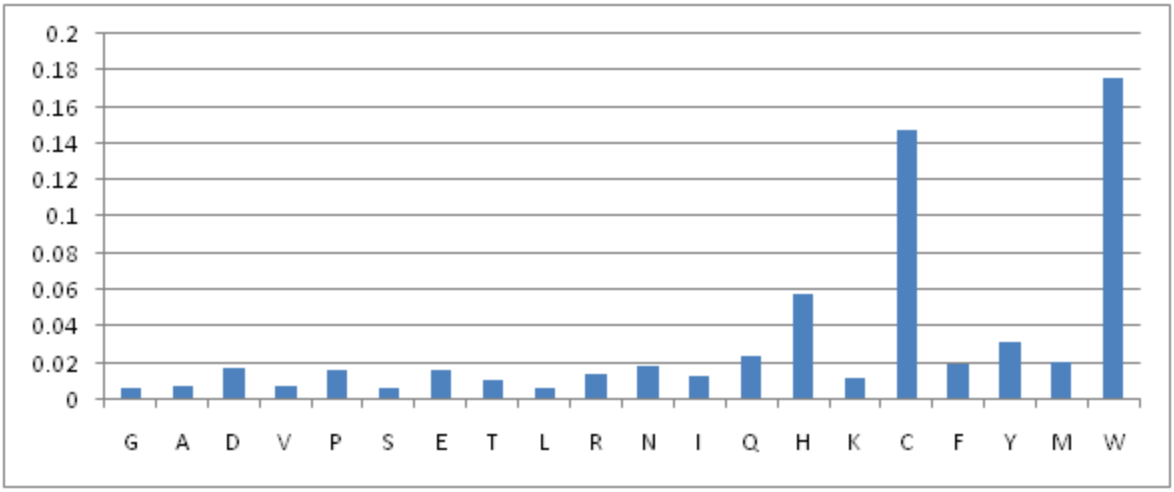


Figure 4.13 Probability of not having a single residue (Swiss-Prot).

The probability of not having the particular residue in the dataset (100% non-redundant sequences) from Swiss-Prot is plotted. Residues on X-axis are ordered by decreasing age and the probability of not having a residue is plotted on Y-axis. Figure 4.13 shows the probability of not having the residue in the Swiss-Prot dataset. Here, the residue tryptophan has the highest probability of not being found in the dataset. As seen in the Figure 4.13, the modern amino acids to the right of the graph have the highest probabilities of not being found in this dataset.

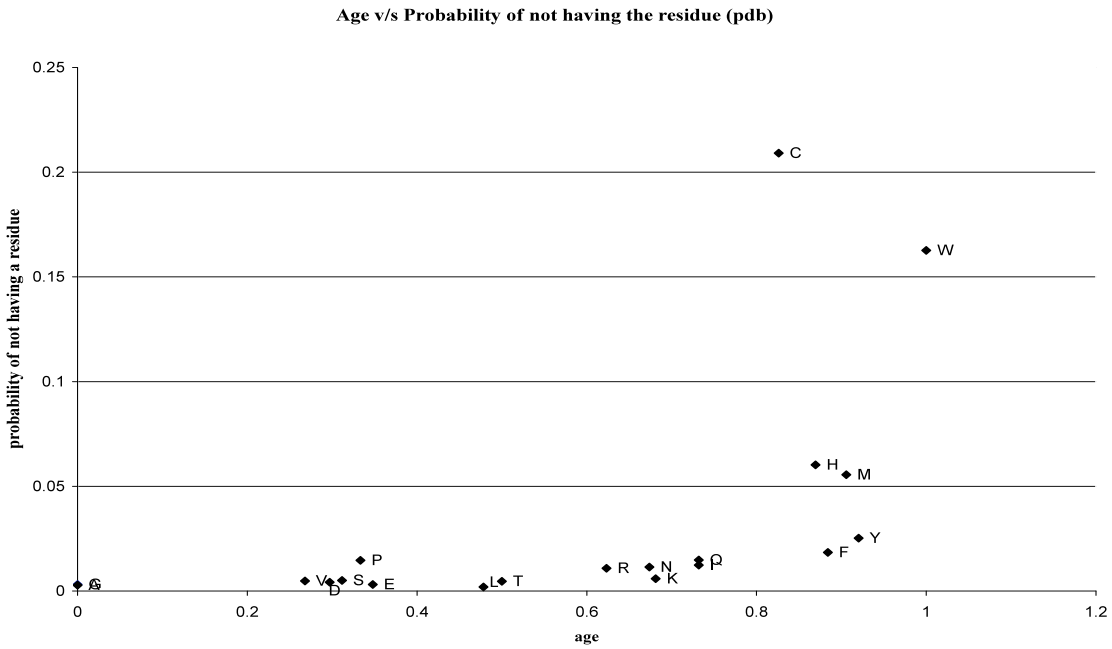


Figure 4.14 Age v/s Probability of not having a single residue (PDB).

The probabilities of not having the residue in the PDB versus the age of the amino acids are plotted. The Age is on x axis and the probability of not having the residue is plotted on y axis. It can be observed that there is a high chance of not finding the modern amino acids in the sequences seen to the right of the Figure 4.14.

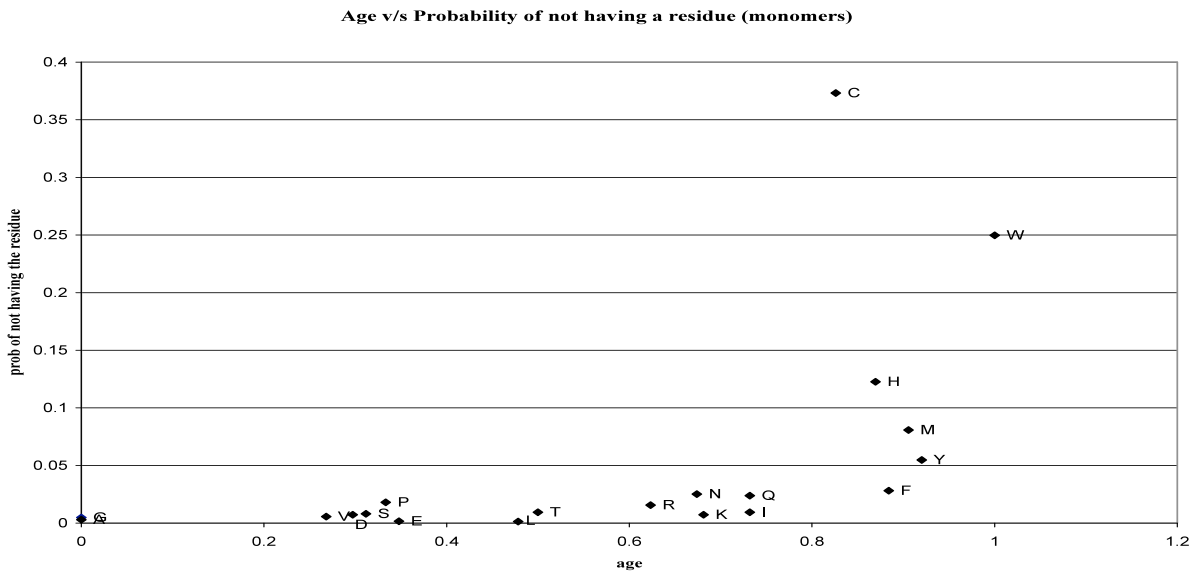


Figure 4.15 Age v/s Probability of not having a single residue (Monomers).

The probabilities of not having the residue in the monomer dataset versus the age of the amino acids are plotted. The same curve as observed in PDB dataset is found in Monomer dataset, but with higher probabilities than PDB as seen in the Figure. 4.15.

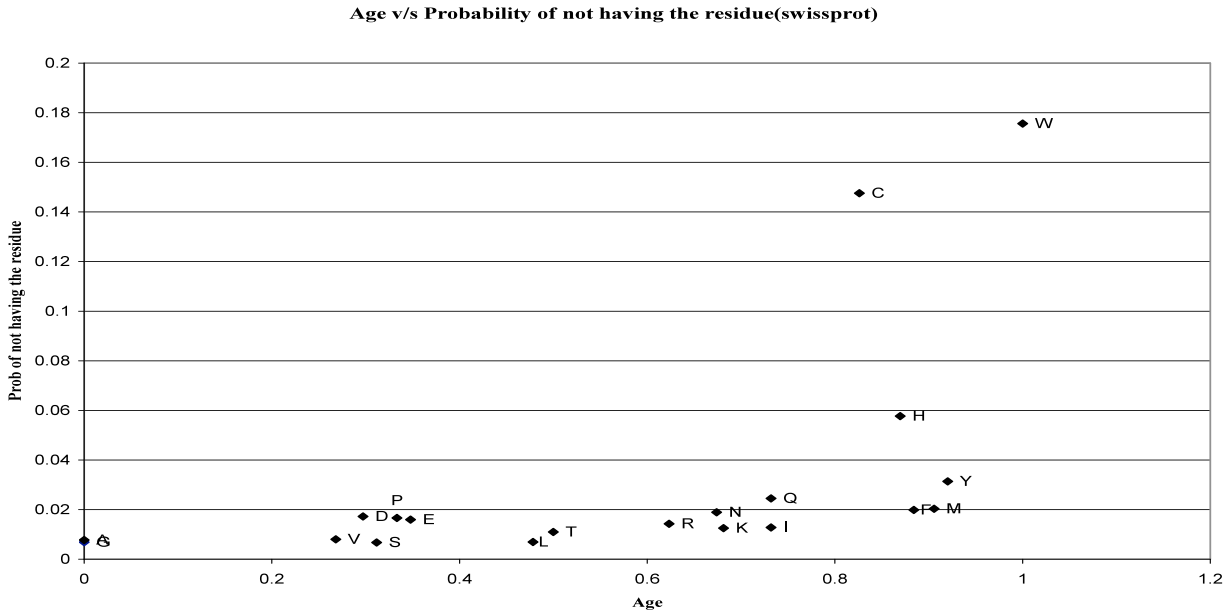
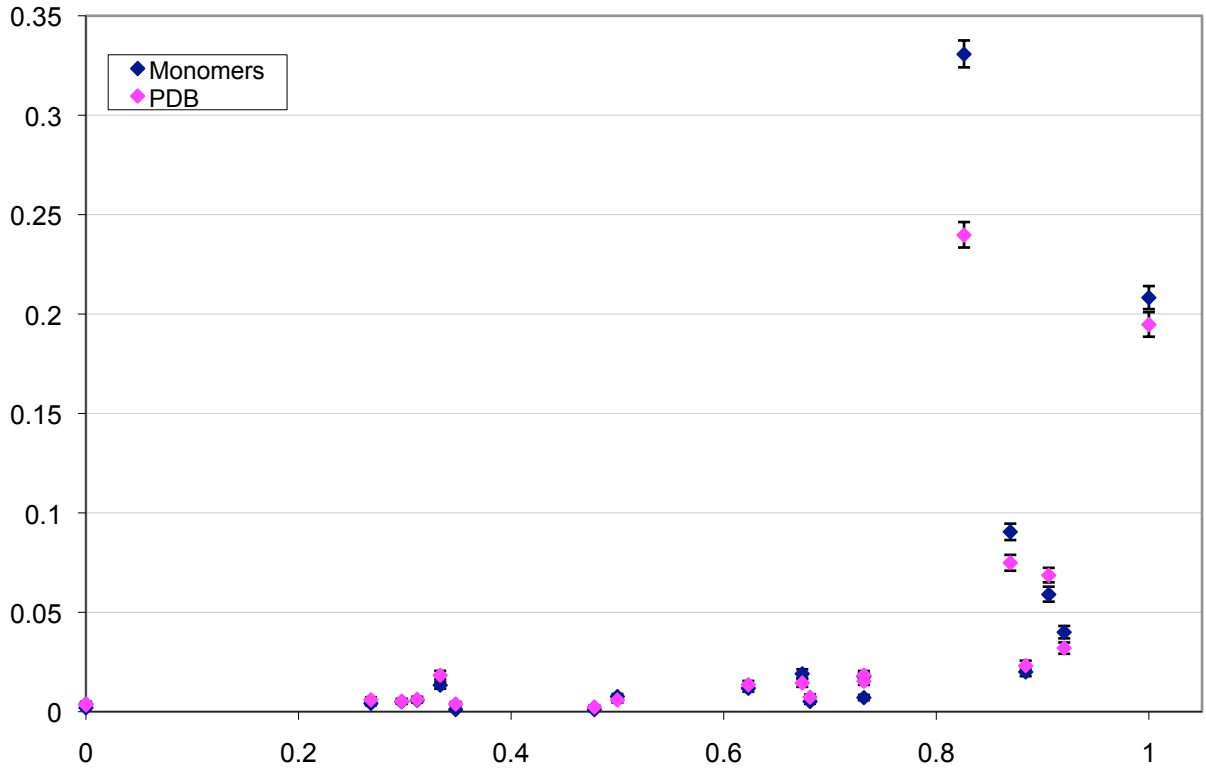


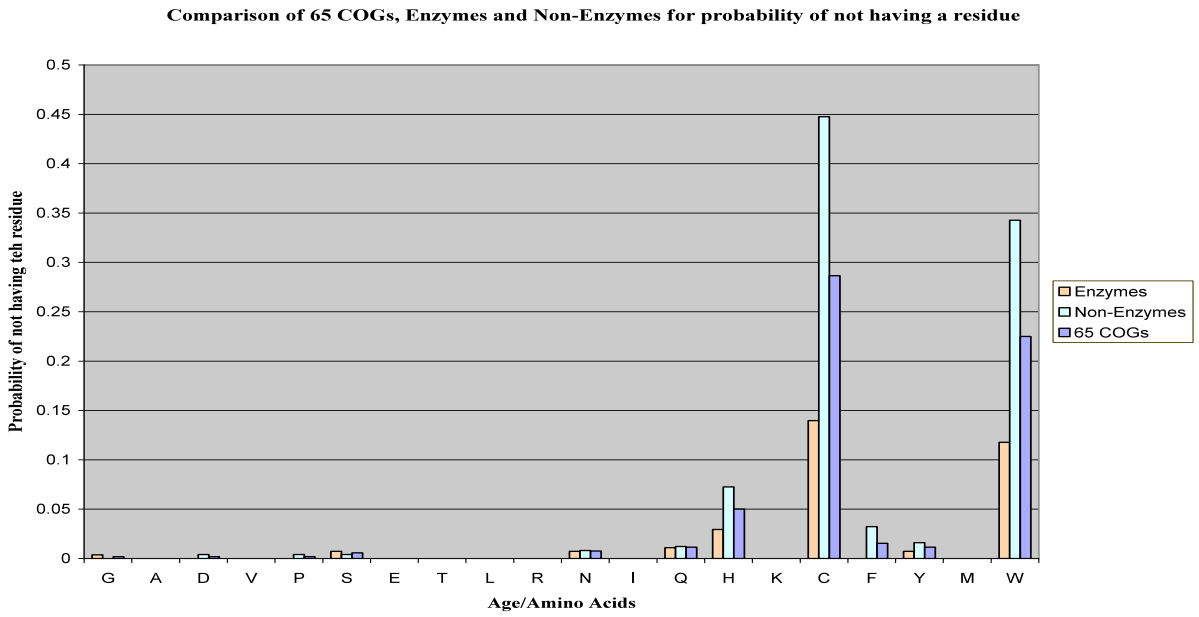
Figure 4.16 Age v/s Probability of not having a single residue (Swiss-Prot).

The Age v/s probability of not having a particular residue for Swiss-Prot data set is plotted. It can be seen that the probability of not having a residue is highest for Tryptophan. The probability of not having the modern residues also increases as seen in the Figure 4.16.



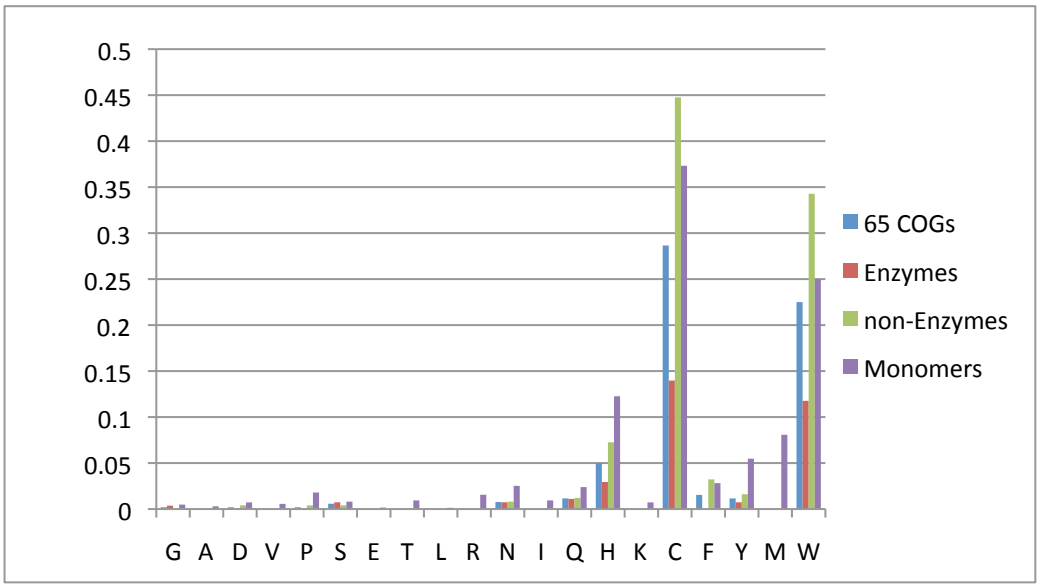
4.17 Result after using Bootstrapping

After removing the length bias in PDB and Monomer dataset using the Bootstrapping approach the age v/s probability of not having the residue in a given dataset was plotted again (see Figure 4.17). The age of the amino acids in the decreasing order (order of appearance of amino acids) is plotted on X-axis and the probability of not having the residue in a sequence is plotted on Y-axis. Even after applying Bootstrapping approach, the probabilities in PDB seem to be lower than in the Monomer dataset. This means that the length bias did not affect the results to a greater extent.



4.18 Comparison of probabilities of not having a residue in different datasets (65 COGs, enzymes and non-enzymes).

Figure 4.18 shows the probability of not having the residue in the data sets: 65 COGs, 34 Enzymes and 31 Non-Enzymes. It can be inferred that there is a highest probability of not finding the modern amino acids in the non-enzymes, which signifies that modern amino acids are lower in content the in non-enzymatic proteins.



4.19 Comparison of probabilities of not having a residue in different datasets (Monomers, 65 COGs, enzymes and non-enzymes).

Figure 4.19 shows the probability of not having the residue in the data sets: 65 COGs, 34 enzymes, 31 non-enzymes and Monomers. It can be inferred that there is a highest probability of not finding the modern amino acids in the non-enzymes from COGs. The previous comparisons of PDB probabilities with Monomers showed that there were higher probabilities for not finding the modern amino acids in Monomers than in PDB. When the probabilities of Monomers are compared with 65 COGs, 34 enzymes and 31 non-enzymes in above graph, it can be inferred that the presence of non-enzymes in the Monomer dataset could be responsible for their higher probabilities of not having the modern amino acids, since the non-enzymes in the Monomer dataset could contain lower number of modern, order-promoting residues.

4.3 Review of Findings

The results clearly indicate that there is a change in the properties of modern proteomes in comparison with the properties if the LUA proteome. As proposed, the enzymes seem to be enriched in the modern, order-promoting residues. The non-enzymes are enriched in the early, disorder-promoting residues. The disorder-promoting residues Serine and Aspartic acid tend to increase in enzymes even though they are disorder promoting. This is explained in the discussion that follows. The second analysis is inconclusive. The sequences from both PDB and Monomer datasets were expected to have higher content of modern amino acids, since both the datasets have structured proteins. However, it was found that the frequency of not having the modern set of amino acids was high in PDB and Monomer dataset. Figures 4.13, 4.14, and 4.15 shows that the probability of not having a residue in the sequence is higher for modern amino acids than for ancient amino acids. The results of Bootstrapping confirm these results even in the absence of sequence length bias. Finally, the results of this same analysis applied to the 65 COGs, enzymes and non-enzymes show that there is a high probability of not finding the modern amino, order-promoting residues in non-enzymes as compared to enzymes. This suggests that since the non-enzymes are not involved in any catalytic functions (unless they bind to their partners), they do not require a rigid structure like enzymes and therefore they have lower content of modern, order-promoting amino acids.

CHAPTER FIVE: DISCUSSION

5.1 Elucidation of Outcomes

The goal of the present work was to test the hypothesis that ancient proteins are much richer in disorder as compared to modern proteins. Even though the results of our analysis support our basic hypothesis, there are some apparent contrary data for certain individual residues such as serine, aspartic acid and glutamic acid. As these three residues are disorder-promoting residues, it was expected that they would be depleted in enzymes. These disorder-promoting residues are proposed to be enriched in non-enzymes, but to the contrary these amino acids tend to be enriched in the enzymes. This could be because glutamic acid and aspartic acid are hydrophilic and have negative charge at different physiological conditions. This negative charge can be utilized in coordination of biologically important metal ions. Serine on the other hand are needed in the process of cell signaling, since signaling is the most important process carried in eukaryotes. A study by Dunker et al. (2000) on proteomes from bacteria, archea and eukaryotes suggests that intrinsic disorder is more prominent in proteins from eukaryotes than bacteria and archea and that these intrinsically disordered proteins take part in signaling, molecular recognition and regulation of transcription which was later supported by Ward et al. (2004). Hence it is obvious that some modern proteins need high amounts of these disorder-promoting residues. It can be inferred from the graphs that there is a considerable increase in the content of modern amino acids in the modern day proteins than the ones in the LUA. This implicates that these residues might have been included into the proteome after the early amino acids entered in to the genetic code; therefore it is obvious that their contents are low in modern proteins. It is observed that some of the positively charged amino acids argnine and lysine are enriched in the non-enzymes.

This is because the non-enzymes from the COGs data set are mostly ribosomal proteins. The RNA in the ribosomal proteins is negatively charged; accordingly, it possesses a high affinity to the positively charged amino acids. Methionine is more abundant in COGs than in the structured proteins. Histidine, which falls into the modern amino acids category, shows no such change in its content. Although there is an increase in its amount, it is not to a greater extent. The modern amino acids such as cystine, phenylalanine, tryptophan and tyrosine seems to be enriched in the enzymes, which is obvious, since they help in proteins structure stability and add to the bulkiness of the ploypeptide chains (Fournier & Gogarten, 2007). These modern, order-promoting residues give the enzymes rigid 3D structures to carry out their catalytic activities. The ancient proteins are enriched in the early, disorder-promoting residues such as glycine, alanine. Both glycine and alanine are weak disorder promoting residues. The ancient proteins had higher content of these disorder-promoting residues that give flexibility rather than rigidity to the structure of proteins. Valine tends to give molten globule structure to proteins in the absence of aromatic. But in the presence of aromatics it is helpful in stabilizing the structure of proteins. In the results, valine was found to be enriched in ancient proteins and aromatics are found to be in lower contents. So, it can be inferred that the ancient proteins were likely to be molten globules. Therefore, some ancient proteins needed a cofactor or a partner to have rigid structure in order to participate in biological processes. When the dataset was split into enzymes and non-enzymes, the non-enzymes were much more enriched in the early, disorder-promoting residues.

An analysis was carried out to find sequences without combinations of the modern amino acids in PDB, Swiss-Prot and monomer datasets. This analysis implies that it is quite common to find sequences without the various combinations of modern amino acids. The

highest frequency for the combination of modern amino acids to be absent in the modern day proteins, especially in the PDB, could be due to fact that these modern amino acids came later into the genetic code and are therefore relatively uncommon. But there is little difference between the three datasets; the reason for this could be the fact that not all ordered proteins are enzymes. Therefore, not all ordered proteins would have the higher contents of modern amino acids, just those that have catalytic function. The higher probabilities in the monomer dataset when compared to the PDB probabilities are also unexpected. The bootstrapping approach used to remove the length bias could not help to remove this bias and explain clearly why there were higher probabilities for monomers as compared to the overall PDB dataset. Since the monomer dataset was a subset of the PDB data set, it would be interesting to determine, what is really influencing the results. Based on the results of our COGs analysis (see Figure 4.18 & 4.19) it can be argued that the presence of non-enzymes in the datasets of PDB, Monomers and Swiss-Prot could have influenced the result. Splitting these datasets into enzymes and non-enzymes could offer us some light in understanding the high probabilities of not finding the modern amino acids in these datasets. It was noticed that other groups of amino acids, for example CHWY and CMWY, had higher frequencies of being absent in the modern day proteins. Finally, the analysis of 65 COGs, enzymes and the non-enzymes revealed that the probability of finding the sequence without the modern day amino acids in the non-enzymes was higher.

5.2 Significance of Results

These results indicate that the order-promoting, modern amino acid residues were more prevalent in the enzymes and less prevalent in the non-enzymes, even in the LUA. Since these modern, order-promoting residues are low in abundance in the modern proteins

from PDB, Monomers and Swiss-Prot, it is likely that they were added into the genetic code more recently. The results on the COGs data set split into enzymes and non-enzymes as seen in Figures 4.18 & 4.19 clearly indicate that the high probabilities of not finding a modern residue in sequences of PDB, Swiss-Prot and Monomers could be due to the presences of non-enzymes and that these non-enzymes are enriched in ancient, disorder-promoting residues as they have more flexible structures rather than a rigid structure like the enzymes have.

5.3 Review of Discussion

In brief, the non-enzymes tend to have higher content of ancient, disorder-promoting residues. Amino acids have different roles in structure and functions of the proteins. For example, the modern, order-promoting residues such as cystine, phenylalanine, tryptophan and tyrosine are needed in proteins to make their structures bulky and to give rigidness. Therefore they are present in higher amounts in the LUA enzymes. Some amino acids have their roles in biological activities such as cell signaling and hence they are found in higher compositions in ordered set proteins, although they are disorder promoting. No obvious trend was noticed with regard to the residues glutamine, threonine, isoleucine and leucine. These amino acids fall in the middle of the order of appearance of the amino acids, and their importance to structure and function. It is known that many ordered proteins contain proline residues that are essential for their rigid structure. In these cases, prolines serve as specific locks, holding appropriate geometry of critical turns. The presence of such essential prolines is known to slow down the rates of protein refolding (Semisotnov et al. 1990). As proline is found to be increased in ancient enzymes, therefore it could be inferred that many of these prolines are essential and also will serve as specific structural locks. Glycine and alanine are

most ancient and tend to give flexibility to the proteins, hence considered weak disordered-promoting residues. They are found to be enriched in non-enzymes. When the proteins from the LUA, 65 COGs were analyzed for the absences of modern amino acids, it was noticed that the non-enzymes had lower contents of the modern, order-promoting residues.

CHAPTER SIX: CONCLUSIONS

6.1 Limitations in this study

Although the results suggest that the proposed hypothesis was true, there were several limitations that cannot be overlooked. The abundance of the individual amino acids in itself can offer little knowledge about the primitive world because of the complex contributions to function made by each amino acid. The abundance of amino acids in groups may be helpful. Many conflicting theories about the entry of amino acids into the genetic code are postulated by different scientists. However there is a consensus that the early amino acids are glycine, alanine, serine, and aspartic acid, which are disorder-promoting residues. Different theories related to the presence and existences of the LUA are presented. There is no consensus, on the proteins present in the LUA. Studies related to the proteomic approach to find the LUA and its nature, use different proteins from different organisms. The present study only limits to the LUA stage of life, it cannot be helpful to make postulations and arguments about the prebiotic stage.

6.2 Future Study

Although the selection of 65 COGs from the three lineages, which were present in the LUA, is justified, it would be interesting to add some organisms to this study and to check the hypothesis as a future study. Only eight organisms were used in this analysis and the number of these model organisms should be increased in a future study. This will require the creation of a new phylogenetic tree. Many researchers have different opinions regarding the LUA; different researchers include different organisms to develop ideas about the LUA and different proteins. Since there are lots of speculations regarding the theory of LUA, it would be important to include more organisms for the study. It would be interesting see if an amino

acid is absent then how it is dependent on other residues, that is if the amino acids are absent in groups of two. Based on the results (See Figure 4.18 & 4.19) the study can be further followed by splitting the datasets of PDB, Monomers and Swiss-Prot into enzymes and non-enzymes.

6.3 Final Summary

Based on the results it can be inferred that the ancient proteins contained a higher contents of disorder-promoting residues, which entered early into the genetic code. As a result, primordial proteins could have been intrinsically disordered to a greater extent than modern proteins. Since the enzymes should have rigid 3D structures to carry out their catalytic functions, they have higher content of the modern amino acids, which are order promoting. Their counterparts, non-enzymes, need not have such rigid 3D structures since they do not act as catalysts. Thus, even today such proteins are highly enriched in the early, disorder-promoting amino acid residues. It is easier to find sequences without the cystine and the aromatics since they are the modern amino acids in evolution. In Swiss-Prot, it was seen that tryptophan has the highest probability of not being in a sequence. The reason for the high probabilities of finding a sequence without order-promoting, modern amino acids or combinations of modern amino acids in the PDB could be due to the presence of non-enzymes in the datasets. Since the non-enzymes are enriched in disorder-promoting residues and have a lower compositions of modern, order promoting residues. All in all, these results suggest that the latest amino acids were selected for their propensity to further stabilize protein structures, a capability that is very useful for creating stable catalytic active site.

REFERENCES

- Adkins JN, Lumb KJ (2002). Intrinsic structural disorder and sequence features of the cell cycle inhibitor p57Kip2. *Proteins,* 46(1), (pp.1-7).
- Anfinsen C. B., Haber E., Sela M. and White F. N. (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Sciences of the USA*, 47, (pp.1309–1314).
- Bairoch A., Apweiler R.(1997). The SWISS-PROT protein sequence database: its relevance to human molecular medical research. *Journal of Molecular Medicine,* 75, (pp.312-316).
- Baskakov IV, Kumar R, Srinivasan G, Ji YS, Bolen DW, Thompson EB (1999). Trimethylamine N-oxide-induced cooperative folding of an intrinsically unfolded transcription-activating fragment of human glucocorticoid receptor. In *The journal of biological Chemistry,* 274(16), (pp.10693-10696).
- Berman H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne (2000). The Protein Data Bank. *Nucleic Acids Research,* 28, (pp.235-242).
- Bienkiewicz EA, Adkins JN, Lumb KJ (2002). Functional consequences of pre organized helical structure in the intrinsically disordered cell-cycle inhibitor p27(Kip1). *Biochemistry,* 41(3), (pp.752-759).
- Bracken C, Iakoucheva LM, Romero PR, Dunker AK (2004). Combining prediction, computation and experiment for the characterization of protein disorder. *Current Opinion in Structural Biology,* 14(5), (pp.570-576).

- Brooks D. J., J.R. Fresco and M. Singh (2004) A novel method for estimating ancestral amino acids composition & its application to proteins of Last Universal Ancestor. *Bioinformatics,* 20(14), (pp.2252-2257).
- Brooks D.J., Jacques R. Fresco, Arthur M. Lesk and Mona Singh (2002) Evolution of amino acids frequencies in proteins over deep time: Inferred order of introduction of amino acids into the genetic code. *The society for molecular biology and evolution,* 19(10), (pp.1645-1655).
- Brown CJ, Takayama S, Campen AM, Vise P, Marshal TW, Oldfield CJ, Williams CJ, Dunker AK (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. *Journal of Molecular evolution,* 55(1), (pp.104-110).
- Cairns-Smith AG (1966). The origin of life and the nature of the primitive gene. *Journal of theoretical Biology,* 10(1), (pp.53-88).
- Campbell KM, Terell AR, Laybourn PJ, Lumb KJ (2000). Intrinsic structural disorder of the C-terminal activation domain from the bZIP transcription factor Fos. *Biochemistry,* 39(10), (pp.2708-2713).
- *Clustalw http://www.ebi.ac.uk/Tools/clustalw2/index.html.*
- Daughdrill, G. W., Chadsey, M. S., Karlinsey, J. E., Hughes, K. T., & Dalquist, F. W. (1997). The C-Terminal Half of the Anti-sigma Factor, FlgM, Becomes Structured When Bound to its Target. *Nature Structural Biology,* 4(4), (pp.285-291).
- Di Giulio M (2001). The universal ancestor was a thermophile or a hyperthermophile. *Gene,* 281(1-2), (pp.11-17).
- Di Giulio M (2003). The universal ancestor and the ancestor of bacteria were hyperthermophiles. *Journal of Molecular Evolution,* 56(6), (pp.721-730).

- Di Giulio M (2007). The universal ancestor and the ancestors of Archaea and Bacteria were anaerobes whereas the ancestor of the Eukarya domain was an aerobe. *Journal of Evolutionary Biology,* 20(2), (pp.543-548).
- Doolittle WF (2000). The nature of the universal ancestor and the evolution of the proteome. *Current Opinion in Structural Biology,* 10(3), (pp.355-358).
- Dunker A. K., Marc S. Cortese, Pedro Romero, Lillia M. Lakooucheva and Vladimir N.Uversky (2005) Flexible nets The role of intrinsic disorder in protein interaction networks. *FEBS,* 272, (pp.5129-5148).
- Dunker AK, Garner E, Guilliot S, Romero P, Albrecht K, Hart J, Obradovic Z, Kissinger C, Villafranca JE.(1998). Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pacific Symposium on Biocomputing,* (pp.473-484).
- Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeve R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Ciu W, Garner EC, Obradovic Z. (2001). Intrinsically disordered protein. *Journal of Molecular Graphics and Modeling,* 19(1), (pp.26-59).
- Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ.(2000).Intrinsic Protein Disorder in Complete Genomes. *Genome Informatics*, 11, (pp.161-171).
- Dunker AK, Obradovic Z.(2001). The protein trinity--linking function and disorder. *Nature Biotechnology*, 19, (pp.805-806).
- Dyson HJ, Wright PE (2002). Coupling of folding and binding for unstructured proteins. *Current Opinion in Structural Biology,* 12(1), (pp.54-60).

- Felsenstein, J. (1993) PHYLIP (Phylogeny inference package) version 35c distributed by the author. Department of genetics, University of Washington, Seattle.
- Flaugh SL, Lumb KJ (2001). Effects of macromolecular crowding on the intrinsically disordered proteins c-Fos and p27(Kip1). *Biomacromolecules,* 2(2), (pp.538-540).
- Fournier GP, Gogarten JP. (2007). Signature of a primitive genetic code in ancient protein lineages.*Journal of Molecualr Evolution,* 65, (pp.425-436).
- Harris JK, Kelley ST, Spiegelman GB, Pace NR (2003). The Genetic Core of the Universal Ancestor. *Genome research,* 13(3), (pp.407-412).
- Hutchinson GE (1964). The determinants and evolution of life. The influence of enviromnent. *Proceedings of the National Academy of Sciences of the USA,* 51, (pp.930-934).
- Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. *Journal of Molecular Biology,* 323(30), (pp.573-584).
- Iakoucheva LM, Kimzey AL, Masselon CD, Bruce JE, Garner EC, Brown CJ, Dunker AK, Smith RD, Ackerman EJ (2001). Identification of intrinsic order and disorder in the DNA repair protein XPA. *Protein Science,* 10(3), (pp.560-71).
- Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Skies JG, Obradovic Z, Dunker AK (2004). The importance of intrinsic disorder for protein phosphorylation. In *Nucleic Acids Research,* 32(3), (pp.1038-1049).
- J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton and D. T. Jones (2004).Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *Journal of Molecular biology*, 337, (pp.635-645).

- Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosciences,* 8, (pp.275-282).
- Joyce GF (2002). The antiquity of RNA-based evolution. *Nature,* 418(6894), (pp.214-221).
- Kolaskar AS, Ramabrahmam V (1982) Obligatory amino acids in primitive proteins. *Bio Systems,* 15(2), (pp.105-109).
- Levy M, Miller SL (1998). The stability of the RNA bases: Implications for the origin of life. *Proceedings of the National Academy of Sciences of the USA,* 95(14), (pp.7933-7938).
- Mayr E (1964). The determinants and evolution of life. The evolution of living systems. *Proceedings of the National Academy of Sciences of the USA,* 51, (pp.934-941).
- Meinschein WG (1965). Carbon compounds in terrestrial samples and the Orgueil meteorite. *Life Sciences and Space research,* 3, (pp.165-181).
- Miller NE (1964). The determinants and evolution of life. Physiological and cultural determinants of behavior. *Proceedings of the National Academy of Sciences of the USA,* 51, (pp.941-954).
- Miller SL, (1953). A production of amino acids under possible primitive earth conditions. *Science,* 117(3046), (pp.528-529).
- Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, Uversky VN. (2006). Analysis of molecular recognition features (MoRFs). *Journal of Molecular Biology,* 362, (pp.1043-1059).

- Nashimoto M (2001) The RNA/Protein Symmetry Hypothesis: Experimental Support for Reverse Translation of Primitive Proteins. *Journal of Theoritical Biology,* 209(2), (pp.181-187).
- Nelson KE, Levy M, Miller SL (2000). Peptide nucleic acids rather than RNA may have been the first genetic molecule. *Proceedings of the National Academy of Sciences of the USA,* 97(8), (pp.3868-3871).
- Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK (2005). Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins,* 61(70), (pp.176-182).
- Oldfield CJ, Cheng Y, Cortese MS, Brown CJ, Uversky VN, Dunker AK (2005). Comparing and combining predictors of mostly disordered proteins. *Biochemistry,* 44(6), (pp.1989-2000).
- Oldfield CJ, Yugong Cheng, Marc S. Cortese, Pedro Romero, Vladimir N. Uversky, and A. Keith Dunker (2005). Coupled folding and binding with helix-forming molecular recognition elements. *Biochemistry,* 44, (pp.12454-12470).
- Orgel L (1994). Origin of Life. A simpler nucleic acid. *Science,* 290(5495), (pp.1306-1307).
- Orgel LE (1994). The origin of life-earth. *Scientific American,* 271(4), (pp.76-83).
- Osawa S, Thomas H.Jukes (1989). Codon Reassignment (Codon Capture) in Evolution. *Journal of Molecular Evolution,* 28, (pp.271-278).
- Peng K, Vucetic S, Radiovojac P, Brown CJ, Dunker AK, Obradovic Z (2005). Optimizing long intrinsic disorder predictors with protein evolutionary information. *Journal of Biocomputing and Computaional Biology,* 3(1), (pp.35-60).

- Ptitsyn O. B. (1995) Molten globule and protein folding. *Advances in Protein Chemistry,* 47, (pp.83–229).
- Radivojac P, Obradovic Z, Brown CJ, Dunker AK (2003). Prediction of boundaries between intrinsically ordered and disordered protein regions.*Pacific Symposium on Biocomputing,* (pp.216-227).
- Rdivojac P, Zoran Obradovic, Celeste J. Brown and A. Keith Dunker (2002). Improving sequence alignments for intrinsically disorder proteins. *Pacific symposium on Biocomputing,* 3 (pp.35-60).
- Rdivojac P, Zoran Obradovic, David K smith, Gaung Zhu, Slobodan Vuetic, Celeste J. Brown, David Lawson and A. Keith Dunker (2004) Protein flexibility and intrinsic disorder. *Protein science,* 13 (1), (pp.71-80).
- Riechmann L, Lavenir I, de Bono S, Winter G (2005). Folding and stability of a primitive protein. *Journal of Molecular Biology,* 348(5), (pp.1261-1272).
- Romero P, Obradovic Z, Kissinger CR, Villafranca JE, Garner E, Guilliot S, Dunker AK. (1998). Thousands of proteins likely to have long disordered regions. *Pacific Symposium on Biocomputing*, (pp.437-448).
- Romero P, Obradovic, Z., Kissinger, C.R., Villafranca, J.E., and Dunker, A.K., (1997) Identifying disordered regions in proteins from amino acid sequences, *Proc. I.E.E.E. International Conference on Neural Networks,* 1, (pp.90-95).
- Romero P, Zoran Obradovic, A. Keith Dunker (2004). Natively disorder proteins functions and predictions. *Appl Bioinformatics,* 3(2-3), (pp.106-113).

- Romero P, Zoran Obradovic, Xiaohong Li, Ethan C. Garner, Celeste J. Brown, A. Keith Dunker (2001). Sequence complexity of disordered protein. *Proteins: Structure, Function and Genetics,* 42(1), (pp.38-48).
- Semisotnovt G. V., V. N. Uversky, I. V. Sokolovsky, A. M. Gutin O. I. Razgulyaev and N. A. Rodionova (1990). Two slow stages in refolding of bovine carbonic anhydrase B are due to proline isomerization. *Journal of Molecular Biology*, 213, (pp.561-568).
- Skrabana R, Sevick J, Novak M (2006). Intrinsically disordered proteins in the neurodegenerative processes: formation of tau protein paired helical filaments and their analysis. *Cellular Molecular Neurobiology,* 26(7-8), (pp.1085-1097).
- Thompson, J.D., Higgins, D.G., and Gibson, T.J.(1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research,* 22, (pp.4673-4680).
- Trifonov E. N. (2000) Consensus temporal order of amino acids and evolution of the triplet code. *Gene,* 261, (pp.139-151).
- Trifonov E. N. (2004) The triplet code from the first principles. *Journal of Biomolecular structure and dynamics,* 22, (pp.1-11).
- Uversky VN (2003). Protein folding revisited. A polypeptide chain at the folding-misfolding-nonfolding cross-roads: which way to go? *Cellular and Molecular Life Sciences,* 60(9), (pp.1852-1871).

- Uversky VN, Christopher J. Oldfield and A. Keith Dunker. (2005).Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *Journal of Molecular Recognition,* 18, (pp.343-384).
- Uversky VN, Gillespie JR, Fink AL. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins*, 41, (pp.415-427).
- Uversky VN, Oldfield CJ, Dunker AK (2008). Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annual Review of Biophysics,* 37, (pp.215-46).
- Uversky VN, Radivojac P, iakoucheva LM, Obradovic Z, Dunker AK (2007). Prediction of intrinsic disorder and its use in functional proteomics. *Methods in Molecular Biology,* 408, (pp.69-92).
- Uversky VN, Roman A, Oldfield Cj, Dunker AK (2006). Protein intrinsic disorder and human papillomaviruses: increased amount of disorder in E6 and E7 oncoproteins from high risk HPVs. *Journal of Proteome Research,* 5(8), (pp.1829-1842).
- Vacic V, Oldfield Cj, Mohan A, Radivojac P, Cortese MS, Uversky VN, Dunker AK (2007). Characterization of molecular recognition features, MoRFs, and their binding partners. *Journal of Proteome Research,* 6(6), (pp.2351-2366).
- Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Freer ST, Rose PW (2003). Simulating disorder-order transitions in molecular recognition of unstructured proteins: where folding meets binding. *Proceedings of the national Academy of Sciences of the USA,* 100(9), (pp.5148-5153).

- Vlassov Av, kazakov SA, Johnston BH, landweber LF (2005). The RNA World on Ice: A New Scenario for the Emergence of RNA Information. *Journal of Molecular Evolution,* 61(2), (pp.264-273).
- Vucetic S, Brown CJ, Dunker AK, Obradovic Z (2003). Flavors of protein disorder. *Proteins,* 52(4), (pp.573-584).
- Wais AC (1986). Archaebacteria: The road to the universal ancestor. *BioEssays,* 5(2), (pp.75-78).
- Weinreb, P. H., Zhen, W., Poon, A. W., Conway, K. A., and Lansbury, P. T. (1996). NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. *Biochemistry,* 35, (pp.13709-13715).
- Williams RM, Obradovi Z, Mathura V, Braun W, Garner EC, Young J, Takayama S, Brown Cj, Dunker AK (2001). The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pacific Symposium on Biocomputing,* (pp.89-100).
- Woese C. (1998). The universal Ancestor. *Proceedings of the National Sciences of the USA,* 95(12) (pp.6854-6859).
- Yang, Z., Kumar, S. and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics,* 141(4), (pp.1641-1650).

Appendices

Amino Acids

1.	Alanine	ALA	A
2.	Arginine	ARG	R
3.	Asparagine	ASN	N
4.	Aspartic Acid	ASP	D
5.	Cystine	CYS	C
6.	Glutamine	GLN	Q
7.	Glutamic Acid	GLU	E
8.	Glycine	GLY	G
9.	Histidine	HIS	H
10.	Isoleucine	ILE	I
11.	Leucine	LUE	L
12.	Lycine	LYS	K
13.	Methionine	MET	M
14.	Phenylalanine	PHE	F
15.	Proline	PRO	P
16.	Serine	SER	S
17.	Threonine	THR	T
18.	Tryptophan	TRP	W
19.	Tyrosine	TYR	Y
20.	Valine	VAL	V

Abbreviations

1. CD- Circular Dichroism.
2. COGs- Clusters of Orthologous Groups of Proteins.
3. DM- Scoring matrix developed based on disordered matrix.
4. DNA- Deoxy Ribonucleic Acid.
5. E- Enzymes from the 65 COGs.
6. ID- Intrinsic Disorder.
7. IDPs- Intrinsically Disordered Proteins.
8. LUA – Last Universal Ancestor.
9. NE- Non-Enzymes from the 65 COGs.
10. PDB- Protein DataBank.
11. Phylip- Phylogeny Inference package.
12. RNA- Ribonucleic Acid.
13. Three Dimensional- 3D.

VITA

Sai Harish Babu Karne

712 A Lockefield ct, Ph: 317-410-1993

Indianapolis, IN-46202. Email: saiharishbabu@gmail.com

A self motivated graduate student of good academic standing coupled with a strong background of Molecular biology and programming skills in C, JAVA & Perl, and an active team member with good communication skills.

Education

Master of Science in Bioinformatics (August'06-08).

- IUPUI, School of Informatics (GPA 3.5/4.0)
- Courses: Machine learning and pattern Recognition, Algorithms in Bioinformatics, Data Integration in Life Sciences, Introduction to Informatics, Introduction to Biostatistics, Introduction to Bioinformatics, Structural Bioinformatics, Translational Bioinformatics Applications.

Bachelor of Technology in Biotechnology (4-Year professional course).

- JBR Engineering College, JNTU (GPA 3.7/4.0)

Computing Skills

Scripting Languages: PERL, JAVA SCRIPT.

Markup Languages: XML, HTML, CSS

Languages : C, SQL, COBOL, JAVA, PHP.

Platforms : UNIX, WINDOWS 98/2000/NT/XP, DOS, LINUX.

Databases : MySQL, SQL, MICROSOFT ACCESS, ORACLE (Novice)

Tools : SPSS, MATLAB, MINTAB, ERWIN.

Experience

- Technical Intern, Molecular Kinetics Inc, Indianapolis.
- Worked under the guidance of Dr. Pedro Romero, Assistant Professor, School of Informatics, IUPUI, for the project Protein evolution.
- Worked under the guidance of Dr. Malika Mahoui Assistant Professor, School of Informatics, IUPUI for the project BioFacets-A new integration system for web based search capability that helps the search engine to search effectively. The project works on JAVA, XML, XSLT, XHTML, AJAX, and MYSQL.