International Journal of Computational Bioinformatics and In Silico Modeling
Vol. 2, No. 4 (2013): 173-179
Review Article
Open Access

ISSN: 2320-0634

Atomic Level Sequence Analysis – A Review

Parul Johri *

Amity Institute of Biotechnology, Amity University Lucknow, Uttar Pradesh, India

*Corresponding author: Parul Johri; e-mail: pjohri_14@yahoo.co.in

Received: 25 June 2013 Accepted: 12 July 2013 Online: 16 July 2013

ABSTRACT

The protein sequence analysis in bioinformatics is done by comparing the sequences residue wise. For pair wise sequence analysis or for multiple sequence analysis, the protein sequences are compared amino acid by amino acid. But the elemental composition of proteins gives the basic level of its organization. All the twenty amino acids are made basically from the five atoms namely – Carbon, Nitrogen, Hydrogen, Sulphur and Oxygen. Amongst all, carbon is the main element that contributes majorly to all hydrophobic reactions. The protein sequence analysis cannot be done by solely comparing their amino acids, but it could be done by going one more step down and comparing the atoms. The present review shows an insight to the upcoming atom level comparison and its potential effects on sequence analysis.

Keywords: Carbon; hydrophobicity; protein sequence analysis

INTRODUCTION

Proteins are essential parts of organisms and participate in virtually every process within a cell. Many proteins are enzymes that catalyze biochemical reactions and are vital to metabolic processes. Proteins also have structural or mechanical functions, such as the actin and myosin in muscle and the proteins in the cytoskeleton, which form a system of scaffolding that maintains cell shape. Other proteins are important in cell signaling, immune responses, cell adhesion, and the cell cycle. Proteins are also necessary in animals' diets, since animals cannot synthesize all the amino acids they need and must obtain essential amino acids from food. Through the process of digestion, animals break down ingested protein into free amino acids that are then used in metabolism. Basically these are organic compounds composed of amino acids arranged in linear chain and folded to a globular form. These building blocks of protein are joined together by peptide bonds between the carboxyl and amino groups of adjacent amino acid residues. The sequence in which amino acids occurs in a particular polypeptide is defined by the sequence of the gene, which is encoded by the genetic code. Depending on the properties of their side chains, the amino acids are classified as hydrophobic and hydrophilic. If the side chain of an amino acid is polar, then it is hydrophilic. If the side is non-polar it is classified as hydrophobic. The hydrophobic residues are found buried deep in the inner core of the protein whereas the hydrophilic residues are found in the outer soluble environment (They can react with water). The distribution of hydrophobic residues in a protein contributes majorly towards protein folding and function. It is by the virtue of these properties that proteins differ in their sequential composition.

The elemental composition of proteins signifies the basic level of biological organization. The key elements in all amino acids are hydrogen, carbon, oxygen, nitrogen and sulphur. These elements are responsible for giving the amino acids the properties of hydrophilicity/hydrophobicity which play an important role in protein interactions. The hydrophobic amino acids characteristically have greater number of carbon atoms as carbon is the main element which contributes to hydrophobic interactions in proteins. On the basis of the hydrophobic/hydrophilic properties of the side chains, each amino acid is assigned a hydropathy index. The higher the index is, the more hydrophobic the amino acid. The presence of carbon contributes to a higher hydropathy index. Thus the overall distribution of carbon in a protein contributes to its hydrophobicity. The structure and activity of proteins are contributed by the presence of Large Hydrophobic residues (LHR) such as Phenylalanine (F), Isoleucine (I), Leucine(L),

Methionine(M), Asparagine(V). Carbon is the main element that contributes to the hydrophobic nature of proteins. It has been observed that proteins need 31.44% of total carbon content for their structure stability and activity [1]. This fraction of carbon can be used as standard and can be applied to various statistical methods pertaining to carbon measurement and comparison. It was proposed that a carbon distribution profile can be applied to calculate and identify the quantity of carbon in proteins. This would pave the way for identification of active sites in proteins as maximum hydrophobic residues are present in active sites and the carbon content of hydrophobic residues is higher relative to hydrophilic residues.

History

Carbon distribution along the protein sequence was studied with help of a dynamic programming approach. Dynamic Programming (DP), an approach to solve a problem which has overlapping sub problems, by breaking them down into simpler problems. It involved the splitting of the amino acid sequence into its respective atomic elements. The atomic sequence is divided into windows of equal size. The carbon percentage is calculated for each window. The windows are then grouped according to the carbon percentage. Then the frequency is calculated for all the grouped windows. A graph was plotted for the carbon percentage against the frequency. This approach was used to calculate the carbon content of globular proteins in various model organisms. It was observed that maximum frequency of carbon occurs at 31.44% [2]. Although mild positive skewness was observed, the graph followed a normal distribution curve (Figure 1).

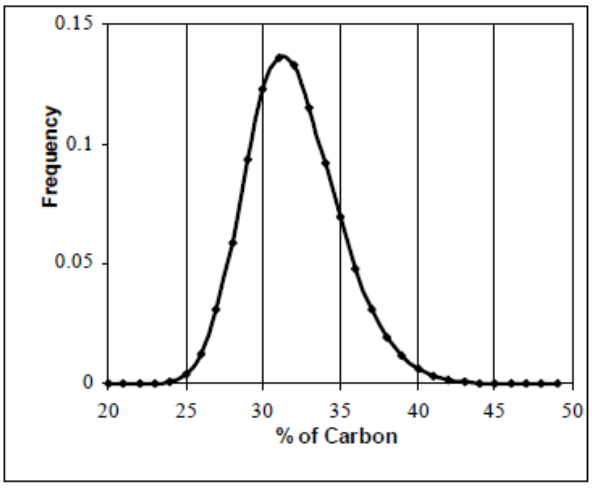

Figure 1. The carbon distribution profile for a length of 100 atoms in a globular protein.

The carbon content analysis of globular proteins in the species of model organisms (human, drosophila, *mus musculus*, cow, chimpanzee, *e.coli,* etc.) revealed that these proteins prefer to have 31.44% of carbon.

Carbon Profile Plot and Hydropathy Plot:

The hydrophobicity and carbon distribution profile of a protein was studied with help of a hydropathy plot. A hydropathy plot of a protein maps the regions of hydrophobicity against the hydropathy indices of the amino acids in the protein. It gives a clear idea of level of hydrophobicity in a protein. . In order to study the carbon distribution in proteins, the human erythrocyte glucose transporter protein was selected and its sequence was retrieved from SWISSPROT database (http://expasy.org/sprot/). ATOMSCAN program, written in C language was used to calculate the carbon distribution in a protein sequence. Using ATOMSCAN it was found that the proteins consist of a fair amount of carbon content (30%-32%) which is characteristic of membrane protein .Weak hydrophilicity is seen in certain regions of the protein. The carbon distribution profile obtained was used to locate the hydrophobic, hydrophilic and also the active sites in the protein. It was compared with the PROTSCALE hydropathy plot (http://expasy.org/cgi-bin/protscale.pl). PROTSCALE is a primary structure analysis tool provided by Expasy Proteomics Server (www.expasy.org). When provided with a protein sequence as input, it calculates the hydrophobicity of the protein based on the hydrophobicity scale provided by Kyte & Do-little (http://gcat.davidson.edu/rakarnik/kyte-doolittle-background.htm#plot). The Kyte & Do-little scale first assigns a score to each amino acid. It chooses a window size and splits the protein sequence into windows. Then it calculates the average hydrophobicity scores for all the windows in the protein sequence. The average scores are then plotted on a graph where the Y-axis represents the hydrophobicity scores and X-axis represents the window number. The Kyte & Do-little hydropathy plots indicate the potential transmembrane and surface regions in a protein. On comparing the carbon distribution profile and the PROTSCALE hydropathy plot (Figure 3), it was found that the profile displays the hydrophobic regions of the protein and can also be used for the identification of active sites. The hydropathy plot does not give information on the active sites of a protein. It was postulated that the carbon distribution profile is a very good alternative to the hydropathy plot [3-4].

Carbon Content and Patterns

Based on the principle that proteins prefer to have 31.44% of carbon for their structure and stability along the protein sequence, a novel pattern was identified. The pattern is defined as a stretch of sequence which has a definite carbon content of 31.44%. A study was conducted on protein sequences of different lengths, where PFIND program was used to retrieve all the patterns in a particular protein sequence, and were further analyzed. The sequences were read in FASTA format and for each amino acid the total number of atoms and carbon atoms were found. For each sequence, the ratio between the number of carbons and total atoms was calculated. If the ratio matched 0.3144 then that stretch of sequence was considered as a pattern. The smallest pattern identified was 7 amino acids long and contained 50 carbon atoms. The patterns

obtained were similar in their atomic sense but differed in the length and sequence at amino acid level. This concluded that the patterns of the proteins may appear as mismatches during protein sequence alignment but are actually same in terms of carbon residues (Table 1). This difference raises the question of accuracy during protein sequence analysis. It is suggested that atomic level comparison can yield better results than sequence level comparison of proteins.

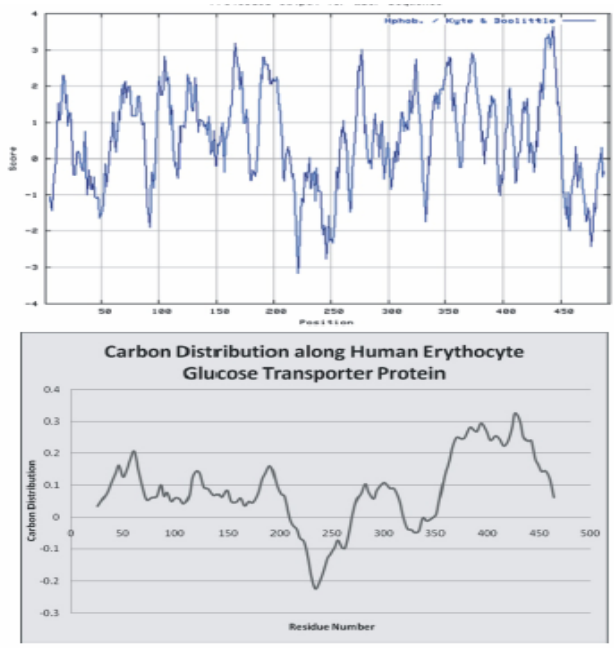

Figure 2. Hydropathy plot and carbon distribution profile compared in human erythrocyte glucose transporter protein.

Table1. Patterns with same number of carbon atoms (50) and total atoms (159) in different length. (Carbon content=50/159=0.3144).

Patterns	No.of Residues	Carbon/ Total atom
RFLRRRW	7	50/159
RLHIKFKE	8	50/159
FNGLEKLLR	9	50/159
IQNPSMLLEP	10	50/159
AQAQREAAAEY	11	50/159
SGRYISAAPGAE	12	50/159
EDSMGGTSGGLYS	13	50/159
EPGEEGPTAGSVGG	14	50/159
GMGGHGYGGAGDASS	15	50/159
GAAGGCGVAGAGADGY	16	50/159

Another observation in case of lengthier patterns was that, for the same length of patterns with 31.44% of carbon, there was a difference of 11 carbon and 35 total atoms (Table 2). This means that the smallest number of carbon and total atoms which will give a ratio of 0.3144 is 11/35.

On evaluating the 3D structure of the patterns that were retrieved, it was observed that they prefer to be folded into a compact stable structure. This suggested that the patterns in proteins are vital for understanding protein stability.

Table 2. Patterns with same length, different number of carbon and total atoms.

Patterns	No.of Residues	Carbon/ Total atom
MKYVAGARPWTHVSNVDIALPCATQ NEVSGDEAKALVASGVKFVAEGANM	50	227/722
GLNIPVILCKNKCDSISNVNANAMVV SENSDDDIDTKVEDEEFIPILMEF	50	238/757
NKIDPELFELRKAVMDTNENEEEKM F RDDTFGKNLNANTNTARLFDDETS	50	249/792

Carbon Content and Active Sites:

The study of carbon distribution profile of proteins was used to suggest the role of carbon content in the identification and development of active sites, protein stability studies, understanding the phylogeny of proteins, gene identification and to study protein – protein and protein-DNA specific and non-specific interactions. Therefore a carbon distribution profile can be used to distinguish the proteins of one species from another. This idea was justified by conducting a study on arenavirus proteins. Protein sequences of 7 arenaviruses were analyzed to show that carbon content in viral proteins is different from normal ones. The four different protein sequences i.e. the RNA dependent RNA polymerase (L), nucleocapsid protein (NP), glycoprotein precursor (GPC) and the zinc binding matrix protein (Z) of 7 arenaviruses were retrieved from NCBI (www.ncbi.nlm.nih.gov). The AACOMP (http://www.mhoenicka.de/software /scisoft/aacomp.html) program was used to calculate the amino acid compositions. AACOMP program provides information on the number of amino acids in the sequence, their relative molecular weight, properties of their side chains etc. The carbon content of the proteins was computed using an online tool CARANA (www.rajasekaran.net.in/tools/carana.html). It accepts the input protein sequence in FASTA format and displays the output an excel file showing the carbon percentage and the atom numbers. It was found that the GPC protein contains higher amount of carbon in the initial portion of the sequence that could hold other interacting molecules for further reactions. This suggested the presence of active sites in the beginning of the sequence which could hold and bind to other molecules to carry out the specific protein interactions. The Z protein showed high amount of carbon content all along the sequence except at the end of the sequence. The carbon distribution of NP protein was similar to that of Z protein and the L protein showed high amount of carbon.

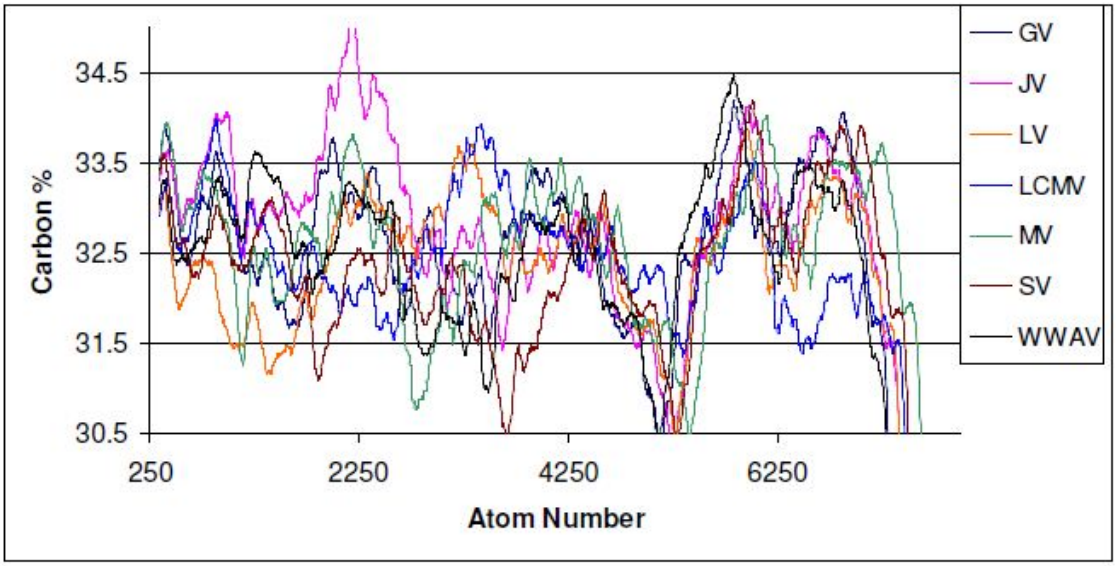

Figure 3. Carbon distribution in the GPC protein of all viruses.

Strand1: 5'-ATGCGGTTGAATTCGTATGCT....-3' (mRNA sequence)
Strand2: 3'-TACGCCAACTTAAGCATACGA....-5' (complementary strand)

Frame No.	Codons	No. of XTX
1	5'-ATG CGG TTG AAT TCG TAT GCT....-3'	2
2	5'-TGC GGT TGA ATT CGT ATG.... -3'	2
3	5'-GCG GTT GAA TTC GTA TGC....-3'	3
4	3'- TAC GCC AAC TTA AGC ATA CGA....-5'	2
5	3'-CGC CAA CTT AAG CAT ACG....-5'	1
6	3'-ACG CCA ACT TAA GCA TAC....-5'	None

Figure 4. Reading of mRNA sequences in different frames.

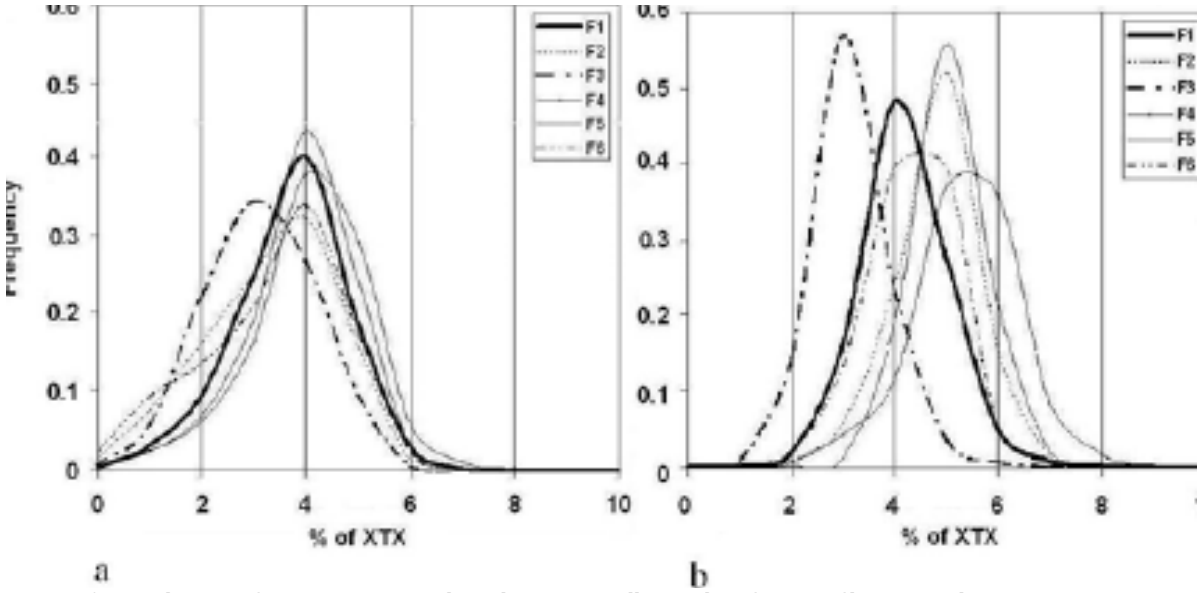

Figure 5. Distribution of mRNA sequences based on XTX in all 6 reading frames of human and yeast

Carbon Content in Thymine

It was concluded from the study done on the proteins of arenaviruses that carbon distribution and content is different in different proteins. It was observed that the functional sites of the proteins displayed higher amount of carbon in comparison to the entire sequence. Overall the proteins showed high carbon content. The LHRs are the major contributors towards higher carbon content. These LHRs are coded by XTX where(X=A, T, G, C). The results also showed that the mRNA sequences of the proteins showed higher amount of Thymine suggesting that the overall AT content in the arenavirus genome is greater than the GC whereas the human genome shows higher GC content [5]. This observation represents a contradiction between the genomic contents of humans and arenaviruses. The importance

of thymine in protein coding frames of mRNA sequences is relevant to protein stability and function [6]. Experiments were conducted on the complete sets of protein coding mRNA sequences of 11 organisms. The sequences were retrieved from NCBI in the FASTA format. Each sequence was divided into six frames. Frames 1, 2 and 3 represented the forward frame and frames 4, 5, 6 represented the reverse frames of the complimentary strand (Figure 4).

Each of the mRNA sequences were read in different frames and the XTX (X=A, T, G, C) was counted for each frame. The fractional quantities of XTX were then grouped based on their thymine content and a graph was plotted with frequency of XTX on the Y-axis and the percentage of XTX on the X-axis. The distribution of thymine in human and yeast mRNA sequences was analyzed and it was reported that in frame 2 and 6, the total amount of thymine in human mRNA sequences is lower than that of yeast (Figure 5).

In general, the probable thymine content of human mRNA sequences was less in comparison to that of plants and fungi but it was also found that the fraction of the total thymine was high in the frame 1 of humans and is reverse in plants and fungus. These observations strongly suggest that the fraction of thymine is important for the production of stable and functional proteins with a definite amount of hydrophobic elements. It was also observed that the variation in the amount of thymine in the DNA sequences of humans was very high as compared to that of fungi and plants. Finally it was concluded that the amount of thymine was reduced in human sequences over the course of evolution. But, the experiments on mRNA sequences on different species showed the importance of thymine for protein function and stability. So it can be inferred that the reduced thymine content is cumbersome to the production of normal functional proteins leading to formation proteins which are unable to function normally or they can be called diseased proteins. This conclusion was in agreement with the hypothesis that proteins have been hydrated during evolution [7].

Carbon and Large Hydrophobic Residues

A study was conducted on the protein sequences of 20 different living systems (human, chimpanzee, dog, cow, mouse, rat, chicken, fruitfly, zebrafish, *Caenorhabdis elegans,* mosquito, honeybee, beetle, yeast, *S.purpuratus, K.lactis, S.Pombe and A.thaliana*) [8]. The fraction of LHRs and the fraction of all amino acids in the proteins of each species were grouped separately. The variation of LHRs in different species was analyzed for its relevance to protein hydration. The length of proteins in heterosexuals was found to increase during evolution (Figure 5). It was also observed that small hydrophobic residues such as glycine, alanine, proline, cysteine and tryptophan were found to compensate for the loss of LHRs. It is due to this compensation by small hydrophobic residues which resulted in the increase in length of protein sequences of animals. It has been noted that the proteins have been hydrated during evolution as there has been a reduction in the amount of LHRs in animals.

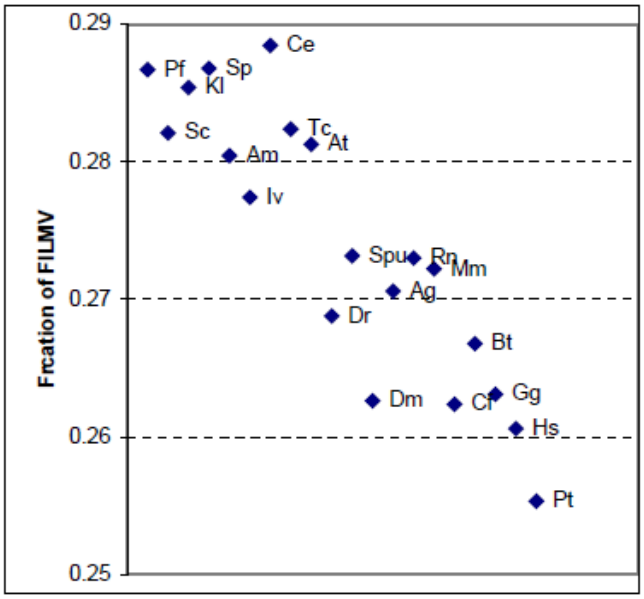

Figure 6. Reduction in the amount of LHR in heterosexual species

In order to examine the role of Large Hydrophobic residues in proteins , the hydrophobic distribution profile of complete sets of protein sequences from different species was studied (Figure 6). The probable amount of LHR in proteins was found to be 27%. The LHR content in animals is less as compared to plants and fungi. It was observed that the amount of LHR was higher near the active sites. Active sites of proteins are vital to the interaction of a protein with other proteins, enzymes and substrates. The presence of LHRs at the active sites contributes to protein interactions (hydrophobic interaction-dominant force); they enhance the ability of a protein to react with another protein and contribute to formation of stable structures for protein interactions. By making use of binary statistics and hydropathy plots it was shown that regions of sequences rich in hydrophobic residues determine the basic topology of proteins. This observation provides evidence of the role of LHR in the protein function and structure. So, the higher amount of LHRs was seen in and around the active sites which is vital for protein function. Around the other regions the content of LHRs was 27% to provide local stability all along the protein [9].

The distribution of LHR in proteins also has implications in protein folding and evolution. The amino acid sequence of any protein shows that hydrophobic residues have a tendency to fold in clusters along the chain. The resulting hydrophobic regions of protein sequences can be identified by the use of hydropathy plots. It has been postulated that the distribution of hydrophobic residues along the chain is random for the majority of membrane and soluble proteins examined. The clustering of hydrophobic residues is a result of random distribution and suggests that functional proteins have evolved from random sequences [10].

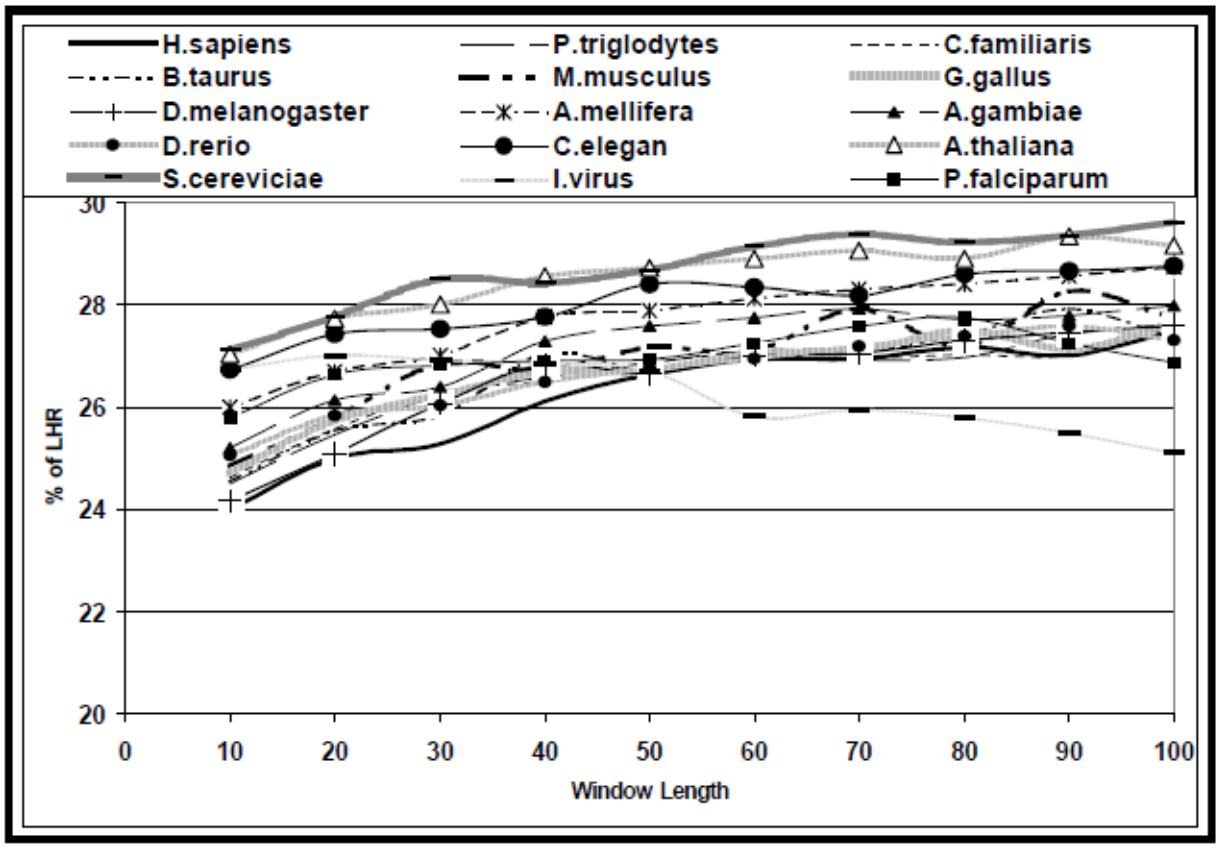

Figure 7. Amount of LHR in different species.

The hydrophobicity of proteins is correlated to their carbon content. Carbon is the major element contributing to hydrophobicity which is the dominant force of biological interactions. The carbon content of the peptide plays a crucial role in evolutionary studies of proteins. The elemental composition of proteins when studied at residue level gives insight into adaptive significance of proteins over time. Evolution of proteins may depend upon the metabolic constraints of amino acids. Thus when carbon or any other nutrient is limited, the system responds by synthesizing proteins which are depleted in that particular nutrient. This is called nutrient sparing and more specifically microbes can respond to nutrient limitation with biases in the atomic compositions of proteins that are highly expressed. The living system adapts to such nutrient limitations and thus metabolic constraints influence the evolution of proteins. It has been postulated that the differential expression of duplicated genes in response to the relevant depleted nutrient could be the cause for such nutrient sparing [11]. Experiments were conducted on *Saccharomyces cereviseae* protein sequences, to demonstrate that protein elemental composition influences evolution by gene duplication Gene expression data between artificially selected strains and unselected ancestral strains of yeast were compared to show that genes with high expression in ancestral strains consist of protein products that have significantly less amount of carbon. The carbon content of amino acids is highly proportional to the amount of energy required for their synthesis. It is a negative correlated. Hydrophobic residues containing high amount of carbon requires more energy to get synthesized. So, proteins which are highly expressed show less amount of carbon in order to conserve energy. Hence selection on the economy of energy and atomic composition would lead to reduction in the carbon content of highly abundant proteins [12-15]. This eventually leads to changes in proteins at residue level. Amino acid composition is used to determine their frequencies in proteins during the course of evolution. The knowledge of amino acid compositions of proteins in ancestral organisms can reveal which amino acids have been favored during the evolution process. The frequency of preferred amino acids has increased over time as compared to that of amino acids which have not been favored. This information can be used to determine the order in which amino acids were introduced into the genetic code. The amino acids with increased frequency are conserved during evolution and the compositions of conserved residues can be determined by the amino acid composition of ancestral sequences [16-17].

Carbon content and Protein Structural Divergent

Comparative analysis on individual protein families showed that the mutations, insertions and deletions in the genes coding for their products have brought about changes in the 3-dimendional structure of proteins. This is because a mutation event can alter an amino acid. This phenomenon affects the ability of the protein to fold in a particular confirmation as the folding of a protein depends upon the properties of the side chains of the amino acids. A systematic comparison of proteins

from eight different protein families was conducted to show that the extent of structural changes is directly proportional to the extent of sequential changes. 25 proteins from 8 different protein families were used which provided 32 pairs of homologous structures [18]. Structures of homologous proteins were divided on the basis of the similarity and differences between the general folds of a polypeptide chains. Using a quantitative procedure, the main chain atoms of the secondary structures were individually superposed and each region that was superposed was extended to contain additional atoms at both ends. The extensions were continued until the deviations in the positions of the atoms reached a threshold value of 3A° .This method defines the segments which contain the same fold in both the proteins. They include peptides which constitute the active sites. Such regions are termed as the common cores. The pairs of sequences with protein identity greater than 50% consist of 90% or more of the residues of the individual structures that fall within the common core. The common cores of both the proteins are then optimally superposed. Then the root mean square differences in the positions of main chain atoms of both the proteins are calculated. This method of finding the common cores between homologous proteins provide information about the degree of structural divergence between the respective proteins [19-21].

Current Scenario

Overall the research that was conducted on the carbon content of proteins gave insight on the implication of carbon distribution along the protein sequence, its role in protein stability, structure, function, evolutionary significance and protein comparison at atomic level. It was observed that the carbon level of proteins in different species was different. This difference can be used for sequence analysis to identify the phylogeny of a protein. Clustering the proteins on the basis of carbon content would give a better an idea of the functional relevance of proteins. The observed carbon contents of different proteins in an organism can be used to classify proteins. Thus the carbon content of proteins is considered to play a vital role in identifying the phylogeny of an organism. Also the atom level comparison of proteins can use carbon content as a major factor for sequence analysis.

REFERENCES

1. Akila A, Rajasekaran E (2009). What might be the difference in viral proteins? International Journal of Bioinformatics Research. 1(2): 1-3.
2. Anandagopu P, Suhanya S, Jayaraj V et al. (2008). Role of thymine in protein coding frames of mRNA sequences. Bioinformation. 2(7): 304-307.
3. Baudouin CP, Schuerer K, Marliere P (2004). Intimate evolution of proteins: Proteomic atomic content correlates with genome based composition. Journal of Biology and Chemistry. 293(5528): 297-300.
4. Bragg GJ, Hyder L (2004). Nitrogen versus carbon use in prokaryotic genomes and proteomes. Biological Sciences. 271: 374-377.
5. Brooks DJ, Fresco JR (2002). Evaluation of amino acid frequencies in proteins over deep time: Inferred order of Introduction of amino acids into the genetic code. Molecular Biology and Evolution. 19(10): 1645-1655.
6. Brooks DJ, Fresco JR (2002). Increased frequency of cysteine, tyrosine and phenylalanine residues since the last universal ancestor. Molecular Cell Proteomics. 1(2): 125-131.
7. Rajasekaran E, Rajadurai M, Vinobha CS et al. (2008). Are Proteins being hydrated during evolution? International Journal of Computational Intelligence in Bioinformatics. 1(2): 115-118.
8. Boer VM, De WJ, Pronk JT (2002). The genome wide transcriptional responses of *Saccharomyces cerevisiae* grown on glucose in aerobic chemostat cultures limited for carbon, nitrogen, phosphorus or sulphur. Journal of Biology and Chemistry. 278: 3265-3274.
9. Chothia C, Lesk AM (1986). The relation between the divergence of sequence and structure in proteins. The EMBO Journal. 5(4): 823-836.
10. Cyrus C, Arthur ML (1986). The relation between the divergence of sequence and structure in proteins. The EMBO Journal. 5(4): 823-826.
11. Dawn J, Jacques R, Arthur M et al. (2002). Evolution of amino acid frequencies in protein over deep time: Inferred order of introduction of amino acid into the genetic code. Molecular Biology and Evolution. 19: 1645-1655.
12. Finch H (2005). Comparison of distance measures in cluster analysis with dichotomous data. Journal of data science. 3: 85-100.
13. Jason GB, Andrew W (2007). Protein carbon content evolves in response to carbon availability and may influence the fate of duplicated genes. The Royal Society 274: 1063-1070.
14. Jayaraj V, Sunhanya R, Vijayasarthy M et al. (2009). Role of large hydrophobic residues in proteins. Bioinformation. 3(9): 409-412.
15. Jayaraj V, Vijayasarathy M, Geerthana R et al. (2009). Pattern Recognition in Proteins based on carbon content. International Journal of Computational Intelligence in Bioinformatics. 2(2): 99-102.
16. Katti MV, Subbu RS, Prabhakar K et al. (2000). Amino acid repeat in protein sequences: their diversity and structural-functional implications. Protein Science. 9: 1203-1209.
17. Li F, Li W, Farzan M et al. (2009). Structure of SARS Coronavirus spike receptor binding domain complexed with receptor. Sciences. 309(5742): 1164-1168.
18. Lynch M, Conery JS (2002). The evolutionary fate and consequences of duplicated genes. Science. 290: 151-1155.
19. Rengananthan S, Rajasekaran E (2009). Comparative analysis of carbon distribution and hydropathy plot. Advanced Biotech 1: 30-31.
20. Stephen H, White, Russell E, Jacobs (1990). Statistical distribution of hydrophobic residues along the length of protein chains. Biophysics Journal. 57: 911-921.
21. Wagner A (2005). Energy constraints on the evolution of gene expression. Molecular Biology and Evolution. 22(6): 1365-1374.

*****