Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2016, 8(5):204-207

Short Communication

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Towards the atomic level protein sequence analysis

Parul Johri[1]*, Mala Trivedi[1], Aditi Singh[1] and Mohammed Haris Siddiqui[2]

[1]Amity Institute of Biotechnology, Amity University Uttar Pradesh,
Lucknow Campus, Malhaur, Gomti Nagar Extension, Lucknow- 226028 (UP)
[2]Department of Bioengineering, Integral University Lucknow (UP)

ABSTRACT

Proteins differ in the arrangements of 20 naturally occurring amino acids. This difference in protein sequence can also be captured at atom level. Carbon is the only element that contributes towards the hydrophobic interactions that drives the protein to carry out its biochemical reactions. Understanding the difference in protein sequence at atomic level could be very useful both to compare sequence from different origin and to define threshold carbon content for bar-coding of proteins. A new methodology for comparing protein sequences at atomic level is proposed. Viral sequences and aquaporins from various plants and animal origin have been analyzed based on the designed algorithm and the demarcation at carbon level was found to be very prominent.

Keywords: Aquaporins, carbon, dynamic programming, viral protein.

INTRODUCTION

Proteins are the essential macromolecules for the physiological function in all biological systems. These are the most vital organic compounds for various biochemical processes like cell cycle, cell signalling, metabolism of carbohydrates, lipids, fats, immune responses and many more. All proteins have a defined structure and specific function. The primary structure of protein comprise of peptide chain formed by the linkage of amino acids in a linear fashion by peptide bonds. Amino acid composition of protein varies significantly among the various taxa [1]. There are twenty different amino acid arranged and rearranged in different manners to give rise to new forms of proteins. All these twenty amino acids are basically made up of five major atoms namely – Carbon (C), Hydrogen (H), Nitrogen (N), Oxygen (O) and Sulphur(S) [2]. Proteins evolve in response to the nature of interaction and stability [3]. These five atoms are the lowest level of biological organization. The signature of natural selection is also visible at this level. Protein sequences can vary greatly in their content of nitrogen, sulphur and carbon atoms and this elemental composition variation in the sequences can be influenced by the process of natural selection.

1.1 Carbon and LHR -

The major force involved in protein folding and acquiring its function is the hydrophobic interaction among the amino acids and their side chains. Large hydrophobic residues (LHR) like phenylalanine, isoleucine, leucine, methionine and valine plays an essential role in protein structure and their activities. Moreover, plants and fungi have higher number of LHRs as compared to the heterosexual animals [4]. On the basis of the hydrophobic residues or hydrophilic residues in the side chain of the amino acids, all the twenty amino acids are assigned with a hydropathy index. Higher index values correspond to more hydrophobic nature of amino acid. This dominant force of hydrophobic interaction arises due to the carbon atom content in the protein [5]. The co relation between the carbon content and the hydrophobicity was studies in detail by comparing a carbon distribution profile and the protscale hydropathy plot of the human erythrocyte glucose transporter protein. 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 [6].

1.2 Threshold of Carbon in proteins -

Recent studies on the total carbon content of proteins states that proteins prefer to have 31.45% of total carbon atom for their stability in structure as well as sequence [7]. Provided this as a standard, many carbon based tools have been developed to study carbon distribution profile [8] and carbon analysis program [9]. However a correct and comprehensive understanding of carbon atom and functionality of protein still needs to be establish. There are many studies which have identified the effect of the distribution of carbon atom in protein sequences. One amongst them showed the higher content of carbon in the active site of toxic shock syndrome toxin from *Staphylococcus aureus* which is destabilized by mutation , hence protein disorder. Another study on the Fragile X mental Retardation 1 protein (FMR1), main cause for the disease Fragile X Syndrome in humans was done that showed the portion of the protein having less than 31.45% of carbon lacks stability [10].

At present it is possible to calculate the overall carbon content of any protein using dynamic algorithm codes and then analyse them statistically for its effect on stability and functioning of that protein. Sericin protein, a hydrophilic protein from *Bombyx mori* was investigated for carbon distribution showed lower amount of carbon percentage hence improving its washing abilities [11].

1.3 Carbon distribution in viruses –

Various viral proteins have been studied which showed that either the total carbon content or the carbon distribution in viruses is different from the normal proteins. Most of the viral protein showed higher content of carbon on them and especially the attachment proteins of viruses. The role of carbon distribution in the proteins of influenza virus showed higher carbon content in surface proteins, optimum in polymerase protein and less in nuclear proteins. These differences in carbon distribution in viral proteome may be correlated to the root of causing disease.

But the only element of complexity is to calculate the carbon content of only the side chains of the protein because they are involved in folding and other interactions. Using the computational tools based on atomic level sequence analysis it will be more easier and immediate to understand the protein characteristics, functioning and mutational analysis. Also the atoms are the lowest level which might bring a new insight for designing tools in bioinformatics.

EXPERIMENTAL SECTION

Two genomes namely – viruses and plants were targeted for their atomic content analysis. The entire work was been divided into two phases. During the first phase, the analysis of the plant genome for aquaporin proteins, a special class of channel protein that controls the precise transport of water molecules across the cell membranes in all the living organisms was done. The aquaporins are thus likely to be of fundamental significance to all facets of plant growth and developments, affected by plant–water relations. A majority of plant aquaporins have been found to share vital structural features with the human aquaporin and show water-transporting ability in various functional assays, and some have been shown experimentally to be of significant importance for plant survival. In the present work, a sample set of 150 aquaporins proteins from Uniprot database was taken and the total carbon atom percentage was calculated using the dynamic programming code (figure 1). The study was done to barcode the aquaporins based on their carbon content, amongst animals and plants.

And in the next phase, virus proteins were analysed for their carbon content. For this the Japanese encephalitis (JE) an enveloped positive-sense single stranded RNA (~ 11 kb in length) virus contains single open reading frame (ORF) encoding a poly-protein that is processed into three structural core protein(C), membrane protein (M) and envelope protein(E) and seven non structural (NS1, NS2B, NS2A, NS3, NS4B, NS4A, and NS5) proteins, flanked by 5′- and 3′-non-translated regions (NTRs) was considered for the study. The study was done to analyze the virulent proteins of JE at atomic level.

The present work was performed with a view to go a step down to analyze these proteins at atom level and find the significance of carbon in them. The virulent proteins were retrieved and studied for the calculation of total carbon percentage in them. The sequences were retrieved from the genome database of NCBI (http://www.ncbi.nlm.nih.gov/genome/). The percentage of carbon was calculated using the dynamic programming code and further analysed using Microsoft excel 2007.

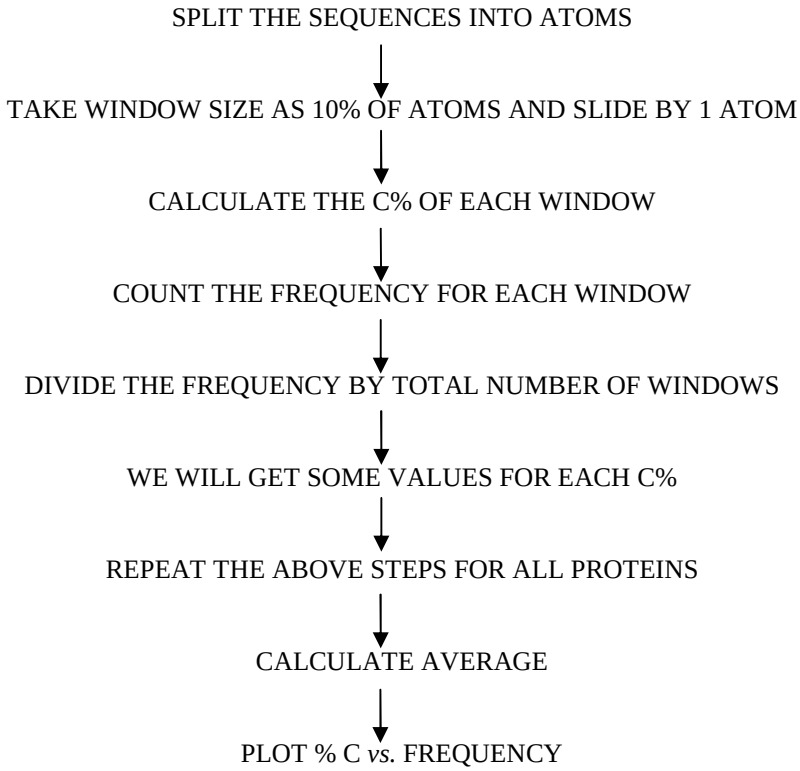

Figure – 1 Dynamic Programming code steps for carbon calculation

RESULTS AND DISCUSSION

The protein taken from *Oryza sativa*, *Zea mays* and *Arabidopsis thaliana* preferred to have carbon percentage of 31.8 to 35, whereas on the other hand sequences taken from *Mus musculus*, *Saccharomyces cerevisiae*, *Homosapiens*, *Bos taurus*, and *Rattus norvegicus* preferred to have carbon percentage of 31 to 33.7. This clearly demarks the carbon range in the aquaporin proteins from plant and animal origin.

The Japanese encephalitis virus genome contains 1880 proteins with 26-27 different types of proteins namely Polyprotein, Envelope protein, prM protein , Capsid protein , E protein , M protein , Matrix protein , Putative envelope protein, GP78 poly protein, JEV poly protein, Nucleoglycosylated protein, Nucleocapsid protein, Pre membrane protein, Protease protein, Viral protein, Genome GEV protein. Out of these proteins, Matrix protein , Putative envelope protein, GP78 poly protein, JEV poly protein, Nucleoglycosylated protein, Nucleocapsid protein, Pre membrane protein, Protease protein, Viral protein, Genome GEV protein are present in a very low amount that there is no variance in Carbon distribution. The results on the other 4 proteins are shown below.

3.1 Carbon content in Polyprotein -

Polyprotein is found in higher amount in JE virus. The number of Polyproteins in JE virus is 1041. And the total amount of carbon in these proteins is 3766775. Over all the range of the Carbon atom in polyprotein is 31.26-31.36.

3.2 Carbon content in Envelope protein -

Envelope protein is a non structural protein. The number of envelope protein in JE virus is 526. And the total amount of Carbon element is 1259620. And the overall range is **29.12 - 35.34**.

3.3 Carbon content in E protein -

There is found that JE virus has total 40 E proteins. And this protein has 100598 Carbon all over. The range of Carbon for this protein is **31.20 – 32.19**.

3.4 Carbon content in Capsid protein –

Capsid protein is also a non structural protein. The total Capsid proteins in JE virus is 11. And it contains 7204 total Carbon. The range of Carbon atom is **30.6 – 30.7**.

CONCLUSION

The present study provides a comprehensive picture of the role of carbon content in viral and plant protein sequence analysis. Carbon being the major element of earth and also our body plays a vital role in sequence analysis. The level of carbon atom in different proteins demarks the influence of atom based sequence analysis. Proteins prefer to have a significant level of carbon for their viral nature. To conclude, carbon distribution is different in viral protein sequences and could be used for a better understanding of their phylogeny, function and stability. Apart than carbon, other atom also needs to be taken into account.

Acknowledgements
Authors are grateful to Dr. A. K. Chauhan, Founder President & Mr. Aseem Chauhan, Chancellor Amity University Haryana & Chairperson AMITY Lucknow for providing necessary facilities and support. We also extend our gratitude to *Maj. Gen. K.K Ohri*, AVSM (Retd.), Pro Vice Chancellor, Amity University, Uttar Pradesh Lucknow Campus for constant support and encouragements.

REFERENCES

[1] I. K. Jordan; F. A. Kondrashov; I. A. Adzhubel; Y. I. Wolf; E. V. Koonin; A. S. Komdrashov; S. Sunyaev. *Letters to Nature*, **2005**, 433, 633.
[2] P. Johri; M. Gokhale, *Journal of Computational Biology*, **2013**, 2(1), 1.
[3] E. Rajasekaran, *Bioinformation*, **2012**, 8(11), 508.
[4] V. Jayaraj; R. Suhanya; M. Vijayasarathy; P. Anandagopu; E. Rajasekaran. *Bioinformation*, **2009**, 3(9), 409.
[5] K. Akila; P. Balamurugan; E. Rajasekaran. *Journal of Bioscience, Biochemistry and Bioinformatics*, **2012**, 2(2), 991.
[6] R. Senthil; E. Rajasekaran, *Advanced Biotech*, **2009**, 8(9), 30.
[7] V. Jayaraj. *Journal of Computational Intelligence in Bioinformatics*, **2009**, 2 (2), 99.
[8] E. Rajasekaran, *Bioinformation*, **2012**, 8 (11):508.
[9] E. Rajasekaran; M. Vijayasarathi, *Bioinformation*, **2011**, 5(10):455.
[10] A B. Jerald. *Annual International Conference on Advances in Biotechnology*, **2012**, 70(2),21.
[11] E. Rajasekaran. *Journal of Advanced Bioinformatics Applications and Research*, **2011**, 2(3),173.