Protein Characterization at atomic level: A Novel approach for sequence analysis

Parul Johri[1*], Mala Trivedi[1], Drishti Srivastava[1], Aman Kumar[1] Singh and Mohammed Haris Siddiqui[2]
[1]Amity Institute of Biotechnology, Amity University Uttar Pradesh,
Lucknow Campus, Malhaur, Gomti Nagar Extension, Lucknow, Uttar Pradesh, India
[2]Department of Bioengineering, Integral University Lucknow Uttar Pradesh, India
* Corresponding Author: pjohri@lko.amity.edu; +91 9838144680

Abstract

Carbon being the most omnipresent element in all the organic compounds is of great importance with regards to its structure and function. The fundamental structure of protein is composed of amino acids arranged in linear chain and folded to a globular form. All twenty amino acids consist of combination of only five different atoms that are Carbon, Nitrogen, Oxygen, Hydrogen and Sulphur. Depending on the property of their side chain the amino acids are classified as hydrophobic and hydrophilic. The allocation of the hydrophobic residue in a protein contributes majorly towards protein folding, protein interaction , active site formation and other biological functions. As carbon is the most important element which contributes to hydrophobic interaction in proteins, the hydrophobic amino acids characteristically have greater number of carbon atoms. In the present study, we have analyzed 4,306 protein sequences of *Escherichia coli,* a gram negative, facultative anaerobic, rod shaped bacterium. All the protein sequences of *Escherichia coli* were scanned to get a profound view on carbon content and its distribution. The sequences were retrieved from proteome section of UniprotKB (http://www.uniprot.org/uniprot/) database and atomic percentages were calculated with the aid of a dynamic programming algorithm and Microsoft Excel. The analysis of the atomic percentages calculated for the proteins revealed that there is a precise range of carbon percentage for all the *Escherichia coli* proteins (30%-33% carbon) .Also the different categories of protein like transport protein, repressor protein , catalytic protein, inhibitory protein etc have discrete range of carbon percentages which could be further linked to their exact activity

Keyword : *Carbon, E. Coli, dynamic programming, proteins.*

Introduction:

Proteins make almost 50% of the dry weight of the cells and are present in profound amount, than any other biomolecules. Proteins are basically organic compounds composed of amino acids arranged in linear chain and folded to a globular form (1). All 20 amino acids consist of combinations of only five different atoms Carbon, Nitrogen, Hydrogen, Sulphur, Oxygen. Depending on the properties of their side chains, the amino acids are classified as hydrophobic and hydrophilic (2; 3; 4). The distribution of hydrophobic residues in a protein contributes majorly towards protein folding, protein interactions, formation of core, active side formation and other biological functions (5).

Materials and Methods:

A hydropathy plot maps the hydrophobic regions against the hydropathy indices of the amino acids in the protein. It gives a clear idea of level of hydrophobicity in a protein. The hydrophobicity and carbon distribution profile of a protein is studied with help of a hydropathy plot (Fig. 1 and Fig. 2). In the carbon distribution profile obtained, it was possible to locate the hydrophobic, hydrophilic and also the active sites in the protein. The hydropathy plot does not give information on the active sites of a protein (6; 7; 8; 9). So it was postulated that the carbon distribution profile is a better alternative to hydropathy plot.

Escherichia coli, a gram negative, facultative anaerobic, rod shaped bacterium was the model organism used by us in our study. The entire project commenced by the retrieval of protein sequences of *Escherichia coli* from the public repository of protein sequences i.e UniprotKB database(http://www.uniprot.org/uniprot/). There were 4,306 protein sequences of *Escherichia coli(*strain K12) present in the proteome section of the database. Each and every sequence was scanned by us using the script written in perl

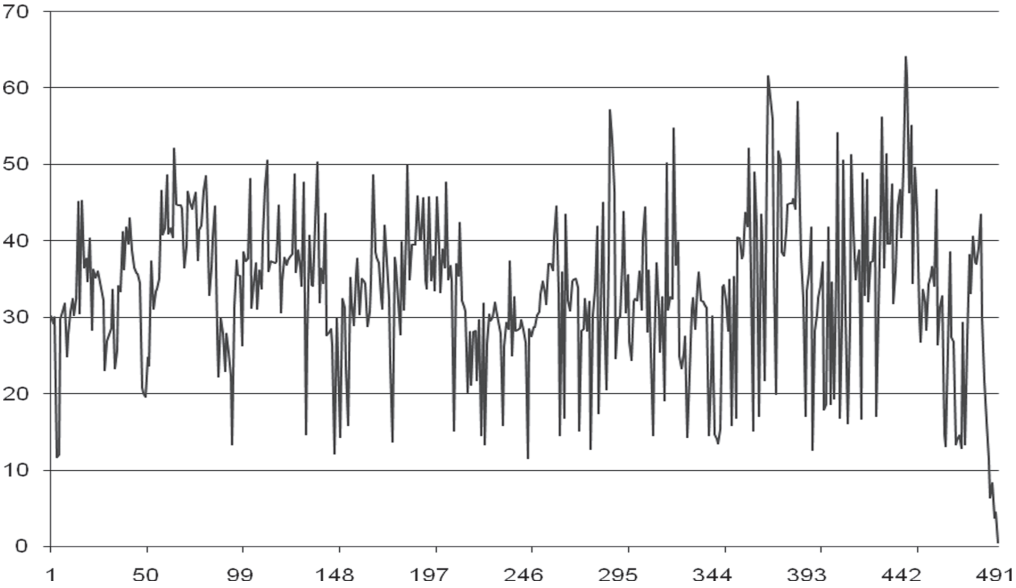

Fig. 1: Carbon distribution plot for GLUT1

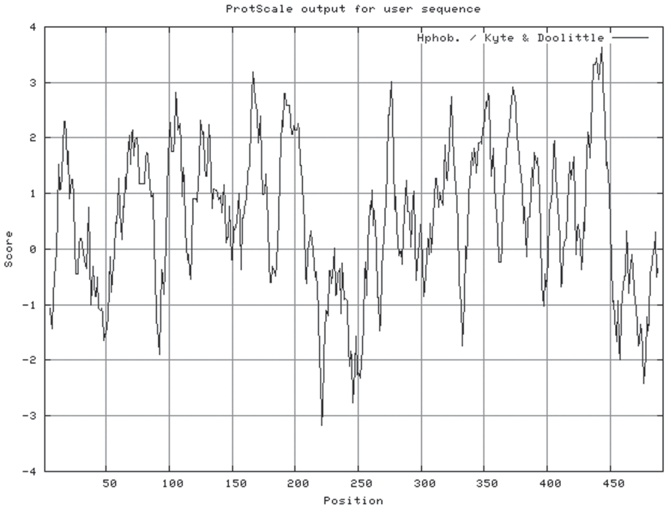

Fig. 2: Hydropathy plot from Protscale for GLUT1

programming based on the dynamic algorithm designed (Fig. 3) (10; 11; 12; 13). So the atomic composition of each and every protein sequence was extracted and were formulated into the excel sheet. After that the respective percentage of carbon was calculated on basis of number of carbon atoms and number of total atoms present in every protein sequences.

Proteins were also classified into broad categories on basis of their function like transport protein, catalytic protein, regulatory protein etc.

SPLIT THE SEQUENCES INTO ATOMS
|
TAKE WINDOW SIZE AS 10% OF ATOMS AND SLIDE BY 1 ATOM
|
CALCULATE THE C% OF EACH WINDOW
|
COUNT THE FREQUENCY FOR EACH WINDOW
|
DIVIDE THE FREQUENCY BY TOTAL NUMBER OF WINDOWS
|
WE WILL GET SOME VALUES FOR EACH C%
|
REPEAT THE ABOVE STEPS FOR ALL PROTEINS
|
CALCULATE AVERAGE
|
PLOT % C vs. FREQUENCY

Fig. 3: Flow chart for the dynamic programming algorithm

The scrutiny of the atomic percentages calculated for the proteins revealed that there is a specific range of carbon percentage for all the *Escherichia coli* proteins that is 30%-33%. Broad classification of proteins also showed that all proteins strictly followed the range.

Results and Discussion:

We got amazing results from our research work. Carbon percentages in all the protein sequences lied between 30%-33% no matter which types of amino acids where present in protein sequences (Table 1). So we can conclude that on the basis of carbon atoms, demarcation can be done. And after analyzing the carbon content in every proteins, percentages of other components can be taken out and studied for further uses. The atomic level analysis of the protein sequence does lead to a new venture of scrutiny of many long stranded questions.

The arenas of molecular biology, pharmacy, drug designing, enzymology, proteomics and many other fields including genetics and phylogenetic are wedged on certain facets which can possibly be answered by the carbon content analysis of protein sequences. The carbon level analysis of protein sites for post translational modification would lead towards new and profound ways of understanding them (14).

Carbon percentage in protein sequence would be the next parameter for prediction of secondary protein structure. Developments in microarray technologies, which would be based on the carbon content of protein, are required. Chromatography techniques which imply carbon as a parameter may also evolve. Drug development and designing must be implied for detection of active site and potential drugs based on carbon content. The development of tool for prediction of ligand binding site in globular proteins, based on average carbon percentage and distribution, was developed for mouse mitochondrial aspartate aminotransferase. This technique has huge potential to further develop for a large spectrum of proteins (15).

Table 1: Carbon percentages in various types of proteins in *E. coli*

S.No	Types of Protein	Range of Carbon (%)
1	Regulatory proteins	29.95%-32.32%
2	Catalytic Proteins	30.82%-32.97%
3	Transport protein	30.37%-32.15%
4	Enzymatic proteins	29.91%-33.65%
5	Transfer proteins	31.13%-32.58%

A new and fascinating concept is of **'Carbon bar coding'.** The concept is to make carbon as the stricture for width of bars. The protein sequence may be converted into its carbon bar code which would be universal. This presentation would be path breaking and very convenient size wise.

Acknowledgement:

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. Johri, P. and Gokhale, M. (2013). A New Perspective for Sequence Analysis – Carbon content. Research and Review: Journal of Computational Biology, 2:1-6.
2. Rajasekaran, E., Akila, K. and Vennila, J.J. (2011). Carbon Contents of H1N1 Proteins. International Conference on Biosciences, Biochemistry and Bioinformatics, 5:27-29.
3. Akila, K. and Rajasekaran, E. (2009). What Might be the Difference in Viral Proteins? International Journal of Bioinformatics Research,1:1-3.
4. Rajasekaran, E., Asha, J. and Klaus, H. (2012). Magnitude of Thymine In Different Frames of Messenger RNAs. International Journal of Bioinformatics Research, 4:273-275.
5. Senthil, R., Sathish, S., Vennila, J.J. and Rajasekaran, E. (2011). Prediction of Ligand Binding Site in Globular Proteins. Journal of Advance Bioinformatics Application and Research, 2:98-99.
6. Akila, K., Kaliaperumal, R. and Rajasekaran, E. (2012).Carbon Distribution to Toxic Effect in Toxin Proteins. Bioinformation Discussion at the Interface of Physical and Biological Science, 8:720-721.
7. Rajasekaran, E. and Vijayasarathy, M. (2011). CARBANA: Carbon Analysis Program for Protein Sequences. Bioinformtaion, 5:455-457.
8. Rajasekaran, E. (2012) CARd: Carbon Distribution Analysis Program for Protein Sequences. Bioinformation, 8:508-512.
9. Chase, M.W. and Fay, M.F. (2009). Barcoding of Plants and Fungi. Science, 325:682-683.
10. Nsimama, P.D., Mamboya, A.F., Amri, E. and Rajasekaran, E. (2012). Corelation Between The Mutated Colour Tunings and Carbon Distributions in Luciferase Bioluminescence. Journal of Computational Intelligence In Bioinformatics, 5:105-112.
11. Rajasekaran, E., Rajadurai, M., Vinobha, C.S. and Senthil, R. (2008). Are The Proteins Being Hydrated During Evolution ? Journal of Computational Intelligence In Bioinformatics,1:115-118.
12. Akila, K., Sneha, N. and Rajasekaran, E. (2012). Study of Carbon Distribution at Protein Regions of Disorder. Journal of Bioscience, Biochemistry and Bioinformatics, 2:58-60.
13. Johri, P., Trivedi, M. and Siddiqui, M.H, and Gokhale, M. (2016). A Study On The Presence and Distribution of Carbon Percentage in and Around the Sites of Glycosylation for Eukaryotic Proteins. Journal of Chemical and Pharmaceutical Research, 8:52-63.
14. Johri, P., Trivedi, M., Singh, A. and Siddiqui, M.H. (2016). Towards The Atomic Level Protein Sequence Analysis. Journal of Chemical and Pharmaceutical Research, 8:204-207.
15. Johri, P. (2013). Atomic Level Sequence Analysis- a Review. International Journal of Computational Bioinformatics and Insilco