Carbon Distribution Accounts a Lot for Patterns in Proteins

E.Rajasekaran* and K.Sheeba

Department of Bioinformatics, School of Biotechnology and Health Sciences, Karunya University, Coimbatore – 641114*,*
Tamil Nadu, India.
**Corresponding author:sekaran@karunya.edu*

Abstract

Background: Pattern is a set of amino acids forms a unit that governs protein structure and activity. So many developments have been made in retrieving these patterns using computational approach. From our recent findings it was concluded that carbon plays an important role in structure formation and binding.

Method: Carbon distribution studies on these patterns are carried out using CARd program. The CARd outputs are plotted for comparison and discussed.

Conclusion: The results reveal that carbon plays an important role in pattern formation. Carbon distribution based patterns account better over the amino acid patterns. That is different amino acid sequences are grouped as same pattern but in carbon distribution. In fact it is hoped for the refinement of pattern databases.

Keywords: Carbon distribution, CARd, Conserved sequences, Domain, Motif, Pattern, Sequence analysis

1. Introduction

In the last two decades there has been considerable work gone into proteins to understand the structure, function and organization. Lots of computational studies leading to reveal the fact behind protein structure formation and relevant activity. In order to understand this buried information in the protein structures, the molecules were visualized at atom level (Rajasekaran, 2012). Carbon is one of the important atoms which determine the protein evolution. In order to maintain carbon content and distribution, the amino acids are organized in some form along the sequences and its blue prints are there in mRNAs, genes and in DNA. Once protein follows the carbon distribution principle, it has become regular synthesis in cells. Otherwise the protein is eliminated from regular synthesis. When the carbon content increases at active sites, the reduction take place at the nearby stretches accordingly. Portion of protein stretches undergo unfolding (minimum carbon site) or form hydrophobic core (carbon rich region) due to carbon distribution. It is hoped that the patterns in proteins form based on carbon distribution. That is the patterns have normal carbon distribution pattern. A stretch of sequence that has this prescribed carbon distribution is considered as a pattern. Some of the reported patterns (Sigrist *et al.,* 2010) are investigated here.

Base distributions in nucleic acids are crucial for translation of proteins with right carbon distribution and passed to next generation. The thymine distribution in particular is important for this purpose. Results reveal that thymine in protein coding sequences are not randomly distributed but with probability (Rajasekaran & Akila, 2011; Anandagopu *et al.,* 2008).

Table 1: *Reference patterns taken from PROSITE for carbon distribution study*

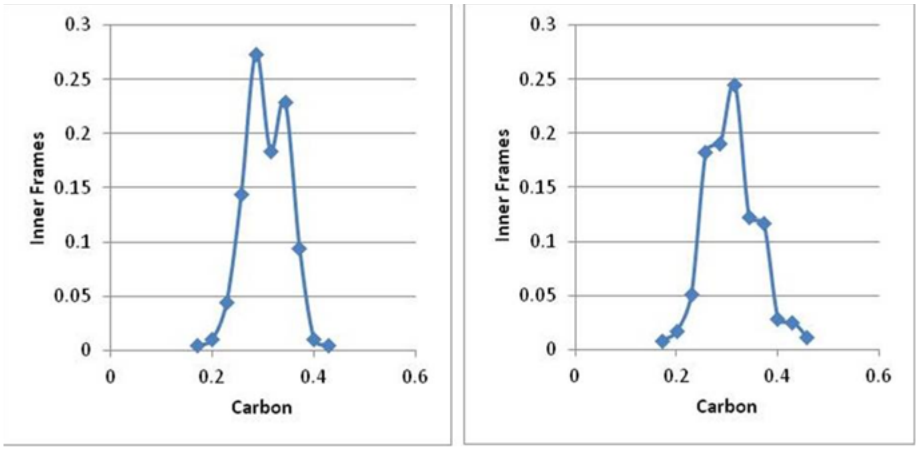
Pattern Name	Pattern	PrositeID	Reference Protein
Serine/Threonine Protein Kinases Active Site Signature	[L I V M F Y C] - { A } - [H Y] - x - D - [L I V M F Y] - [R S T A C] - { D } - { P F } - N - [L I V M F Y C] (3)	PS00109	YVHRDLAARNVLV 999 – 1011 in Human Tyrosine-protein kinase JAK1 **(**P23458**)**
LBP / BPI / CETP family signature (Available to all)	[P A] - [G A] - [L I V M C] - x (2) - R - [I V] - [S T] - x (3) - L - x (5) - [E Q A V] - x (4) - [L I V M] - [E Q K] - x (8) - P	PS00400	PGIKARITQRALDYGVQAGMKMI 28- 50 in Human BPI fold-containing family C protein (Q8NFQ6)

2. Method

The following pattern details and sequence are taken from prosite (Sigrist *et al.,* 2010) pattern database. For entry PS00109, there are more than 500 crystal structures available in PDB. Carbon distribution investigations are done using CARd program (Rajasekaran, 2012). The inner and outer lengths of 35 and 255 are generally used in all calculation. Here the outer window length of 200 and 350 are used for sequence 1 (PS00109) and sequence 2 (PS00400) in order to cover entire length of patterns. That is the pattern lengths are 13aa and 23aa respectively. The CARd outputs are plotted for comparison and discussed.

3. Results and Discussion

Figure 1*: Carbon distribution plot for patterns, PS00109 and PS00400.*



The carbon distribution plots for both example patterns are plotted in figure 1. The plot is a carbon fraction versus amount of inner frames. That is the fraction of inner frames in different carbon compositions is shown. Normally it is expected to a normal distribution curve for a stable stretch. Both patterns show close to normal curve. It also expected that the distribution maximum is at 0.3145 for normal hydrophobic character. If it is greater than 0.3145, it is a hydrophobic region. If shift is in lower side, then hydrophilic in character. Here both patterns exhibit maximum at 0.3145. Again it is stable. From this it can be observe that both distribution curve and maxima are favor towards stability. That mean carbon plays an important in pattern formation. The average, statistical mean, median and mode of this carbon distribution curve are computed. The values are given in table 2.

Table 2*: Statistical values of carbon distribution curve.*

Average C	Statistical Mean	Median	Mode
0.3000	0.3059	0.3143	0.2857
0.3029	0.3080	0.3143	0.3143

The median seem to be close to the expected value. The average and statistical mean under value here in these patterns. Mode generally expected at 0.3145.In pattern1, it is exception. Rectifying these under value might further strengthen the pattern stability. There are sequences of these patterns with different amino acid composition might be following it. In fact the different combinations and permutation given in patterns need not follow this carbon distribution, except those experimentally available sequences. Based on the carbon distribution curve, one can retrieve all available patterns accurately. The large hydrophobic residues are reduced during evolution in animals (Jayaraj *et al.,* 2009). It is a major worrisome in human. In order to maintain the carbon content in place of large hydrophobic residues, the small hydrophobic residues are added along the sequences (Vinobha & Rajasekaran, 2011). Because of these reason the patterns remain same but different amino acid combinations. That is the functionality is maintained though different protein sequences. The corresponding mRNAs are also altered accordingly and passed to next generation (Rajasekaran & Anandagopu, 2010).

4. Conclusion

Carbon distribution based investigation on known patterns reveals that carbon plays an important role in pattern formation. These patterns are vital for understanding the structure, activity and evolution of proteins. Carbon distribution (CARd) program can be better utilized for pattern retrieval. This method is hoped for the refinement of pattern databases. Large hydrophobic residues are the major contributors for carbon in proteins. As part of evolution the large hydrophobic residues are reduced in many proteins. It is very much concern in animals and in particular, the human. In order to maintain the carbon content in those places of large hydrophobic residues, other small hydrophobic residues are added along the sequences. Because of these reason the patterns remain same but different amino acid combinations. That is the functionality is maintained though different protein sequences. The corresponding mRNAs altered accordingly and passed to next generation.

5. References

1• Anandagopu P, Suhanya S, Jayaraj V, Rajasekaran E (2008). Role of thymine in protein coding frames of mRNA sequences, Bioinformation, 2:304-307.

2• Jayaraj V, Suhanya R, Vijayasarathy M, Anandagopu P, Rajasekaran E (2009). Role of large hydrophobic residues in proteins, Bioinformation, 3:409-412.

3• Rajasekaran E (2012). CARd: Carbon distribution analysis program for protein sequences, Bioinformation, 8:508-512.

4• Rajasekaran E, and Akila K (2011). Adenine in viral mRNAs manipulate the carbon in proteins, International Journal of Bioscience, Biochemistry and Bioinformatics, 1:249-252.

5• Rajasekaran E, and Anandagopu P, (2010). Reduction of thymine in mRNA sequence of tumor protein, J. Advanced BioTech, 9:9-10.

6• Sigrist CJA, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N. (2010). *PROSITE, a protein domain database for functional characterization and annotation,* Nucleic Acids Res, 38:161-166.

7• Vinobha CS, and Rajasekaran E, (2011) Comparative analysis on large hydrophobic residues and small hydrophobic residues in different organisms, International Journal of Bioinformatics Research, 3:115-117.