*Bioinformation* **Volume 5** *open access*

*www.bioinformation.net* Issue 10 **Software**

# CARBANA: Carbon analysis program for protein sequences

**Ekambaram Rajasekaran*, Marimuthu Vijayasarathy**

Department of Bioinformatics, School of Biotechnology and Health Sciences, Karunya University, Coimbatore – 641114 Tamil Nadu, India;
Ekambaram Rajasekaran - Email: ersekaran@gmail.com; *Corresponding author

Received November 01, 2010; Accepted January 29, 2011; Published February 15, 2011

**Abstract:**
There are lots of works gone into proteins to understand the nature of proteins. Hydrophobic interaction is the dominant force that drives the proteins to carry out the biochemical reactions in all living system. Carbon is the only element that contributes towards this hydrophobic interaction. Studies find that globular proteins prefer to have 31.45% of carbon for its stability. Taking this as standard, a carbon analysis program has been developed to study the carbon distribution profile of protein sequences. This carbon analysis program has been made available online. This can be accessed at www.rajasekaran.net.in/tools/carbana.html. This new program is hoped to help in identification and development of active sites, study of protein stability, evolutionary understating of proteins, gene identification, ligand binding site identification, and to solve the long-standing problem of protein-protein and protein-DNA interactions.

**Keywords:** carbon distribution; CARBANA analysis; hydrophobicity; carbon profile; hydropathy plot;

**Background:**
There is lot of work gone into proteins to understand the ultimate truth of real information **[1-3].** Hydrophobic interaction is the dominant force that comes from presence of carbon. Recent studies reveal that proteins prefer to have 31.45% of carbon in its structure and in sequence **[2]**. To understand the buried information further in proteins this work has been taken up.

**Methodology:**
The idea behind this method is visualising the molecule on actual basis. That is the basic units of proteins are elements such as carbon, sulphur, nitrogen, oxygen and hydrogen. In this method the amino acid sequences are converted into atomic sequences. Example is given in supplementary material.

It is also hoped that a protein sequence with 100 amino acids should have about 1555 atoms in the atomic sequence. Further the percentage of carbon in the first 500 atoms are computed and marked as carbon percentage at the point of 250. Residue number that carries the 250th atom is taken as reference point. Next the group of 500 atoms is taken from 5 to 505. Again the carbon percentage computed is assigned to reference point of 255. This way by a shift of 5 atoms all 500 groups are computed with carbon percentage and assigned to corresponding reference point. This shift by 5 atoms can increased or decreased depends upon the resolution required. Similarly the window length 500 atoms (~32 amino acids) can changed for different calculations. A plot of carbon percentage versus the reference point is plotted to indentify the carbon distribution profile along the sequence. A C program has been written to carry out all these calculation. A sample input, output **(Table 1 see Supplementary material)** and plot **(Figure 1)** are given and discussed.

**Discussion:**
The program reads protein sequences and converts it into array of elements. The percentage of carbon is computed for a group of atoms is assigned to reference point residue. Normally the shift value of 5 is used. It can be increased or decreased depends upon the resolution required. Reduction in shift value creates too many points and makes the plot congested. A shift value of 17 may be optimum. This value is half of the smallest unit (35 atoms) that is producing 31.45% of carbon. Further improvements in having all amino acids (including first and last 17 residues) represented in the output and in figures are underway. Also the computation of carbon percentage at alpha carbon position will be implemented for mutational studies and for other applications.

There is window length of 500 atoms taken for carbon percentage calculation **(Figure 1).** This value may be increased or decreased depends upon required resolution. This can be from 35 to 1000 atoms length. The 35 atom length is chosen because the smallest unit which can produce 31.45 is 35 with 11 carbons in it. Carbon accumulation in active site or in core can be easily identified at length of 500. So by default a length of 500 atoms is taken for general carbon profile study. To identify the residue contributing to the stabilization or destabilization factors, one can reduce this length. For mutational study a length value of 50 atoms may be appropriate. A sample input and output are given below for length of 500 atoms and shift size of 17 atoms.

**Input:**
>gi|110833718|ref|YP_692577.1| hypothetical protein ABO_0857 [Alcanivorax borkumensis SK2]

```
MRHVMKRKATTLMATAISALILSGCGGEQAATPVSGIEPKVYTDSL
FAVMNADRTNYTKLIIGRLGPAGADSIKPHEYWEDLENGAPLPAQ
MFRYGAESVSEMTSEFSYSLQSLWPINGQNEPKTGLEKEGLQYIVD
NPGENFYGEEKLGDVTYYTAVYPDVAVAAPCVACHNNHKDSPKT
DFELGDVMGGVVIRVPM
```

So the input is protein sequence and the output is the residue number and corresponding carbon percentage. This output can be plotted in XY plot for better visualisation as shown in **Figure 1.**

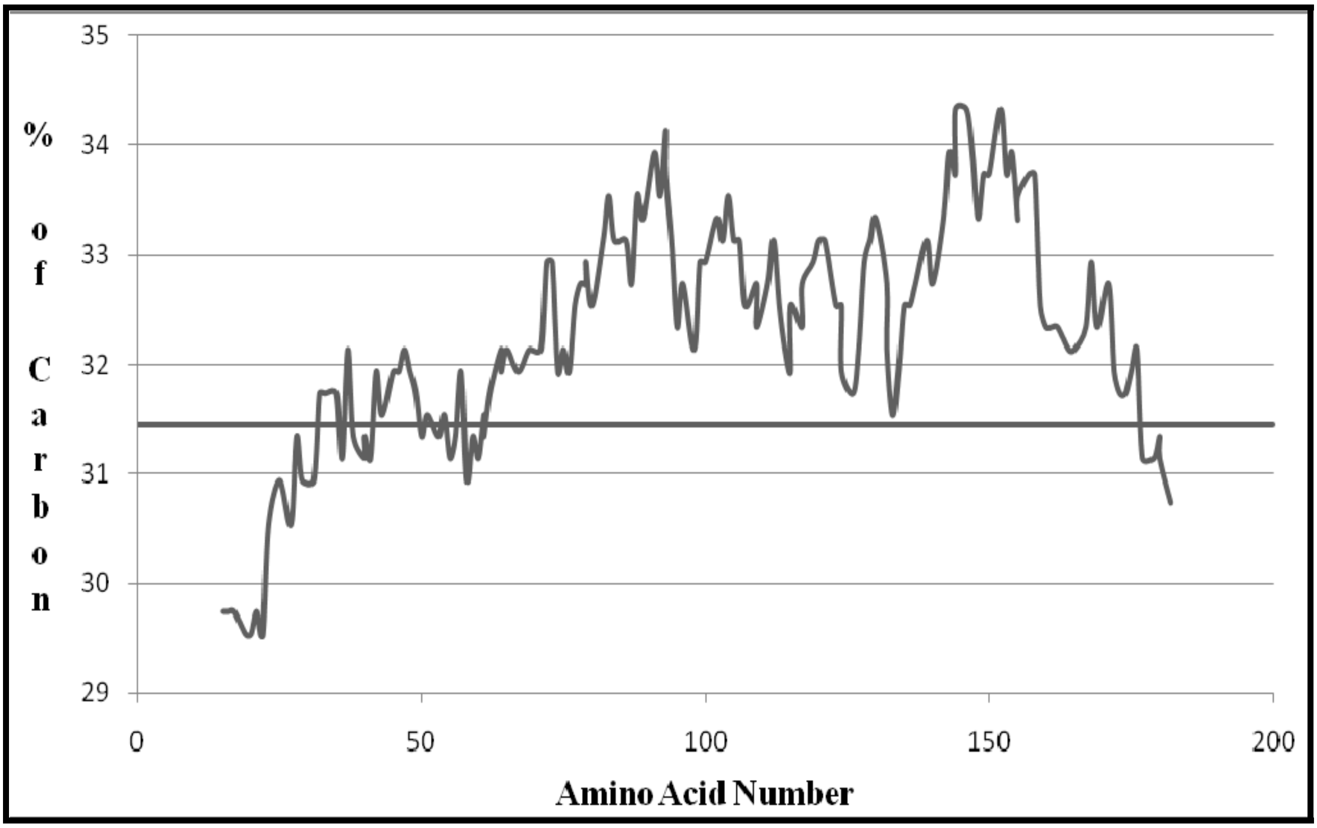

**Figure 1:** Plot of Carbana output. Points above 31.45% are hydrophobic regions.

**Conclusion:**

Carbon profile analysis software [CARBANA] has been developed and presented here. This program is capable of locating the carbon accumulated site in proteins. It can clearly identify the hydrophobic and hydrophilic regions along the sequence. It can also pinpoint an amino acid which is causing instability. Atomic level representation of proteins can yield better results. This carbon analysis program is available online. This new program is hoped to address several biological problems based on hydrophobicity. Particularly, it can help in identification and development of active sites, address the proteins in diseased and healthy state, characterize the disordered proteins, address the role of carbon in half of proteins and understand patterns and repeats in proteins.

**References:**

**[1]** V Jayaraj *et al. Bioinformation* (2009) **3:** 409 [PMID: PMC2732037]
**[2]** E Rajasekaran *et al. IACSIT-SC, IEEE* (2009) 452
**[3]** E Rajasekaran *et al. J Comput Intelli. Bioinfo* (2008)**1**:115

**Edited by IR Konieczna,**
**Citation: Rajasekaran & Vijayasarathy,** Bioinformation 5(10): 455-457 (2011)

## Supplementary material:

For example sequence MATAISALIVE… etc. is converted into atomic sequence as follows.

CCCCCSNOHHHHHHHHHHCCCNOHHHHHCCCCNOOHHHHHHHHCCCNOHHHHH…etc
M A T A …etc

**Table 1**: The output of the Carbana program showing amino acid number and percentage of carbon

| 15 | 29.7405 | 41 | 31.1377 | 64 | 32.1357 | 88 | 33.5329 | 112 | 33.1337 | 134 | 31.9361 | 158 | 33.7325 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 16 | 29.7405 | 42 | 31.9361 | 64 | 31.9361 | 89 | 33.3333 | 113 | 32.5349 | 135 | 32.5349 | 159 | 32.5349 |
| 17 | 29.7405 | 43 | 31.5369 | 65 | 32.1357 | 91 | 33.9321 | 114 | 32.1357 | 136 | 32.5349 | 160 | 32.3353 |
| 19 | 29.5409 | 45 | 31.9361 | 67 | 31.9361 | 92 | 33.5329 | 115 | 31.9361 | 137 | 32.7345 | 161 | 32.3353 |
| 20 | 29.5409 | 46 | 31.9361 | 69 | 32.1357 | 93 | 34.1317 | 115 | 32.5349 | 139 | 33.1337 | 162 | 32.3353 |
| 21 | 29.7405 | 47 | 32.1357 | 71 | 32.1357 | 93 | 33.7325 | 117 | 32.3353 | 140 | 32.7345 | 164 | 32.1357 |
| 22 | 29.5409 | 48 | 31.9361 | 72 | 32.9341 | 94 | 33.1337 | 117 | 32.7345 | 142 | 33.3333 | 165 | 32.1357 |
| 23 | 30.5389 | 49 | 31.7365 | 73 | 32.9341 | 95 | 32.3353 | 119 | 32.9341 | 143 | 33.9321 | 167 | 32.3353 |
| 25 | 30.9381 | 50 | 31.3373 | 74 | 31.9361 | 96 | 32.7345 | 120 | 33.1337 | 144 | 33.7325 | 168 | 32.9341 |
| 27 | 30.5389 | 51 | 31.5369 | 75 | 32.1357 | 98 | 32.1357 | 121 | 33.1337 | 144 | 34.3313 | 169 | 32.3353 |
| 28 | 31.3373 | 53 | 31.3373 | 76 | 31.9361 | 99 | 32.9341 | 123 | 32.5349 | 146 | 34.3313 | 171 | 32.7345 |
| 29 | 30.9381 | 54 | 31.5369 | 77 | 32.5349 | 100 | 32.9341 | 124 | 32.5349 | 147 | 33.9321 | 172 | 31.9361 |
| 31 | 30.9381 | 55 | 31.1377 | 78 | 32.7345 | 102 | 33.3333 | 124 | 31.9361 | 148 | 33.3333 | 173 | 31.7365 |
| 32 | 31.7365 | 56 | 31.3373 | 79 | 32.7345 | 103 | 33.1337 | 126 | 31.7365 | 149 | 33.7325 | 174 | 31.7365 |