Journal of Computational Intelligence in Bioinformatics
ISSN: 0973-385X Volume 2 Number 2 (2009) pp. 99–102

http://www.ripublication.com/jcib.htm

Pattern Recognition in Proteins Based on Carbon Content

Veerasamy Jayaraj[1], Marimuthu Vijayasarathy, Rajendran Geerthana, Renganathan Senthil and Ekambaram Rajasekaran*

Departments of Biotechnology and [1]Computer Application, Periyar Maniammai University, Thanjavur – 613403, Tamil Nadu, India.
**E-mail: ersekaran@gmail.com*

Abstract

A novel pattern based on carbon content is reported here. The principle behind this pattern is that proteins prefer to have 31.44% of carbon in its structure for stability. A stretch of sequence that has this prescribed carbon content is considered as a pattern. The smallest pattern identified is 7 residues length that contain 50 carbon and 159 total atoms. Patterns with same number of carbon and total atoms are found in different lengths and sequence. The 3D structures of these patterns are evaluated. The patterns prefer to be in folded and compact 3D structure. These patterns in protein are vital for understanding the proteins stability and for sequence alignment. This newly identified pattern encourages a new way of sequence comparison at atomic level. This is because of patterns of same atomic profile differs in sequence and length.

Keywords: Carbon content; Patterns; Sequence analysis; Protein stability; Protein comparison.

Introduction

Proteins are large organic compounds made of amino acids arranged in a linear fashion. The side chains of these amino acids are chemically different from one another in some respect that can be classified broadly in two ways i.e., hydrophobic and hydrophilic. Atomic details in these side chains make the amino acid different. The atoms include carbon, nitrogen, oxygen, sulphur and hydrogen. Carbon is the only atom contributes towards the hydrophobic interaction. These carbons added to proteins by largely due to large hydrophobic residues such as F, I, L, M and V. The contribution of these large hydrophobic residues in protein is understood so far [1-5]. As a next step the carbon content analysis in the proteins reveals that the proteins

prefer to have 31.44% of carbon along the chain and in total. This study finds a stretch of sequences, the so called patterns which has 31.44% of carbon.

Methodology

Using home made program (PFIND) all available patterns in a given protein is searched and retrieved. That is given the carbon content of 31.44%; different patterns of different length are retried. The sequences are read in Fasta format and assigned atomic details to each amino acids. For example alanine has 3 carbon atoms and 10 total atoms. For a given stretch of sequence the ratio between number of carbon and total atoms are calculated. The ratio matches with 0.3144 then taken as pattern.

Results and Discussion

From our earlier studies on proteins it is found that proteins prefer to have 31.44% carbon in its structure and all along the sequence to maintain stability. Here the patterns of different length and sequence that contain 31.44 carbons are retrieved and analysed. The smallest pattern identified is 7 amino acids long and contain 50 carbon atoms. The patterns that contain the 50 carbon atoms and 159 total atoms are retrieved and listed in table1. These patterns contain same atomic profiles but differ in length and sequence. Note the length variation from 7 to 16 amino acids. In a sequence alignment programs these patterns may appear as mismatches but atomically same. It is suggested that atom based alignment might yield better results than the sequence level comparison.

Table1: Patterns with same number of carbon (50 no.) and total atoms (159 no.) 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

Patterns with same length and different sequence that contain 31.44% of carbon are given here.

IWTARKIVS; NIKSELKYV; LIFSKMKET; RNLIYLATI; VSVWSKVLR; KALQERDYI; KEFRKPSDL; RYELAQQLQ; RGRVISLWE; MFNQLMKQV.

These patterns are similar in atomic sense but shows dissimilarity in the sequence alignment. This again raises a question on the sequence alignment programs. Atomic level representation of protein sequences might yield better results than that of residue level.

Further, it is observed in lengthier patterns that for same length of patterns with 31.44% of carbon a difference of 11 carbon and 35 total atoms are observed. The table 2 illustrates this. That is the smallest number of carbon and total atoms, which will give 0.3144 is 11/35.

Table2: Patterns with same length with different number of carbon and total atoms.

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

This newly identified patterns raises question on sequence alignment programs and vital for understanding the protein stability and for sequence comparison. The charges and atom types in these patterns are not taken into account. The 3D structures of these patterns are evaluated. The patterns prefer to be in folded and compact 3D structure.

4 Conclusion

A novel pattern based on carbon content is reported here. That is a stretch of sequence that has 31.44% of carbon considered as a pattern. The smallest pattern identified is 7 residues long that contain 50 carbon and 159 total atoms. Patterns with same carbon atoms and total number of atoms are found with different length. Similarly patterns of same length with different amino acid sequences show a same atomic profile. Though it is similar atomically, it shows dissimilarity in the sequence alignment. This needs to be taken into account in protein comparison. Atomic level representation of protein sequences might yield better results than that of residue level. The 3D structures of these patterns reveal that the patterns prefer to be in folded and compact structure.

These patterns in protein are vital for understanding the proteins stability and for comparison.

5 References

[1] P.Anandagopu, S.Suhanya, V.Jayaraj and E.Rajasekaran, Role of thymine in protein coding frames of mRNA sequences, Bioinformation, (2008) **2:** 304

[2] V.Jayaraj, R. Suhanya, M. Vijayasarathy, P. Anandagopu and E. Rajasekaran, Role of large hydrophobic residues in proteins, Bioinformation (Under revision).

[3] E.Rajasekaran, M.Rajadurai, C.S.Vinobha and R.Senthil, Are the proteins being hydrated during evolution, J. Comp. Intelligence. in Bioinformatics, (2008) **1**: 115.

[4] M.V.Katti, R.Sami-Subbu, P.K.Ranjekar, and V.S.Gupta, Amino acid repeat patterns in protein sequences: Their diversity and structural-functional implications, Protein Sci., *(*2000) **9**: 1203.

[5] P.Baudouin-Cornu, Y.Surdin-Kerjan, P.Marlière and D.Thomas, Molecular evolution of protein atomic composition**,** Science, (2001) **293**: 297.