ACADEMIA ROMÂNĂ

Revue Roumaine de Chimie

http://web.icf.ro/rrch/

Rev. Roum. Chim.,
2012, *57*(1), 45-48

USING CHAOS GAME REPRESENTATION FOR SIMILARITY STUDIES ON ALBUMIN AMINO ACIDS SEQUENCES

Cristina STAN,[a,*] Constantin P. CRISTESCU[a] and Dana Andreea NEACSU[b]

[a] Department of Physics I, Faculty of Applied Sciences, Politehnica University of Bucharest, 313 Spl. Independentei, Roumania
[b] "Ilie Murgulescu" Institute of Physical Chemistry of the Roumanian Academy, 202 Spl. Independentei, P.O.Box. 12-194, 060021 Bucharest, Roumania

Received March 25, 2011

In this work we propose a relatively simple method of emphasizing similarity between albumin protein sequences from different organisms, based on chaos game representation. The degree of similarity is established by comparison of the distribution of points in the chaos game patterns generated from the amino acids sequences.

INTRODUCTION

Similarity studies on biological sequences are given high attention in recent literature.[1,2] Serum albumins found in blood plasma are often known as transport proteins since they act as carriers for numerous exogenous and endogenous compounds.[3,4] Studies on the chemical and optical properties of albumin and on the interaction of different acids with such protein are of great interest in medical and pharmaceutical applications since the amino acid compositions of the albumin proteins, in particular human and bovine albumin are very similar.[5-7]

In this paper we propose a simple method of identifying similarity between albumin protein sequences from different organisms, using the Chaos Game Representation (CGR).[8] The degree of similarity is defined on the basis of the distribution of points in the CGR patterns generated from the Amino Acid (AA) sequences. The twenty different kinds of amino acids are divided into four classes, according to the HP classification (non-polar, negative polar, uncharged polar and positive polar).[9] By this procedure the protein sequence is transformed into a four letter alphabet sequence.

MATERIAL AND METHOD

Fig. 1 presents the general structure of an amino acid. It consists of a central (C_α) carbon atom to which the following groups are attached: (1) an amino group (NH_3^+); (2) a carboxyl group (COOH); (3) a hydrogen atom (H); and (4) a radical group, designated R. The groups (1-3) are common to all 20 amino acids and the parts of the R group are specific to each one.

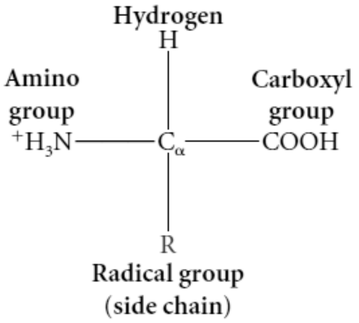

Fig. 1 – The general structure of an amino acid.

We consider the coding sequence of albumin amino acid sequence for eight different species: Homo sapiens (human) albumin [EMBL-Bank: CAA00606], Bos taurus (cattle) serum albumin

[*] Corresponding author: cstan@physics.pub.ro

[EMBL-Bank: AAN17824.1], Felis catus (cat) [EMBL-Bank: CAA59279], Rattus norvegicus (rat) [EMBL-Bank: AAH85359], Macaca mulatta (monkey) serum albumin [EMBL-Bank: AAA36906], Canis lupus familiaris (dog) [EMBL-Bank: CAB64867], Sus scrofa (pig) albumin [EMBL-Bank: AAT98610] and Oryctolagus cuniculus (rabbit) serum albumin [EMBL-Bank: AAB58347].[10]

METHODS AND ALGORITHM OF CGR

The CGR is obtained in a square [0,1]x[0,1], where the four vertices correspond to a conventional denomination. For example, let us consider a sequence of four letters. The corners of the square are denoted by the four letters of the alphabet. In the CGR plot, the first letter of the sequence of interest is plotted as a point halfway between the center of the square and the vertex representing this letter; the second one is plotted by a point halfway between the vertex denoted as the second letter and the previous point, etc.[3] As the process continues, to each new step corresponds a division of the accessible area by 2×2 factor. After *n* successive steps the unit square is divided into $2^n \times 2^n$ square cells. Consequently, the number of points in a particular cell of a $2^n \times 2^n$ partition represents the frequency of a selected "*n*-letter word" in the whole analyzed sequence.

In the case of biological sequences, the chaos game pattern will consist of a number of points equal to the number of amino acids in the sequence.

The present investigation is based on the HP (Hydrophobicity-Polarity) model. In this model, the 20 amino acids are associated in the following four groups: the first group includes A, I, L, M, F, P, W, V, the second includes D and E, the third one consists of N, C, Q, G, S, T and Y and the fourth the remaining four amino acids R, H and K. In this work, we denote the four groups by Λ, Π, Σ and Ω respectively.

The abundance of the various amino acids according to the HP classification in the albumins of the eight organisms analyzed in this study is shown in Fig. 2. We observe that the four groups are not uniformly represented in any of the sequences of interest, however, the frequencies in each group for different albumins are distributed in a small interval.

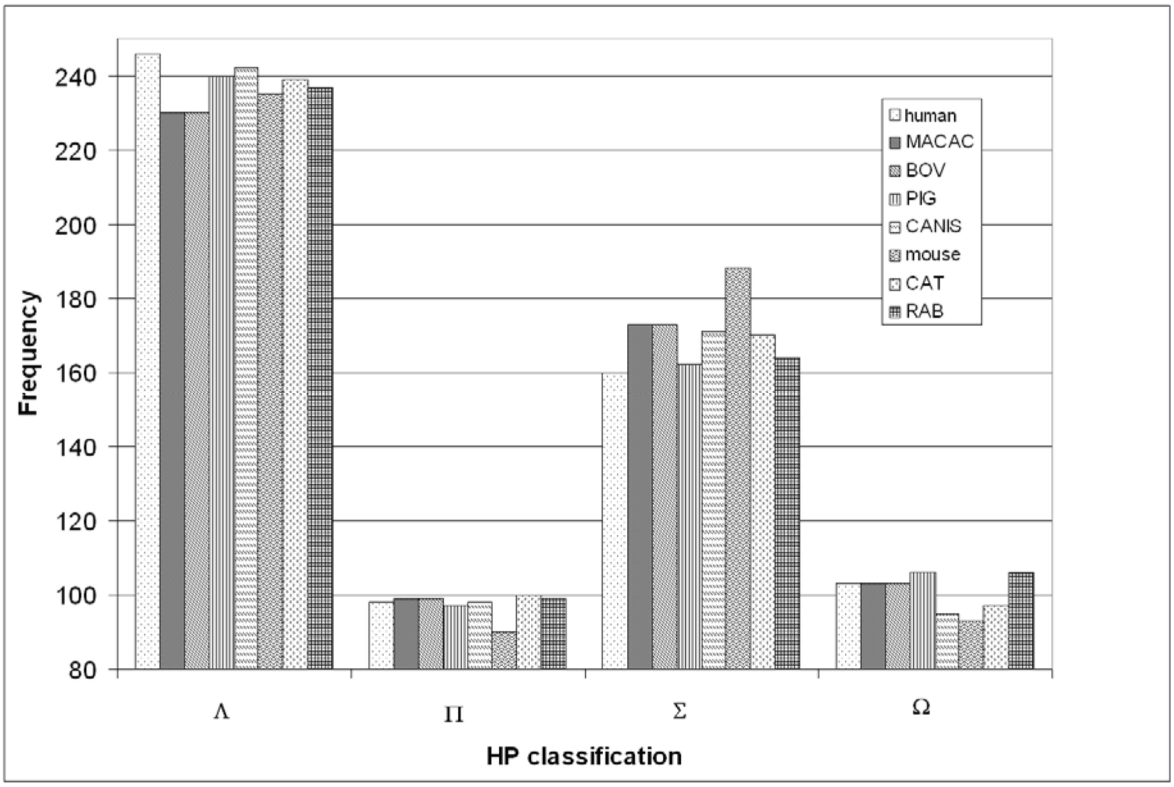

Fig. 2 – The frequency of the amino acids according to the HP classification for the sequences of interest as shown in the legend.

The number of points in a particular cell represents the frequency of the corresponding word in the sequence.[11] We generate a matrix whose elements are the frequencies associated to various n-letter words in the sequence. The similarities are emphasized by the comparison between the matrices of the AA sequences of albumin from different organisms. Considering the human albumin as reference seven other organisms are classified based on a newly defined similarity measure.

By subsequently counting the points in each cell, we construct a square $2^n \times 2^n$ matrix; the frequency matrix. Next, the difference matrix D_n of any two sequences to be compared is computed as:

$$(d_n)_{jk} = (a_n)_{jk} - (b_n)_{jk} \qquad (j,k = 1,\cdots,2^n) \qquad (1)$$

where $(a_n)_{jk}$ and $(b_n)_{jk}$ are the elements of the matrices A_n and B_n associated to the sequences to be compared. The degree of similarity is quantified by means of the difference of word frequencies between two different albumin sequences.

In this study we present results of the analysis of 3-letter words (e. g. "ΛΠΣ", "ΛΣΩ", "ΠΣΩ", etc.). A 3-dimensional representation of the difference frequency matrix gives a local as well as a global visualization allowing a quick estimate of the similarity/dissimilarity of the two sequences.

This is shown in Fig. 3 for four cases.

It is to be observed that high values of the elements of the difference matrix occur in the neighborhood of the corners representing the groups with the greater abundance, in agreement with the situation shown in Fig. 2. The height of the bars, Δ*N* (which generically denotes the elements of the difference frequency matrix *D*$_n$) can be either positive or negative. In order to show the sign, we marked the $\Delta N = 0$ contour by a solid line. The plot in Fig. 3 corresponds to the following situations: (a) human-monkey; (b) human-cattle; (c) human-rat and (d) human-rabbit.

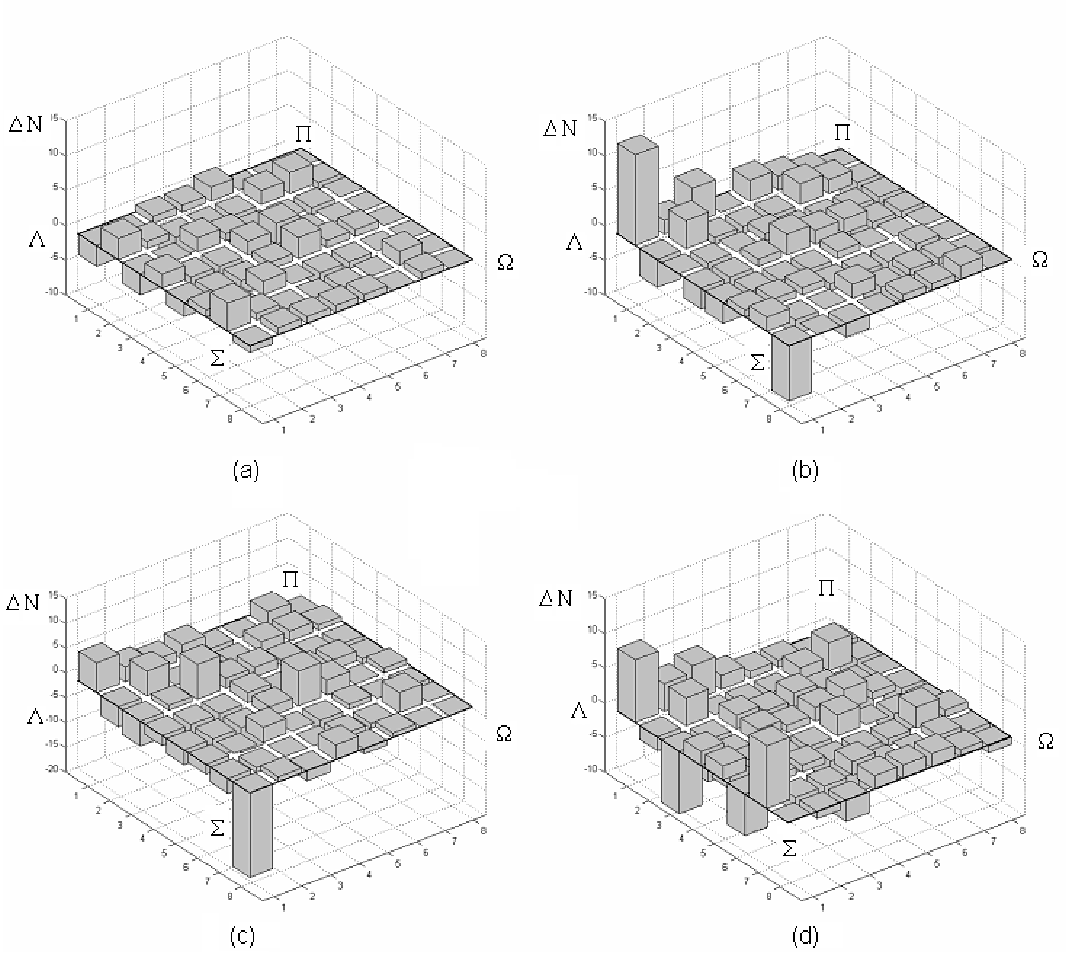

Fig. 3 – 3D representation of the difference of the frequency matrices:
(a) human-monkey; (b) human-cattle; (c) human-rat; (d) human-rabbit.

By visual inspection it is observed that the highest global similarity corresponds to the human-monkey albumin and the highest dissimilarity is presented by the human-rabbit case. The cases (b) and (c) illustrate intermediate similarity degrees. Additionally, one can easily identify the 3 letter words with most similar/dissimilar frequency, by looking at individual cells.

We also propose a rigorous quantitative analysis. We define a "3-letter word similarity measure" as the ratio between the number of elements of the difference matrix in the interval $0, \pm 1$, N_3 and the total number of elements N, of the same matrix:

$$S_3 = \frac{N_3}{2^3 \times 2^3} \tag{2}$$

and a similarity percentage measure

$$\eta_3(\%) = 100 S_3 . \tag{3}$$

The values of the similarity percentage measure computed according to equation (3) are given in Table 1.

Table 1

The values of the similarity measure for 3-letter words, η_3, for the analyzed albumin sequences

measure	human-monkey	human-cattle	human-pig	human-dog	human-rat	human-cat	human-rabbit
η_3	65.6	57.8	46.9	43.7	42.2	39.1	36.5

The analysis of similarity/dissimilarity is not restricted by the length of the biological sequence, the only requirement being the approximate equality of the length of the compared sequences.

CONCLUSIONS

The proposed method is characterized by reduced complexity, requires low computation effort and can be used for data classification. This kind of treatment is particularly valuable in looking for evolutionary relationships between organisms and identifying functionally conserved sequences.

The results of the present study are in good agreement with the conclusions of similar study on albumin based on DNA sequences.[12]

REFERENCES

1. S. Wang, F. Tian, Y. Qiu and X. Liu, *J. Theor. Biol.*, **2010**, *256*, 194-201.
2. J.Y. Yang, Z.L. Peng, Z.G. Yu, R.J. Zhang, V. Anh and D. Wang, *J. Theor. Biol.*, **2009**, *257*, 618-626.
3. D.C. Carter and J. X. Ho, "Serum albumin. Structure, Advances in Protein Chemistry", vol. 45, New York: Academic Press, 1994.
4. T. Peters, "Serum albumin. Advances in protein chemistry", vol. 37, New York: Academic Press, 1985.
5 A. Varlan and M. Hillebrand, *Rev. Roum. Chim.*, **2010**, *55*, 69-77.
6 V. Chiosa and V. E. Sahin, *Rev. Roum. Chim*, **2007**, *52*, 89-91.
7. M. Olteanu, S. Geadau, A. Zarna and T. Constantinescu, *Rev. Roum. Chim.*, **2000**, *45*, 369-374.
8. C.P. Cristescu, C. Stan and E. Scarlat, *Physica A.*, **2009**, *388*, 4845-4855.
9. Z.G Yu, V. Anh and K. S. Lau, *J. Theor. Biol.*, **2004**, *226*, 341-348.
10. http://www.ebi.ac.uk/embl/
11. S. Nair Achuthsankar, V. Nair Vrinda, K.S. Arun, K. Kant and A. Dey, "Bio-sequence Signatures Using Chaos Game Representation in Bioinformatics: Applications in Life and Environmental Sciences", Ed. M.H. Fulekar, Springer, 2009.
12. C Stan, C.P. Cristescu and E. Scarlat, *J. Theor. Biol.*, **2010**, *267*, 513-518.