Similarity analysis for DNA sequences based on chaos game representation. Case study: The albumin

Cristina Stan, Constantin P. Cristescu, Eugen I. Scarlat

www.elsevier.com/locate/yjtbi

Similarity analysis for DNA sequences based on chaos game representation. Case study: The albumin

Cristina Stan, Constantin P. Cristescu, Eugen I. Scarlat

PII: S0022-5193(10)00501-1
DOI: doi:10.1016/j.jtbi.2010.09.027
Reference: YJTBI6167

To appear in: *Journal of Theoretical Biology*

Received date: 11 June 2010
Revised date: 16 September 2010
Accepted date: 17 September 2010

Cite this article as: Cristina Stan, Constantin P. Cristescu and Eugen I. Scarlat, Similarity analysis for DNA sequences based on chaos game representation. Case study: The albumin, *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2010.09.027

Similarity analysis for DNA sequences based on chaos game representation. Case study: the albumin

Cristina Stan, Constantin P. Cristescu, Eugen I. Scarlat

Department of Physics I, Faculty of Applied Sciences, Politehnica University of Bucharest, 313 Spl. Independentei, RO-060042, Bucharest, Romania

Abstract. Using chaos game representation we introduce a novel and straightforward method for identifying similarities/dissimilarities between DNA sequences of the same type, from different organisms. A matrix is associated to each CGR pattern and the similarities result from the comparison between the matrices of the sequences of interest. Three different methods of analysis of the resulting difference matrix are considered: a 3-dimensional representation giving both local and global information, a numerical characterization by defining an *n*-letter word similarity measure and a statistical evaluation. The method is illustrated by implementation to the study of albumin nucleotides sequences from eight mammal species taking as reference the human albumin.

Keywords: similarity analysis, chaos game representation, DNA sequence, albumin

1. Introduction

The nucleotide composition and the distribution along a DNA sequence are known to play a vital role in the determination of the functions of various biological structures. In characterizing biological sequences, one can choose the specific set of string patterns of a given length that could play a certain role in regulating biological properties of the sequence.

Comparison between DNA sequences of the same type from different organisms represents a subject of continuous interest for biology research. The results of similarity studies can be valuable in looking for evolutionary relationships between organisms, identifying functionally conserved sequences and inferring homology.

In recent years many different graphical representations of DNA/protein sequences have been reported (Gates, 1986; Roy *et al*., 1998; Li *et al*., 2006; Randić *et al*., 2007; González-Díaza *et al*., 2009; and the references therein).

An important graphical method extensively used in various fields, including biology, is the Chaos Game Representation (CGR) (Jeffrey, 1990; Yu *et al*., 2004; Randić, 2008; Cristescu *et al*., 2009a; Rasouli *et al.*, 2010).

The advantage of graphical representations in biology is that they allow visual inspection of data, helping in recognizing major differences among similar sequences. They also create the possibility of numerical characterization to obtain a quantitative measure of the degree of similarity/dissimilarity of different DNA/protein sequences (Deschavanne *et al*., 1999; Randić *et al*., 2003; Liao and Wang, 2004; Wang *et al*., 2010).

Using CGR, we propose a novel and relatively simple method of identifying similarity/dissimilarity between different DNA sequences of the same type. The degree of similarity is identified by comparison of the distribution of points in the CGR patterns. A matrix is associated to each pattern and the similarities are identified by comparison between the matrices of the sequences of interest. A frequency difference matrix computed using matrices of two selected sequences is defined. We consider three different methods of analysis of the resulting matrix: a 3-dimensional representation, a numerical quantification based on a newly defined "*n*-letter word similarity measure" and a statistical evaluation.

Unlike the existing methods of sequence comparison, the proposed method defines a new similarity measure for words of selected length. The resulting algorithm is characterized by reduced complexity, requires low computation cost and can be used for data classification.

The method is applied to the study of nucleotides albumin sequences from eight mammal species considering as reference the human albumin.

2. Similarity evaluation derived from CGR

The chaos game representation was proposed as a scale-independent representation for genomic sequences by H.J. Jeffrey in 1990. This method of visual analysis was later extended and generalized for arbitrary time-series by Peak and Frame in 1994. In this case, the process begins with the sorting of the data in increasing order from the minimum to the maximum value and then, the string is divided by quartiles. A new series is generated consisting of only four entries. Details on the procedure can be found in (Cristescu *et al*. 2009b). This iterative technique can be used to create a pattern that helps to visualize the structure present in the respective time series. Departure from a uniform distribution is evidence for the possibility of the presence of a deterministic mechanism.

In the case of CGR for DNA sequences, the process is carried out in a unit square, the four vertices of which are labeled by the nucleotides names A, C, T, G. Here A, C, T and G stand for adenine, cytosine, thymine and guanine, respectively. The first nucleotide of the sequence is plotted as a point halfway between the center of the square and the vertex representing this nucleotide; the second one is plotted by a point halfway between the vertex with the second nucleotide name and the previous point, etc.

Figure 1

Figure 1 shows the CGR representation for a sequence containing the word TGCA using as seed (initial point) the center of the square. The image demonstrates that irrespective of the nucleotide preceding this word, the ending point is always localized in a particular cell of the $2^4 \times 2^4$ partition of the square. We consider four possible situations denoted 1 if the nucleotide preceding the selected structure is C, 2 if this is T, 3 if it is G and 4 if it is A. The arrows show the evolution of the CGR construction in each case. Whenever the chosen word appears in the analyzed sequence, a point in the corresponding cell is generated. Consequently, the number of points in a particular cell of the $2^4 \times 2^4$ partition represents the frequency of a selected four letter word in the whole analyzed sequence. Said it differently, the whole set of frequencies of the four letter words found in a given genomic sequence are displayed in the form of an image in which each cell is associated with a specific word. The chosen length of the word imposes the resolution of the cell partition. For example, if the word of interest has six letters (e.g. ACGTAT) then, the appropriate partition must be $2^6 \times 2^6$.

The chaos game representation image gives local as well as global visual information which can be identified as both qualitative and quantitative expressions of the order, regularity, structure, and complexity of DNA sequences (Zbilut *et al*. 2002; Yang *et al*., 2009).

A problem of interest to the biology research is the similarity of the same type of sequences characteristic of different species. Most of the existing methods for similarity identification are based on some kind of graphical representation (Randić *et al*., 2003; Liao and Wang, 2004; Wang *et al*., 2010).

In this work we present a method to evaluate similarities between images corresponding to different DNA sequences based on CGR patterns. Each CGR image is described by a numerical matrix generated using the versatile tool CGREx presented by Achuthsankar *et al* in 2009. Analysis based on this type of associated matrices considered as generalized scale-

independent Markov probability tables was previously used in the development of new statistical analysis for DNA sequences (Almeida *et al*., 2001).

Comparison between two DNA sequences is achieved by manipulating the associated matrices.

In order to mathematically process the CGR patterns, we divide the unit square into a $2^n \times 2^n$ cell partition where n is the length of the word of interest. The number of points in a particular cell represents the frequency of the corresponding word in the sequence.

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 is computed as:

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

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 sequences of the same type.

We find useful to consider a 3-dimensional representation of the difference frequency matrix. It gives local as well as global visualization allowing a quick estimate of the similarity/dissimilarity of the two sequences.

We also propose a rigorous quantitative analysis based on a "*n*-letter word similarity measure" defined in terms of the difference matrix. A particular cell of this matrix is empty when the frequencies of a certain structure (word) in the two compared sequences are equal. This is considered as a similarity of the two sequences. A global similarity can be expressed as the ratio of the number of cells that corresponds to the similarity situation to the total number of cells for the particular word-length.

Here, we are facing the problem of the correctness of the frequency count in the CGR patterns of the two sequences. We have to take into consideration the following factors: first it is possible that the length of the two sequences representing the same type of protein is slightly different. This means that the two CGR patterns do not have exactly the same number of points. Second, in the process of counting, some points can be lost. Due to the limited resolution of the algorithm they can be confused as points of the lines that separate the cells.

These factors of uncertainty in the frequency counts show that consideration of the empty cells of the difference matrix as the only situation corresponding to similarity is not rigorously justified. Cells with $(d_n)_{jk} = \pm 1, \pm 2, \ldots$ could also correspond to similarity. This situation imposes the definition of a threshold which we choose to call "*n*-letter-word relevance threshold" ζ_n. In

order to be widely applicable we define it as function of the average frequency of the *n*-letter word

$$\bar{f}_n = \frac{N}{2^n \times 2^n} \tag{2}$$

where N is the total length of the sequence.

This is realized by introducing the fractional coefficient γ such that

$$\zeta_n = \gamma \bar{f}_n. \tag{3}$$

If the word-length n is very large, then it is possible that too many of the cells of the $2^n \times 2^n$ partition will contain no points (too many elements of the A_n and B_n matrices will be vanishing). For reliable results this situation has to be avoided. In order to satisfy the condition of reduced number of empty cells we require that the mean frequency in the A_n and B_n matrices should be significantly larger than 1, $\bar{f}_n > 1$.

The condition $f_n > 1$ is equivalent to $n < \frac{\ln N}{2 \ln 2}$, where again "$< (\cdot)$" means significantly smaller than $(\cdot)$. We propose as the maximum value of n:

$$n_{\max} = \mathrm{int}\left(\frac{\ln N}{2 \ln 2} - 1 \right) \tag{4}$$

and we consider as reasonable the value $\zeta_n = 1$ for the maximum value of n, i.e.

$$\zeta_{n_{\max}} = \gamma \bar{f}_{n_{\max}} = 1 \tag{5}$$

which gives:

$$\gamma = \frac{1}{\bar{f}_{n_{\max}}} = \frac{1}{N} 2^{2\left[\mathrm{int}\left(\frac{\ln N}{2 \ln 2} - 1 \right) \right]}. \tag{6}$$

We can also estimate similarity for shorter words. In order to be able to compare the results for words of different length we have to use the same value of the fractional coefficient γ.

For words with length $n = n_{\max} - 1$ we have:

$$\zeta_{n_{\max}-1} = \gamma \bar{f}_{n_{\max}-1} = \gamma \frac{N}{2^{2n_{\max}-2}} = 2^2 \zeta_{n_{\max}} = 2^2. \tag{7}$$

Similarly,

$$\zeta_{n_{\max}-2} = 2^4 \zeta_{n_{\max}} = 2^4 \tag{8}$$

which can be generalized for words of length $n_{\max} - k, \quad k < n_{\max}$ as:

$$\zeta_{n_{\max}-k} = 2^{2k}\zeta_{n_{\max}} = 2^{2k}. \quad (9)$$

This algorithm is applicable for sequences of any length, *N*.

Now, we define an "*n*-letter word similarity measure" as the ratio between the number of elements of the difference matrix in the interval $\pm\zeta_n$, N_{ζ_n} and the total number of elements of the same matrix:

$$S_n = \frac{N_{\zeta_n}}{2^n \times 2^n}, \quad (10)$$

and a "similarity percentage measure"

$$\eta_n(\%) = 100 S_n. \quad (11)$$

The introduction of the relevance threshold creates the possibility of comparison of the results of the computation of the similarity measure for words of different lengths $n \le n_{\max}$.

The definition of a maximum value of the word length is not absolutely restrictive. Similarity computation for words of larger length is possible, however the results cannot be related (compared) to the results of computation for shorter words as practicable for $n \le n_{\max}$.

An alternative way of looking at the difference matrix is the generation of a series of $2^n \times 2^n$ elements by concatenating the columns (or rows) of the matrix and performing a statistical study on this series, mainly a histogram analysis. The standard deviation can be considered as a measure of the similarity of the two compared sequences for a given word length. It is expected that the larger the similarity measure, the smaller the standard deviation of the corresponding histogram, such that, for high similarity, the values of the elements of the difference matrix are concentrated in the neighborhood of 0. This analysis is relevant only for $n \ge 3$ because one cannot perform a relevant statistical computation on a set of much less than 100 elements.

3. Case study: albumin nucleotide sequences

We illustrate the proposed method of analysis by applying it to the study of albumin sequences. We consider the coding sequence of albumin from eight different species:Homo sapiens (human) albumin [EMBL-Bank: CAA00606], Bos taurus (cattle) serum albumin [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].

Although the sequences have slightly different length, from 1830 bases (human albumin) to 1803 bases (rhesus monkey), the relative error ($\Delta N / N$) in the implementation of our proposed method for similarity analysis is about 1.5%.

Figure 2

As an illustration of the CGR patterns used in the analysis, Fig.2 shows the case of human albumin. A simple visual examination can give information on the frequencies of various substructures. In this situation, the slightly lower density of points along the GC diagonal line corresponds to a relative scarcity of the group cytosine following guanine in the data sequence.

Figure 3

Figure 3 presents the 3-dimensional representation of the difference frequency matrices measured by the height of the bars (Δd) corresponding to the cells in the partition with n=3 (3 - letter words). In the computation, we consider A_n as the human matrix and B_n as a matrix of another particular mammal. Accordingly, Δd (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 $\Delta d = 0$ contour by a solid line. The plot in Fig.3 corresponds to the following situations: (a) human - monkey; (b) human - cattle; (c) human - pig and (d) human - rat. By simple 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-rat 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. For example, in the case (d), the highest dissimilarity is for words "AAA", irrespective of the preceding or of the following letter in the sequence.

Table 1

Table 2

Table 3

The quantitative investigation is based on the measure defined by equations (10) and (11). We apply this analysis for words of length n=2, 3 and 4. The Tables 1, 2 and 3 give the similarity percentage measure between the human albumin and the albumins of various other mammals, with a relevance threshold (defined by equation (3)) for the same value of the fractional coefficient $\gamma = 0.14$ (approximating 0.139890..., computed using equation (6) for N=1830).

It is interesting to notice that the classification that can be introduced on the basis of the value of the similarity measure computed for words of different length is the same.

Clearly, the computation is not restricted to sequences of a particular length as long as the condition implied by (4) is satisfied.

Figure 4

Finally, we consider the statistical analysis based on the series obtained by concatenating the columns of the frequency difference matrix. We restrict the computation for the case n=3 and n=4 for the reasons discussed in section 2. The histograms corresponding to the case n=3 illustrated in Fig.3 are shown in Fig.4. On each graph, the best fitting Gaussian distribution is drawn.

The standard deviations for all the situations under discussion are included in Tables 2 and 3, respectively. It is observed that the larger the similarity measure, the smaller the standard deviation of the corresponding histogram. This shows that, for high similarity, the values of the elements of the difference matrix are concentrated in the neighborhood of 0.

The consistency of the results obtained by the three types of analysis can be considered as validation of the proposed method.

4. Conclusion

Using chaos game representation we introduce a novel and straightforward method of analysis of the similarities between DNA sequences of the same type, from different organisms. The degree of similarity is established by comparison of the point distribution in the CGR of the investigated sequences as reflected in the corresponding frequency matrices. We define a frequency difference matrix computed using matrices of two selected sequences. The similarity/dissimilarity is emphasized considering three different methods of analysis: a 3-dimensional representation giving both local and global information, a numerical

characterization introduced by defining an *n*-letter word similarity measure and a statistical evaluation. The method is applied to the study of nucleotides albumin sequences from eight mammal species taking as reference the human albumin.

We stress upon the fact that the proposed method of similarity/dissimilarity analysis is not restricted by the length of the DNA sequence, the only requirement being the approximate equality of the length of the compared sequences. The definition of a maximum value of the word length is not absolutely restrictive. Similarity computation for words of larger length is possible, however the results cannot be quantitatively related to the results of computation for shorter words because in this case, a common value of the fractional coefficient cannot be defined.

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.

Acknowledgments: The work was partially supported by UEFISCSU-Ro Research Grant ID#1556 No.836/2009.

References

Achuthsankar S. Nair, Vrinda V. Nair, Arun K.S., Kant K., Dey A., Bio-sequence Signatures Using Chaos Game Representation in Bioinformatics: Applications in Life and Environmental Sciences, Editor M.H. Fulekar, Springer, 2009, cap 6, 62-76.

Almeida J.S., Carriço J. A., Maretzek A., Noble P.A., Fletcher M, 2001. Analysis of genomic sequences by Chaos Game Representation. Bioinformatics. 17, 429-37.

Cristescu, C.P., Stan, C., Scarlat I.E., 2009a. Modeling with the chaos game (i). Simulating some features of real time series. U.P.B. Sci. Bull., Series A, 71, 95-100.

Cristescu, C.P., Stan, C., Scarlat, E., 2009b. The dynamics of exchange rate time series and the chaos game. Physica A. 388, 4845-4855.

Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G., Fertil, B., 1999. Genomic Signature: Characterization and Classification of Species Assessed by Chaos Game Representation of Sequences. Mol. Biol. Evol. 16, 1391–1399.

Gates, M.A., 1986. A simple way to look at DNA. J. Theor. Biol. 119, 319-328.

González-Díaza, H., Pérez-Montoto, L.G., Duardo-Sanchez, A., Paniagua E., Vázquez-Prieto S., Vilas, R., Dea-Ayuela, M.A., Bolas-Fernández, F., Munteanu, C.R., Dorado, J., Costas, J., Ubeira F.M., 2009. Generalized lattice graphs for 2D-visualization of biological information. J. Theor. Biol. 261, 136-147.

Jeffrey, H.J., 1990. Chaos game representation of gene structure. Nucleic Acids Res. 18, 2163–2170.

Li, C., Tang, N., Wang, J., 2006. Directed graphs of DNA sequences and their numerical characterization, J. Theor. Biol. 241, 173-177.

Liao, B., Wang, T.M., 2004. Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation. Chem. Phys. Lett. 388, R195–R200.

Peak, D., Frame, M., 1994. Chaos under control: the art and science of complexity, Freeman, New York.

Randić, M., 2008. Another look at the chaos-game representation of DNA. Chem. Phys. Lett. 456, 84-88.

Randić, M., Vračko, M., Lerš, N., Plavšić, D., 2003. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. Chem. Phys. Lett. 371, R202–R207.

Randić, M., Zupan, J., Vikic-Topic, D., 2007. On representation of proteins by star-like graphs. J. Mol. Graph. Model. 26, 290-305.

Rasouli, M., Rasouli, G., Lenz, F.A., Borrett, D., Verhagen, L., Kwan, H.C., 2010. Chaos game representation of human pallidal spike trains, J. Biol. Phys. 36, 197-205.

Roy, A., Raychaudhury, C., Nandy, A., 1998. Novel techniques of graphical representation and analysis of DNA sequences – a review. J. Biosci. 23, 55-71.

Wang, S., Tian, F., Qiu, Y., Liu, X., 2010. J. Theor. Biol. doi:10.1016/j.jtbi.2010.04.013.

Yang, J.Y., Peng, Z.L., Yu, Z.G., Zhang, R.J., Anh, V., Wang, D., 2009. Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation. J. Theor. Biol. 257, 618-626.

Yu, Z.G., Anh, V., Lau, K.S., 2004. Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses. J. Theor. Biol. 226, 341-348.

Zbilut J.P., Sirabella P., Giuliani A., Manetti C., Colosimo A., Webber, Jr. C.L., 2002. Review of nonlinear analysis of proteins through recurrence quantification. Cell Biochem. and Biophys. 36, 67-87.

List of captions

Fig.1 CGR representation for a sequence containing the word TGCA with different preceding letters (1 -...CTGCA; 2 -... TTGCA; 3 -...GTGCA; 4 -...ATGCA)

Fig. 2 CGR plot for the human albumin

Fig.3 3-dimensional representation of the difference frequency matrices for *n*=3, for 4 of the studied cases: (a) human - monkey; (b) human - cattle; (c) human - pig; (d) human - rat. The correspondence between the corners of the CGR rectangle and the 4 nucleotides is the same in all plots and is shown only on (a).

Fig. 4 Statistical analysis of the difference matrices for *n*=3 for the cases: (a) human - monkey; (b) human - cattle; (c) human - pig; (d) human - rat

Table 1 Similarity percentage measure for words with length *n*=2

Table 2 Similarity percentage measure and standard deviation for words with length *n*=3

Table 3 Similarity percentage measure and standard deviation for words with length *n*=4

Table 1

measure	human-monkey	human-cattle	human-pig	human-rabbit	human-cat	human-dog	human-rat
η_2	100	81.6	77.5	69.8	64.6	52.1	31.8

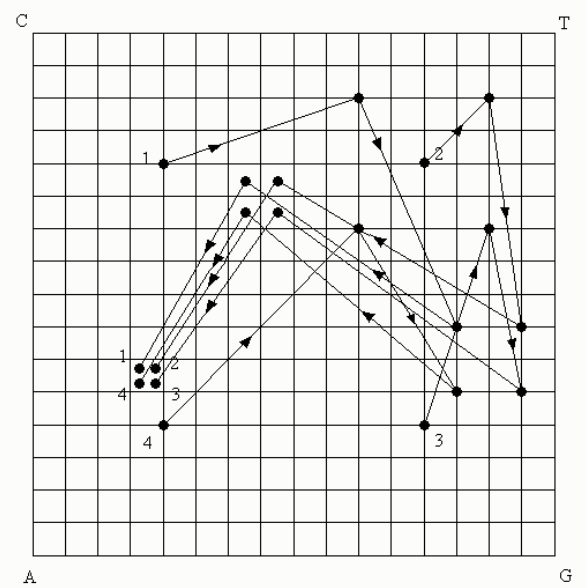
Table 2

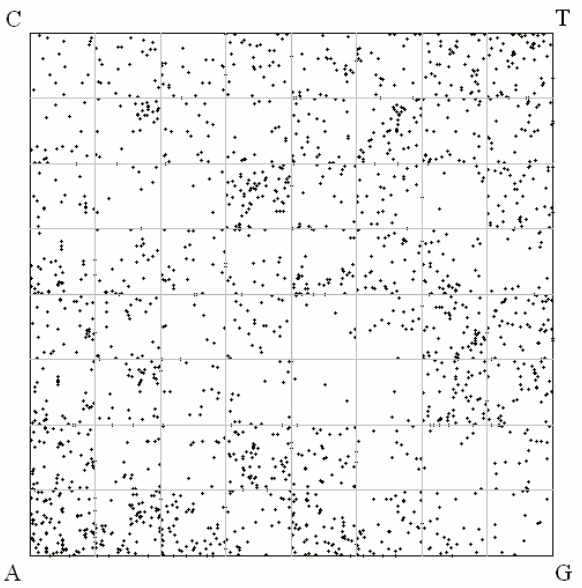
measure	human-monkey	human-cattle	human-pig	human-rabbit	human-cat	human-dog	human-rat
η_3	81.2	57.1	55.3	52.8	49.4	36.2	29.2
standard deviation	2.98	5.76	6.14	6.47	6.73	8.48	10.8

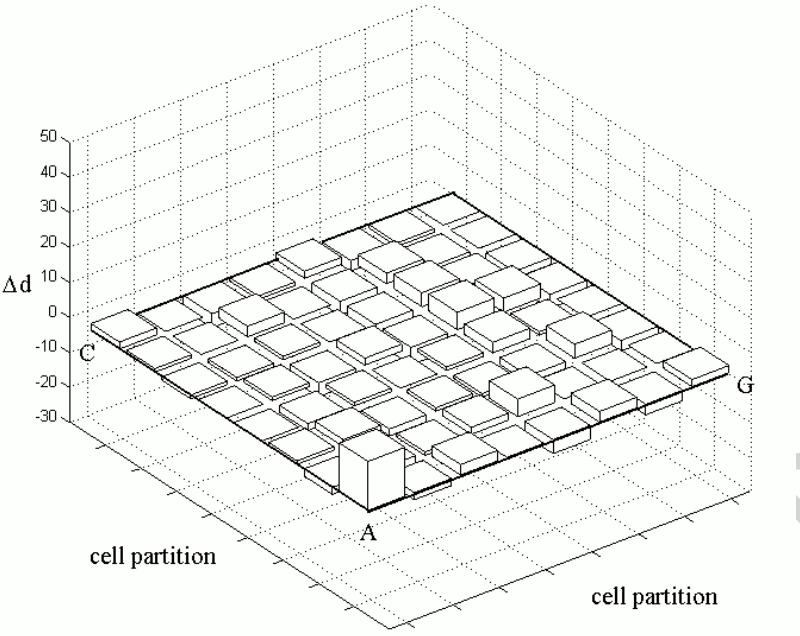
Table 3

measure	human-monkey	human-cattle	human-pig	human-rabbit	human-cat	human-dog	human-rat
η_4	70.7	49.6	46.9	43.7	42.1	40.3	34
standard deviation	1.62	2.92	3.13	3.21	3.29	3.41	4.12

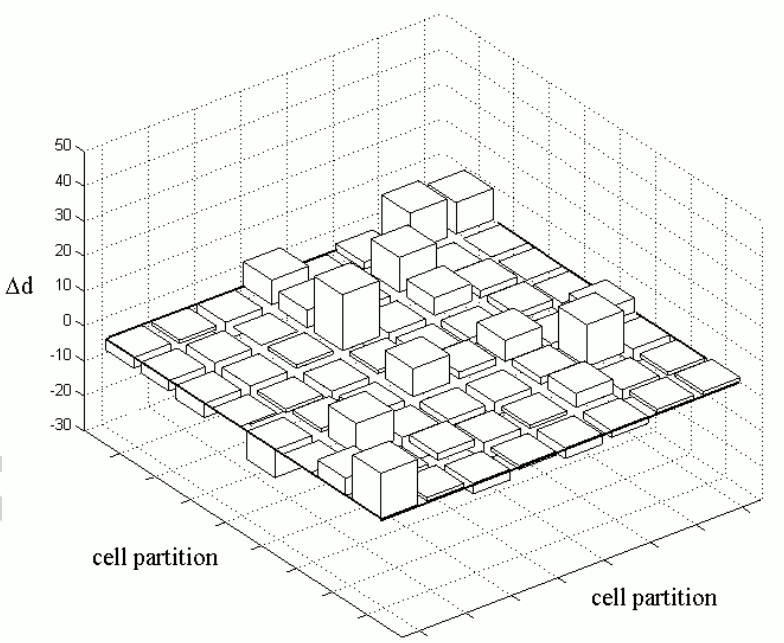
Figure1



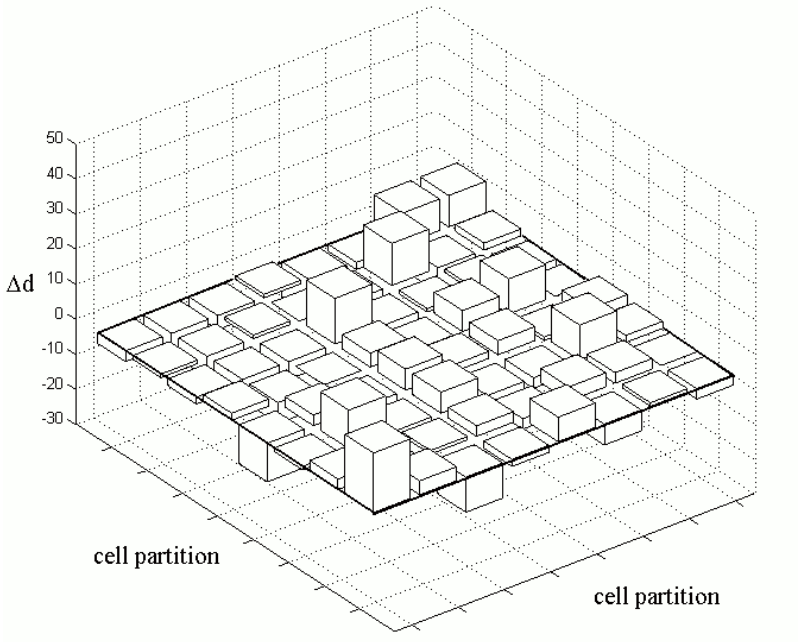




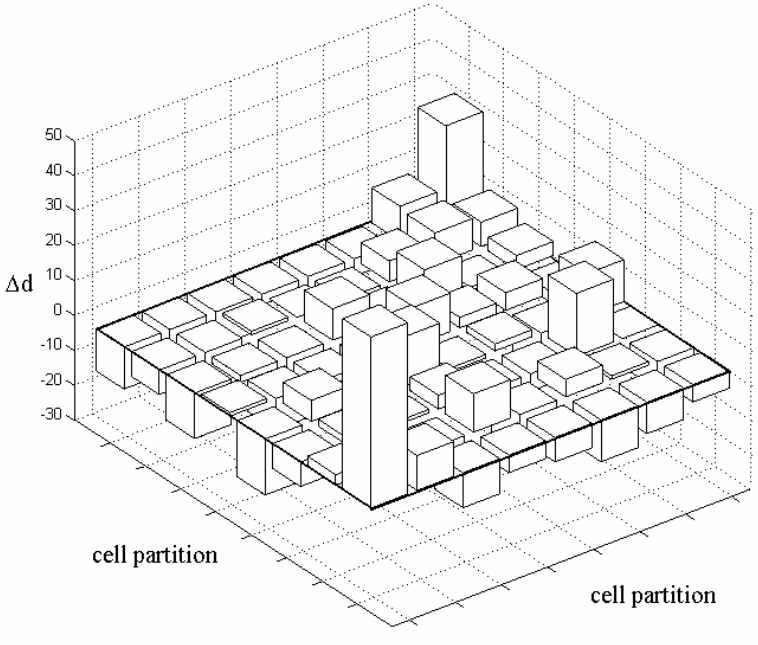
(a)



(b)



(c)



(d)