CMST 20(3) 93-100 (2014) DOI:10.12921/cmst.2014.20.03.93-100

A New Method for Symbolic Sequences Analysis. An Application to Long Sequences

B. Kozarzewski

University of Information Technology and Management
ul. H. Sucharskiego 2, 35-225 Rzeszów, Poland
E-mail: bkozarzewski@wsiz.rzeszow.pl

Received: 28 April 2014; revised: 31 July 2014; accepted: 31 July 2014; published online: 28 August 2014

Abstract: The method for symbolic sequence decomposition into a set of consecutive, distinct, non-overlapping strings of various lengths is proposed. Representation of the sequence as a set of words allows one to use set theory notions. The main result is a quite new definition of the similarity between any two sequences over a given alphabet. No prior sequence alignment is necessary. In the present paper two applications of a set of words are described. In the first a similarity measure is applied to prepare centroids for K-means algorithm. It results in a high performance grouping method for long DNA sequences. The other application concerns the statistical analysis of word attributes. It is shown that similarity, complexity and correlation function of word attributes across sequences of digits of fractional parts of some irrational numbers support the suggestion that the sequences are instances of a random sequence of decimal digits.

Key words: similarity and distance measures, clustering, DNA sequences, irrational numbers

I. INTRODUCTION

Very long sequences, which occur in many fields, when directly represented by vectors are usually difficult to analyse or compare, e.g. genomes or share prices. Decreasing the size of a vector in the course of the preprocessing procedure has numerous advantages. It reduces significantly the base of the recorded sequence for further retrieval. The size of a vector can be reduced by mapping a sequence onto a set of subsequences (strings, words). The set of words results from suitable decomposition (parsing) of a sequence. The first useful decomposition was proposed by Lempel and Ziv [1] in order to define the quantitative measure of symbolic sequence complexity. Ke and Tong [2] pointed out that the Lempel-Ziv algorithm leads to a complexity measure which some artificial, regular sequences reckon as complex.

They developed an algorithm for parsing a sequence of binary symbols, which is free of such faults. The algorithm provides a better quantitative measure of complexity which is defined as the total number of strings. In my paper [3] the Ke and Tong algorithm was generalized to arbitrary sequences over a finite alphabet. It was also shown that the whole set of strings (not only their number) is a very rich source of information on symbolic sequences.

The most fundamental task in sequence analysis is to discover any relationships (e.g. similarity) between two experimental sequences. To compare any of two sequences a certain measure is required that can determine whether and how similar (or distant) they are. There are several definitions of similarity, [4-7]. Neither of them provides a simple, satisfactory measure of global similarity between two arbitrary symbolic sequences over the same alphabet. An alternative similarity measure based on decomposition of sequences into a set of specific distinct words was proposed in [3]. The similarity between two sequences is related to the number of common words in the decomposition of two sequences. It is very convenient, as it does not need previous alignment of the sequences.

Another task in the set of sequences analysis is sequences clustering. Its aim is to assign a set of sequences into groups

so that the objects in the same group are more similar to each other than to those in other groups. Clustering symbolic sequences is more challenging than clustering numeric data because there is no natural measure of similarity between symbolic sequences. There are many clustering methods and algorithms. Among partitioning methods, K-means is a commonly used algorithm. The algorithm needs initialising a set of cluster centres, which are often difficult to find, and some measure of clustering quality. In the present paper the hierarchical method and new similarity between sequences is used to find initial cluster centres. As the clustering quality measure, the property of an intra-cluster sum of squared new specific distances is used.

Every word following from sequence decomposition has several numerical attributes, like length (number of symbols) and composition (numbers of particular letters and their distribution). The relation between the values of these attributes in different segments of the sequence distribution can be detected with the use of word correlation functions.

The fractional parts of the base of natural logarithm and $\sqrt{2}$ numbers are considered as symbolic sequences over the decimal digits alphabet. It is shown in the present paper that similarity and correlation functions of word attributes of both sequences, their randomised counterparts and pseudorandom sequence of decimal digits are similar with high accuracy.

The aim of the present paper is twofold: firstly, to show that the set of words, the similarity and distance measures make the K-means algorithm of clustering symbolic sequences more robust, and secondly, to support the hypothesis saying that very long sequences of decimal digits of irrational numbers are random sequences.

The paper is organized as follows: in section 2 symbolic sequence decomposition into a set of consecutive, distinct strings is recollected. Next, the similarity, distance measure between sequences, and correlation functions of word attributes are defined. In section 3.2 the clustering algorithm is applied to a set of 400 long DNA sequences, and in section 3.3 the analysis of three real very long numerical sequences and their randomised counterparts is made.

II. METHODS

II. 1. Parsing algorithm

The naive decomposition of a symbolic sequence into a set of short strings of predefined length (called k-mers) has a limited application. Strings of the same physical meaning may differ in length and can exist in different places of the sequence. Successful sequence decomposition can be done with the use of the Ke and Tong algorithm generalised to any sequence over a finite alphabet. The details of the algorithm are as follows. Suppose there is a primary sequence C of symbols c_1, c_2,..., c_n. Suppose S_t is a set of words obtained so far and the first symbol of the new word w is c_i. The word is formed as a result of a specific procedure of appending symbol c_i by the following symbols in three steps.

Step 1. String $Q = c_i$ is neither periodic nor chaotic because there is only one symbol in it. So it has to be appended by the next symbol. Appending is continued until some symbol c_{i+j+l} repeats one of the symbols, say k-th, in the string $Q = c_i, \ldots, c_{i+j}$.
Step 2. Let $P = c_k$ and $R = c_{i+j+1}$, so far they are equal. Both strings are appended $P = c_k c_{k+1}$, $R = c_{i+j+1}c_{i+j+2}$ and so on, until they become different. Then string Q found in Step 1 is appended by string R, and the new string is $Q = QR$.
Step 3. Set S_t of words is searched for the presence of string Q. If string Q is found, it is appended by the following (next to the last symbol of Q) symbol of C becoming $Q = Qc_{i+j+k+1}$. Appending is continued until some string $Qc_{i+j+k+l}$ does not replicate any word from S_t. The string $w = Qc_{i+j+k+l}$ becomes the new word of the spectrum representing sequence C. It may happen that several last symbols of C cannot be processed by the above replication, they make a new word.

The code of the parsing algorithm is available on request. The result of the sequence decomposition is a set of ordered, distinct and non-overlapping words which will be called the word spectrum of the primary symbolic sequence C.

II. 2. Similarity

Measuring the similarity between symbolic sequences is essential in many data (numeric as well) analyses. So far, due to the lack of natural geometrical interpretation of the symbolic sequence the resemblance (similarity) measure between two sequences was difficult to define, mainly because of the necessity of prior sequences aligning [6, 7].

When spectra S_1 and S_2 of two sequences C_1 and C_2 are known, several similarity measures can be defined. Among them, the set theory intersection of S_1 and S_2 against the total number of words in both spectra

$$\sigma(C_1, C_2) = \frac{2l(\mathrm{int}(S_1, S_2))}{l\,(S_1) + l\,(S_2)}$$

seems to be the most natural. Here, $\mathrm{int}(S_1, S_2)$ is a set of words that the two spectra share (intersection of S_1 and S_2), and $l\,(S)$ means the length (number of words) in set S. The similarity measure $\sigma(C_1, C_2)$ has the form of the Dice coefficient [10], except that the symbols have different meaning. The $\sigma(C_1, C_2)$ function varies between 0 when the spectra are disjoint sets and 1 when sequences S_1 and S_2 are the same.

II. 3. Distance

It has been found that the distance between sequences is very useful when it comes to identifying the number of clusters that best fits a given dataset. Let C be a set of n symbolic sequences, the distance between the sequences can be established with the use of similarities between them. They form a symmetric table with ones along the main diagonal, which

is called similarity matrix Σ. An element σ_{ij} of the matrix depends on sequences S_i and S_j but not on the other sequences of the set. Every sequence S_i can be uniquely mapped on i-th row (or i-th column) of similarity matrix Σ

$$(\sigma_{i1}, \ldots, \sigma_{in}) \rightarrow \mathbf{x}_i,$$

where $\mathbf{x}_i$ is n-dimensional numeric vector representing i-th sequence. All rows span specific, linear n dimensional vector space. Now, a distance between any two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ can be defined, in the following it will be Euclidean distance

$$d_{ij} = |\mathbf{x}_i - \mathbf{x}_j| = \sqrt{\sum_{k=1}^{n} (\sigma_{ik} - \sigma_{jk})^2}.$$

The d_{ij} is a context dependent distance between the corresponding sequences S_i and S_j. The lower bound on the distance can be shown to be equal to $\sqrt{2}(1 - \sigma_{ij})$, which means that

$$d_{ij} \geq \sqrt{2}(1 - \sigma_{ij}).$$

There is no monotonic relationship between distances and corresponding similarities. The set of differences $d_{ij} - \sqrt{2}(1 - \sigma_{ij})$, looks like a random variable of the Weibull distribution. There is an advantage in using distance in some applications because distance is an additive measure contrary to the similarity.

II. 4. Correlation function of word attributes

One of the goals of every analysis of any set of objects is to find relations between the attributes of its elements across the sequence. Parsing the sequence into a set of words (word spectrum) reduces significantly the set of objects representing the sequence. However, words are more complex objects than single symbols. A word is characterised by attributes such as length, a set of symbols and their order. The spectrum can be considered as a discrete function of many variables (attributes of the word). The aim of the advanced analysis of symbolic sequence is to find relations between the variables. The relations are given by various correlation coefficients. Let $S(w_1, w_2, \ldots, w_n)$ be the word spectrum of sequence C over alphabet $a_1, a_2, \ldots$. Any word w_i is a string of letters $a_1, a_2, \ldots, a_l$. The numerical attributes of the word are, for example, l – length of the word, number of a_1 instances, set of integers indicating positions of instances a_1, number of 'a_1a_2' strings in the word, and so on. The simplest one is the words length correlation function

$$\mathrm{cor}(i;;l) = \frac{\sum_{j=1}^{n} (l(w_j) - m)(l(w_{j+i}) - m)}{\sum_{j=1}^{n} (l(w_j) - m)^2},$$

where $m = 1/n \sum_{j=1}^{n} l(w_j)$.

The correlation functions can be used as a test for sequence randomness. If the correlation function exhibits small rapid changes of its value, there are no correlations between distant parts of the symbolic sequence.

III. EXAMPLE AND APPLICATIONS

III. 1. Example

To explain the main notions introduced in section 2, a simple example is presented. A set of four small gene sequences is decomposed into words and the similarity and distance matrices of the set are found. A set of short fragments (63 bases each) of small subunits S16 rRNA gene sequences from four organisms: (1)-Human, (2)-Saccharomyces cerevisiae, (yeast), (3)-Zea mays, (corn) and (4)-Escherichia coli downloaded from the GenBank database (http://www.ncbi.nlm.nih.gov). The partition algorithm yields word spectra of length from 13 to 15 words. The shortest one is the spectrum of E. coli: 'GTGC', 'CAGCA', 'GCCG', 'CGGT', ,'AAT', 'ACGGA', 'GGGT', 'GCAAG', 'CGTTA', 'ATCGGA', 'ATTA', 'CTGGGC', 'GTAAAG'. The intersection of the human and yeast spectra consists of 13 words, while the intersection of human and E. coli is only 5 words long. So the similarity between the human and yeast sequences is $2 \cdot 13/(15+15) = 0.867$, and between the human and E. coli sequences it is $2 \cdot 5/(15+13) = 0.357$.

III. 2. Clustering mitochondrial DNA sequences of 400 species

In clustering, the task is to group a data set into a set of disjoint classes of objects, so that each class is assigned to a unique cluster. The data within each cluster are supposed to be similar to each other but different from members of other clusters. Therefore the basic objective in clustering is to discover a natural set of clusters based on some similarity or distance measures. An important problem of clustering procedure initialisation arises at the beginning of cluster analysis. A common question is about the number of clusters present in a given dataset and their initial central vectors which may not necessarily be members of the data set. Actually, cluster initialisation remains the biggest drawback of the partition-based clustering algorithms. In the well known K-means cluster algorithm (see, for example [11] and quotes therein), if there is no prior knowledge on initial clusters, it is difficult to obtain good results. Determining the initial cluster central vectors centroids is an optimization problem. It needs some measure of quality of the centroids as well as a number of final clusters.

In [12] a method for sequence clustering that is based on a representative set of strings was published. In the present paper a more efficient and reliable method for sequence clustering than that presented in [12] is proposed. It uses the similarity measure for the preparation of input centroids to

K-means algorithm and a distance measure for the detection of the optimal set of centroids and clusters. The number of clusters and the initial set of centroids need not be known a priori.

The clustering method proposed in the present paper is a hierarchical agglomerative in the beginning with the aim of finding k initial cluster centroids. After it has been done the standard K-means algorithm is used to partition all the sequences. The centroids of the clusters become input parameters which are inserted into the K-means clustering algorithm to partition all the sequences. The output of K-means algorithm is a set of k clusters. The procedure of joining and validation of the number of clusters completes the clustering process. The approach appears to be very efficient.

Suppose we want to classify a set S_N consisting of N sequences. First of all, the word spectra of sequences and the similarity matrix Σ of the set have to be found. Next, the sequences are mapped onto corresponding rows of Σ matrix and the distance matrix as defined above is computed. In the agglomerative clustering, the hierarchy of clusters is built starting with each spectrum as an individual cluster. The two closest spectra are then grouped, giving one cluster of two sequences. The remaining clusters still consist of spectra of a single sequence. To find the distance between two clusters the Unweighted Pair Group Method Using Arithmetic Mean is used. Within the cluster, the distance is defined as an arithmetic mean of distances between all pairs of sequences that are members of the clusters. Besides, a method for stopping the clustering process, i.e. determining the best number of temporary clusters is needed. Several methods are used to determine the number of clusters. Unfortunately, neither of them can be called the perfect one. In the present research it is proposed that the difference (knee point) of the sum of intra-cluster sum of squared distances (ssd) of all clusters detects precisely enough the optimal set of cluster. The rapid increase of the knee point variation signals that the number of clusters is too high. The sum of squared distances of the cluster is defined as

$$s_k = \frac{1}{n_k^2} \sum_{i,j}^{n_k} |\mathbf{x}_i - \mathbf{x}_j|^2$$

where n_k is size of the k-th cluster and $\mathbf{x}_i$ is a vector representing the sequence belonging to the cluster. The sum of squared distances of the set of clusters is given by $s_{sc} = \sum_k s_k$. The knee point defined as $kp\,(n) = ssd\,(n-1) + ssd\,(n+1) - 2ssd(n)$ versus n – the number of pairs (or paired clusters) is calculated and the recommended set of the best pairs is the one at which absolute knee point values are still relatively small.

At the initial step the set of mean vectors of the selected pairs is considered as a crude approximation for centroids. In the next step the selected pairs are joined into fours. The best set of fours is selected in the same manner as described above, providing better candidates for centroids. Joining the sets is continued until the number of centroids becomes close to the expected number of clusters or all knee point variations of ssd are large. Now the K-means clustering algorithm with centroids as input parameters is used for partitioning of all sequences. The resulting number of clusters in the partition based clustering methods equals the number of centroids and usually exceeds the optimal number of clusters. Some of them have to be joined. Validation of the number of clusters is done with the use of knee point of the sc as the score function measuring the quality of clustering.

The data set used to test the algorithm consists of mitochondrial genomes of four groups (birds, fishes, mammals and reptiles) each of 100 species, all downloaded from the GenBank database (http://www.ncbi.nlm.nih.gov). The set of best pairs selected according to low kp rule consists of 150 mitochondrial sequences, as shown in Fig. 1.

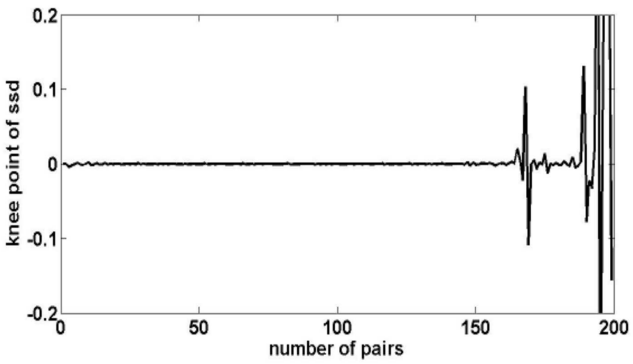

Fig. 1. Knee point of s_k versus number of pairs

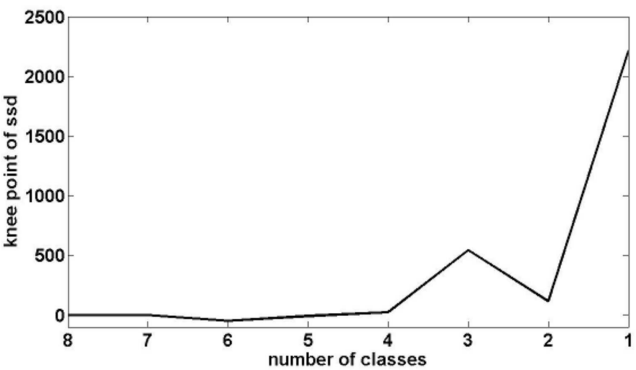

Fig. 2. Knee point of s_{sc} versus number of clusters

The best set of fours includes 60 sequences, and so on. The last is a set of eight groups, each consisting of 16 mitochondrial sequences represented by corresponding vectors (rows of the similarity matrix). The mean vector of each group is used as the centroid starting location. The K-means clustering algorithm returns eight clusters quoted in the supplementary material. The knee point of

sc of the set of clusters is low, indicating that eight is not the optimal number of clusters. Joining the closest clusters into pairs followed by the knee point calculation yields the plot in Fig. 2. It suggests that four cluster is the best result of the clustering process. The resulting clusters are presented in the supplementary material. 382 (96%) MtDNA sequences were grouped properly. Incorrect clustering results refer to six sequences of fish species assigned to the bird family, four sequences of reptile species assigned to mammals and eight sequences of mammals species assigned to fishes.

III. 3. Fractional part of irrational number represented as sequence of digits

Any sequence of digits can be considered as a symbolic sequence over ten digits alphabet. An example is the fractional part of any irrational number. Let as recall that the irrational number is one that cannot be written as a ratio of any two integers. If it is written in the decimal form, it goes on forever without repeating. Some irrational numbers can be expressed as a polynomial with rational coefficients. They are called algebraic numbers. However, much more irrational numbers are not algebraic, and they are called transcendental numbers (see, for example, [13] and quotes therein). Many sequences representing mathematical constants in the decimal base are thought to be normal. In the present context "normal" means that any single digit in a sequence of length n occurs approximately $n/10$ times. All we know on normality follows from E. Borel theorem: almost every real number is a normal number. Another property that irrational numbers is supposed to have is random distribution of digits across the sequence. In what follows it is shown that the word spectrum of the sequence provides some arguments for the conjecture.

In recent years new algorithms based on Bailey-Borwein-Plouffe (BBP) [14] have been discovered and high-performance computing tools have been developed. They made possible computation digits of very long sequences of several irrational numbers. It follows from them that the finite length sequences of commonly known mathematical constants satisfy the normality condition. The paper [15] (and references within) presents a compendium of the set of BBP-type formulas for various mathematical constants.

The fractional part of any irrational number is considered as a symbolic sequence over the decimal digits alphabet. The sequence can be parsed into a set of distinct words (a spectrum) according to the method presented in section 2.1. The spectrum allows for the calculation of complexity, entropy of sequences and comparison of two such sequences with the use of the similarity measure. What is more, it can be done for subsets of different length in order to discover how complexity, entropy and similarities depend on the subset location or length of the subset. In general, some correlations within a sequence may exist. They can be removed after the elements of the sequence are thoroughly shuffled. The shuffling algorithm for the sequence n symbols long consists of the selection of some random permutation of the set of integers from 1 to n and then rearrangement of the elements of the sequence according to the permutation. When it is repeated, for example, 100 times, any correlations disappear.

Nemiroff and Bonnell [16] computed and made available several long sequences of digits in decimal base, including two million long sequences of base of natural logarithm and square root of 2. The authors declare the sequences have been checked, they also suggest the sequences are random and have no unexpected correlations, and ask for testing these properties.

In what follows a set of five two million long sequences of decimal digits is considered, they are fractional parts of: original sequences; oe – base of natural logarithm, os – square root of 2, their shuffled (randomised) counterparts – se, ss, and pseudorandom normal sequence of decimal digits – r. In order to discover tendencies within sequences, from each of them the first twenty-two fragments of length from 1000 to 2 million digits was selected.Let us first discuss the normality property of e and $\sqrt{2}$ sequences.

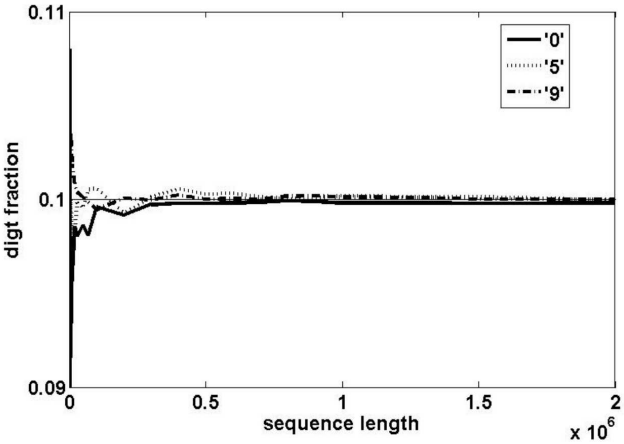

Fig. 3. Share of selected digits ("0", "5", and "9") in sequence of e number

Fig. 3 shows that the longer sequence e, the closer to normal it becomes, and the same is true for $\sqrt{2}$ sequence. It can be assumed that it is true for the arbitrarily long e and $\sqrt{2}$ sequences.

The spectra of all 5×22 sequences were found and used to calculate the complexity – length of the spectrum and similarities between the sequences of the same length. Detailed results are included in Tab. 1 of the supplementary material. In general, the complexity per 1 000 digits slowly decreases with increasing sequences length. Plots of all five complexities are hardly distinguishable. The variance of the five results averaged over subsequence length amounts 10^{-6} for complexity. One can come to a conclusion that the longer sequence e and square root of 2, the closer they become to an instance of the random sequence. Let us now discuss the correlation functions for e and $\sqrt{2}$ sequences. The two attributes of the word are considered for correlation of $\sqrt{2}$ sequence, they are

length and number of "0" digits in the word. Fig. 4 shows the noise-like dependence of both correlations on distance.

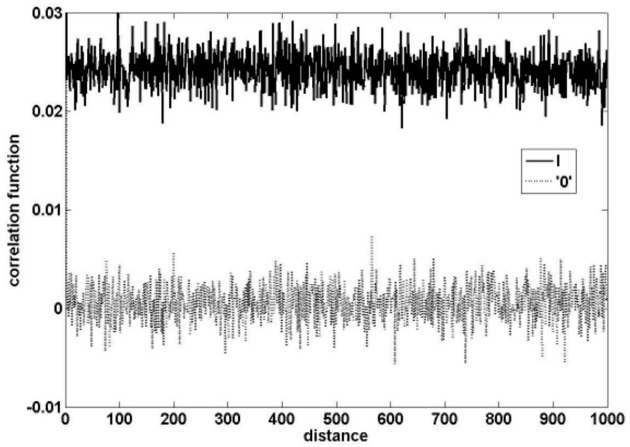

Fig. 4. Correlation functions of word length and score of digit "0" instances in a word for the sequence of $\sqrt{2}$ number

It does not change when distance of tens of thousands is considered. The same is true for e sequence.

There are five independent pairs of sequences of every length. The similarity of each pair is calculated and Fig. 5 shows the similarity of (oe, os) pairs of sequences versus sequence length. The same plots for other pairs are almost indistinguishable from the pair (oe, os). The corresponding average variance is of the order 10^{-5}.

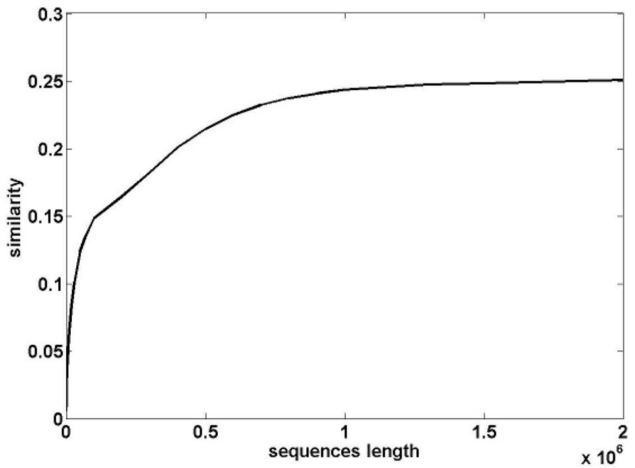

Fig. 5. Similarity of digit sequences of e and $\sqrt{2}$ versus length of subsequences

Again it can be concluded that all five sequences behave like different instances of a random sequence. The word spectra of any two normal random sequences $1\,000$ digits long have on average 1.6% of words in common while sequences two million digits long have approximately a quarter of words in common. The increase in similarity per $1\,000$ digits slows down from 10^{-2} at the beginning to 10^{-6} for sequence two million digits long. The conclusion following from the experiment is rather obvious: all five sequences are statistically indistinguishable. So both sequences of digits being the fractional part of square root of 2 and base of natural logarithm of length not exceeding 2 million are probably random sequences. Another conclusion also follows that the similarity between any random, two million digits long sequences is about 0.25.

For comparison it is worth considering Chapernowne's number $0.12345678910111213141 5\ldots$, which is constructed by concatenating "0" digit and digits of consecutive natural numbers as the fractional part. The number is assumed to be transcendental. The set of two million long sequences of decimal digits is considered. As follows from Fig. 6, the plot of digit fractions suggests that the sequence is not normal.

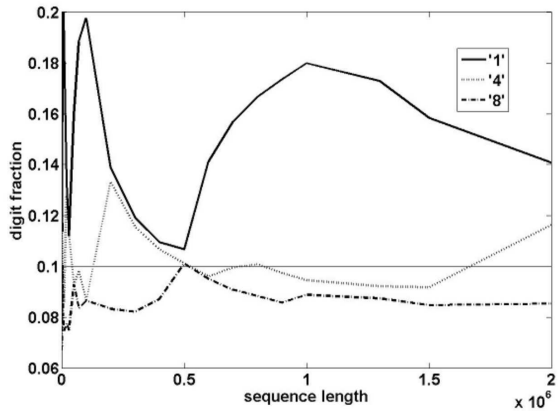

Fig. 6. Share of selected digits in sequence of Chapernowne's number

After decomposition of the original and shuffled sequences into sets of words, the complexity of both sequences was found. The detailed results are presented in the supplementary material. As before, the length and number of '0' digits in the word are two attributes considered for correlation. Fig. 7 clearly shows that the sequence is not random.

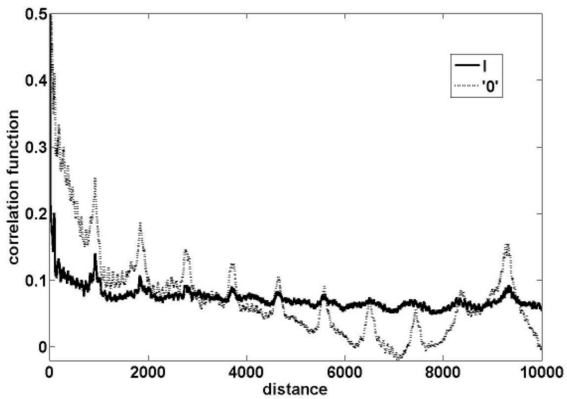

Fig. 7. Correlation functions of word length and number of digit "0" instances in word in sequence of Chapernowne's number

The correlation function of words length $\mathrm{cor}(l;i)$ displays maxima at various distances i, so does the function $\mathrm{cor}("0";i)$. The correlations are very long, of order 10^5. It can be concluded that a set of digits of Chapernowne's number sequence has quite different global properties from sequences of e or $\sqrt{2}$ numbers.

IV. CONCLUSIONS

The main objective of the present paper was to demonstrate that suitable parsing of an arbitrarily long (in principle) symbolic sequence into a set of strings called words has the ability to discover several properties of a single symbolic sequence and some relations between many sequences over the same alphabet. What is particularly interesting is the correlations of attributes of words across a sequence and the algorithm dividing a set of sequences into a small number of relatively homogenous subsets on the basis of their similarity and specific distance between them. Demonstrations were performed on very reliable data; mitochondrial DNA sequences and sequences of digits of irrational numbers. Rather satisfactory results open the door to more advanced investigations. One of them is the identification of hypothetical ancestral sequences of the protein families of extant species. Another one is the analysis of time series generated by dynamical systems, based on their specific features, which is an important issue in diverse areas, including physiological data time series. There is some evidence that the symbolic rather than numerical analysis can be more fruitful in the search for the characteristics of non stationary time series.

Acknowledgement

The present author thanks the anonymous reviewer for their useful comments.

References

[1] A. Lempel, J. Ziv, *On the complexity of finite sequences*, IEEE Trans. Inform. Theory **22**, 75-81 (1976).

[2] D.-G. Ke, Q.-Y. Tong, *Easily adaptable complexity measure for finite time series*, Phys. Rev. E **77**, 066215 (2008).

[3] B. Kozarzewski, *A method for nucleotide sequences analysis*, CMST **18** (1), 5-10 (2012).

[4] M-S. Yang and K-L. Wu, A *Similarity-Based Robust Clustering Method*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **2** (4), 434-448 (2004).

[5] Y. Liu, *The Numerical Characterization and Similarity Analysis of DNA Primary Sequences*, Internet Electronic Journal of Molecular Design **1**, 675-684 (2002).

[6] J. Wen, C. Li, *Similarity analysis of DNA sequences based on the LZ complexity*, Internet Electronic Journal of Molecular Design **6**, 1-12 (2007).

[7] A. Kelil, S. Wang, Q. Jiang, R. Brzezinski, *A general measure of similarity for categorical sequences*, Knowl. Inf. Syst., **24**, 197-220 (2010), (DOI 10.1007/s10115-009-0237-8).

[8] S. Kumar, A. Filipski, *Multiple sequence alignment: In pursuit of homologous DNA* positions, Genome Research **17**, 27-135 (2007).

[9] S. Vinga, J. Almeida, *Alignment-free sequence comparison – a review*, Bioinformatics **19**, 513-523 (2003).

[10] L.R. Dice, *Measures of the Amount of Ecologic Association Between Species*, Ecology **26** (**3**), 297-302 (1945).

[11] T. Kanungo, N.S. Netanyahu, A.Y. Wu, *An Efficient k-Means Clustering Algorithm: Analysis and Implementation,* IEEE Trans. Pattern Analysis and Machine Intelligence, **24**, (7), 881-892 (2002).

[12] B. Kozarzewski, A *representative set method for symbolic sequence clustering*, CMST **19** (2), 35-47 (2013).

[13] G.P. Dresden, Three *Transcendental Numbers from the Last Non-Zero Digits of , F_n, and $n!$*, Mathematical Magazine, **81** (2), 96 (2007).

[14] D. Bailey, P. Borwein, S. Plouffe, *On the rapid computation of various polylogarithmic constants*, Mathematics of Computation, vol. 66, (218), 903-913.

[15] D.H. Bailey, *A Compendium of BBP-type Formulas for Mathematical Constants*, http://crd-legacy.lbl.gov/~dhbailey/dhbpapers/bbp-formulas.pdf (2013).

[16] R. Nemiroff and J. Bonnell, http://apod.nasa.gov/htmltest/rjn_dig.html, http://apod.nasa.gov/htmltest/gifcity/e.2mil, http://apod.nasa.gov/htmltest/gifcity/sqrt2.2mil.

SUPPLEMENTARY MATERIALS

Supplementary materials are available on the web page: http://cmst.eu/articles/a-new-method-for-symbolic-sequences-analysis-an-application-to-long-sequences/

Bohdan Kozarzewski received PhD degree in physics (1965) at the Jagiellonian University in Cracow. Habilitated in 1975 in solid state theory. Since 1989 Professor at the Institute of Physics Technical University Cracow, sine 2006 at the University of Information Technology and Management, Rzeszow. Present research activity in computer modeling of nonlinear dynamical systems and time series analysis. Author and co-author of about 50 scientific publications.

DOI:10.12921/cmst.2014.20.03.93-100