COMPUTATIONAL METHODS IN SCIENCE AND TECHNOLOGY **18(1)**, 5-10 (2012)

A Method for Nucleotide Sequence Analysis

Bohdan Kozarzewski

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

(Received: 22 February 2012; revised: 29 May 2012; accepted: 7 June 2012; published online: 21 June 2012)

Abstract: Symbolic sequence decomposition into a set of consecutive, distinct subsequences (mers) is presented. Several statistical distributions of nucleotide subsequences are defined and analysed. Sequence entropy and similarity between sequences in terms of mer lengths distribution are defined. An alignment-free method of phylogenetic tree construction is proposed.

Key words: sequence parsing, similarity measure, phylogenetic tree

I. INTRODUCTION

A DNA sequence is a string consisting of four types of bases. It is an example of a symbolic sequence over the four letters alphabet. The length of DNA strings creates an urgent need for efficient methods of statistical analysis of long symbolic sequences. The approaches known so far have not been robust enough. Among them there is the distribution of the frequency of occurrence of subsequences of consecutive symbols, called oligonucleotides or *k*-mers. For example, the distribution of *k*-mers lengths is believed to be related to the complexity of a symbolic sequence. However, the question arises which *k*-mer selection and sizes have to be considered. The aim of the present paper is twofold: to develop the method of sequence decomposition into a set of distinct, non-overlapping *k*-mers (hereinafter called mers) of size greater than 2, and to give some examples of how useful that set can be. The paper is organized as follows: in the first section, entropy, similarity and examples of statistical measures of symbolic sequences are defined and shortly discussed. As an illustration, the mer spectra of two short sequences of nucleotides are obtained and their entropy and similarity are calculated. In the second section, the *Saccharomyces cerevisiae* chromosome M is analysed. Distribution of selected sequences of nucleotides between mers of all lengths is calculated. The long-distance statistical correlation between bases loci is shown. In the third section, mer spectra of complete mitochondrial DNA sequences of 13 human closest relatives are found and their similarity matrix which has been used in the following for construction of evolutionary tree for that set of species is calculated.

II. DECOMPOSITION OF SYMBOLIC SEQUENCE

Characterization of symbolic sequences by an ordered set of subsequences has numerous advantages. However, the question arises how to perform sequence decomposition to obtain suitable subsequences. The first parsing algorithm was developed by Lempel and Ziv [1], in order to define quantitative measure of symbolic sequence complexity. However, the Lempel-Ziv algorithm has at least one defect, their complexity measures randomness mainly. There was an attempt [2] to exploit the parsing algorithm by Lempel and Ziv to sequences comparison, but they used too intricate and not unique distance measure. Ke and Tong [3] proposed some substantial modification of the Lempel-Ziv algorithm by adding replication operation. Recently, Kása [4] has considered similar decomposition of the symbolic sequences into a set of substrings called *d*-substrings. However, Kása was interested in some special sequences and did not provide a general decomposition algorithm.

Both mentioned approaches claim that the total number (different in the two cases) of substrings is supposed to be a measure of sequence complexity. The relation between symbolic sequence entropy and complexity is discussed in [5]. On the other hand, a distance measure between symbolic sequences based on the Lempel-Ziv complexity is presented in [6]. In all papers quoted above only the number of subsequences was considered as important. The most interesting result of the present paper is the statement that the whole set of mers (not only their number) is a very rich resource of information on a symbolic sequence.

Sequence decomposition leading to characterizations of numeric sequences by utilizing the subsequence representation has been introduced in [7]. Subsequences (mers) arise as a result of a specific parsing algorithm applied to the sequence of interest. There the Ke and Tong decomposition algorithm was used after some modifications. Subsequence *s* arises as a result of appending some symbol c_1 from the primary (the sequence of nucleotides) sequence by the following symbols. After each step subsequence *s* is checked. In the beginning, whether it is chaotic, if it is not then *s* is checked whether it is periodic. If it is not periodic, the whole set of the subsequences obtained so far (the set of all *s*'s) is searched for the presence of subsequence *s*. If subsequence *s* has been found, it is appended by the following symbol c_2 and so on. Appending stops when *s* has not been found, and it becomes a new distinct subsequence (mer). The code of the presently applied parsing algorithm is available on request.

After the parsing procedure is completed we are in possession of a set of distinct mers representing the primary sequence, and the mers can be consecutively enumerated. The enumerated set of mers is called mer spectrum *S* of the primary symbolic sequence. Once the mer spectrum is found the nonparametric measure of several interesting quantities can be defined. In the present paper the mer spectrum is considered as an ensemble of strings. The ensemble can be used to find the probability distribution of various quantities. An example is mer lengths distribution which is used in the following to define Shannon entropy of the primary sequence. A common set of mer spectra of two sequences (their intersection) is proposed as a measure of similarity between the sequences. Then the alignment-free method of sequence comparison is possible as well.

III. MER SPECTRUM

The mer spectrum can be used to identify several properties of a symbolic sequence. The large group of properties includes statistical distributions of any set of symbols present in mers. Another one is the entropy of sequence and the similarity measure between two sequences. When the mer spectrum is considered as an ensemble of strings it is straightforward to calculate the string lengths distribution. Let $p(l)$ be the probability to find string of length *l* in the mer spectrum, then

$$H = -\sum_{l} p(l)\log_2\left(p(l)\right)$$

can be considered as Shannon entropy of a symbolic sequence. In the present context *H* is the configuration entropy. Unfortunately entropy (including Shannon entropy) has various interpretations. Brissaud [8] suggests that entropy measures freedom of the system. However, it is not clear what exactly the freedom of a nucleotide sequence means.

Estimating the degree of similarity between a set of sequences is an important problem in molecular sequence analysis. The similarity helps to discover relationships within a set of sequences which is a prerequisite to comparative genomic analyses. For example, similarity matrix of β-globin sequences approximately 100 bases long is (among others) used to construct a species phylogenetic tree. The similarity measure is dual to the often used distance between sequences. There are numerous measures of distance between symbolic sequences in use. For example, the distance between two sequences is related to the minimum number of events required to convert one sequence (or its segments) into another. For comparison of very long sequences, e.g. whole genomes, an approach based on the frequency of *k*-mers that appear in a set of sequences is used [9]. The approach consists in counting occurrences of *k*-mers in a sequence, for *k* typically ranging from 2 to 8 and applying different statistical methods to *k*-mer distribution.

The similarity measure proposed in the present paper has the following advantages: it works well for pairs of sequences of different and arbitrary length, and it does not depend on any distance measure. With the use of spectrum, the similarity between nucleotide sequences C_1 and C_2 can be defined simply as

$$\mathrm{sim}(C_1, C_2) = \frac{d(\mathrm{int}(S_1, S_2))}{\sqrt{d(S_1)d(S_2)}}$$

where $\mathrm{int}(S_1, S_2)$ is a one-column vector of mers common for the two spectra (intersection of S_1 and S_2), and $d(S)$ means dimension (length) of spectrum *S*. Similarity measures is normalized to fall into the range between null and one. Similarity is a symmetric function of its arguments, so for the set of sequences, the similarity matrix is symmetrical $\mathrm{sim}(C_j, C_i) = \mathrm{sim}(C_i, C_j)$.. It can also be useful to know which regions of the two sequences are similar, and which are different. As a measure which shows a similar region (common mer numbers) between two sequences the vector $\text{int}(S_1, S_2)$ of $d(\text{int}(S_1, S_2))$ rows can be appended by two columns. As a result, we get a three-column matrix. The first element of each row shows a mer common for the two sequences, the second one is the index of the mer in S_1 spectrum, and the last one shows the index of the same mer in S_2 spectrum.

To demonstrate how the introduced measures work, two short sequences piRNAof *H. sapiens* are considered. They were downloaded from GenBank, http://www.ncbi.nih.gov/. The accession number of the first one is DQ601956.1, and of the second one – DQ601954.1, they are 29 and 26 bases long, respectively. Partition algorithm yields the following mer spectrum for the first sequence S_1 = (TAGTG, ATGCTTC, ATGGA, CAAG, GCTTG, GCA), it is $d(S_1) = 6$ mers long. Probability $p(n)$ to find n 'A' symbols in any mer is $p(0)$ = 1/6, $p(1)$ = 3/6, $p(2)$ =2/6. Probability $p(l)$ to find 'GC' sequence among mers of length l is $p(2) = p(4) = 0$, $p(3) = p(5)$ =1/4, $p(7)$ = 1/4, $p(2)$ =2/4. Mer length probabilities are $p(3) = p(4) = p(7)$ = 1/6, $p(5)$ = 3/6, $p(6)$ = 0. Therefore, sequence entropy is given by

$$H(s_1) = -\frac{3\log_2(1/6) + 3\log_2(3/6)}{6} = 1.79.$$

The mer spectrum of the second sequence S_2 = (TAGTG, ATGAC, ATTG, TGGA, CAAG, CTGC) is also 6 mers long and the corresponding entropy is 0.92. The intersection of the two spectra consists of 2 mers int(S_1, S_2) = (TAGTG, CAAG), so similarity between the two sequences is

$$\text{sim}(s_1, s_2) = \frac{2}{\sqrt{6*6}} = \frac{1}{3}.$$

Besides, common mer pairs are $(S_1(1), S_2(1))$ and $(S)_1(4)$, $S_2(5))$.

IV. ANALYSIS OF *SACCHAROMYCES CEREVISIAE* CHROMOSOME M

The increasing amount of DNA data resulted in many new problems of statistical nature. Mer spectra can help to cope with large data sets. To provide closer demonstration of mer spectrum ability a relatively long sequence of nucleotides is considered. *S. cerevisiae* chromosome M was downloaded from GenBank, and the accession number of the sequence is NC_001224.1. The sequence is 85 779 bases long, its 'A', 'C', 'G' and 'T' symbols content is 36 169, 6863, 7813 and 34 934, respectively. Sequence decomposition results in the mer spectrum consisting of 10 312 mers of length from 3 to 19 bases. Once the spectrum is known the entropy and many statistical characteristics can be obtained. The simplest are three probabilities. The first of them answers the question about probability $p(l)$ that a mer picked up randomly is l bases long. Another two give probability $p_{A/T}(l)$ (or $p_{C/G}(l)$) that in a randomly selected mer of length l at least one A or T (C or G) base can be found.

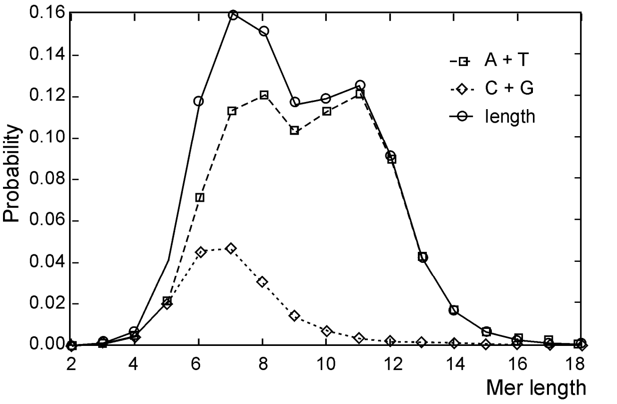

Fig. 1. Probability distributions of 'A'+'T' and 'C'+'G' content found in the mer spectrum of the sequence NC_001224 (GenBank)

Figure 1 shows $p(l)$ and weighted probabilities $c_{A|T}\,p_{A|T}(l)$ and $(1 - c_{A|T})p_{C|G}(l)$, where $c_{A|T} = 0.83$ is ratio of $C + T$ bases. The $p(l)$ probability is the sum of the other two $p(l) = c_{A|T}\,p_{A|T}(l) + (1 - c_{A|T})p_{C|G}(l)$. From Fig. 1 it follows that distributions of A and T bases (which are very similar) are qualitatively different from that of C and G bases. It is worth noting that the shape of base distribution like that of $p_{A/T}(l)$ has not been observed elsewhere so far. When the primary sequence is thoroughly shuffled to remove any correlation between bases loci both distribution functions become quite similar, as it follows from Fig. 2.

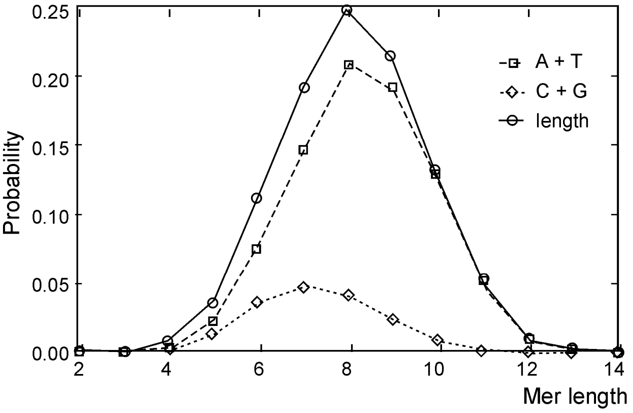

Fig. 2. Probability distributions the same as in Fig. 2 but for a shuffled sequence

One may ask whether the above probabilities can be fit to any known partial distribution function. The relatively good fit to weighted sum of two Weibull [10] distribution functions was found. The Weibull partial distribution function

$$p(x) = \frac{b}{a}\left(\frac{x - x_0}{a}\right)^{b-1} e^{\left(\frac{x-x_0}{a}\right)^b}$$

is defined for $x \geq x_0$, the scale $-a$ and shape $-b$ parameters are positive numbers. Figure 3 shows decomposition of $p_A(l)$ (which is almost the same as $p_T(l)$) into two Weibull distribution functions with $x_0 = 2$ and parameters W1:(a_1 = 9.05, b_1 = 5.01), and W2:(a_2 = 5.41, b_2 = 4.98), and ratio of W1 Weibull distribution function is 0.70.

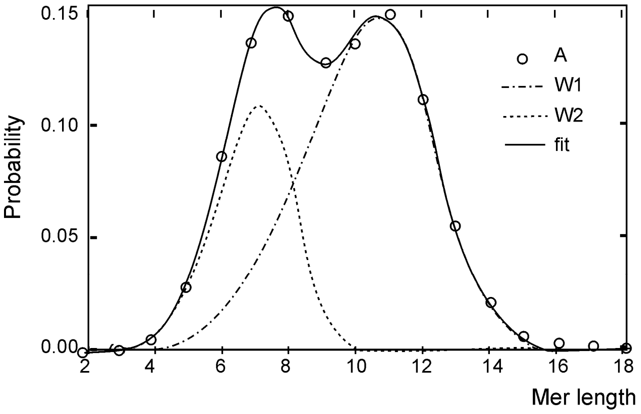

Fig. 3. Fit of ‘A’ content probability distributions to two Weibull distributions found in the mer spectrum of the sequence NC_001224 fitted to the weighted sum of two Weibull distri bution functions

For shuffled sequence probability $p_A(1)$ is rather well approximated by single Weibull distribution function of parameters a = 6.85 and b = 4.57 as Fig. 3 shows.

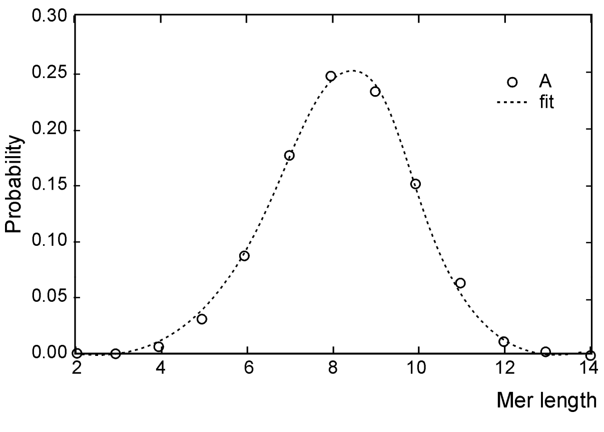

Fig. 4. Fit of ‘A’ content for shuffled sequence to single Weibull distribution

One can also be interested in distribution of particular base among mers, for example in the probability $p_A(l)$ that there are l A bases in a randomly selected mer of arbitrary length. Distribution of particular short bases sequence among mers can be found as well as correlation functions between position of two (or more) bases within the same (or different) mers. The entropy of a primary sequence equals 3.13 and is significantly larger than the shuffled one, which is 2.71. Sliding window analyses of the mentioned as well as other distributions are also possible.

V. EVOLUTIONARY TREE

In this section, a phylogenetic tree is constructed from complete mitochondrial DNA sequences for 13 human closest relatives. Mitochondrial DNA in mammals has a faster mutation rate than nuclear DNA sequences. The faster rate of mutation produces more variance between sequences and is an advantage when studying closely related species. Usually the control region is used for various applications, due to higher overall mutation rate. However, local differences in the mutation rate among nucleotide positions within control region exist. Therefore, for more accurate inference control region sequence data are usually supplemented with some coding region data. However, the best choice would probably be to rely on a complete mitochondrion sequence. The set of 13 mitochondrion sequences was downloaded from the GenBank database. With the use of a parsing algorithm, mer spectra of all sequences were obtained. For example, *Homo sapiens* mitochondrion genome (accession number AF347015.1) includes 15 571 bases, its spectrum consists of 2741 mers. Then similarity matrix between them was calculated and is presented in Table 1.

The algorithm for the phylogenetic tree that has been used is modified Unweighted Pair Group Method Using Arithmetic Mean because of its simplicity. It consists of the following steps performed on a collection of all sequence pairs represented by their similarities. Find the closest (of maximal similarity) pair of species. Join leaves to get node which becomes the end of branch representing hypothetical most recent common ancestor. Similarity between any species and the ancestor is assumed to be the average of similarities between the species and the two descendants. Update collection of species, deleting both leaves and replacing them by the branch of hypothetical ancestor, and correspondingly similarities between branches. Continue until there is no sequence left. The result is shown in Fig. 5.

Table 1. Similarity matrix between 13 selected mitochondrion sequences based on the mer spectra analysis

Speciesname	Accession #	Similarity												
		1	2	3	4	5	6	7	8	9	10	11	12	13
H. sapiens (Yoruba)	AF347015.1	1.00	0.59	0.58	0.54	0.53	0.54	0.54	0.53	0.52	0.53	0.52	0.51	0.51
H. sapiens (Cambridge)	NC_012920.1		1.00	0.57	0.54	0.54	0.55	0.54	0.53	0.52	0.53	0.52	0.51	0.52
H. sapiens neanderthal.	NC_011137.1			1.00	0.55	0.53	0.54	0.53	0.54	0.52	0.52	0.51	0.52	0.52
P. troglodytes verus	X93335.2				1.00	0.60	0.58	0.54	0.54	0.53	0.51	0.52	0.52	0.52
P. troglodytes	NC_001643.2					1.00	0.56	0.53	0.54	0.51	0.51	0.50	0.51	0.50
P. paniscus	GU189661.1						1.00	0.53	0.54	0.52	0.52	0.52	0.51	0.52
G. gorillagorilla	NC_011120.1							1.00	0.59	0.52	0.53	0.52	0.51	0.51
G. gorilla	NC_001645.1								1.00	0.53	0.53	0.51	0.52	0.50
P. pygmaeus	NC_001646.1									1.00	0.52	0.51	0.51	0.50
H. lar	NC_002082.1										1.00	0.52	0.52	0.52
M. mulatta	NC_005943.1											1.00	0.55	0.52
M. sylvanus	NC_002764.1												1.00	0.52
P. hamadryas	NC_001992.1													1.00

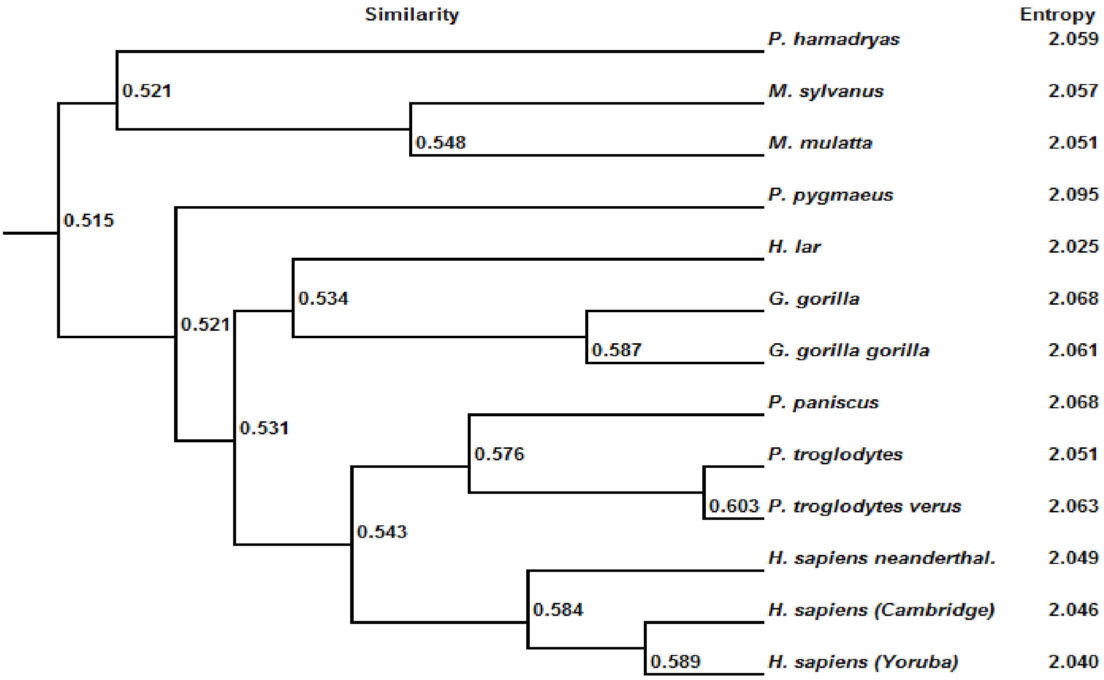

Fig. 5. Phylogenetic tree for 13 human closest relatives

The number to the right of the node indicates the similarity between corresponding branches, while the number to the right of the species name is the entropy value of the corresponding sequence. Noteworthy is the decreasing entropy when moving up within each clade with one exception of *H. sapiens neanderthalensis*, aswell as the decreasing similarity when moving from down right to up left of the tree. The tree is in agreement with the last updated phylogenetic tree by M. van Oven except for the position of *H. lar* which is grouped with gorillas and *P. pygmaeus* which is grouped with Hominidae instead of Hominінae according to van Oven's tree [11].

V. CONCLUSIONS

The main objective of this paper was the presentation of substantially new method for symbolic sequence analysis. It is relied on efficient and unique decomposition of primary sequence into distinct subsequences (spectrum of mers). Spectrum makes possible evaluation of various distribution functions, long-distance correlation between bases and particular bases patterns. Besides, distribution of mer lengths is related to the Shannon entropy of the primary sequence. Spectrum allows for simple, practical measure of global similarity (or distance) between two symbolic sequences over the same alphabet.The measure proposed does not require any previous sequence alignment and works well for sequences of arbitrary length. The utility of the measure for phylogenetic tree construction based on whole mitochondrial genomes was demonstrated. The parsing algorithm and similarity measure work for protein sequences as well.

Acknowledgments

Author thanks two anonymous reviewers 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] H.H. Out, K. Sayood, *A new sequence distance measure for phylogenetic tree construction*. Bioinformatics 19, 2122-2130 (2003).

[3] D.-G. Ke, Q.-Y. Tong, *Easily adaptable complexity measure for finite time series*. Phys. Rev. E77, 066215 (2008).

[4] Z. Kása, *On the d-complexity of strings*. http://arxiv.org/abs/1002.2721v1.

[5] C. Adami, N.J. Ceref, 1999. *Physical complexity of symbolic sequences*. arxiv: adap-org/9605002v3

[6] J. Wen, C. Li, *Similarity analysis of DNA sequences based on the LZ complexity*. Internet Electron. J. Mol. Des. 6, 1-12 (2007).

[7] B. Kozarzewski, *Multilevel time series complexity*. Journal of Applied Computer Science 19, 2, 61-71 (2011).

[8] J.-B. Brissaud, *The meaning of entropy*. Entropy 7, 68-96 (2005).

[9] Y.-H. Chen, S.-L. Nyeo, C.-Y. Yeh, *Model for distribution of k-mers in DNA sequences.* Physical Review E72, 011908 (2005).

[10] W.K. Brown, K.H. Wohletz, *Derivation of the Weibull distribution based on physical principles and its connection to the Rossin-Rammler and lognormal distributions*. Journal of Applied Physics 78, 2758-2763 (1995).

[11] M. van Oven, http://www.phylotree.org (2009).

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.