ONCONET
UDC: 61:651.2:681.31:615.07

On direct sequential analysis of heart rate variability signals

Dragana Bajić[1], Tatjana Lončar-Turukalo[1], Olijandra Šibarević[2]

[1]Department of Communications and Signal Processing, University of Novi Sad, [2]Faculty of Technical Sciences, Univ. of Novi Sad; Address correspondence to: Dragana Bajić, Department of Communications and Signal Processing, University of Novi Sad, 21000 Novi Sad, Serbia & Montenegro, E-mail: lmcdra@eunet.yu; The manuscript was received: 27.10.2004, Provisionally accepted: 10.11.2004, Accepted for publication: 10.01.2005

ABSTRACT

Heart rate variability analysis represents one of the most promising and the most commonly used quantitative measures of the cardiovascular autonomic regulatory system. The analysis includes traditional statistical analytical tools and a number of new methods based on nonlinear system theory, recently developed to give better insight into complex HR. This paper introduces a direct sequential analysis.

KEY WORDS: *Heart Rate; Sequence Analysis: Entropy; Data Interpretation, Statistical; Models, Statistical; Models, Theoretical*

INTRODUCTION

Heart rate variability (HRV) has become the conventionally accepted term to describe variations of interval between consecutive heartbeats, as well as the oscillations between consecutive instantaneous heart rates. It represents one of the most promising and the most commonly used quantitative measures of the cardiovascular autonomic regulatory system. The analysis include traditional statistical analytical tools both in time and spectral domain and a number of new methods based on nonlinear system theory, recently developed to give better insight into complex HR dynamics, such as fractal correlation properties, the slope of the power law relation and approximate entropy (ApEn). These methods might reveal malign abnormalities at early stage that may not be uncovered by traditional measures. However, the significance and meaning of these different measures of HRV are more complex than generally appreciated, and there is a potential for incorrect conclusions and for excessive or unfounded extrapolations. Besides, in spite of general opinion that HRV time series is easy to obtain, visual inspection and careful manual editing after automatic extraction is absolutely necessary. For 24 hours' Holter signal it presents a cumbersome task [1-4].

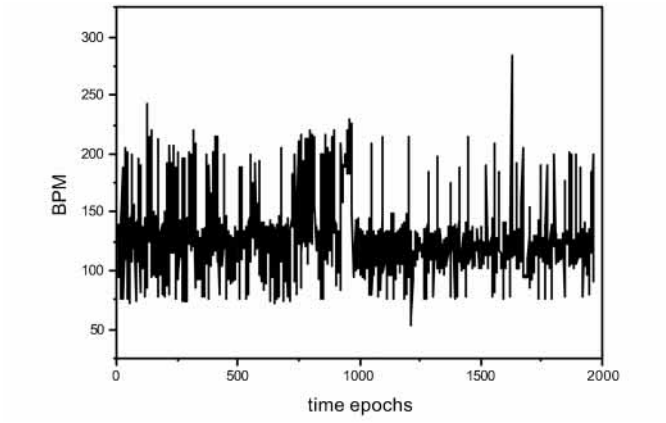

Figure 1. Sample of HRV time series

This paper discusses the possibility of application of a recently developed direct sequential analysis to HRV signals, based upon the sequence matching. The subsequent section gives a brief review of some other methods based upon a template (sequence) matching, with illustrative examples. The third section gives a theoretical approach to the new direct sequence evaluations. The method is illustrated using analog ECG Holter signals of clinically healthy children recorded at Children's hospital (Tirsova) are digitalized and HRV extracted at Faculty of Technical Sciences, Novi Sad. A part of one HRV signal (in beats per minute, BPM) is shown in Fig. 1.

SEQUENCES AND TEMPLATES

The HRV signal samples form a vector of length L_S, denoted by y = [y(j)], j=1,...,L_S. An analytical approach based upon the sequence (template) matching tries to unveil if the similar set of samples is followed by other similar set of samples. Therefore, a "sequence" (or a "template" of length N is defined as a short vector $x_N(i)$ = [y(i+k-1)], k = 1,...,N, i=1,...,L_S-N+1 that is a part of a long time series. For each pair of sequences a distance $d(x_N(i), x_N(j))$, i,j=1,...,L_S-N+1 is defined. It can be maximal absolute distance, mean square distance or any other distance suitable for the current investigation.

Approximate entropy and its modifications

ApEn can be defined as a "regularity statistic" that quantifies the unpredictability of fluctuations in a time series. Intuitively, one may reason that the presence of repetitive patterns of fluctuation in a time series renders it more predictable than a time series in which such patterns are absent. A time series containing many repetitive patterns has a relatively small ApEn; a less predictable (i.e., more complex) process has a higher ApEn. Therefore, ApEn estimates likelihood that patterns of certain length that are close one to another would remain close if the pattern length increases. The procedure for its evaluation from a time series y is simple: number of N-tuples (and N+1-tuples) for which $d(x_N(i),x_N(j))$, i,j=1,...,L_S-N+1 is within a specified distance r are counted an processed:

$$C_i^N(r) = \frac{1}{L_s - N + 1} \cdot \sum_{j=1}^{L_s - N + 1} z_j$$

$$z_j = \begin{cases} 0, d(x_N(i), x_N(j)) \geq r \\ 1, d(x_N(i), x_N(j)) < r \end{cases} \qquad (1)$$

where estimates the probability that any sequence $x_N(j)$ is within a distance r from the template $x_N(i)$. Then the approximate entropy can be estimated as:

$$ApEn(N,r,L_s) = \frac{1}{L_S - N + 1} \cdot \sum_{i=1}^{L_s-N+1} \ln\left[C_i^N(r)\right] - \frac{1}{L_S - N} \cdot \sum_{i=1}^{L_s-N} \ln\left[C_i^{N+1}(r)\right] \quad (2)$$

Sets of ApEn curves, as well as sets of curves that correspond to its modification (an unbiased one (SampEn) [4] and sliding window-by-window approach [5]) are shown in Figures 2-4.

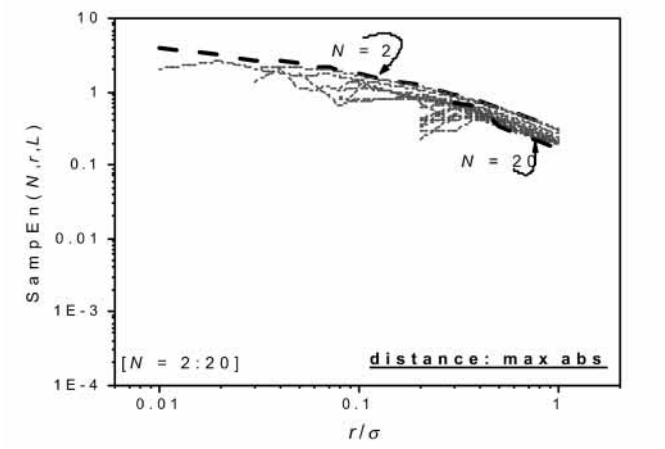

Figure 2. Approximate entropy for child "L"

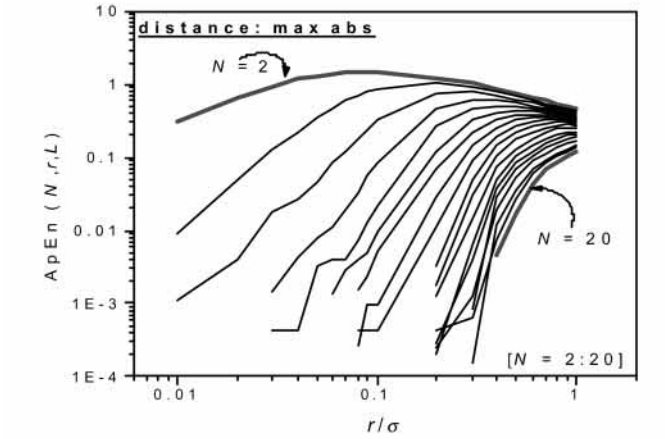

Figure 3. Sample entropy for child "L"

Correlation dimension approach - D_C

Correlation dimension estimates fractal dimension of an attractor from a time series. An attractor dimension itself shows a statistical measure of the self-similarity of the geometry of the sets of points in the phase space, i.e. the number of degrees of freedom necessary to describe a process. In our case, each sequence (N-tuple) is a point in an N-dimensional phase space.

DC is estimated from a time series using the correlation integral $C_N(r)$ that measures the number of points correlated with each other in a sphere of radius r around the point $x_N(i)$ [6]:

$$C_N(r) = \frac{1}{\left(L_s - N + 1\right)\cdot\left(L_s - N\right)} \cdot \sum_{j=1}^{L_s-N} \sum_{j=i+1}^{L_s-N+1} z_{ij}$$
$$z_{jj} = \begin{cases} 0, d(x_N(i), x_N(j)) \geq r \\ 1, d(x_N(i), x_N(j)) < r \end{cases} \quad (3)$$

This quantity is similar to probability estimate (Eq (1)). The differences are a) triangular rather than rectangular summation; b) squared instead of max. absolute distance; and c) for an ApEn approach, standard deviation of each separate signal is used as a scaling factor for r.

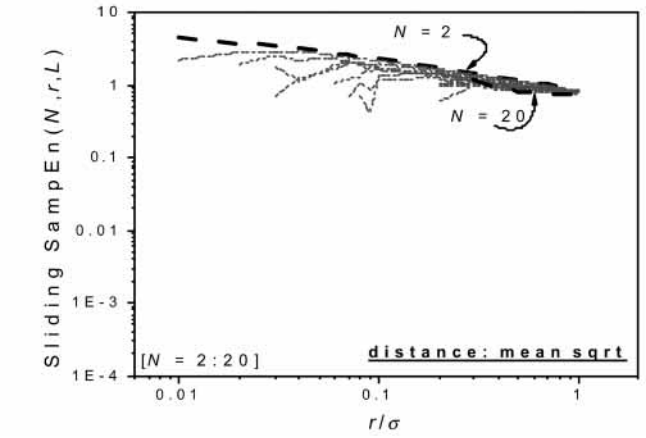

Figure 4. Sliding entropy for child "L"

The correlation dimension is evaluated as:

$$D_C = \lim_{r \to 0} \frac{ld(C_N(r))}{ld(r)} \quad (4)$$

Its value can be obtained by plotting $ld(C_N(r))$ vs. ld(r). The slope of the resulting straight lines, for different N, tends to constant value D_C, as explained in [6].

The D_C analysis is closely related to multifractal property of HRV signal. Contrary to the monofractal signals that are homogeneous, multifractal signals can be decomposed into many subsets. The statistical properties of the different subsets are characterized by local Hurst exponents h that shows the local singular behavior and can be quantified by the function D(h) - fractal dimension of the subset of the time series [7].

Illustrative examples and a note on stationarity

As previously mentioned, HRV time-series are obtained from the children's ECG, known to be extremely non-stationary and mutually different. Our sample signals were no different; yet, the children's cardiologist claims all of them to be healthy [Dr med. sci. Nina Žigon: consultations]. After the HRV extraction, a cumbersome task of visual inspection of all time series was performed, to correct the 5% errors that software for extraction makes. At last, from each 24h signals two 15 min HRV sub-series were chosen, the ones that seemed (by mere visual inspection) to be stationary. To verify the assumption, a stationarity test is performed [8].

Each HRV series is divided into K intervals. For each interval a mean value mi and variance σ^2_i are estimated, i = 1,...,K. Then the following sums are evaluated, both for mi and σ^2_i:

$$A = \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} a_{ij}$$
$$a_{ij} = \begin{cases} 1, m_i > m_j (or \sigma_i^2 > \sigma_i^2) \\ 0, m_i \leq m_j (or \sigma_i^2 > \sigma_i^2) \end{cases} \quad (5)$$

The acceptance region for the stationarity hypothesis at an α level of significance was considered done by:

$$\left[A_{K;1-\alpha/2} < A < A_{K;\alpha/2}\right]$$

The corresponding border values $A_{K;1-\alpha/2}$ and $A_{K;\alpha/2}$ are obtained from discrete distribution function:

$$\Pr\{A=i\},\ i=0,...,K\cdot(K-1)/2 \quad (6)$$

This probability can be obtained knowing that probability that exactly b elements exceeds the

value of (K-n)th element, given that m elements in total exceeds the value of (K-n)th one:

$$\Pr\{b/K-n,m\} = \frac{\binom{b}{n}\cdot\binom{K-n-1}{m-b}}{\binom{K-1}{m}} \quad (7)$$

$$b=0,...,n; n=1,...,K-1; m=b,...,K-n-1+b$$

$$\Pr\{A\} = \sum_{m,b1+b2+..+bK-1=A} \Pr\{b_1/K-1,m\} + \Pr\{b_2/K-2,m\} + ... + \Pr\{b_{K-1}/1,m\} \quad (8)$$

Table 1. Values of A for m and σ^2; for α=0.05 and K=10 values should be 11<A<33

	K=10 m	σ^2
A1	5	29
A2	18	20
C1	7	27
C2	9	39
D1	25	38
D2	24	21
L1	23	12
L2	30	15
S1	39	10
S2	25	20

Although the series were chosen to be, according to the visual inspection, "the most stationary ones", only 5 of them passed both tests with α=0.05 level of significance.

Figure 5 shows the correlation dimension - dashed lines for non-stationary series. It is interesting to note that it was not possible to extract DC - neither automatically, nor manually, for one subject. Fractal dimension, again with dashed lines representing the non-stationary data, is shown in Fig. 6. Thick lines show the examples that fits the best and the worst the stationarity hypothesis.

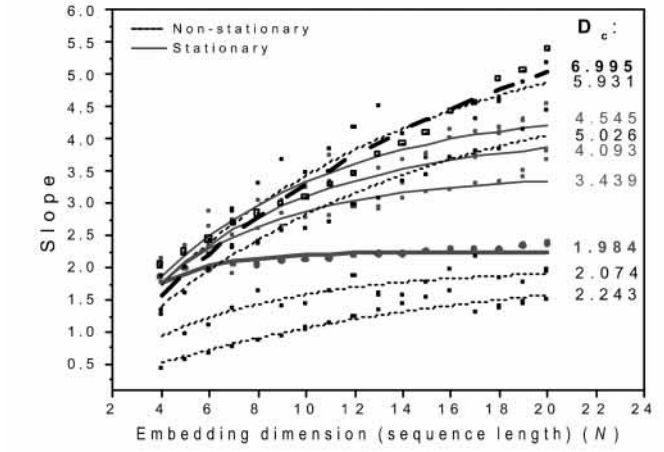

Figure 5. Correlation dimension D_C

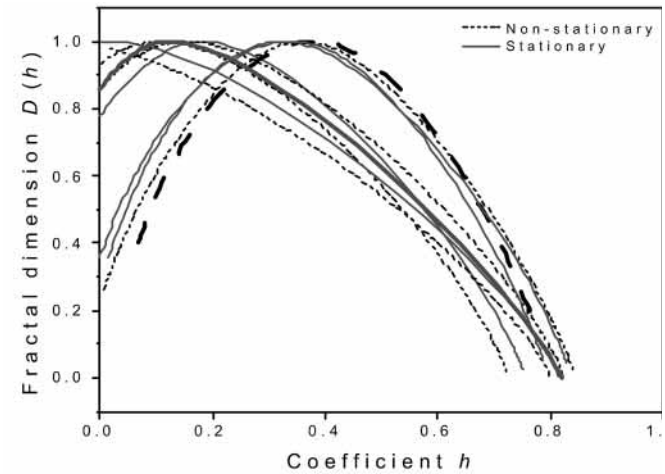

Figure 6. Fractal dimension D(h)

It is known that children's HR is extremely variable. From the above figures no conclusions can be done. The fact that the children are healthy is according to the children's cardiologist opinion.

DIRECT SEQUENTIAL ANALYSIS

Direct sequential analysis deals with the analysis of time parameters of the observed signals - the expected value of time units (number of samples) between the predefined set of M sequences (templates), and expected number of time units from the random starting position until one of the predefined M sequences is found. The first time distance is of the first type, the other one of the second type. The set of M sequences (templates) under consideration are the ones that match a certain criterion. For introductory explanation of this new analytical approach, the criterion for forming the sequence set would be the criterion of "smoothness" - the expected number of "rough" samples between the "smooth" intervals.

Time parameters and empirical series

Suppose that the HRV signal is shown in Fig. 7 and the smooth intervals are the ones for which the absolute value of sample do not exceed certain level.

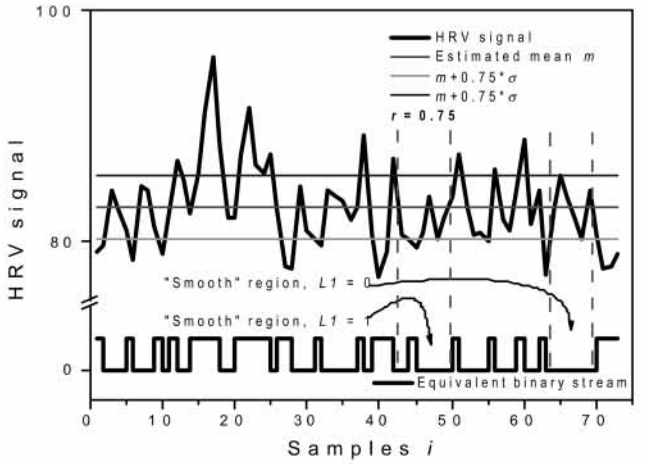

Figure 7. HRV signal, "smooth regions" and binary signal

"Smooth" sequence of length N is the one that does not contain more than L excursions beyond this specified level. If L is the number of allowed excursions, there would be exactly M different types of predefined sequences:

$$M = \sum_{n=0}^{L}\binom{N}{n} \quad (9)$$

Similarly to (1) and (3), it is defined

$$z_j = \begin{cases} 0, d(x_N(i+j-1),m) \le r \\ 1, d(x_N(i+j-1),m) > r \end{cases}, j=1,...,N \quad (10)$$

and

$$z_i = \begin{cases} 1, \sum_{j=1}^{N} z_{ij} \le L \\ 0, \sum_{j=1}^{N} z_{ij} > L \end{cases}, i=1,...,L_S - N + 1 \quad (11)$$

Then a sample time span between "smooth" intervals can be obtained as:

$$B_{1i}^{N}(r) = \sum_{z_j,z_k \ne 1, i<k<1}^{z_j} 1 \quad (12)$$

for the time distance of the first type, and for the time distance of the second type, the same

sample time would be:

$$B_{1i}^{N}(r) = \sum_{z_j, z_k \neq 1, i<k<j}^{z_j} 1 \tag{13}$$

Averaging the sample times does the estimate of the corresponding mean times. If i_{max} is the position of the last "smooth" interval within the time series, and

$$z_{\max} = \sum_{i=1}^{L_S - N + 1} z_i \tag{14}$$

number of "smooth" intervals, then the corresponding estimates of mean times of type 1 and 2 are:

$$\bar{t}_1(r) = \frac{1}{i_{\max} - N + 1} \cdot \sum_{i=1}^{i_{\max} - N + 1} B_{1i}^{N}(r) \tag{15}$$

$$\bar{t}_2(r) = \frac{1}{z_{\max} - 1} \cdot \sum_{k=1}^{z_{\max} - 1} B_{2k}^{N}(r) \tag{16}$$

Theoretical approach

The equations (15) and (16) might be just another set of measures that could be extracted from an empirical series and compared to the artificially generated data stream. However, values of time distance can be obtained theoretically, given that the estimate of probability that the sample of time series would exceed the specific border. It would be equivalent to generating the binary series from HRV data in a following way:

$$b(i) = \begin{cases} 0, m - r \cdot \sigma < y(i) \le m + r \cdot \sigma \\ 1, otherwise \end{cases}, i = 1, ..., L_S \tag{17}$$

The estimate of the probability of "rough" sample in HRV series, i.e. the probability of one in binary series is:

$$p = \frac{1}{L_S} \cdot \sum_{i=1}^{L_S} b(i) \tag{18}$$

In order to derive analytically the expected value of time distance of the first type, a set if M sequences are described using two recently introduced terms. The first one is a set of cross-bifix indicators, and the second one is the set of tails [9-14].

A cross-bifix is a subsequence of length n≤N that is a suffix of ith sequence and a prefix of jth sequence, i,j=1,...,M. The corresponding cross-bifix indicator equals to 1 if cross-bifix of length n exists, e.g. binary sequences P_i=0**001** and P_j=**001**1 have a 3-bit cross-bifix, while obviously. If i=j, denotes classical bifix indicator introduced in [12]. The default values for cross-bifix indicators are:

$$h_{ij}^{(0)} = 1, h_{ij}^{(N)} = \begin{cases} 0, i \neq j \\ 1, i = j \end{cases}, i, j = 1, ..., M \tag{19}$$

A tail is a suffix of length n of a sequence no. j. Its value equals to a product of probabilities of the last n symbols of a sequence. By default, and a tail of length N equals to the probability of a sequence itself.

The expected value of number of tests necessary to find one of the specified M sequences equals to:

$$E\{t_1\} = T = 1 - N + \left[\sum_{i=1}^{M} r_i^{(N)} \right]^{-1} \cdot \sum_{i=1}^{M} S_i \cdot C_i \tag{20}$$

Terms Ci depend upon the cross-bifix indicators and upon the tails:

$$C_{ij} = \sum_{m=1}^{N} r_j^{(m-1)} \cdot h_{ij}^{(N-m+1)}$$

$$C_i = \sum_{j=1}^{M} C_{ij} = \sum_{j=1}^{M} \sum_{m=1}^{N} r_j^{(m-1)} \cdot h_{ij}^{(N-m+1)} \tag{21}$$

Terms Si can be evaluated by solving a set of M linear equations:

$$\begin{bmatrix} C_{11} \cdot \left[r_1^{(N)}\right]^{-1} - C_{12} \cdot \left[r_2^{(N)}\right]^{-1} & \cdots & C_{M1} \cdot \left[r_1^{(N)}\right]^{-1} - C_{M2} \cdot \left[r_2^{(N)}\right]^{-1} \\ C_{11} \cdot \left[r_1^{(N)}\right]^{-1} - C_{13} \cdot \left[r_3^{(N)}\right]^{-1} & \cdots & C_{M1} \cdot \left[r_1^{(N)}\right]^{-1} - C_{M3} \cdot \left[r_3^{(N)}\right]^{-1} \\ \vdots & \cdots & \vdots \\ C_{11} \cdot \left[r_1^{(N)}\right]^{-1} - C_{1M} \cdot \left[r_M^{(N)}\right]^{-1} & \cdots & C_{M1} \cdot \left[r_1^{(N)}\right]^{-1} - C_{MM} \cdot \left[r_M^{(N)}\right]^{-1} \\ 1 & \cdots & 1 \end{bmatrix} \cdot \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_{M-1} \\ S_M \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \tag{22}$$

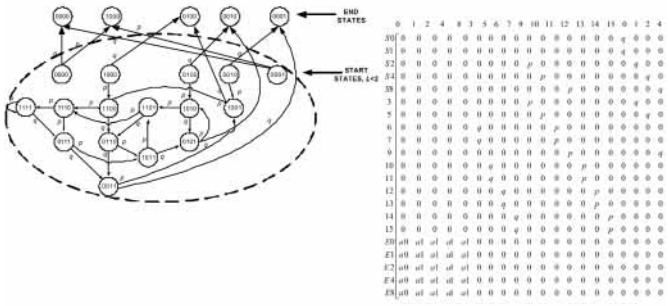

Figure 8. State transition matrix and diagram for N = 4 and L = 1; transitions from end to start states are not drawn for the sake of clarity

Considering the expected value of type 2-time distance, it could be evaluated using the state-transition diagram and first passage time evaluation method described in [15,16]. If N=4 and L=1, the state-transition diagram with the corresponding state-transition matrix P is shown in Figure 8. States with 0 and 1 ones are "smooth" states. The probabilities a0 and a1 are the conditional probabilities that, if the search has started from the smooth state, it would be a state with 0 (1) ones:

$$a0 = \frac{q}{1 + 3 \cdot p}, a1 = \frac{p}{1 + 3 \cdot p} \tag{23}$$

The total number of states in Figure 8 is $S=2^N+M$. The first M states are start states, the last M ones are ending states, and the remaining ones are transient states. The corresponding state selection probabilities $\pi = [\pi_1\ \pi_2\ \pi_3\ ...\ \pi_{S-2}\ \pi_{S-1}\ \pi_S]$ can be obtained by solving a set of equations $\pi \cdot P = \pi$, with the constraint:

$$\pi \cdot 1^T = 1, 1 = [11...11] \tag{24}$$

The expected value of type 2-time distance equals to

$$E\{t_2\} = \frac{1 - \sum_{i=2^N+1}^{2^N+M} \pi_i}{\sum_{i=2^N+1}^{2^N+M} \pi_i} \tag{25}$$

EXAMPLES AND DISCUSSION

For an illustrative example, some results for subject C2 are shown in Figure 10. Distance r is, through this investigation, normalized by standard deviation (the same as suggested for ApEn approach). Both the theoretical results and the results of empirical series are plotted. Distance values for measured data are constrained by series length L_S: if its value is, e.g., 3000, then mean value cannot be greater.

In Figure 11 the measured and theoretical data of type 1 time distance are compared. The figures plot

$$\frac{E\{t_1\} - \bar{t}_1}{E\{t_1\}} \cdot 100 \tag{26}$$

Difference of 100% is obtained in cases when the theoretical value for time distance exceeds the length of time series. Most of the subjects had a ppositive relative error (26).

i.e. empirical values of mean time necessary to reach the smooth region are shorter for empirical sequences.

Figure 12 compares theoretical and measured values of type 2-time distance, plotting the values

$$\frac{E\{t_2(N+1)\}}{E\{t_2(N)\}}, \frac{\bar{t}_2(N+1)}{\bar{t}_2(N)} \tag{27}$$

The measured data are in perfect accordance with the theoretical ones (this was not the case with time of type 1). The only exception is subject D1, who possess some other interesting properties: it was not possible to extract the correlation dimension, neither automatically nor manually. Besides, its value for distance type 1 has an interesting shape, a shown in Figure 9.

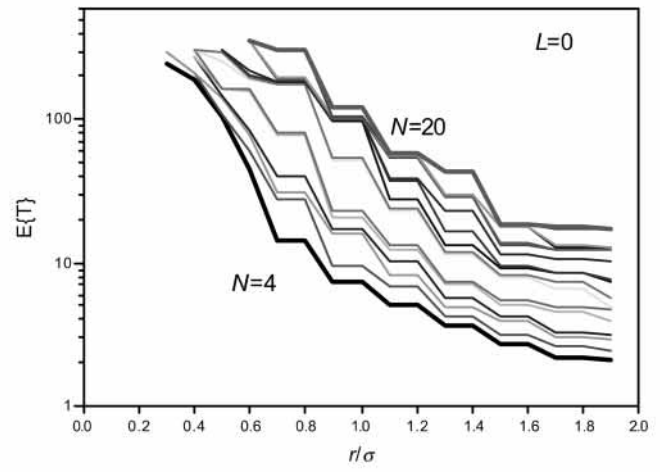

Figure 9. Mean time type 1 estimate for subject D1

The results are promising, although no firm conclusion can be made: the children's data are variable and there was neither diagnosis, nor a medicine applied, influence of which could be observed within our data. Therefore, future investigation would be based either upon patients (adults) with applied treatment, or, which are more probable, laboratory rats.

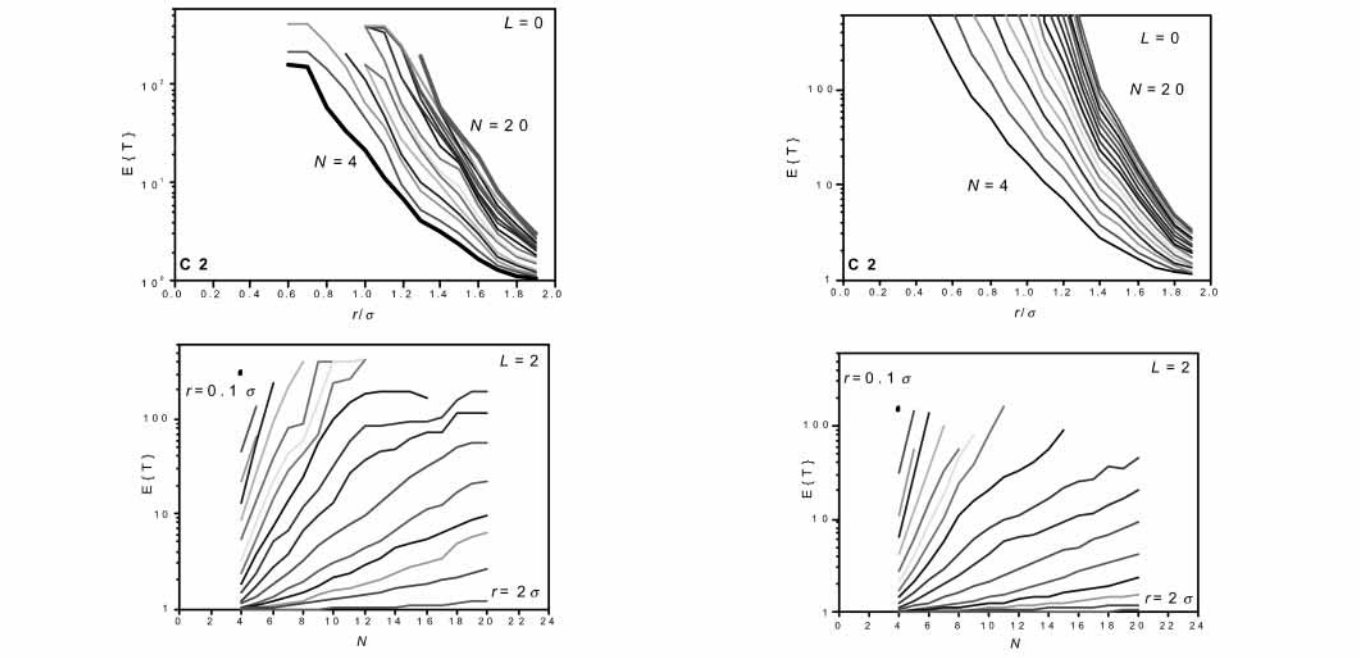

Figure 10. Time values, theoretical and empirical - Subject C1

REFERENCES

1. Pincus SM. Approximate entropy as a measure of system complexity. Proc Natl Acad Sci USA 1991;88:2297-301.
2. Goldberger A., Pincus SM. Physiological time-series analysis: What does regularity quantify? Am J Physiol 1994;266:H1643-H1656 (Heart Circ Physiol)
3. Lipsitz LA. Dynamics of StaBility. J Gerontology 2002;57A(3):B115-B125.
4. Richman JS, Moorman JR. Physiological time-series analysis using approximate entropy and sample entropy. Am J Physiol Heart Circ Physiol 2000;278(6):H2039-H2049.
5. Lončar T, Bajić D, Žigon N. On information-theoretic approach to HRV time-series analysis, 48. Konferencija ETRAN-a, Čačak, June, 2004.
6. Karvajal R et al. Dimensional Analysis of HRV in Hypertrophic Cardiomyopathy Patients. IEEE Engineering in Medicine and Biology 2002;21(4):71-8.
7. Ivanov PCh et al. Multifractality in human heartbeat dynamics. Nature 1999;399:461-5.
8. Bendat, Piersol. Random data analysis and measurement procedures. New York: Wiley Interscience; 1986.
9. Bajić D, Drajić D. Duration of search for a fixed pattern in random data: Distribution function and variance. Electronics Letters 1995;31(8):631-2.
10. Bajić D, Stojanović J, Lindner J. Multiple Window-sliding Search. Proceedings of 2003 IEEE International Symposium on Information Theory ISIT-2003, Yokohama, Japan, June 2003. p. 249.
11. Bajić D, Stojanović J. An Analysis of the Search Process for Different Patterns in Random Data. Proceedings of IEEE Region 8 EUROCON 2003 Ljubljana, Slovenia, September, 2003.
12. Bajić D, Stojanović J. Distributed Sequences and Search Process. IEEE International Conference on Communications - ICC2004, Paris, France, June, 2004.
13. Bajić D, Stefanović Č, Vukobratović D. On probabilistic approach to bifix indicators, (in press).
14. McConell T. The Expected Time to Find a String in a Random Binary Sequence. [Available from http://barnyard.syr.edu/cover.pdf], January 2001.
15. Bajić D. New simple method for solving first passage time problem. Electronics Letters 1991;27(16):1419-21.
16. Bajić D, Stojanović J. Analytical approach to the duration of search for different patterns. SympoTIC04, Bratislava, Slovakia, October, 2004

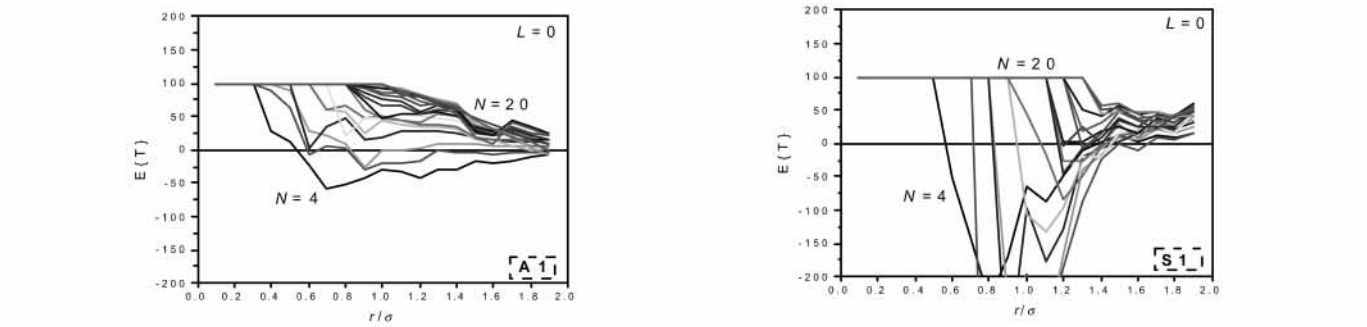

Figure 11. Examples of fistance type 1 - analytical vs. emprical values

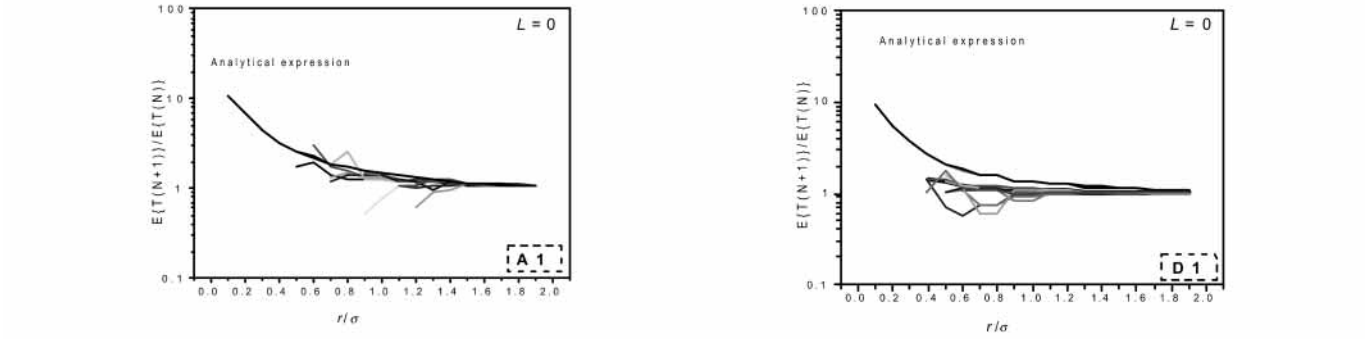

Figure 12. Examples of distance type 2 - analytical vs. emprical values