JP Journal of Biostatistics **1**(1) (2007), 53-78

A SIMILARITY MEASURE FOR BIOLOGICAL SIGNALS: NEW APPLICATIONS TO HRV ANALYSIS

MIRKO DEGLI ESPOSTI[1], **CHIARA FARINELLI**[1]
MARCO MANCA[2] **and ANDREA TOLOMELLI**[1]

[1]Department of Mathematics
University of Bologna, Piazza di Porta S. Donato 5
I-40127 Bologna, Italy
e-mail: desposti@dm.unibo.it

[2]G. C. Descovich Atherosclerosis and
Metabolic Diseases Research Center
University of Bologna, Via Massarenti 9
40138 Bologna, Italy

Abstract

In this paper we introduce and discuss a similarity distance function between symbolic strings recently introduced in [33]. Application to phylogenetic tree construction and HRV analysis are considered.

1. Introduction and Motivations

The aim of this paper is to introduce and describe a recent approach to heuristic estimation of similarity between symbolic sequences.

While these methods have been already applied to various classes of sequences (such as DNA sequences and literary texts), our purpose is to suggest the possibility of using this approach also to extract information and classify heart rate variability (HRV) sequences.

2000 Mathematics Subject Classification: 94A17, 94A24, 92B05, 92D20, 92C50.

Keywords and phrases: similarity measure, Lemple-Ziv complexity, entropy methods, DNA sequences, HRV analysis.

Received December 5, 2006

ECG signal constitutes a paradigmatic example of non-linear, non-stationary noisy process. This one dimensional time series reflects the net results of an enormous number of interactions among the cardiovascular system, the autonomous nervous system and the external environment; nevertheless they still contain *valuable information* concerning the clinical/pathological state of the *source* [15].

Various and quite sophisticated techniques are presently available for *extracting useful information* out of the ECG signal. These approaches range from non-linear methods developed in the realm of the theory of finite dimensional dynamical systems to time-domain and frequency-domain spectral analysis [39], as shown in [27]. Recently, also interesting tools out of linguistic analysis have been used to study human heartbeat [5, 41].

The heart rate, defined in terms of the number of myocardial contractions, is a complex entity and lays under a plethora of regulatory factors (i.e., autonomic nervous system, endocrine setting, circuitry resistance, cell membrane plasticity, etc. [4, 11, 12, 22]) some of which act even during chaotic functional states like fibrillation [40].

Even though physiology and medicine have investigated for long time the dynamics behind heart rate functionality and "behavior", the relevance of each single factor is still unknown [42]. It is then interesting to discuss the nature of a widely known investigation method: Heart Rate Variability (HRV), the analysis of variations in the instantaneous heart rate time series using the beat-to-beat RR-intervals.

HRV accounts for a large portion of the homeostatic efforts of the individual, it is an essential part of stress: it is quickly changing to grant adaptation to every life-compatible circumstances/stimuli, and furthermore it also shows long-range correlations [24].

Nevertheless while HRV analysis, introduced in clinical practice in the late 'sixties in obstetrics' has been used in thousands studies to investigate the most different functional parameters in human beings (even the effects of geomagnetism on health [32]) and has yield to the development of a huge number of analytical methods [15] dedicated to its investigation, little light has been shed on its foundation. HRV has proved to be a relevant tool to evaluate autonomic system function [13], in

particular when it is studied through a quite elementary use of the Fourier Analysis [15], and it has proven to be an independent index of life expectancy [37], when undergone an even simpler elaboration (i.e., statistical evaluation of the time series distribution) [15]. Unfortunately when it comes to investigate deeper properties of an individual, the real effectiveness of HRV is still unclear, and this posed a challenge to many researchers to built a more and more complex and efficient methods for extracting *valuable* information out of HRV data [23, 28].

Here we would like to discuss a more general and fundamental approach to information extraction out of ECG's signals.

Before going into some more details, let us discuss the main ideas and considerations underlying our work. First of all, this method relies on the existence of a *distance function* between any two given finite strings out of a common alphabet. This distance is obtained by a proper normalized estimate of (some kind of) *conditional complexity* of one string with respect to the other. The founding quantity of this construction is the so called *Lempel and Ziv complexity* [26] $c(S)$ associated to any finite string S.

Roughly speaking, $c(S)$ is just a positive integer number counting the number of certain *new words* (*substrings*) produced along the string itself.

Little adaptation of this general idea, combined with suitable *coding* of the words leads to some of the very well-known compression programs used daily for *zipping* files on the computer.

In fact, various *similarity metrics* based on compression algorithms are now available and have been used for *clustering* various kinds of data: from DNA sequences to literary texts, from music melodies to proteins, always with encouraging but also disputed results [8, 9, 25, 41]. It is important to note that the distance used in our experiments and introduced in [33] is in fact similar but not identical to the ones cited above. Even if the experimental results are basically identical in all the cases we have studied, we believe that the one used here has clever mathematical aspects and more elementary defining properties, as we will thoroughly discuss in [10].

While we leave the details for the next section and references therein, we stress here that this complexity and its related distance rely only on a general parsing procedure that is completely independent from the eventual *process/grammar* used in the production of the string. This parsing procedure, namely the subdivision of the string, is able to capture periodic structures or more general motif in the sequence. When this is applied to a concatenation of two different sequences, various kinds of relative statistical properties and common similarities are naturally detected.

The associated distance function turns this observation into a quantitative estimate but it is important to stress that while mathematical results are available in specific theoretical setups, such as stationary ergodic sources in the limit when the length of the sequences goes to infinity, *no* rigorous conclusions can be drawn when this distance is applied to specific categories of finite sequences arising from applications. In this case we can just rely on a *try and see* approach.

Before discussing the two main classes of biological signals we are interested in, namely ECG signals and also DNA sequences, it is instructive to comment on a completely different context where these similarity distances have been successfully applied: the so called *author's attribution* problem.

Authorship attribution is a long standing problem: the challenge is to identify the author of an unknown text by identifying a *stylistic fingerprint* characteristic of the writer. In practice, one has a corpus of classified literary works and tries to identify the writer (*source*) that likely *produced/emitted* the unknown text. Various and different techniques have been developed for this task: they range from pure statistical analysis of words to detailed linguistic analysis based on grammar and syntactic structures.

Successful approaches to this problem have been developed with the help also of similarity metrics based on compression algorithms [3, 8].

Few years ago a specific competition for a systematic comparison between the different methods in authorship attribution has been

established [1] and it can now be considered a solid benchmark. The problems in the competition are very different: they range from short essays from a group of students to Latin authors and French ones. For each problem one has a certain number of sample texts from known authors and few unknown ones that must be either attributed or recognized not to belong to the given set of known writers. The Ad-hoc Author Attribution Competition (AAAC) was hosted in June 2004 and the results are described in [1]. The most surprising conclusion (at least for us) that can be raised from the final results of the contest is the following: *ad-hoc* linguistic methods based on the semantic or grammar structure of literary texts tend to perform worst than *generic* methods that have no link with the detailed structure of the string, such as the methods based on compression algorithms or on the frequencies of N-grams [1].

The distance presented here and introduced in [33] has been tested on the material of the AAAC contest and also on the corpus of authors used in [3] with satisfactory and competitive results (not presented here).

Motivated by these results we argued that this distance could also be competitive with respect to ad-hoc and more sophisticated methods created for analyzing both DNA sequences and ECG time series.

More precisely, in this paper we address this issue and bring strong indications (probably not conclusive) that also in these cases more *fundamental* and *elementary* approaches to the investigation and comparison of the information content in complex biological strings or signals can be at least as efficient, if not better than more sophisticated and computational demanding *ad-hoc* methods.

The paper is organized as follows: in the first section we briefly introduce the main definition and mathematical properties of the Lempel-Ziv complexity and related distance. Then we discuss some recent and also new applications to genetic sequences. The third section is devoted to the description of our approach to ECG signals, whereas the corresponding experimental results and some remarks about future developments are described in the last section.

2. L-Z Complexity and Related Distance

Now we give some of the formal mathematical definitions and features of the distance proposed in [33] and essentially already contained in [26]. We will then discuss how this turn to be useful in biological sequences comparison.

We start by fixing some notations. Always $\mathcal{A}$ will denote a finite alphabet, $\mathcal{A}_n$ all possible words (i.e., arbitrary strings) of length exactly n and we let $\mathcal{A}^* = \bigcup_{n\in\mathbb{N}} \mathcal{A}_n$ be the set of all possible finite strings over the given alphabet. In the following we denote by S, Q and R arbitrary finite sequences definite over $\mathcal{A}$: $S, Q, R \in \mathcal{A}^*$. Without loss of generality, we might assume $\mathcal{A} = \{0, 1\}$, but sometimes it will be useful to consider the ASCII alphabet or the DNA nucleotides set $\mathcal{A} = \{A, C, G, T\}$.

$L(S)$ denotes the length of S, $S(i)$ is the i-th element of S, whereas $S(i, j)$ is the substring of S composed by the elements of S between positions i and j (included). The *dictionary* of S, $Dic\{S\}$, denotes the set of all substrings of S.

Lempel and Ziv already in the '70s [26] proposed to monitor whenever part of the string can be produced just by sequentially copying a portion of the past string or if, on the contrary, new information is produced while reading the sequence: following [26], we shall now recall a mechanism of generating a nonnull string S from some proper prefix Q, called *exhaustive parsing procedure*, and associate an integer number $c(S)$ to it. The *complexity* $c(S)$ shall then be used to define a distance between finite strings, along the lines outlined in [33].

More formally, a string S is *reproducible* from its proper prefix $Q = S(1, j)$ if $S = QR$ and $R = S(j+1, n) \in Dic\{S(1, n-1)\}$. A string S is *producible* from its proper prefix $Q = S(1, j)$ if $S(1, n-1)$ is reproducible from Q. In addition to the pure copying involved by the reproduction process, the production allows for a single-symbol innovation at the end of the copying.

Any finite, nonnull string S may be represented as the final product

of an iterative production process $\mathcal{P}$, where the first step is the production of $S(1, 1)$ from the empty string Λ and any following step is such that $S(1, h_m)$ is the result of a production from $S(1, h_{m-1})$. Finally, the original string S is parsed in t substrings:

$$S(1, n) = S(1, h_1)S(h_1 + 1, h_2)\cdots S(h_{t-1}, n),$$

where t is the complexity of the production process and is denoted by $c_{\mathcal{P}}(S) \le n$.

A production process where each step (with the only possible exception of the last one) is a production step, but not a reproduction step is called *exhaustive process* and gives rise to the *exhaustive parsing* of S. We denote by $c(S)$ the complexity of the exhaustive parsing of S.

The exhaustive parsing of S is unique and represents the least possible number of production steps in which S may be generated (Theorem 1 of [26]).

We can now use basic properties of this complexity to define a quite natural *similarity distance* between two arbitrary finite strings.

Given two sequences Q and S, consider the sequence SQ and its exhaustive process. By definition, the number of components needed to build Q when concatenated to S is $c(SQ) - c(S)$. This number will be lower than or equal to $c(Q)$ and this, in turn, would reduce the number of exhaustive components. Given a third string R, if R is *more similar* to Q than to S, then we would expect $c(RQ) - c(R)$ to be smaller than $c(SQ) - c(S)$.

For example, let $S = 11010110$, $R = 00110110$ and $Q = 1110010$. Then the exhaustive histories of these sequences would be:

$$H_E(S) = 1 \cdot 10 \cdot 1011 \cdot 0, \quad H_E(R) = 0 \cdot 01 \cdot 10 \cdot 110,$$
$$H_E(Q) = 1 \cdot 110 \cdot 01 \cdot 0$$

yielding $c(S) = c(R) = c(Q) = 4$. The exhaustive histories of the sequences SQ, and RQ would be:

$$1 \cdot 10 \cdot 1011 \cdot 0111 \cdot 001 \cdot 0, \quad 0 \cdot 01 \cdot 10 \cdot 110111 \cdot 0010$$

respectively. Note that it took two steps to build Q in the production

process of SQ. On the other hand, we used only one step to generate Q in the production process of RQ. The reason it took more steps in the first case is because Q is *closer* to R than S. In this example we can observe this by looking at the pattern 110 which Q and R share. We can formulate the number of steps it takes to generate a sequence Q from a sequence S by $c(SQ) - c(S)$. Thus, if R is closer to Q than S, then we would expect $c(RQ) - c(R)$ to be smaller than $c(SQ) - c(S)$ as is the case in the above example. Based on this idea of closeness, various distances can be defined [33]. Here we just recall the two normalized distances that we have been using in the numerical experiments: given any two finite sequences $S,\ Q \in \mathcal{A}^*$, we define the function $d(S,\ Q)$ as

$$d(S,\ Q) = \frac{c(SQ) - c(S) + c(QS) - c(Q)}{\frac{1}{2}\left(c(SQ) + c(QS)\right)}.$$

Another possible choice is the following:

$$d_1(S,\ Q) = \frac{\max\{c(SQ) - c(S),\ c(QS) - c(Q)\}}{\max\{c(S),\ c(Q)\}}.$$

As shown in Appendix of [33], these two functions do in fact satisfy all the properties required for a valid (evolutionary) distance metrics.

Concerning the practical implementation of these distances, we have performed several experiments on various kinds of data and signals using both distance functions. Here for the sake of clarity, we report only the numerical results obtained using the first distance $d(\cdot,\ \cdot)$. The results obtained through the use of the second distance are usually very similar; only few times they turn out to be slightly worst.

The implementation required suitable computer programs to automate and manage the results. In particular, we build a C program to calculate the distance between two arbitrary files and a VB program to built and manage the matrix distance associated to any given set of files. The VB application uses the C program as a component and applies it to all possible pairs of files inside a specific folder; the calculated values are then used to produce the matrix distance. The VB application comes with an interface that allows the user to choose among different kind of analysis and options and it is available upon request. In addition, well-known free tools have been used to built and visualize different kinds of trees associated to the data [35, 36].

Now we turn to the discussion of possible applications of this distance to two different kind of biological signals: DNA sequences and ECG signals.

2.1. Phylogenetic tree construction

One application of this new similarity distance concerns DNA sequences and it has been already explored in [33] and also in [2]. We now briefly recall their results and describe some new experiments we have performed. It is our opinion that both sets of experiments, even if they do not bring new concrete advances in genomic, shade some lights on the fundamental properties and potentiality of this method.

Using this metric, we can construct a symmetric distance matrix for any given set of DNA sequences and from this we can apply the very well- known classical methods for phylogenetic tree construction [20]. It is important to remark that this gives a tool which is alternative to all the methods based on the best fitness between different tree topologies, such as for example *Parsimony* and *Maximum Likelihood* methods.

While these last methods all rely on multiple alignment and on some sort of evolutionary model, the method based on the distance arising from LZ complexity is so general and fundamental that does not share these ambiguity on the choice of the alignment cost criteria and on the evolutionary assumptions. As a very important consequence, this distance method can also be applied for phylogenetic tree analysis of complete genomic sequences, where the multiple alignment based methods cannot be used due to the enormous computational costs.

As it has been nicely discussed in [33], this distance is able to perform very well with respect to classical phylogenetic tree methods and also with respect similar (but a little more involved) distance-based methods, such as the ones using various compression algorithms [7].

For example, in [33] a well-known debate concerning the phylogeny of eutharian order has been explored with the use of complete mtDNA mitochondrial sequences of various Rodents, Ferungulates and Primates (see [33] for the detailed list of species used in the experiment). As shown there, the resulting tree perfectly agrees with the one in [6]. For completeness, we recall that also a second experiment has been performed

with an extended data set obtained by the addition of non-murid rodents (squirrel, dormouse and guinea pig) and more ferungulate sequences. Also in this case, the final consensus phylogeny is in agreement with the existing literature (see Figure 2, Figure 3 in [33] and discussion therein).

It is important to stress again that these experiments have been performed using the whole genome sequences, without relying just on the coding sequences or individual proteins.

Stimulated by these results and in order to continue the exploration of the use of this distance to DNA sequences, we have performed two additional experiments that confirm the validity of the method.

First of all, we have repeated another experiment already performed in [7] with seven complete genomes from [31]:

Archaea Bacteria *Archaeoglobus fulgidus* (**NC_000917**)[1], *Pyrococcus abyssi* (**NC_000868**) and *Pyrococcus horikoshii* OT3 (**NC_000961**)

Bacteria *Escherichia coli* K-12 MG1655 (**NC_000913**), *Haemophilus influenzae* Rd (**NC_000907**), *Helicobacter pylori* 26695 (**NC_000915**) and *Helicobacter pylori* strain J99 (**NC_000921**).

Also in this case, our distance is able to reproduce the same genome tree originated through either alignment methods or method based on local DNA mutations: compare Figure 1 with Figure 2 in [7].

Finally, we tested our distance on very short DNA sequences corresponding to single genes and we get again quite surprising (at least for us) results: contrarily to what expected, our LZ distance is able to reproduce the correct evolutionary tree also on *short single genes*.

Here we present only the results we have obtained using exactly the same data as in [7]: single 16S rRNA genes (18S rRNA for Eukaryotes) have been selected from GenBank [31].

More specifically:

Archaebacteria *H. butylicus* (**X99553**) and *Halobaculum gomorrense* (**L37444**)

[1] Accession number.

Eubacteria *Aerococcus urina* (**U64456**), *M. glauca* strain B1448-1 (**X94705**) and *Rhodopila globiformis* (**D86513**)

Eukaryotes *Urosporidium crescens* (**U47852**), *Labyrinthula sp. Nakagiri* (**AB022105**).

The corresponding tree is shown in Figure 2 and again it must be compared with the similar one obtained in [7], Figure 2.

2.2. Symbolic ECG analysis

ECG signals constitutes a formidable example of noisy, non-stationary time series, and as discussed in the introduction a large amount of techniques have been proposed to analyze the information contained in the signals [15, 39].

The similarity distance introduced here has been used against three main elementary tasks: *clustering* of healthy ECG's signals with respect to age, *discrimination* between normal signals and various congestive heart failures and finally *classification* of NYHA classes [15].

In practice we considered 24 h. Holter ECG signals, obtained from [14] and also from [27].

The necessary coding procedure is the following: first of all we extract RR interval sequence (R_j) from the full ECG signals, disregarding in particular *all* the information contained between these two events, such as the P wave, QRS complex, ST segment, etc. Then we perform an elementary binary coding by looking at the sign of variability, i.e., we construct a new 0, 1 sequence by setting $w_j = 0$ if $R_j - R_{j-1} > 0$, and $w_j = 1$ otherwise.

After this process, the original information has been tremendously reduced: an original 24 hours ECG signal of few Mega byte (Mb) is reduce to a binary file of about 100 Kb.

In our opinion it is remarkable that even after this tremendous and elementary reduction of the signals, our distance is still capable of capturing *common features* in the signals as our positive and consistent results in all the classification experiments shown.

We believe that this is a feature of our method that deserves future investigations.

Clearly this binary coding based on the decreasing or increasing of two consecutive RR intervals is just the most elementary one.

More refined coding, for example based on the magnitudes of variability explored in the PNNx statistics [30] are currently under investigation. A small increasing of the vocabulary used in coding the RR sequences should bring to a better performance, while still keeping the computational complexity into practical affordable limits.

We now turn to describe and discuss the experiments performed.

3. HRV Experimental Results

In this section we describe the numerical experiments performed with our method on several ECG's from different groups of subjects, where the ECG's signals have been coded as previously described. The data come from two main sources: the Physionet archive [14] and also from the signals used in [27] and available upon request.

Initially, we have analyzed and used the data obtained from the group of research from Gdansk University (Poland), with the aim of comparing our results with the ones obtained by them using multifractal analysis [27]. In particular, they performed classification tasks by calculating the local exponents of the spectrum associated to the RR series with the use of the Wavelet Transform Modulus Maxima Method and also with the use of the Multifractal Detrended Fluctuation Analysis ([21, 29, 34] and references in [27]).

As described in [27] and repeated here, two main groups of patients have been used:

nk group made of 90 patients hospitalized during 2001-2004 in the 1st Department of Cardiology of Medical University in Gdansk, Poland (9 women, 81 men, average age: 57 ± 10) in whom the reduced left ventricular systolic function was recognized by echocardiogram due to the low left ventricular ejection fraction (LVEF $\leq$ 40%, mean LVEF = 30, 2 ± 6, 7%);

gk group made of 40 healthy individuals (4 women, 36 men, average age: 52 ± 8) without past history of cardiovascular disease, with both

echocardiogram and electrocardiogram in normal range. The left ventricle ejection fraction was normal (mean LVEF = 68, 0 ± 4, 7%).

Our first simple task is to detect if the distance is able to discriminate between the two groups. In order to investigate this issue, we computed the entire matrix distance between all the patients using the full 24 hours HRV and performed several experiments.

For example, we randomly chose n patients from each group $(5 \leq n \leq 10)$ and used them to define the two classes, then for any other patient we computed the average distance from the two groups and classified accordingly.

From the results we can clearly and consistently see that patients belonging to group **nk** are on average closer to patients of group **nk** than to patients of group **gk**, and viceversa: the number of successes range between 75% and 85%.

More precisely, the number of success in classifying patience from the **gk** group is consistently higher than from the **nk** group: 80%-85% of successes in the first case and 75%-80% in the second. One of such experiments is reported in Table 1.

After realizing that our method works quite well and consistently on the full 24 hours ECG files, we then tested the distance on much shorter portion of the signals. In particular, using again the data from [27], we repeated the same experiments on the two 5 hours portion of the signals that correspond in general to the patients being awake (**nk_w**, and **gk_w** groups) or asleep (**nk_s**, and **gk_s** groups), respectively.

The results obtained using the signals belonging to the wake state groups (**nk_w**, **gk_w**) are basically identical (if not better) to those obtained previously using the whole signal (see Table 2 where the subjects are the same of those in Table 1). We remark that the binary sequences corresponding to these 5 hours interval are very small (few K bytes), which clearly implies a very limited set of words created during the parsing rule. In our opinion, the good performance of the method even on these short sequences represents a clever indication of its consistency and efficacy.

To visualize the *clustering* property of our distance, we show in Figure 4 the tree generated by the distance matrix computed using the group **gk_w** and 38 patience randomly chosen from the other group. It is important to remark that we use the tree to present our data only for exposition purposes and qualitative preliminary considerations. Quantitative and statistical features of our method have been in fact directly extracted from the numerical values of the distance matrix.

Similar results on a smaller set of data are shown in Table 3 where we compare signals of subject in **gk_w** group with those in **nk_w**.

In order to visualize better the ability of the distance to *cluster* patients of the same class, we show in Figure 3 a tree constructed out of this matrix distance.

According to our expectations, the same experiments on the *sleeping* part of the data (**nk_s**, **gk_s** groups) give sometimes worst results, confirming that the wake part of the signal is clearly the most significant one.

One might suspect that in some of these experiments the individual complexity of the signal can be enough for achieving a good classification. This supposition turned out to be completely wrong: by calculating the single complexity of each file (not reported here), both the classifications **gk** v.s. **nk**, or *weak* v.s. *sleep* fails consistently.

Namely, the single complexity is far from being able to discriminate between the **nk** and **gk** groups, i.e., the specific properties of the distance, arising from the conditional parsing of one string with respect to the other, are really necessary.

Thanks to the data in [27], we have performed an additional similar experiment. In particular, we used the following two sets of data:

nsr_wake group: these are 13 healthy subjects belonging to [19] from which the wake parts of the signals have been extracted.

chf_wake group: wake part of the signals corresponding to 13 subjects with congestive heart failure [16].

As a common feature to all the experiments, also in this case we can note significant differences in the distances starting from the second digit. Furthermore, we can also remark that the gap in the second digits

becomes smaller when relative distances between healthy subjects (**nsr**) are computed, rather than relative distance between patients of the second group (**chf**). In other words, healthy patients signals are more *similar* to one another than patients with past cardiac events.

Moreover, also for these experiments we note that the whole 24 hours signals of both groups could be substituted by the 5 hours corresponding to the wake period without degrading the final results.

The final outcome of this experiment are shown in Table 4 and in Figure 5.

A second kind of experiment that we have performed consists in clustering old patients from young patients, again by measuring the relative distances between the binary HRV coding extracted from the ECG signals. We have considered two databases:

old group 13 healthy subject belonging to **gk** previously described.

young group 13 healthy and rather young people (age between 20-40 years). These patients (3 men, 10 women) show no significant arrhythmias. The corresponding ECG recordings are available from the Physionet archive [20][2].

Also in this case, we consistently got correct results: a single young (old) patient has an averaged distance from the whole group of young (old) consistently smaller than the other group.

In order to give a visual presentation of part of the results, we again show a table with the average distances from the two groups (Table 5).

We have then repeated the same experiment with the data downloaded from [18], where we have chosen some individuals of age from 60 to 68, and other from 28 to 40 years. The results are shown in Table 6 and Figure 6. Again, even if the statistical significance of the results can be disputed due to the limited numbers involved, the portion of successes is very high and strongly support the validity of the method.

In the last experiment we tried to recognize the NYHA class of individual patients with classified congestive heart failure. Various Holter ECG's files were downloaded from [17] and consist of patients

[2] The original physionet database consists of 18 nsr records, from these we have removed patients with age greater than 40 years.

belonging to classes I, II and III of the NYHA classification. After the usual binary coding, we repetitively chose few signals out of each class and use them as reference data for classification. We then picked randomly other unknown HRV strings and used the minimum averaged distance from the previously defined sets to attribute the NYHA classification. Our method is able to distinguish quite well the subjects in classes I and III, whereas quite often it consistently attributes to class III patients that were classified in class II. Because of the small number of trials, the numerical results are not statistical significant and are not presented here. More numerical experiments on bigger corpus of data are currently performed with preliminary positive indications. In any case this classification is quite subjective, being related to the general conditions of the patience, and it is not surprising if it will turn out to be difficult to detect a sharp boundary between classes II and III of the NYHA classification using our or others statistical methods.

Acknowledgments

We are very grateful to the authors of [27] for giving us the opportunity of using their data for our numerical experiments and also for useful comments and suggestions. We also thank S. Graffi and A. Gaddi for interesting discussions and comments.

References

[1] Ad-hoc Authorship Attribution Competition, http://www.mathcs.duq.edu/juola/authorship-contest.html

[2] D.-R. Bastola, H.-H. Otu, S. Doukas, K. Sayood, S. H. Hinrichs and P. Iwen, Utilization of the relative complexity measure to construct a phylogenetic tree for fungi, Mycological Research 107 (2003), 1-10.

[3] D. Benedetto, E. Caglioti and V. Loreto, Language Tree and Zipping, Phys. Rev. Lett. 88(4) (2002).

[4] A. J. Camm and L. Fei, Chronotropic incompetence-Part I: Normal regulation of the heart rate, Clin Cardiol. 19(5) (1996), 424-428.

[5] C. Cammarota and E. Rogora, Testing independence in time series via universal distributions of permutations and word, Int. J. Bif. Chaos 15 (2005), 1-9.

[6] Y. Cao, A. Janke, P. J. Waddell, M. Westerman, O. Takenaka, S. Murata, N. Okada, S. Paabo and M. Hasegawa, Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders, J. Mol. Evol. 47 (1998), 307-322.

[7] X. Chen, S. Kwong and M. Li, A compression algorithm for DNA sequences and its applications in genome comparison, The Tenth Workshop on Genome Informatics (GIW'99), pp. 51-61, Tokyo, Japan, 1999.

[8] R. Cilibrasi and P. M. B. Vitányi, Clustering by compression, IEEE Transaction on Information Theory 51(4) (2005), 1523-1545.

[9] R. Cilibrasi and P. M. B. Vitányi, Similarity of Objects and the Meaning of Words, http://arxiv.org/abs/cs.CV/0602065, (2006).

[10] M. Degli Esposti, C. Farinelli and G. Menconi in preparation (2006).

[11] M. Fahim, Cardiovascular sensory receptors and their regulatory mechanisms, Indian J Physiol Pharmacol. 47 (2003), 124-146.

[12] S. Fazio, E. A. Palmieri, G. Lombardi G and B. Biondi, Effects of thyroid hormone on the cardiovascular system, Recent Prog. Horm Res. 59 (2004), 31-50.

[13] R. Freeman, Assessment of cardiovascular autonomic function, Clin Neurophysiol. (2006).

[14] A. L. Goldberger, et al. PhysioBank, PhysioToolkit, and PhysioNet:components of a new research resource for complex physiologic signals, Circulation, 101(23) (2000), e215-e220.

[15] Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. Task force of the European Society of Cardiology and the North American Society of Pacing and Electro-physiology, Circulation 93 (1996), 1043-1065.

[16] http://www.physionet.org/physiobank/database/chfdb/

[17] http://www.physionet.org/physiobank/database/chf2db/

[18] http://www.physionet.org/physiobank/database/nsr2db/

[19] http://www.physionet.org/physiobank/database/nsrdb/

[20] A. Isaev, Introduction to mathematical methods in bioinformatics, Springer Universitext, 2004.

[21] P. Ch. Ivanov, M. G. Rosenblum, C.-K. Peng, J. E. Mietus, S. Havlin, H. E. Stanley and A. L. Goldberger, Scaling and universality in heart rate varibability distributions, Physica A 249 (1998), 587.

[22] J. R. Jennings JR and M. W. van der Molen, Cardiac timing and the central regulation of action, Psychol. Res. 66(4) (2002), 337-349.

[23] J. Kalda, M. Sakki, M. Vainu and M. Laan, Non-linear and scale-invariant analysis of the Heart Rate Variability, Arxiv.org physics/0303041 (2003).

[24] J.-W. Kantelhardt, Y. Ashkenazy, P.-Ch. Ivanov, A. Bunde, S. Havlin, T. Penzel, J.-H. Peter and H.-E. Stanley, Characterization of sleep stages by correaltions in the magnitude and sign of heartbeat increments, Physical Revie E 65 (2002), 1-6.

[25] N. Krasnogor and D. A. Pelta, Measuring the similarity of protein structures by means of the universal similarity metric, Bioinformatics 20 (2004), 1015-1021.

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

[27] D. Makowiec, R. Galska, A. Dudkowska, A. Rynkiewicz and M. Zwierz, Long-range dependencies in heart rate signals-revisited, Physica A: Statistical and Theoretical Physics 369 (2006), 632-644.

[28] N. Marwan, N. Wessel, U. Meyerfeldt, A. Schirdewan and J. Kurths, Recurrence Plot Based Measures of Complexity and its Application to Heart Rate Variability Data, Arxiv.org physics/0201064 (2002).

[29] M. Meyer and O. Stiedl, Self-affine fractal variability of human heart-beat interval dynamics in health and disease, Eur. J. Appl. Physiol. 90 (2003), 305-316.

[30] J. E. Mietus, C. K. Peng, I. Henry I, R. L. Goldsmith and A. L. Gold-Berger, The pNNx files: re-examining a widely used heart rate variability measure, Heart 88(4) (2002), 378-380.

[31] National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov

[32] K. Otsuka, S. Murakami, Y. Kubo, T. Yamanaka, G. Mitsutake, S. Ohkawa, K. Matsubayashi, S. Yano, G. Cornelissen and F. Halberg, Chronomics for chronoastrobiology with immediate spin-offs for life quality and longevity, Biomed Pharmacother. 57 (2003), 1-18.

[33] Hasan H. Otu and Khalid Sayood, A new sequence distance measure for phylogenetic tree construction, Bioinformatics 19(16) (2003), 735-741.

[34] C. K. Peng, S. Havlin, H. E. Stanley and A. L. Goldberger, Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series, Chaos 5 (1995), 82.

[35] PHYLIP (Phylogeny Inference Package), Cladistics 5 (1989), 164-166.

[36] PHYLODRAW, http://pearl.cs.pusan.ac.kr/phylodraw/

[37] A. K. Reyners, B. P. Hazenberg, W. D. Reitsma and A. J. Smit, Heart rate variability as a predictor of mortality in patients with AA and AL amyloidosis, Eur Heart J. 23(2) (2002), 157-161.

[38] J. Rocha, F. Rosselló and J. Segura, The Universal Similarity Metric does not detect domain similarity, arXiv:q-bio.QM/0603007 v1 6 March (2006).

[39] M. C. Teich, S. B. Lowen, B. M. Jost, K. Vibe-Rheymer and C. Heneghan, Heart Rate variability: measures and models, nonlinear biomedical signal processing, Vol. II, ed., M. Akay, Dynamic Analysis and Modeling, 2001, pp. 159-213.

[40] D. G. Wyse, Rate control vs rhythm control strategies in atrial fibrillation, Prog Cardiovasc Dis. 48(2) (2005), 125-138.

[41] A. C.-C. Yang, S.-S. Hseu, H.-W. Yien, A. L. Goldberger and C.-K. Peng, Linguistic analysis of the human heartbeat using frequency and rank order statistics, Phys. Rev. Lett. 90(10) (2003), 1-4.

[42] M. E. Young, The circadian clock within the heart: potential influence on myocardial gene expression, metabolism, and function, Am J Physiol Heart Circ Physiol. 290 (2006), 1-16.

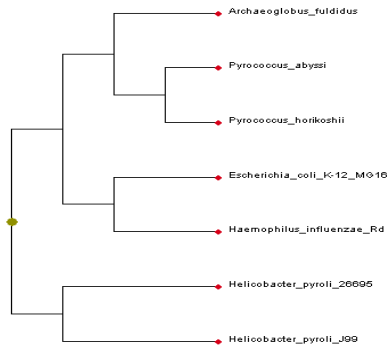

Figure 1. Full genome tree.

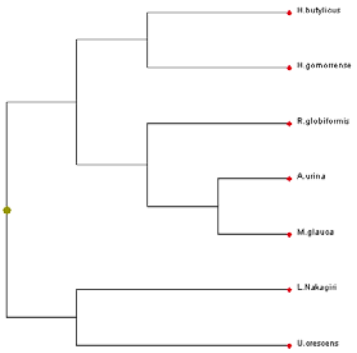

Figure 2. rRNA tree.

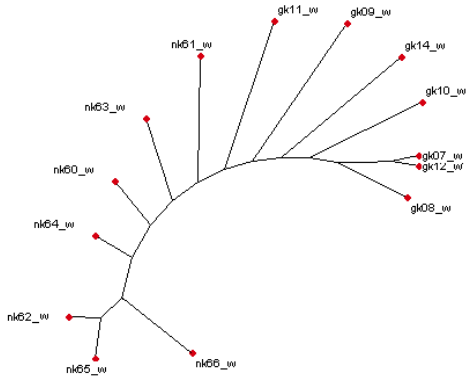

Figure 3. Tree of distances of subjects in **gk_w** group and **nk_w** group. The patients are the same as those in Table 3.

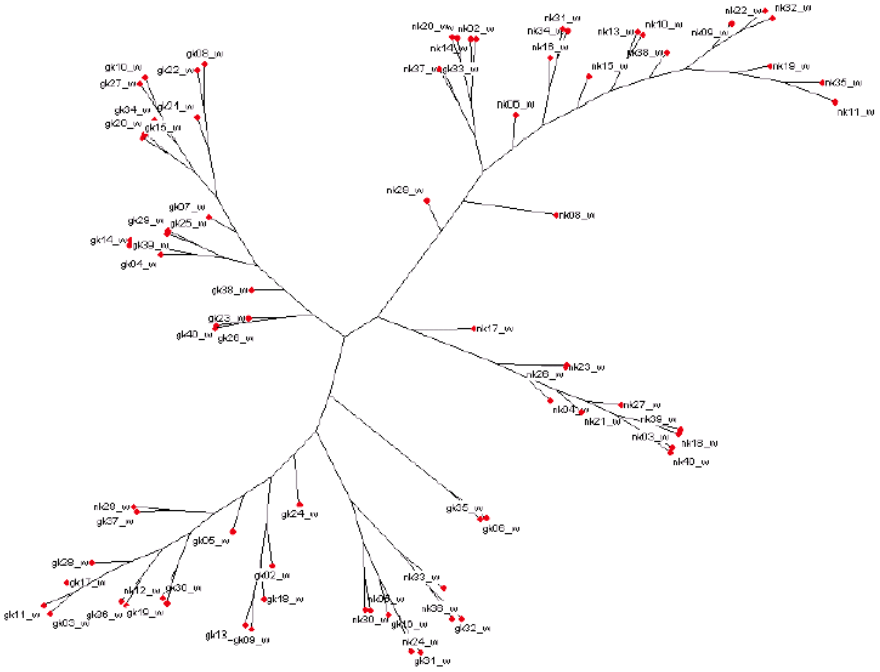

Figure 4. Tree generated by the distance matrix computed using the group **gk_w** and 38 patients randomly chosen from the **nk_w** group.

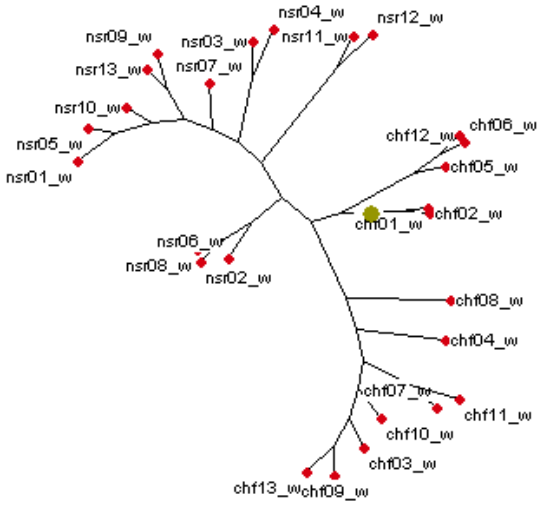

Figure 5. Distance tree out of subjects from the **chf** and **nsr** groups.

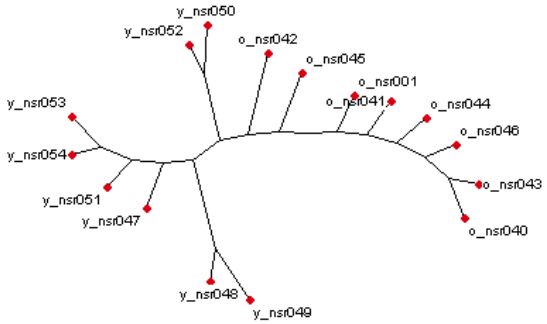

Figure 6. Distance tree based on two groups of old and young subjects. The patients are the ones of Table 6.

Table 1. Averaged distances of each patient from **gk_group** = (**gk**30_nn, **gk**31_nn, …, **gk**39_nn) and **nk_group** = (**nk**30_nn, **nk**31_nn, …, **nk**39_nn), respectively. Wrong classifications are marked in *red*

	gk_group	**nk_group**
gk02_nn	0,950977	0,955649
gk03_nn	0,9512	0,959749
gk04_nn	0,951591	0,957155
gk05_nn	0,949889	0,953167
gk06_nn	0,949679	0,958141
gk07_nn	0,951273	0,962977
gk08_nn	0,951308	0,962828
gk09_nn	0,949684	0,95644
gk10_nn	0,950085	0,959365
gk11_nn	0,949688	0,954517
gk13_nn	0,94936	0,95906
gk14_nn	0,949817	0,957204
gk15_nn	0,951751	0,964054
gk16_nn	0,949499	0,952967
gk17_nn	0,950058	0,956208
gk18_nn	0,951352	0,958267
gk19_nn	0,950012	0,957825
gk20_nn	0,953429	0,965333
gk21_nn	0,950678	0,959302
gk22_nn	0,950278	0,958852
nk10_nn	0,953073	0,952105
nk11_nn	0,955284	0,950414
nk12_nn	0,951612	0,954686
nk13_nn	0,955527	0,950697
nk14_nn	0,95358	0,958575
nk15_nn	0,952657	0,950346
nk16_nn	0,95545	0,952969
nk17_nn	0,975155	0,969354
nk18_nn	0,976497	0,964703
nk19_nn	0,952482	0,950202
nk20_nn	0,960154	0,955664
nk21_nn	0,960711	0,95591

nk22_nn	0,956478	0,95132
nk23_nn	0,961284	0,959017
nk24_nn	0,949156	0,956412
nk25_nn	0,959659	0,957893
nk26_nn	0,966242	0,958213
nk27_nn	0,960459	0,952844
nk28_nn	0,950147	0,953585
nk29_nn	0,953256	0,953296
nk30_nn	0,95347	0,953471

Table 2. Averaged distances of each patient from **gk_group** = (**gk**30_w, **gk**31_w, …, **gk**39_w) and **nk_group** = (**nk**30_w, **nk**31_w, …, **nk**39_w), respectively. Wrong classifications are marked in *red*

	gk_group	**nk_group**
gk02_w	0,944999	0,949697
gk03_w	0,942169	0,949849
gk04_w	0,94477	0,949449
gk05_w	0,946066	0,947472
gk06_w	0,943874	0,953748
gk07_w	0,945075	0,960126
gk08_w	0,94387	0,955866
gk09_w	0,943006	0,951416
gk10_w	0,941327	0,954052
gk11_w	0,942418	0,945749
gk13_w	0,940751	0,948664
gk14_w	0,942632	0,954633
gk15_w	0,943504	0,956356
gk16_w	0,94459	0,947752
gk17_w	0,940355	0,949688
gk18_w	0,944521	0,950204
gk19_w	0,942666	0,946773
gk20_w	0,944984	0,960437
gk21_w	0,943947	0,955633
gk22_w	0,944009	0,95303
nk10_w	0,94555	0,94192
nk11_w	0,950804	0,942961
nk12_w	0,94292	0,943463
nk13_w	0,950983	0,941804
nk14_w	0,949428	0,952428
nk15_w	0,947493	0,944664
nk16_w	0,950896	0,944168
nk17_w	0,970349	0,962885
nk18_w	0,964134	0,948842
nk19_w	0,946231	0,942469
nk20_w	0,948818	0,946029
nk21_w	0,966554	0,953678
nk22_w	0,953546	0,942596
nk23_w	0,960295	0,954839
nk24_w	0,943758	0,952151
nk25_w	0,953903	0,949325
nk26_w	0,961058	0,949702
nk27_w	0,966069	0,951091
nk28_w	0,944156	0,947916
nk29_w	0,949629	0,948452
nk30_w	0,945028	0,942885

Table 3. Matrix Distance of patients belonging to **gk_w** group with those in **nk_w** group. In the last row the averaged distances of each patient from the other patients in the gk_w and nk_w group respectively are shown. All the patience are correctly classified.

	gk07_w	gk08_w	gk09_w	gk10_w	gk11_w	gk12_w	gk14_w	nk60_w	nk61_w	nk62_w	nk63_w	nk64_w	nk65_w	nk66_w
gk07_w	0,001435	0,938174	0,946253	0,939033	0,947803	0,936606	0,942831	0,959717	0,955122	0,970542	0,956847	0,966968	0,978436	0,968842
gk08_w	0,938174	0,001412	0,946301	0,935012	0,945584	0,936529	0,934704	0,950728	0,95718	0,968602	0,953539	0,959136	0,974991	0,963785
gk09_w	0,946253	0,946301	0,001437	0,943899	0,938236	0,941878	0,946592	0,952453	0,943381	0,955213	0,944322	0,952932	0,964428	0,956702
gk10_w	0,939033	0,935012	0,943899	0,001368	0,941446	0,934417	0,933499	0,950982	0,9518	0,963374	0,94747	0,956438	0,968409	0,959216
gk11_w	0,947803	0,945584	0,938236	0,941446	0,001298	0,946182	0,944444	0,944265	0,944391	0,954103	0,941266	0,942231	0,953527	0,952257
gk12_w	0,936606	0,936529	0,941878	0,934417	0,946182	0,001472	0,935685	0,953846	0,952128	0,967961	0,952266	0,959454	0,973272	0,96475
gk14_w	0,942831	0,934704	0,946592	0,933499	0,944444	0,935685	0,0013	0,948952	0,954016	0,966076	0,947011	0,959655	0,97334	0,962134
nk60_w	0,959717	0,950728	0,952453	0,950982	0,944265	0,953846	0,948952	0,001432	0,943066	0,953047	0,942749	0,939516	0,951759	0,944677
nk61_w	0,955122	0,95718	0,943381	0,9518	0,944391	0,952128	0,954016	0,943066	0,001661	0,946108	0,941405	0,94639	0,953541	0,943809
nk62_w	0,970542	0,968602	0,955213	0,963374	0,954103	0,967961	0,966076	0,953047	0,946108	0,001311	0,942139	0,948552	0,962209	0,948499
nk63_w	0,956847	0,953539	0,944322	0,94747	0,941266	0,952266	0,947011	0,942749	0,941405	0,942139	0,001594	0,936979	0,955556	0,940503
nk64_w	0,966968	0,959136	0,952932	0,956438	0,942231	0,959454	0,959655	0,939516	0,94639	0,948552	0,936979	0,001619	0,94726	0,938837
nk65_w	0,978436	0,974991	0,964428	0,968409	0,953527	0,973272	0,97334	0,951759	0,953541	0,962209	0,955556	0,94726	0,001502	0,949249
nk66_w	0,968842	0,963785	0,956702	0,959216	0,952257	0,96475	0,962134	0,944677	0,943809	0,948499	0,940503	0,938837	0,949249	0,002286

gk	0,941783	0,939384	0,94386	0,937884	0,943949	0,93855	0,939626	0,951563	0,951145	0,963696	0,94896	0,956688	0,969486	0,961098
nk	0,965211	0,961137	0,952776	0,956813	0,947434	0,960525	0,958741	0,945802	0,94572	0,950092	0,943222	0,942922	0,953262	0,944262

Table 4. Averaged distances obtained by comparing non healthy patients (**chf**), with healthy subjects (**nsr**). For any single individual, the average is calculated with respect to the remaining patient in each group. Wrong classifications are marked in *red*

	chf	**nsr**
chf01_w	0,988736	0,993407
chf02_w	0,992512	0,994858
chf03_w	0,971186	0,996126
chf04_w	0,980403	0,991931
chf05_w	0,980736	0,992299
chf06_w	0,979843	0,9914
chf07_w	0,974151	0,993553
chf08_w	0,994647	0,99748
chf09_w	0,969402	0,994815
chf10_w	0,966486	0,992431
chf11_w	0,979891	0,99794
chf12_w	0,981962	0,992295
chf13_w	0,973136	0,996432
nsr01_w	0,994181	0,925976
nsr02_w	0,993675	0,928663
nsr03_w	0,993803	0,923911
nsr04_w	0,994018	0,935523
nsr05_w	0,994254	0,925418
nsr06_w	0,994561	0,930583
nsr07_w	0,993325	0,922587
nsr08_w	0,994585	0,938982
nsr09_w	0,994489	0,923555
nsr10_w	0,994857	0,926272
nsr11_w	0,994628	0,92443
nsr12_w	0,994004	0,931252
nsr13_w	0,994587	0,923272

Table 5. Averaged distances obtained by comparing young patients [18] (recorded with numbers) against old individuals (gk). For any single individual, the average is calculated with respect to the remaining patients in each group. Wrong classifications are marked in *red*

	Young	Old
16272	0,948901182	0,94897375
16273	0,949356636	0,952352833
16420	0,951251364	0,95512975
16483	0,957391273	0,95428375
16539	0,955175818	0,9531345
16773	0,954099182	0,952232333
16786	0,951502545	0,953977667
16795	0,949578455	0,9527505
17052	0,950846545	0,955799333
17453	0,949297455	0,950581
18177	0,951735364	0,953157417
18184	0,954342545	0,951167333
gk11_nn	0,952366167	0,949516818
gk12_nn	0,951272333	0,949272273
gk13_nn	0,950989583	0,9476455
gk14_nn	0,950517167	0,9487745
gk15_nn	0,955528917	0,9496472
gk16_nn	0,948811833	0,9510059
gk17_nn	0,95274075	0,9486855
gk18_nn	0,955450167	0,9518555
gk19_nn	0,953471583	0,9502753
gk20_nn	0,95707775	0,9512193
gk21_nn	0,952414667	0,9481727
gk22_nn	0,95289925	0,9486232

Table 6. Distance matrix out of old patients (from nsr001 to nsr046) and young individuals (from nsr047 to nsr054). Data are downloaded from [20]. Wrong classifications are marked in red.

	old_nsr001	old_nsr040	old_nsr041	old_nsr042	old_nsr043	old_nsr044	old_nsr045	old_nsr046	young_nsr047	young_nsr048	young_nsr049	young_nsr050	young_nsr051	young_nsr052	young_nsr053	young_nsr054
old_nsr001	0.000367	0.952024	0.95061	0.945933	0.949736	0.950958	0.948364	0.954146	0.949841	0.954005	0.955413	0.948662	0.953394	0.950865	0.9557	0.955683
old_nsr040	0.952024	0.00034	0.953121	0.951928	0.946163	0.949815	0.953287	0.949094	0.958728	0.962704	0.960268	0.955893	0.962963	0.956731	0.969789	0.96495
old_nsr041	0.95061	0.953121	0.000377	0.951346	0.949566	0.949914	0.950161	0.955327	0.955474	0.958248	0.958666	0.952243	0.955644	0.952609	0.958549	0.953429
old_nsr042	0.945933	0.951928	0.951346	0.000345	0.949109	0.951131	0.946419	0.951499	0.946238	0.951136	0.949203	0.947774	0.947583	0.949594	0.951523	0.952853
old_nsr043	0.949736	0.946163	0.949566	0.949109	0.000378	0.948612	0.951932	0.947139	0.954896	0.957349	0.957155	0.95058	0.96005	0.951108	0.964736	0.960688
old_nsr044	0.950958	0.949815	0.949914	0.951131	0.948612	0.000348	0.950339	0.951358	0.956038	0.958926	0.959561	0.95272	0.958239	0.954904	0.962013	0.959726
old_nsr045	0.948364	0.953287	0.950161	0.946419	0.951932	0.950339	0.000404	0.954754	0.949637	0.956214	0.953575	0.95223	0.950636	0.952009	0.953416	0.956672
old_nsr046	0.954146	0.949094	0.955327	0.951499	0.947139	0.951358	0.954754	0.000412	0.9564	0.960209	0.956874	0.954809	0.960467	0.955471	0.967215	0.96261
young_nsr047	0.949841	0.958728	0.955474	0.946238	0.954896	0.956038	0.949637	0.9564	0.000335	0.948902	0.945155	0.950148	0.946676	0.951359	0.948919	0.950762
young_nsr048	0.954005	0.962704	0.958248	0.951136	0.957349	0.958926	0.956214	0.960209	0.948902	0.000327	0.947119	0.948525	0.952556	0.949741	0.951433	0.950291
young_nsr049	0.955413	0.960268	0.958666	0.949203	0.957155	0.959561	0.953575	0.956874	0.945155	0.947119	0.000367	0.952881	0.949172	0.951173	0.951686	0.951869
young_nsr050	0.948662	0.955893	0.952243	0.947774	0.95058	0.95272	0.95223	0.954809	0.950148	0.948525	0.952881	0.000302	0.955884	0.945132	0.954209	0.951154
young_nsr051	0.953394	0.962963	0.955644	0.947583	0.96005	0.958239	0.950636	0.960467	0.946676	0.952556	0.949172	0.955884	0.000359	0.953737	0.948845	0.951921
young_nsr052	0.950865	0.956731	0.952609	0.949594	0.951108	0.954904	0.952009	0.955471	0.951359	0.949741	0.951173	0.945132	0.953737	0.000331	0.95448	0.951443
young_nsr053	0.9557	0.969789	0.958549	0.951523	0.964736	0.962013	0.953416	0.967215	0.948919	0.951433	0.951686	0.954209	0.948845	0.95448	0.00038	0.947724
young_nsr054	0.955683	0.96495	0.953429	0.952853	0.960688	0.959726	0.956672	0.96261	0.950762	0.950291	0.951869	0.951154	0.951921	0.951443	0.947724	0.00032
olds	0.950253	0.950776	0.951435	0.949623571	0.948893857	0.950303857	0.950750857	0.951902429	0.9534065	0.957348875	0.956339375	0.951863875	0.956122	0.952911375	0.960367625	0.958326375
youngs	0.952945375	0.96150325	0.95560775	0.949488	0.95707025	0.957765875	0.953048625	0.959256875	0.948845857	0.949795286	0.949865	0.951133286	0.951255857	0.951009286	0.951042286	0.950737714

■