Linguistic Analysis of the Human Heartbeat Using Frequency and Rank Order Statistics

Albert C.-C. Yang,[1,2] Shu-Shya Hseu,[1,2] Huey-Wen Yien,[1,2,*] Ary L. Goldberger,[3] and C.-K. Peng[3]

[1]*Department of Anesthesiology, Taipei Veterans General Hospital, Taipei, Taiwan*

[2]*School of Medicine, National Yang-Ming University, Taipei, Taiwan*

[3]*Cardiovascular Division and Margret and H. A. Rey Institute for Nonlinear Dynamics in Medicine, Beth Israel Deaconess Medical Center/Harvard Medical School, Boston, Massachusetts 02215*

(Received 9 May 2002; published 13 March 2003)

Complex physiologic signals may carry unique dynamical signatures that are related to their underlying mechanisms. We present a method based on rank order statistics of symbolic sequences to investigate the profile of different types of physiologic dynamics. We apply this method to heart rate fluctuations, the output of a central physiologic control system. The method robustly discriminates patterns generated from healthy and pathologic states, as well as aging. Furthermore, we observe increased randomness in the heartbeat time series with physiologic aging and pathologic states and also uncover nonrandom patterns in the ventricular response to atrial fibrillation.

DOI: 10.1103/PhysRevLett.90.108103 PACS numbers: 87.19.Hh, 05.40.–a, 05.45.Tp, 89.75.Kd

Physiologic systems generate complex fluctuations in their output signals that reflect the underlying dynamics. Therefore, finding and analyzing hidden dynamical structures of these signals are of both basic and clinical interest. Here, we detect and quantify such temporal structures in the human heart rate time series using tools from statistical linguistics.

Human cardiac dynamics are driven by the complex nonlinear interactions of two competing forces: sympathetic stimulation increases and parasympathetic stimulation decreases heart rate. For this type of intrinsically noisy system, it may be useful to simplify the dynamics via mapping the output to binary sequences, where the increase and decrease of the interbeat intervals are denoted by 1 and 0, respectively [1]. The resulting binary sequence retains important features of the dynamics generated by the underlying control system, but is tractable enough to be analyzed as a symbolic sequence.

Consider an interbeat interval time series, $\{x_0, x_1, x_2, \ldots, x_N\}$, where x_i is the ith interbeat interval. We can classify each pair of successive interbeat intervals into one of the two states that represents a decrease in x, or an increase in x. These two states are mapped to the symbols 0 and 1, respectively [Fig. 1(a)]:

$$I_n = \begin{cases} 0, & \text{if } x_n \le x_{n-1}, \\ 1, & \text{if } x_n > x_{n-1}. \end{cases} \quad (1)$$

In this study, we map $m+1$ successive intervals to a binary sequence of length m, called an m-bit "word." Each m-bit word, w_k, therefore, represents a unique pattern of fluctuations in a given time series. By shifting one data point at a time, the algorithm produces a collection of m-bit words over the whole time series. Therefore, it is plausible that the occurrence of these m-bit words reflects the underlying dynamics of the original time series. Different types of dynamics thus produce different distributions of these m-bit words.

In studies of natural languages, it has been observed that different authors have a preference for the words they use with higher frequency [3]. To apply this concept to symbolic sequences mapped from the interbeat interval time series, we count the occurrences of different words [Fig. 1(b)], and then sort them according to descending frequency. The resulting rank-frequency distribution [Fig. 1(c)], therefore, represents the statistical hierarchy of symbolic words of the original time series. For example, the first rank word corresponds to one type of fluctuation which is the most frequent pattern in the time series. In contrast, the last rank word defines the most unlikely pattern in the time series.

To define a measurement of *similarity* between two signals, we plot the rank number of each m-bit word in the first time series against that of the second time series (see Fig. 2). If two time series are similar in their rank order of the words, the scattered points will be located near the diagonal line. Therefore, the average deviation of these scattered points away from the diagonal line is a measure of the "distance" between these two time series. Greater distance indicates less similarity and vice versa. In addition, we incorporate the likelihood of each word in the following definition of a weighted distance, D_m, between two symbolic sequences, S_1 and S_2.

$$D_m(S_1, S_2) = \frac{\sum_{k=1}^{2^m} |R_1(w_k) - R_2(w_k)|\, p_1(w_k)\, p_2(w_k)}{(2^m - 1) \sum_{k=1}^{2^m} p_1(w_k)\, p_2(w_k)}. \quad (2)$$

Here $p_1(w_k)$, and $R_1(w_k)$ represent probability and rank of a specific word, w_k, in time series S_1. Similarly, $p_2(w_k)$ and $R_2(w_k)$ stand for probability and rank of the same m-bit word in time series S_2. The absolute difference of ranks is multiplied by the normalized probabilities as a weighted sum [4] and divided by the value $2^m - 1$ to keep the D_m value in the same range of [0, 1].

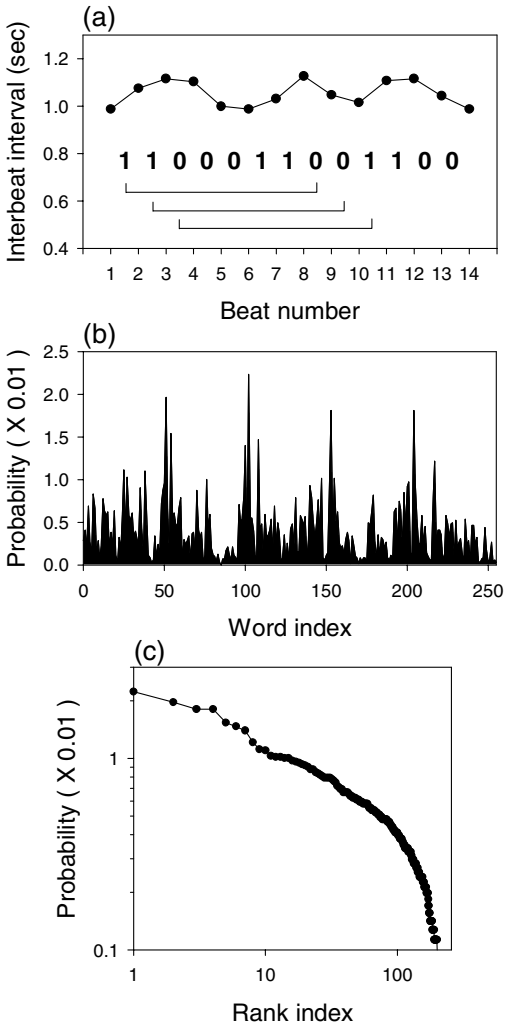

FIG. 1. (a) Schematic illustration of the mapping procedure for 8-bit words from part of a 2 h heartbeat time series. (b) Probability distribution of every 8-bit word. The word index ranges from 0 to $2^m - 1$ for m-bit words. For our example of $m = 8$, there is a total of 256 possible words. The word (11000110) shown in (a) is labeled as word 198 ($= 128 + 64 + 4 + 2$). (c) Rank ordered probability plotted on a log-log scale. The linear regime (for rank ≤ 50) is reminiscent of Zipf's law for natural languages [2].

We apply the distance measurement to address the following questions: (i) Are there unique dynamical patterns associated with each individual? (ii) Are there characteristic patterns that describe the dynamical structures of different physiologic/pathologic states? (iii) As a healthy system changes with disease and aging, can we see quantifiable changes in the dynamical patterns related to the degradation of the integrative control systems? We investigate these questions with databases [5] that include 40 ostensibly healthy subjects with subgroups of young (ten females and ten males, average 25.9 years) and elderly (ten females and ten males, average 74.5 years), a group of subjects ($n = 43$) with severe congestive heart failure (CHF) (15 females and 28 males, average 55.5 years), and a group of nine subjects with atrial fibrillation (AF). All subjects in the healthy and AF groups had 2 h recordings. The CHF group had longer data sets (16 to 24 h for each subject) [6]. Here we present only the analysis for the case $m = 8$; however, similar results were obtained for $m = 4$ to 12.

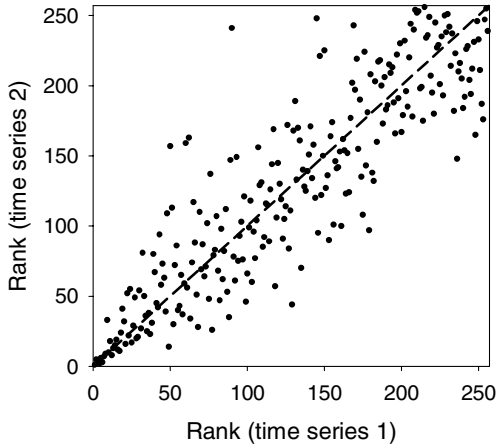

FIG. 2. Rank order comparison of two cardiac interbeat interval time series from the same subject. For each word, its rank in the first time series is plotted against its rank in the second time series. The dashed diagonal line indicates the case where the rank order of words for both time series is identical.

First, we examine whether the distance for subsets of the time series from the same subject is closer than that for the time series from different subjects under similar physiologic states. We divide each subject's time series into two subsets and measure the distance between these two subsets. We also calculate the distance between each pair of subjects who belong to the same group (Table I). Our results show that the intrasubject distances are indeed smaller than the intragroup distances. The small intrasubject distances indicate that there are unique dynamical patterns associated with each individual at small time scales. The only exception is for the AF group. It is difficult to distinguish one AF subject from another based on our rank order distance. This result is consistent with previous studies showing that, on small time scales (≤ 200 s), heart rate fluctuations of AF subjects do not exhibit consistent structures [7].

Next, we measure the average distance between subjects across different groups. We notice that distances between subjects across groups are typically greater than distances between subjects within a group. This result supports the notion that there are dynamical patterns for different physiologic/pathologic states. However, there is overlap among groups. To simplify the picture, we define the intergroup distance of groups A and B as the average distance between all pairs of subjects where one subject is from group A and the other subject is from group B. We calculate the intergroup distances among all groups of our databases as well as a group of

TABLE I. Distance measurement of 8-bit words. The intrasubject results are average distances measured between two subset time series from the same subject. The intragroup results are the average distances between two different subjects from the same group. Values are given as mean ± standard deviation. The intrasubject distances are significantly lower ($p < 10^{-4}$ by t test) than the intragroup distances in all groups except for the atrial fibrillation group ($p = 0.8$).

	Intrasubject	Intragroup
Healthy young	0.056 ± 0.050	0.161 ± 0.106
Healthy elderly	0.077 ± 0.052	0.209 ± 0.110
Congestive heart failure	0.053 ± 0.047	0.100 ± 0.062
Atrial fibrillation	0.046 ± 0.015	0.045 ± 0.012

100 artificial time series of uncorrelated noise (white noise group).

The method for constructing phylogenetic trees [8] is a useful tool to present our results since the algorithm arranges different groups on a branching tree to best fit the pairwise distance measurements. In Fig. 3 we show the result of a rooted tree for the case of $m = 8$ [9]. We note that the structure of the tree is consistent with the underlying physiology: the farther down the branch the more complex the dynamics are. The groups are arranged in the following order (from bottom to top shown in Fig. 3): (i) The time series from the healthy young group represents dynamical fluctuations of a highly complex integrative control system. (ii) The healthy elderly group represents a slight deviation from the "optimal" youthful state, possibly due to the decoupling (or dropout) of components in the integrative control system [10]. (iii) Severe damage to the control system is represented by the CHF group. These individuals have profound abnormalities in cardiac function associated with pathologic alterations in both the sympathetic and the parasympathetic control mechanisms that regulate beat-to-beat variability [11]. (iv) The AF group is an example of a pathologic state in which there appears to be very limited external input on the heartbeat control system. (v) The artificial white noise group represents the extreme case in that only noise and no signal is present.

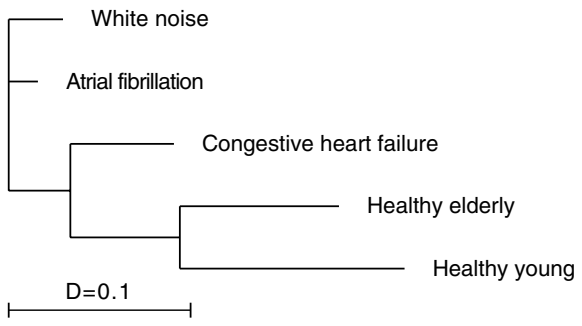

FIG. 3. A rooted phylogenetic tree generated according to the distances between different groups. White noise indicates a simulated uncorrelated random time series. The distance (D) between any two groups is the summation of the horizontal lengths along the shortest path on the tree that connects them. For example, the distance is 0.211 between healthy young and healthy elderly groups.

A further application of the rank order distance concept is to quantify the degree of nonrandomness. To this end, we generate the surrogate time series by random shuffling of the original time series. Random shuffling of the data yields exactly the same distribution of the original interbeat intervals sequence, but destroys their sequential ordering. The distance defined in Eq. (2) between an interbeat interval time series and its randomized surrogate provides an index of the nonrandomness of the time series. Here we present some intriguing results by applying this nonrandomness index. Figure 4(a) illustrates a heartbeat interval time series from a healthy subject showing complex variability. In contrast, a time series from a CHF subject [Fig. 4(b)] shows less variability. For healthy subjects, the rank map between each original signal and its randomized surrogate shows prominent scatter [Fig. 4(c)]. The nonrandomness index measured here is 0.31. In contrast, heart rate dynamics with CHF show rank maps with relatively narrow distributions [Fig. 4(d)] indicating that fluctuations in CHF are closer to random (nonrandomness index = 0.10).

Next, we calculate nonrandomness distances that correspond to different word lengths, m, ranging from 2 to 12. Figure 5 shows the result for $m = 8$. For healthy and

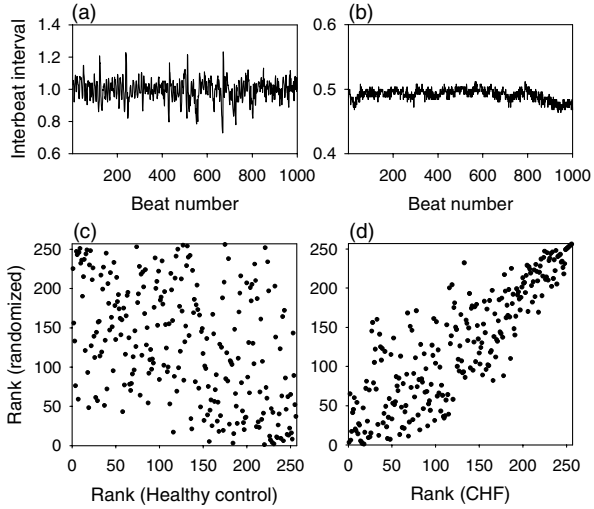

FIG. 4. Representative interbeat time series for (a) a healthy subject, and (b) a subject with congestive heart failure (CHF). (c) Rank order comparison of the time series in (a) and its randomized surrogate. (d) Rank order comparison of the time series in (b) and its randomized surrogate. The narrower scattering of (d) compared to (c) implies that heartbeat fluctuations in the congestive heart failure subject are more comparable to random than that of the healthy subject. Both (c) and (d) are results for the case $m = 8$.

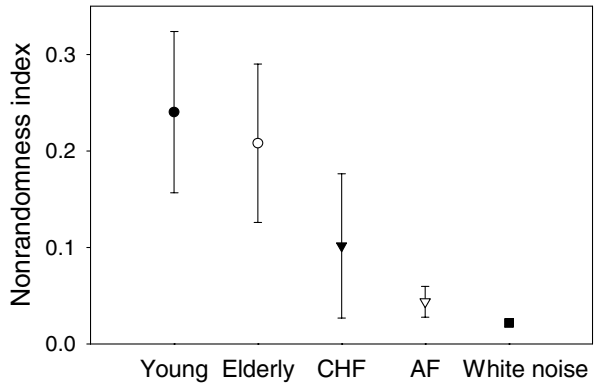

FIG. 5. Nonrandomness index ($m = 8$) of the interbeat interval time series derived from healthy young and elderly subjects, subjects with congestive heart failure (CHF), subjects with atrial fibrillation (AF), and artificial time series of uncorrelated noise. Values are given as mean ± standard deviation.

CHF subjects, there are significant differences ($p < 10^{-4}$) in this nonrandomness index over the entire range of m studied. However, the nonrandomness distance of the healthy young group is only significantly higher than that of the healthy elderly group at the scale $m = 3$ ($p < 0.05$), suggesting a preservation of most of the nonrandom features of heart rate dynamics with physiologic aging. Subjects with AF also show significantly higher values of the nonrandomness index than white noise over the range of $5 \leq m \leq 9$ ($p < 10^{-4}$). Therefore, even on small time scales, our method can effectively discriminate certain data sets of the AF group from white noise, whereas conventional methods have not been successful in this regard.

Another attractive feature of rank order statistics is that the method is useful in examining the details of the underlying dynamics. For example, our nonrandomness test indicates a significant difference between AF and uncorrelated noise. We can further analyze the rank numbers of the "words" that contribute to this difference from white noise. The assumption is that if a word dramatically changes its rank after randomization (shuffling), the fluctuations mapped by this word may not be random and could contain relevant structural information. After systemically reviewing all AF recordings, the words that are significantly different from random sequences, occurring in a subset of these subjects, are given as (00100100), (00110001), (00101000), and (01000100). This finding suggests hidden structural organization in the short-term variation of the ventricular rhythm in AF. These sequences need further systematic analysis, in conjunction with information from intracardiac electrophysiologic studies, to elucidate the mechanism of the ventricular response to AF in different settings.

In summary, we introduce a quantitative metric to define distances among symbolic sequences. In application to the heart rate time series, this approach provides new quantitative information that is not measured by conventional heart rate variability techniques [12]. The method can be easily adapted to other physiologic and physical time series provided that a meaningful mapping to symbolic sequences can be obtained. Finally, this new linguistic-type method is potentially useful because of its ability to take into account both macroscopic structures and the microscopic details of the dynamics.

We thank L. Glass, S. Havlin, J. M. Hausdorff, C.-K. Hu, I. C. Henry, J. E. Mietus, and V. Schulte-Frohlinde for valuable discussions. We gratefully acknowledge the support from the NIH/NCRR (P41-RR13622), the NIH/NIA OAIC, the G. Harold and Leila Y. Mathers Charitable Foundation, the Centers for Disease Control and Prevention, the National Science Council of Taiwan (NSC90-2314-B-075-127), and the Academia Sinica (Taipei).

*Corresponding author.
Email address: hwyiin@vghtpe.gov.tw

[1] J. Kurths *et al.*, Chaos **5**, 88 (1995); Y. Ashkenazy *et al.*, Phys. Rev. Lett. **86**, 1900 (2001).
[2] G. K. Zipf, *Human Behavior and the Principle of "Least Effort"* (Addison-Wesley, New York, 1949).
[3] F. Mosteller and D. L. Wallace, *Applied Bayesian and Classical Inference: The Case of the Federalist Papers* (Springer-Verlag, New York, 1984), 2nd ed.; D. I. Holmes, Comput. Humanities **28**, 87 (1994); S. Havlin, Physica (Amsterdam) **216A**, 148 (1995); F. J. Tweedie and R. H. Baayen, Comput. Humanities **32**, 323 (1998).
[4] Other weighting functions may be used in Eq. (2).
[5] Databases are available at http://www.physionet.org/. See A. L. Goldberger *et al.*, Circulation **101**, E215 (2000).
[6] Our conclusions remain the same with shorter (2 hour) recordings for the CHF group.
[7] F. Hayano *et al.*, Am. J. Physiol. **273**, H2811 (1997); K. M. Stein *et al.*, Am. J. Physiol. **277**, H452 (1999).
[8] W. M. Fitch and E. Margoliash, Science **155**, 279 (1967); J. Felsenstein, Cladistics **5**, 164 (1989); R. D. Page, Comput. Appl. Biosci. **12**, 357 (1996).
[9] We obtained a similar tree structure for m in the range of 4 to 12. Larger m values require longer length data sets to be statistically reliable.
[10] A. L. Goldberger, C.-K. Peng, and L. A. Lipsitz, Neurobiol. Aging **23**, 23 (2002).
[11] H. R. Middlekauff, Curr. Opin. Cardiol. **12**, 265 (1997); H. V. Huikuri, A. Castellanos, and R. J. Myerburg, N. Engl. J. Med. **345**, 1473 (2001).
[12] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, Circulation **93**, 1043 (1996). We have calculated 12 conventional measurements of heart rate variability discussed therein. The nonrandomness index is not strongly correlated with any of these conventional analyses, including those that relate to parasympathetic regulation (e.g., pNN50).