Statistical physics approach to categorize biologic signals: From heart rate dynamics to DNA sequences

C.-K. Peng
Margret and H. A. Rey Institute for Nonlinear Dynamics in Medicine, Division of Interdisciplinary Medicine and Biotechnology, Beth Israel Deaconess Medical Center/Harvard Medical School, Boston, Massachusetts 02215

Albert C.-C. Yang
Taipei Veterans General Hospital, Taipei, Taiwan

Ary L. Goldberger
Margret and H. A. Rey Institute for Nonlinear Dynamics in Medicine, Division of Interdisciplinary Medicine and Biotechnology, Beth Israel Deaconess Medical Center/Harvard Medical School, Boston, Massachusetts 02215

(Received 6 September 2006; accepted 21 February 2007; published online 30 March 2007)

We recently proposed a novel approach to categorize information carried by symbolic sequences based on their usage of repetitive patterns. A simple quantitative index to measure the dissimilarity between two symbolic sequences can be defined. This information dissimilarity index, defined by our formula, is closely related to the Shannon entropy and rank order of the repetitive patterns in the symbolic sequences. Here we discuss the underlying statistical physics assumptions of this dissimilarity index. We use human cardiac interbeat interval time series and DNA sequences as examples to illustrate the applicability of this generic approach to real-world problems. © *2007 American Institute of Physics*. [DOI: 10.1063/1.2716147]

Human cardiac interbeat interval fluctuates in a complex manner that reflects the dynamical regulation of the underlying control system. Understanding the content of information being expressed by dynamical processes constitutes a critical step to comprehend the underlying processes. Deciphering these information conveyers, therefore, becomes a major scientific objective in many research endeavors. However, to achieve this goal, certain knowledge of the underlying process is needed. Typically, it is not feasible to develop generic algorithms that can decipher every possible type of information without essential knowledge about the specific system being studied. A more modest and realistic task is to analyze certain aspects of the information and to group them into different categories without detailed understanding of the content of the information. From a practical point of view, the ability to categorize different types of signals can serve as a useful tool in real-world applications, where monitoring a system's status is important. In this paper, we propose a generic approach to address the challenge of categorizing dynamical signals based on some fundamental assumptions of statistical physics and information theory.

I. INTRODUCTION

We focus our efforts here on developing an understanding of symbolic sequences generated by biological processes. Examples of biological symbolic sequences include neural spike trains, DNA codes, and human language. This emphasis of symbolic sequences is necessary for the analytical basis of our computational method. However, this emphasis does not limit our approach since most signals of continuous variables can be mapped to symbolic sequences while retaining essential information of the original signals. A good example is the heartbeat time series. Numerous studies have shown that symbolic sequence representation of heart rate time series can often reveal many hidden dynamical patterns that have clinical significance.[1–7]

Consider the case where a specific symbolic sequence is generated by an underlying process. This symbolic sequence may be created as a carrier to transmit information, such as the neural spike activities propagating from one neuron to another. Alternatively, a symbolic sequence could primarily be a by-product (biomarker) of the dynamical system that produced it. One such example is the sleep stage transitions during an overnight rest. In this case, the dynamics of the entire sleep period can be represented by a symbolic sequence of sleep stages. Often, mixtures of both factors discussed above can affect the creation of these sequences. For instance, DNA sequences are modified by stochastic mutation processes while simultaneously being selected by evolutionary pressures according to the information they carried. In all three scenarios, the information contained in the symbolic sequences can shed some light on the underlying dynamics that generate these sequences. The goal of this paper is to develop a generic algorithm to categorize different types of symbolic sequences, regardless of their origins and functions.

A common observation of complex symbolic sequences is that different dynamical signals appear to exhibit different "styles" of behaviors. This style or *signature* of the dynamics often can be identified by certain complex but repetitive patterns in the symbolic sequences. Therefore, we are interested

1054-1500/2007/17(1)/015115/6/$23.00

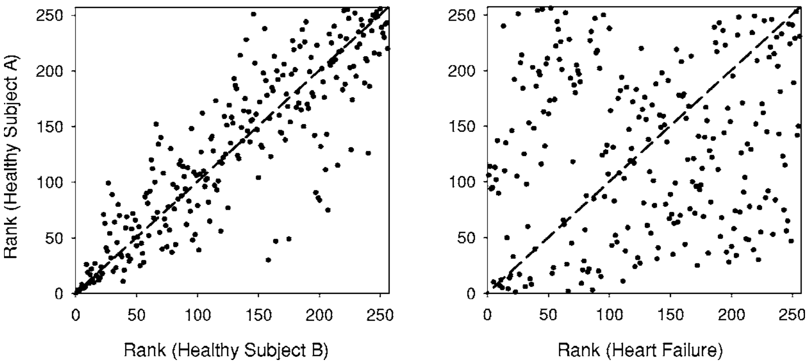

FIG. 1. Rank order comparison of two symbolic sequences mapped from cardiac interbeat interval time series from different subjects. See text for the exact mapping procedure. For each word, its rank in the first symbolic sequence is plotted against its rank in the second symbolic sequence. The dashed diagonal line indicates the case where the rank-order of words for both time series is identical. Note the greater scattering of data points (dissimilarity) when the data from the subject with heart failure is compared to the healthy subject.

in the following questions: Can we identify relevant "structures" of these repetitive patterns? Furthermore, can we quantitatively compare one style of repetitive patterns to another?

To address these challenges, we discuss a systematic approach to categorize different types of information encoded in symbolic sequences. The traditional approach is to develop a set of indices that can characterize the dynamics and classify signals in terms of these indices. The alternative approach that we take involves pairwise comparisons of signals in order to calculate an abstract *distance* or "dissimilarity" between the signals. In our analysis, the similarity between two different symbolic sequences is measured by the comparability of the usage of repetitive patterns. Our approach is based on the concept that the information content in any symbolic sequence is primarily determined by the repetitive usage of its basic elements. The specific goal of our algorithm, therefore, is to quantify the similarity between symbolic sequences based on statistical comparisons of the rank and frequency of repetitive elements.

At the core of our algorithm is the definition of a measure for dissimilarity between two symbolic sequences. This dissimilarity index forms the quantitative basis of our information categorization process. We will discuss the basic theoretical considerations for deriving this dissimilarity index in the next section.

II. MEASUREMENT OF DISSIMILARITY

To define a measurement of *similarity* between two symbolic sequences, we carry out the following procedures. First, we applied a sliding window (observation box) of size m to one symbolic sequence. In each window, the segment of the sequence can be identified as a "word" of length m. By sliding the window down the entire sequence, we can count the occurrence of each m-bit word. Then, we sort these m-bit words according to their frequencies of occurrence. The most frequently occurring word is ranked number 1, and so on. Then, we perform the same procedure on the other symbolic sequence. Note that for any given m-bit word, its rank order can be different in these two sequences. Therefore, we can plot the rank number of each m-bit word in the first symbolic sequence against that of the second symbolic sequence (see Fig. 1). If two symbolic sequences 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 symbolic sequences. Greater distance indicates less similarity and vice versa. The advantage of using rank is that it is a nonparametric quantity that is less perturbed by noise. Thus, Havlin introduced a measure of "distance" based on the rank difference alone.[8] However, we think it is reasonable to incorporate the likelihood of each word in the definition of the dissimilarity. Therefore, we proposed the following definition of a weighted distance, D_m, between two symbolic sequences, Ψ_1 and Ψ_2,

$$D_m(\Psi_1,\Psi_2) = \frac{1}{L}\sum_{k=1}^{L} |R_1(s_k) - R_2(s_k)|F(s_k), \tag{1}$$

where $F(s_k)$ is the weighting factor for the word s_k. The incorporation of this factor into the definition provides a way to take into account that different symbols have different contributions to the overall measure. The simplest view is that $F(s_k)$ should be proportional to the probability of occurrences for s_k in sequences Ψ_1 and Ψ_2, denoted as $p_1(s_k)$ and $p_2(s_k)$, respectively. This definition was successfully used in our previous studies[6,9] as a special case of Eq. (1). Recently, we proposed another definition for the weighting factor,[10,11] i.e.,

$$F(s_k) = [-p_1(s_k)\log p_1(s_k) - p_2(s_k)\log p_2(s_k)]/Z. \tag{2}$$

Here $p_1(s_k)$ and $R_1(s_k)$ represent probability and rank of a specific word, s_k, in symbolic sequence Ψ_1. Similarly, $p_2(s_k)$ and $R_2(s_k)$ stand for probability and rank of the same m-bit word in symbolic sequence Ψ_2. In summary, the absolute difference of ranks is multiplied by the normalized probabilities as a weighted sum by using Shannon entropy[12] as the weighting factor. Finally, the sum is divided by the value L to keep the D_m value in the same range of [0 1]. The normalization factor Z in Eq. (2) is given by

$$Z = \sum_{k=1}^{L} [-p_1(s_k)\log p_1(s_k) - p_2(s_k)\log p_2(s_k)]. \tag{3}$$

Although, empirically, the definition of Eq. (2) provides better results in various applications when compared to using other proposed definitions, it is unclear if there is a generic optimal weighting factor that can provide the best classification across all types of symbolic sequences. Therefore, it is important to derive the formula for the weighting function from fundamental assumptions as we will carry out in the following pages.

In order to better understand the potential and limitations of this approach, we need to examine the assumptions underlying this categorization algorithm. Here we present a rationale of how the weighting function, $F(s_k)$, in Eq. (2) can be derived from fundamental assumptions. Basically, we consider complex signals generated by a biological system as information carriers. Further, the information is directly related to the underlying dynamical state of the system. The entire symbolic sequences (signals) can be considered as comprised of many distinct short sequences (building blocks) that correspond to different *microstates* of the system.

Consider a dynamical process that jumps between l microstates. It is reasonable to expect that the signal generated by this dynamical process contains l distinct patterns: $\{s_1, s_2, \ldots, s_l\}$. In a given signal, a pattern s_i will appear repeatedly n_i times if the corresponding microstate is visited n_i times.

We assume that for each pattern s_i there is an associated *energy*, denoted as ε_i, corresponding to the underlying microstate. Low-energy states are more stable and can be visited more often, while transitions to higher-energy states consume more energy and are less likely to occur. Introducing the concept of energy as a *hidden variable* has the advantage of mapping the original problem of distribution to a fundamental statistical physics problem where analytical techniques for solving it are more accessible.

Furthermore, to simplify this problem, we assume initially that each pattern appears independently of other patterns in the sequence. In other words, each microstate can be visited independent of the history of the dynamical system. This assumption is clearly invalid in many dynamical processes where correlations among microstates are important. Nevertheless, this assumption can be treated as a first-order approximation, i.e., as a starting point for approaching the much more challenging problem of signals with correlations and hysteresis.

One important question concerns how often does the dynamical system visit each microstate as a function of its energy? We can find the answer to this question in the following way:

The total number of microstates being visited, $\mathcal{N}$, and the total energy, $\mathcal{E}$, dissipated by the dynamical process are given by

$$\mathcal{N} = \sum_{i=1}^{l} n_i \quad \text{and } \mathcal{E} = \sum_{i=1}^{l} n_i \varepsilon_i . \tag{4}$$

For a given set of $\{n_1, n_2, \ldots, n_l\}$, the number of different combinatorial configurations of visiting these microstates is

$$\Omega = \frac{\mathcal{N}!}{\Pi(n_i!)} . \tag{5}$$

Obviously, the greater the Ω, the higher the probability that the given set of $\{n_1, n_2, \ldots, n_l\}$ will be observed. Therefore, it is reasonable to assume that, given the constraints of the total length $\mathcal{N}$ and total energy $\mathcal{E}$, the observed n_i actually comes from the distribution that leads to the maximal Ω value. Thus, the fundamental question is what type of distribution of $\{s_i\}$ will provide the maximal number of configurations, Ω. This problem is equivalent to finding the distribution for the maximal value of $\ln \Omega$, since $\ln \Omega$ is a monotonic function of Ω. The derivative of $\ln \Omega$ is zero at the maximal value. By introducing two Lagrangian multipliers to take into account the two constraints of Eq. (4), we can write the condition as follows:

$$\frac{\partial \ln \Omega}{\partial n_i} - \alpha \frac{\partial \sum_i n_i}{\partial n_i} - \beta \frac{\partial \sum_i n_i \varepsilon_i}{\partial n_i} = 0 . \tag{6}$$

Substituting Eq. (5) into Eq. (6) and applying the Stirling formula for large $\mathcal{N}$ value, $\ln \mathcal{N}! \simeq \mathcal{N} \ln \mathcal{N} - \mathcal{N}$, we obtain

$$\ln n_i = -\alpha - \beta \varepsilon_i \quad \text{or} \quad n_i = \exp(-\alpha - \beta \varepsilon_i) . \tag{7}$$

From this we find that the maximal value of Ω is given by the Boltzmann distribution; i.e., the probability of finding the pattern s_i is

$$p(s_i) = \frac{n_i}{\mathcal{N}} = \frac{1}{Q} \exp(-\beta \varepsilon_i) , \tag{8}$$

where $Q = \Sigma_i \exp(-\beta \varepsilon_i)$ is the *partition function*. It is also necessary to show that the fluctuations around the maximal Ω are very small. In other words, Ω has a very sharp peak at the maximum. Small deviations from the Boltzmann distribution will lead to significant drops in the value of Ω. (We omit this proof due to page limitations.) Therefore, it is consistent with our assumption that the actual distribution of $\{n_i\}$ appears at the condition with maximal Ω, since other possible distributions are all negligible when compared to that at the maximal Ω.

From the above derivation, one interpretation of the dissimilarity index that we defined in Eq. (2) becomes evident based on a simple connection between the microstate energy and the Shannon entropy used in Eq. (2),

$$n_i \varepsilon_i \propto -p(s_i) \log p(s_i) . \tag{9}$$

Therefore, in this case, the Shannon entropy of a specific repetitive pattern s_i is proportional to the total amount of energy represented by repetitive transitions to the corresponding microstate. This connection is important since it links one fundamental concept of information theory (Shannon entropy) to that of a dynamical system (energy of microstates). Further work is needed to validate this connection in different classes of dynamical systems.

III. APPLICATIONS TO BIOLOGICAL SIGNALS

The information categorization method developed here provides a novel tool to classify complex signals. The generic features of the method make it applicable to different types of complex systems, ranging from complex dynamics under physiologic control to DNA evolution. However, due to the generality of the core method, customized modifications are needed for different applications. In this section, we will discuss how to apply this method to categorize different types of heart rate time series and DNA sequences.

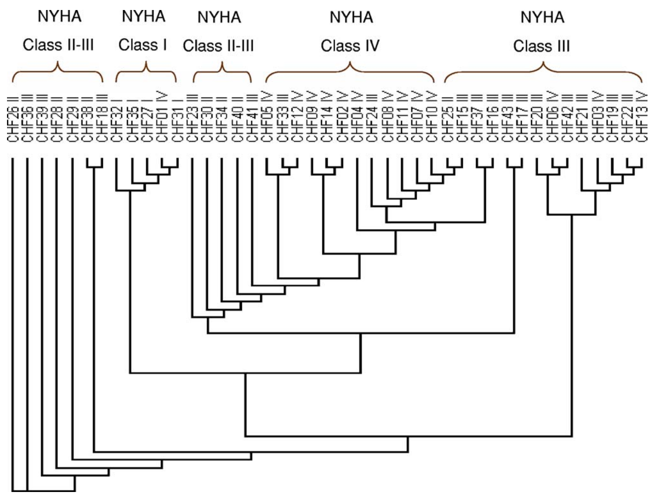

FIG. 2. A rooted phylogenetic tree generated according to the distances between the different subjects with congestive heart failure (CHF). To accommodate all 40 subjects on this graph, the scale of the tree is not proportional to the actual distance. Each subject is identified by its subject number and the New York Heart Association (NYHA) classification. Class I–IV are patients with increasing degree of symptoms.

A. Cardiac interbeat interval time series

Human heartbeat time series are not symbolic sequences; therefore, it is necessary to map the continuous variable of interbeat intervals to a set of symbols. Let us consider a binary mapping rule,[1,5,6] i.e., to map each pair of successive interbeat intervals to the symbols 0 and 1, corresponding to a decrease or increase in the interbeat interval, respectively. This mapping rule has the advantage of being simple enough for practical purposes, and at the same time retaining important dynamical characteristics of the original time series reflecting, in part, the complex nonlinear interactions of competing neuroautonomic control forces relating to sympathetic and parasympathetic stimulations.

Next, we map $m+1$ successive intervals to a binary sequence of length m, i.e., the m-bit word discussed in the previous section. Then, we can calculate the dissimilarity index defined in Eq. (1) between different heart rate time series. After all the pairwise distances (as measured by the dissimilarity indices) are obtained for all data sets, we can use some standard techniques of categorization to present our results. Here, we use the phylogenetic tree algorithm as an example. The method for constructing phylogenetic trees[13–15] 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.

By applying the new weighting function defined in Eq. (2), we reanalyzed the database reported in our previous study.[6] The database includes 40 healthy subjects with subgroups of young (10 females and 10 males, average 25.9 years) and elderly (10 females and 10 males, average 74.5 years), a group of 43 subjects 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). Compared to previous studies, the new result has less overlap between groups. Furthermore, we noticed that this new analysis can better distinguish classes corresponding to the severity of the subjects in the CHF group. In Fig. 2 we show the result of a rooted tree for the case of $m=8$ for heart rate time series of CHF subjects. Different classes of CHF, as defined by the New York Heart Association (NYHA) classification, can be roughly separated on this tree.

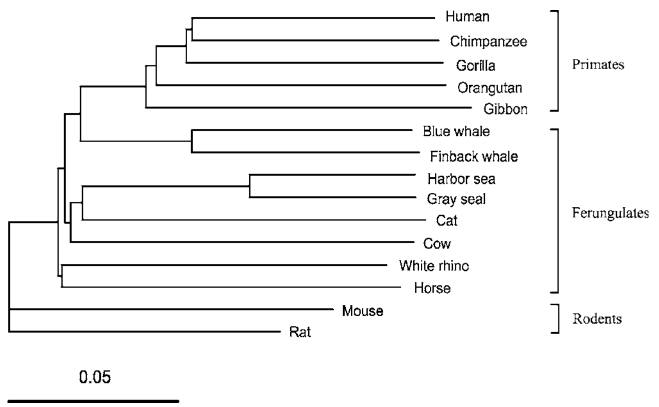

FIG. 3. Phylogenetic tree of 15 mammalian mitochodiral DNA (mtDNA) sequences consisting of three main groups of placental mammals: rodents, ferungulates, and primates. The scale of dissimilarity=0.05 is plotted for reference.

B. DNA sequence comparisons

To demonstrate that our dissimilarity measurement, as defined in Eqs. (1) and (2), can be easily applied to very different types of biological signals, we present results for DNA sequence comparison analysis.

Figure 3 shows a phylogenetic tree of 15 mammalian mitochondrial DNA (mtDNA) sequences consisting of three main groups of placental mammals: rodents, ferungulates, and primates. Each mtDNA genome was analyzed by the method described above using 5-tuples (segments of 5 nucleotides) as "words." The tree shows an evolutionary trend from rat to human and is comparable to the consensus of clustering of these three lineages using conventional methods.[16]

IV. DISCUSSION

First, we want to emphasize that our dissimilarity index does not fulfill rigorous mathematical criteria of a distance measure.[9] A mathematical distance measure has to meet two criteria: (1) $d(A,B)=0$ if, and only if $A \equiv B$; and (2) to obey the triangular inequality

$$d(A,B)+d(B,C) \geqslant d(A,C), \ \forall A,B,C.$$

Initially, we considered two different approaches to developing a similarity index based on the frequency distribution of repetitive patterns. The first approach was to use a mathematically rigorous distance definition. For example, each symbolic sequence can be presented as an n-dimensional vector of unit length such that its projection on each axis is proportional to the frequency of the corresponding pattern. Thus, the distance among symbolic sequences can be defined as the Euclidean distance between the pattern frequency vectors. However, this type of distance weights the impact of each repetitive pattern equally, and we found that it cannot correctly classify symbolic sequences mapped from heart rate time series discussed in the previous section.

An alternative approach, presented in this paper, is to define an intuitively meaningful distance that considers the impact of each repetitive pattern differently based on its frequency and rank order statistics. This approach, while not always meeting the rigorous mathematical criteria of a distance measurement, only violates them under rare situations in real-world applications. To use this type of empirically defined distance we need to examine the extent to which our similarity index will be "well behaved." To this end, we checked all combinations of triangular inequalities for the databases we used to understand when and why our measurement will violate those mathematical criteria in practical use. Overall, we found that this empirical definition behaves well in real-world databases. For example, when applied to symbolic sequences mapped from heart rate signals, the violations of triangular inequality almost never occurred (about 0.8%) for pairwise distances among individual records using 8-tuple repetitive patterns. Generally, we found these violations occurred in the cases where at least one of the heart rate time series involved in the triangular relationship had qualitatively different dynamics than the others (i.e., sinus rhythm in young subjects versus atrial fibrillation).

We have also used this algorithm to successfully investigate literary authorship problems from different languages and different genres.[10] We found that this algorithm did not violate triangular inequality in distances among literary texts of the same language.

When applying this dissimilarity index to heart rate databases, violations of triangular inequality only appeared when comparing extremely different types of heart rate signals. This type of situation is analogous to measuring the distance between computer binary codings of unrelated languages (e.g., English versus Chinese texts). Under this type of comparison, our similarity index is no longer a meaningful distance measure, thus resulting in some violations of the triangular inequality. However, since these problematic cases typically occur when the distance is very large, the objects will be placed on very different branches of the phylogenetic tree. Therefore, these types of violations do not alter the topology of the phylogenetic tree. We also note that the phylogenetic tree algorithm (such as the neighbor-joining method) is an optimization procedure. Therefore, the solution of the tree is more likely to be a local minimum rather than a true solution (global minimum). In this case, the error generated by the phylogenetic tree algorithm is likely to be greater than the uncertainty introduced by our similarity index.

The key idea of the analysis presented in this paper is the connection between dynamical patterns of the output signal and the underlying dynamical (micro)states. However, to find the optimal selection of dynamical patterns that represent the microstates without *a priori* knowledge of the dynamical system is a difficult task. Our current method assumes that all microstates are represented by patterns of equal length. This simplistic approach has been successful, but further refinements will be necessary for future applications. For example, one can try the implementation of selecting dynamical patterns with different lengths as building blocks of the whole sequence. To this end, it is necessary to adapt techniques developed in data compression methods,[17,18] where identifying repetitive patterns of various length is of fundamental importance.

A further challenge is to take into account that the transitions between microstates are not independent. The analytical derivation can be quite complicated if we consider the correlation between microstates. One possible implementation will be to consider consecutive multiple microstate transitions as new states, and to study the dynamics of these new states accordingly. Another approach is to incorporate transition probabilities into our theoretical consideration.

In summary, we derived an empirical measurement of dissimilarity between two symbolic sequences. This derivation is based on generic statistical physics assumptions and, therefore, can be applied to a wide range of problems. This information dissimilarity index, defined by our formula, is closely related to the Shannon entropy and rank order of the repetitive patterns in the symbolic sequences. With this simple measure of dissimilarity, we can categorize different types of symbolic sequences by using standard clustering algorithms. This classification of symbolic sequences may provide very useful information about the underlying dynamical processes that generate these sequences.

ACKNOWLEDGMENTS

We thank M. Costa and J. E. Mietus for valuable discussions. We gratefully acknowledge support from the NIH/NCRR (Grant No. P41-RR13622), NIH/NIA OAIC (Grant No. P60-AG08814), NIH/NICHD (Grant No. R01-HD39838), NIH/NINDS (Grant No. R01-NS45745), the Defense Advanced Research Projects Agency (Grant No. HR0011-05-1-0057), the Ellison Medical Foundation, and the G. Harold and Leila Y. Mathers Charitable Foundation.

[1] J. Kurths, A. Voss, P. Saparin, A. Witt, H. J. Kleiner, and N. Wessel, Chaos **5**, 88 (1995).

[2] A. Voss, J. Kurths, H. J. Kleiner, A. Witt, N. Wessel, P. Saparin, K. J. Osterziel, R. Schurath, and R. Dietz, Cardiovasc. Res. **31**, 419–433 (1996).

[3] D. Cysarz, H. Bettermann, and P. van Leeuwen, Am. J. Physiol. **278**, H2163–2172 (2000).

[4] N. Wessel, C. Ziehmann, J. Kurths, U. Meyerfeldt, A. Schirdewan, and A. Voss, Phys. Rev. E **61**, 733–739 (2000).

[5] Y. Ashkenazy, P. C. Ivanov, S. Havlin, C.-K. Peng, A. L. Goldberger, and H. E. Stanley, Phys. Rev. Lett. **86**, 1900 (2001).

[6] A. C. C. Yang, S. S. Hseu, H. W. Yien, A. L. Goldberger, and C.-K. Peng, Phys. Rev. Lett. **90**, 108103 (2003).

[7] S. Guzzetti, E. Borroni, P. E. Garbelli, E. Ceriani, P. Della Bella, N. Montano, C. Cogliati, V. K. Somers, A. Malliani, and A. Porta, Circulation **112**, 465–470 (2005).

[8] S. Havlin, Physica A **216**, 148–150 (1995).

[9] A. C. C. Yang, S. S. Hseu, H. W. Yien, A. L. Goldberger, and C.-K. Peng, Phys. Rev. Lett. **92**, 109802 (2004).

[10] A. C. C. Yang, C.-K. Peng, H. W. Yien, and A. L. Goldberger, Physica A **329**, 473–483 (2003).

[11] A. C. C. Yang, A. L. Goldberger, and C.-K. Peng, J. Comput. Biol. **12**, 1103–1116 (2005).

[12] C. E. Shannon, Bell Syst. Tech. J. **27**, 379–423 (1948).

[13] J. Felsenstein, Phylogeny Inference Package PHYLIP 3.5c (Department of Genetics, University of Washington, Seattle, 1993).

[14] N. Saitou and M. Nei, Mol. Biol. Evol. **4**, 406–425 (1987).

[15] R. D. M. Page and E. C. Holmes, *Molecular Evolution: A Phylogenetic Approach* (Blackwell Science, Cambridge, 1998).

[16] Y. Cao, N. Okada, and M. Hasegawa, Mol. Biol. Evol. **14**, 461–464 (1997).

[17] J. Ziv and A. Lempel, IEEE Trans. Inf. Theory **23**, 337–343 (1977).

[18] *Data Compression and Error Control Techniques with Applications*, edited by V. Cappellini (Academic, London, 1985).