Advances in Adaptive Data Analysis
Vol. 1, No. 1 (2009) 61–70
© World Scientific Publishing Company

ADAPTIVE DATA ANALYSIS OF COMPLEX FLUCTUATIONS IN PHYSIOLOGIC TIME SERIES

C.-K. PENG*, MADALENA COSTA† and ARY L. GOLDBERGER‡

Margret & H.A. Rey Institute of Nonlinear Dynamics in Physiology and Medicine
Division of Interdisciplinary Medicine and Biotechnology
Beth Israel Deaconess Medical Center
Harvard Medical School
330 Brookline Ave., Boston, MA 02215, USA
**cpeng@bidmc.harvard.edu*
†mcosta3@bidmc.harvard.edu
‡agoldber@bidmc.harvard.edu

We introduce a generic framework of dynamical complexity to understand and quantify fluctuations of physiologic time series. In particular, we discuss the importance of applying adaptive data analysis techniques, such as the empirical mode decomposition algorithm, to address the challenges of nonlinearity and nonstationarity that are typically exhibited in biological fluctuations.

Keywords: Time series; complexity; entropy.

1. Introduction

One of the great challenges of contemporary biomedical science is to understand more fully the dynamics of living systems in health and disease. The importance of this challenge is highlighted by headlines announcing unexpected, life-threatening side effects of once-promising drugs, as well as the serendipitous discoveries deriving from "outside the box" approaches to major public health problems, for example, in heart disease and cancer biology. The basis of such unexpected findings, both negative and positive, is the extraordinary complexity of physiologic systems, which exceeds that of the most challenging systems in the physical world. These systems defy understanding based on traditional mechanistic models and conventional biostatistical analyses.

The overall aim of this paper is to develop a deeper understanding of the dynamics underlying *healthy* biological systems and what occurs when these systems lose their robustness due to aging or disease. We will address these fundamental questions from data analysis perspective. Specifically, why novel adaptive data analysis techniques essential to understand these important issues are. However, because of the nonlinear complexity of these biological systems, it is unrealistic to achieve this

goal purely by a traditional engineering (reductionist) approach in which one disassembles the system into its constituent pieces, studies each component in detail, and finally puts them back together, recreating the original entity. Even in rare cases where this type of reductionist program can be accomplished, the integrative system's behavior typically surprises expectations based solely on the information gathered through analyzing each component in isolation. In everyday parlance, this well-known effect is referred to as *the whole being different than the sum of the parts*. In the language of complex systems, it is known by the term "emergent properties." In nonlinear systems, the composite or group behavior (of molecules, cells, organs, individuals, and even societies) cannot be fully understood by simply "adding up" the components. Instead, one needs rigorous, new approaches to model, measure and analyze a system's integrative behavior.

2. Complex System Approaches

Central to this enterprise are computational tools and models that usefully represent the behavior of the intact system. These system-level measurements and models also need to capture certain generic and robust properties of complex biological systems, such that they have a wide range of applications across many disciplines. To this end, we have focused on studying the output signals generated by complex biological systems. The dynamical fluctuations of these signals in health and disease provide a unique window into the free-running behavior of the integrative systems.

To identify system-level behaviors that are critical to our understanding of healthy dynamics and of pathological disturbances, we pursued investigations under the framework of three complementary hypotheses:

1. The complexity of a biological system reflects its ability to adapt and function in an ever-changing environment.
2. Biological systems need to operate across multiple scales of space and time, and hence their complexity is also multiscale and hierarchical.
3. A wide class of disease states, as well as aging, appear to degrade this biological complexity and reduce the adaptive capacity of the system. Thus, *loss of complexity* may be a generic, defining feature of pathologic dynamics, and the basis of new diagnostic, prognostic, and therapeutic approaches.

To investigate the above hypotheses by studying the dynamical fluctuations of output signals generated by complex biological systems. We developed some innovative approaches in recent years. These system approaches and their associated computational tools promise to provide insights into a wide range of biomedical problems. Examples include forecasting catastrophic events such as epileptic seizures and sudden cardiac arrest, studying gene evolution, searching and categorizing large biomedical and other types of databases, and screening for drug toxicity and efficacy, to name but a few. These diverse applications are strong indications of the potential of these new approaches to advance the science of complex systems.

3. The Origin of Physiologic Variability

Dynamical fluctuations in the output of complex biological systems with multiple interacting components often exhibit remarkably complicated patterns. Such fluctuations have long been ignored by conventional analyses. Indeed, the presence of these fluctuations is often assumed to simply reflect the fact that biological systems are being constantly perturbed by external and intrinsic noise. However, recent findings by our group and others clearly indicate that these complex fluctuations exhibit interesting structures that were not previously anticipated.[1–6] More importantly, these fluctuations may also contain useful information about the emerging complexity of the systems.[7–13] Here we develop a dynamical system perspective to understand the origin of these fluctuations.

3.1. *State space representation*

In dynamical systems research, it is common to describe a system by a set of variables. If defined properly, these so-called *state variables* can uniquely determine the *state* of the system and the time course of its revolution (see Fig. 1).

Assuming that how a system changes in time is purely deterministic, then the goal of the state space approach is to find *equations of motion* for the underlying dynamics in order to understand, predict, and control the system.

However, for biological systems, this approach is not feasible due to two intrinsic difficulties. First, the state space is of very high dimensionality, and not all variables can be measured. For example, to fully describe the state of human physiology, one might need to monitor hundreds of variables (including heart rate, blood pressure, body position, muscle tone, oxygen and multiple hormones level in the blood, etc). Although macroscopic variables can be used as state variables to reduce the dimensionality of the state space, it is unclear what the proper macroscopic variables are in this case. Furthermore, biological systems are not purely deterministic, many stochastic factors constantly influence them. Although these two considerations significantly limit the application of tools developed in dynamical system analysis to biological systems, the state space representation is still a useful picture conceptually.

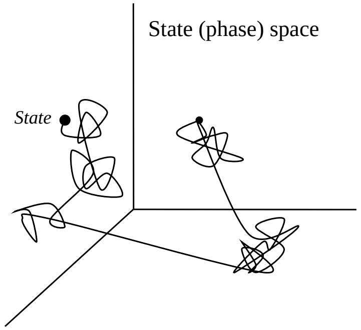

Fig. 1. A schematic illustration of 3-D state space. In this example, a system is fully described by 3 state variables. At any given moment, the system is represented as a point (state) in this space. The trajectory of the system traces out the time evolution of changes of the system's state.

3.2. *System complexity as a measure of adaptability*

As we discussed previously, a meaningful quantification of the complexity of a biological system should be related to the system's capacity to adapt and function in an ever changing environment. The system that can adapt to the most external challenges (stresses) will have the best advantage for survival. Therefore, we propose that biological systems have been evolving to increase their dynamical capacity (complexity). As a result, biological systems we observed today are highly complex since they are the products of a very long evolutionary process. We also hypothesized that aging and disease will degrade a systems complexity, since they represent a less adapted system.

Using the state space concept, an external perturbation (challenge or stress) to a biological system requires the system to move from one location to a different area of the state space in order to adapt to the perturbation. A healthy system should be able to easily move from one area to another, while a diseased system has a very limited ability to adapt, and thus cannot move to other regions of the state space.

Complexity is a measure of a system's capacity to adapt, therefore, it should be related to the total available volume of the state space. Theoretically, we can measure the size of the available state space by either observing the system's trajectory for a very long time (asymptotically, the underlying dynamical system will visit all available state space), or by perturbing the system with all possible stresses and calculate the volume of the state space being covered. However, both implementations are not realistically feasible. Therefore, we proposed an alternative way to derive the desirable information as will be discussed in the following sections.

3.3. *Analogy of Brownian motion*

In 1905, Einstein published several important papers that took physics into a completely new world. In addition to his famous papers on special relativity and photoelectric effect, his paper on Brownian motion also had a great impact. In that paper, he concluded that the same random forces which cause the erratic Brownian motion of a particle suspended in fluid would also cause drag (viscosity) if the particles were pulled through the fluid. In other words, by measuring the spontaneous fluctuation of the particle at rest, one can know how much dissipative frictional force one must do work against, if one tries to perturb the system in a particular direction.

This derivation between spontaneous fluctuations without external perturbation and the system's response to perturbation is of fundamental importance. It is later generalized as the fluctuation dissipation theorem.[14] It motivated the investigation of fluctuating phenomena in statistical physics of the 20th century.

We hypothesized that the same principle can be applied to the state space representation. If our assumption is true, then we can simply measure the spontaneous fluctuations of a system in the state space when it is under free-running condition, and use that information to predict the ability that a system can adapt when encounters a challenge. Similar to Einstein's finding for Brownian particle, the greater the spontaneous fluctuation, the easier for it to move (lower viscosity) in that space when external perturbation is applied.

This assumption dramatically simplifies our task of defining a system's complexity. Next, we will discuss how to construct a surrogate state space when there is only limited information on state variables.

3.4. *Surrogate state space*

In the past several years, we have successfully developed an innovative algorithm to probe the state space indirectly. The goal was to overcome the barrier that in real-world condition, one can only monitor a very limited set of physiologic signals (as state variables). Effectively, we are observing a low-dimensional projection of a trajectory embedded in the much higher dimension of state space. Therefore, it is critical to extract as much information as possible from any single physiologic variable to gain some insight into the high dimensional state space.

For deterministic dynamical systems, there are rigorous approaches, such as the Poincaré map, to study a high dimensional trajectory in a low dimensional subspace. Similarly, in chaos theory, recurrence plots[15] and phase-space portraits[16] are frequently used techniques for this purpose. However, physiologic systems do not meet the criterion (e.g., deterministic and periodic) for applying these analyses. Off-the-shelf usage of those tools to biological time series may lead to misleading conclusions.

Our approach was to take advantage of the fact that an integrative physiologic system will have complex coupling between different components of the system. In biological systems, these couplings often exhibit different spatial and temporal scales. Therefore, by investigating any given signal at various time scales, we can probe the other dimensions of the abstract state space.

By combining these concepts discussed in this section, we have implemented some useful computational algorithms to quantify features related to complexity of biological systems from fluctuating time series of physiologic variables. Our definition of a system's complexity also ensures that our index closely reflect the general health status of the system. In the next section, we will briefly discuss the algorithms we have developed.

4. Quantifying a System's Complexity

For practical purposes, it is useful to quantify the degree of complexity of a biological system by examining its dynamical fluctuations. Such metrics have potentially important applications both with respect to evaluating dynamical models of biological systems and to clinical monitoring. Substantial attention, therefore, has been focused on defining a quantitative measurement of complexity.[9–13, 17–21] However, no consensus has been reached on this issue. We have used an alternative view, as discussed in previous sections, to look at these biological variabilities to derive some useful measurements of how complex a system is.

Over the past several years, our group have developed quantitative algorithms to probe some of the generic features of complex systems and applied these computational tools to the understanding of the underlying system dynamics. For example, we have introduced *fractal scaling*,[22, 23] *multiscale entropy* (MSE)[24, 25] and *time irreversibility*[26] analysis techniques and applied them to the study of the cardiac dynamics of healthy subjects and patients with different types of pathologies. The former technique quantifies the information content of a signal across multiple time scales and the latter quantifies the degree of temporal irreversibility over multiple time scales. Time irreversibility is a property related to the unidirectionality of the energy flow across the boundaries of a living system, which utilizes free energy to evolve to more hierarchically ordered structural configurations and less entropic states in comparison with the surrounding environment.

Based initially on the analysis of the cardiac rhythm[24, 25] (under neuroautonomic control) and gait dynamics,[27] we have shown that healthy systems, those with the highest capacity to adjust to continuous (and often unpredictable) changes of internal and external variables, generate the most physiologically complex and the most time irreversible signals. We have shown further that both multiscale variability and time irreversibility properties degrade with aging and disease. These results challenge traditional mechanisms of physiologic control based on classical homeostasis (single steady state dynamics) and are of interest from a number of other perspectives, including basic modeling of regulatory systems and practical bedside applications.

5. Technical Challenges and Adaptive Signal Analysis

In this section, we will briefly discuss the importance of applying adaptive signal analysis techniques, in conjunction with the complexity related methods described above, to obtain more accurate quantitative measurements of complex biological systems.

5.1. *Problem of nonstationarity*

The quantitative tools we have developed, such as the multiscale entropy (MSE) analysis, for the analysis of complex physiologic time series are based on generic concepts that are fundamental to biological systems. As a result, these tools are readily applicable to many different biomedical problems. However, since physiologic time series are typically nonstationary, there are important technical issues that need to be addressed in order to obtain reliable results.

For example, the MSE analysis was derived from stationary processes. In practice, time series need not to be strictly stationary according to the mathematical definition to yield meaningful results. However, nonstationarities appearing on scales larger than those considered for MSE analysis may substantially affect our measurements. Such nonstationarities need to be taken care of prior to performing the MSE analysis. Our study of postural sway time series[28] indicates that by properly detrending the time series on scales greater than those being measured by the MSE, the analysis provides robust and consistent results. The empirical mode decomposition (EMD) technique[29] is a very adequate candidate for pre-processing the data, since it provides a systematic way to detrend the data without *a priori* assumptions of what type of trend the data may possess.[30]

5.2. *Nonlinear dynamical coupling among components of system*

A fundamental question about complex biological systems is how does the observed complex dynamics, as quantified by our complexity related measurements, emerge from integrated system functions. Understanding possible mechanisms of healthy complexity is important both on the basic scientific level and on the practical level, where clinical interventions can be proposed to maintain or restore this dynamical complexity. By observing the degradation of dynamical complexity in disease and aging, one realizes that life-threatening pathologic conditions are typically accompanied by either complete de-coupling between sub-components of the whole system, or a strong "mode-locking" among them. In contrast, a healthy biological system usually displays *intermittent* coupling between its sub-systems. Each component of the system may engage and then dis-engage with other components of the system. This type of on-and-off "cross-talk" between different parts of a complex system (reminiscent of how different instruments are integrated together in a symphony orchestra) seems to be a prominent characteristic of healthy biological function. As a result, quantifying the coupling among different sub-system components is critical to our understanding of the complex system as a whole. From a data analysis point of view, one should be able to characterize the coupling between the two components of a system by simultaneously collecting the signals that represent those components. However, technically, quantifying the coupling is not an easy task. The main difficulties are due to the fact that both signals are often nonstationary, and the coupling between them is usually nonlinear and intermittent. To quantify the intermittency, the analysis method has to separate any local variation and collate the different scales of the intermittent processes separately and cleanly in both temporal and scale variables. Here the recently developed Ensemble EMD[31] has the potential to offer great help.

Therefore, it is essential to apply adaptive data analysis techniques to address the nonlinear and nonstationary challenges as demonstrated by recent works of our group and others.[32–34] For example, we have applied the EMD algorithm to study the role of coupling between blood pressure and cerebral flood flow in cerebral autoregulation. Cerebral autoregulation is a mechanism that involves dilatation and constriction of arterioles to maintain relatively stable cerebral blood flow in response to changes of systemic blood pressure. Traditional assessments of cerebral autoregulation use Fourier-based techniques, such as transfer function analysis, that fail to yield robust and consistent results in typical clinical settings. The EMD method substantially improves our ability to accurately quantify the dynamical interactions between blood pressure and cerebral blood flow.[32–34] Furthermore, since the EMD can provide phase and frequency information on instantaneous basis, analysis of its dynamical feature (i.e., how do these interaction change over time) becomes feasible. Future work along this direction may have clinical importance and also provide mechanistic understanding toward the theory of dynamical complexity we proposed.

6. Discussion

We have developed a generic framework for extracting "hidden information" in time series generated by complex biological systems. Specifically, we discussed the underlying assumptions that make it possible to probe the behavior on the system level via examining the dynamical fluctuations of a single variable. We also proposed meaningful measurements of complexity for biological systems that are based on the framework we developed. We have used those complexity measures to study the outputs of cardiac heartbeat regulatory system,[25] gait dynamics,[27] and postural control.[28] Briefly, we found that, under free-running conditions, the dynamics of healthy systems are the most complex, as measured by the multiscale entropy and time irreversibility methods, and that complexity breaks down with aging and disease. We also studied the effects of a noise-based therapeutic intervention designed to improve postural balance[28] on the overall complexity of the postural sway dynamics. We found that there is an increase in multiscale complexity during the application of this intervention. This finding supports the notion of using *dynamical biomarkers* for assessing noise-based and other types of therapeutic interventions. However, one needs to be aware of potential technical issues when applying these new measures to physiologic time series. In this paper, we discussed how to utilize the EMD technique to overcome the problems when data are not "well-behaved." Thus the EMD approach constitutes an essential step of complex physiologic signal analysis.

Acknowledgments

We thank Tim Buchman, Norden Huang, Yanhui Liu, Men-Tzung Lo, Ben Mann, Joe Mietus, and Albert Yang for their valuable discussions. We gratefully acknowledge the support from the NIH/NIBIB and NIGMS (U01-EB008577), the NIH/NIA OAIC (P60-AG08814), the NIH/NICHD (R01-HD39838), the NIH/NINDS (R01-NS45745), the Defense Advanced Research Projects Agency (DARPA grant HR0011-05-1-0057), the Ellison Medical Foundation, the James S. McDonnell Foundation, and the G. Harold and Leila Y. Mathers Charitable Foundation.

References

1. T. G. Buchman, The community of the self, *Nature* **420** (2002) 246–251.
2. B. Suki, A. M. Alencar, M. K. Sujeer, K. R. Lutchen, J. J. Collins, J. S. Andrade Jr., E. P. Ingenito, S. Zapperi and H. E. Stanley, Life-support system benefits from noise, *Nature* **393** (1998) 127–128.
3. C.-K. Peng, J. Mietus, J. M. Hausdorff, S. Havlin, H. E. Stanley and A. L. Goldberger, Long-range anti-correlations and non-Gaussian behavior of the heartbeat, *Phys. Rev. Lett.* **70** (1993) 1343–1346.
4. J. M. Hausdorff, S. L. Mitchell, R. Firtion, C.-K. Peng, M. E. Cudkowicz, J. Y. Wei and A. L. Goldberger, Altered fractal dynamics of gait: Reduced stride interval correlations with aging and Huntington's disease, *J. Appl. Physiol.* **82** (1997) 262–269.
5. C.-K. Peng, J. E. Mietus, Y. Liu, C. Lee, J. M. Hausdorff, H. E. Stanley, A. L. Goldberger and L. A. Lipsitz, Quantifying fractal organization of respiratory dynamics: Age and gender effects, *Ann. Biomed. Eng.* **30** (2002) 683–692.
6. N. Iyengar, C.-K. Peng, R. Morin, A. L. Goldberger and L. A. Lipsitz, Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics, *Am. J. Physiol.* **271** (1996) 1078–1084.
7. A. L. Goldberger, C.-K. Peng and L. A. Lipsitz, What is physiologic complexity and how does it change with aging and disease? *Neurobiol. Aging* **23** (2002) 23–26.
8. A. L. Goldberger, L. A. N. Amaral, J. M. Hausdorff, P. Ch. Ivanov, C.-K. Peng and H. E. Stanley, Fractal dynamics in physiology: Alterations with disease and aging, *Proc. Natl. Acad. Sci. (USA)* **99**(Suppl 1) (2002) 2466–2472.
9. S. Pincus, Approximate entropy as a measure of system complexity, *Proc. Natl. Acad. Sci. USA* **88** (1991) 2297–2301.
10. Y. Bar-Yam, *Dynamics of Complex Systems* (Addison-Wesley, 1992).
11. S. Pincus and B. Singer, Randomness and degree of irregularity, *Proc. Natl. Acad. Sci. USA* **93** (1996) 2083–2088.
12. J. S. Richman and J. R. Moorman, Physiological time-series analysis using approximate entropy and sample entropy, *Am. J. Physiol. Heart Circ. Physiol.* **278** (2000) H2039–H2049.
13. S. Pincus, Assessing serial irregularity and its implications for health, *Ann. NY Acad. Sci.* **954** (2001) 245–267.
14. H. Nyquist, Thermal agitation of electric charge in conductors, *Phys. Rev.* **32** (1928) 110–113.
15. J. P. Eckmann, S. O. Kamphorst and D. Ruelle, Recurrence plots of dynamical systems, *Europhys. Lett.* **5** (1987) 973–977.
16. F. Takens, Detecting strange attractors in turbulence, in *Dynamical Systems and Turbulence*, in Lecture Notes in Mathematics, Vol. 898 (Berlin, 1981), pp. 366–381.
17. T. Schurmann and P. Grassberger, Entropy estimation of symbol sequences, *Chaos* **6** (1996) 414–427.

18. N. Marwan, N. Wessel, U. Meyerfeldt, A. Schirdewan and J. Kurths, Recurrence-plot-based measures of complexity and their application to heart-rate-variability data, *Phys. Rev. E* **66** (2002) 026702.
19. M. A. Jimenez-Montano, W. Ebeling, T. Pohl and P. E. Rapp, Entropy and complexity of finite sequences as fluctuating quantities, *Biosystems* **64** (2002) 23–32.
20. C. Bandt and B. Pompe, Permutation entropy: A natural complexity measure for time series, *Phys. Rev. Lett.* **88** (2002) 174102.
21. C. Adami, What is complexity? *Bioessays* **24** (2002) 1085–1094.
22. C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley and A. L. Goldberger, Mosaic organization of DNA nucleotides, *Phys. Rev. E* **49** (1994) 1685–1689.
23. 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) 82–87.
24. M. Costa, A. L. Goldberger and C.-K. Peng, Multiscale entropy analysis of complex physiologic time series, *Phys. Rev. Lett.* **89** (2002) 068102.
25. M. Costa, A. L. Goldberger and C.-K. Peng, Multiscale entropy analysis of biological signals, *Phys. Rev. E* **71** (2005) 021906.
26. M. Costa, A. L. Goldberger and C.-K. Peng, Broken asymmetry of the human heartbeat: Loss of time irreversibility in aging and disease, *Phys. Rev. Lett.* **95** (2005) 198102.
27. M. Costa, C.-K. Peng, A. L. Goldberger and J. M. Hausdorff, Multiscale entropy analysis of human gait dynamics, *Physica A* **330** (2003) 53–60.
28. M. Costa, A. A. Priplata, L. A. Lipsitz, Z. Wu, N. E. Huang, A. L. Goldberger and C.-K. Peng, Noise and poise: Enhancement of postural complexity in the elderly with a stochastic resonance-based therapy, *Europhys. Lett.* **77** (2007) 68008.
29. N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung and H. H. Liu, The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, *Proc. R. Soc. Lond. Ser. A* **454** (1998) 903–995.
30. Z. Wu, N. E. Huang, S. R. Long and C.-K. Peng, On the trend, detrending and the variability of nonlinear and nonstationary time series, *Proc. Natl. Acad. Sci. USA* **104** (2007) 14889–14894.
31. Z. Wu and N. E. Huang, Ensamble Empirical Mode Decomposition: A noise assisted data analysis method, *Adv. Adaptive Data Analy.* **1** (2009) 1–41.
32. V. Novak, A. C. C. Yang, L. Lepicovsky, A. L. Goldberger, L. A. Lipsitz and C.-K. Peng, Multimodal pressure-flow method to assess dynamics of cerebral autoregulation in stroke and hypertension, *Biomed. Eng. Online* **3** (2004) 39.
33. K. Hu, C.-K. Peng, N. E. Huang, Z. Wu, L. A. Lipsitz, J. Cavallerano and V. Novak, Altered phase interactions between spontaneous blood pressure and flow fluctuations in type 2 diabetes mellitus: Nonlinear assessment of cerebral autoregulation, *Physica A* **387** (2008) 2279–2292.
34. K. Hu, C.-K. Peng, M. Czosnyka, P. Zhao and V. Novak, Nonlinear assessment of cerebral autoregulation from spontaneous blood pressure and cerebral perfusion pressure fluctuation, *Cardiovasc. Eng.* 2008 (in press).

Multiscale Entropy Analysis of Complex Physiologic Time Series

Madalena Costa,[1,2] Ary L. Goldberger,[1] and C.-K. Peng[1]
[1]Cardiovascular Division, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts 02215
[2]Institute of Biophysics and Biomedical Engineering, Faculty of Science of the University of Lisbon,
Campo Grande, 1749-016 Lisbon, Portugal

(Received 26 March 2002; published 19 July 2002)

There has been considerable interest in quantifying the complexity of physiologic time series, such as heart rate. However, traditional algorithms indicate higher complexity for certain pathologic processes associated with random outputs than for healthy dynamics exhibiting long-range correlations. This paradox may be due to the fact that conventional algorithms fail to account for the multiple time scales inherent in healthy physiologic dynamics. We introduce a method to calculate multiscale entropy (MSE) for complex time series. We find that MSE robustly separates healthy and pathologic groups and consistently yields higher values for simulated long-range correlated noise compared to uncorrelated noise.

DOI: 10.1103/PhysRevLett.89.068102 PACS numbers: 87.19.Hh, 05.40.Ca, 05.45.Tp

Quantifying the "complexity" of physiologic signals in health and disease has been the focus of considerable attention [1–4]. Such metrics have potentially important applications with respect to evaluating both dynamical models of biologic control systems and bedside diagnostics. For example, a wide class of disease states, as well as aging, appear to degrade physiologic information content and reduce the adaptive capacity of the individual. Loss of complexity, therefore, has been proposed as a generic feature of pathologic dynamics [1,3].

Traditional entropy-based algorithms quantify the regularity (orderliness) of a time series. Entropy increases with the degree of disorder and is maximum for completely random systems. However, an increase in the entropy may not always be associated with an increase in dynamical complexity. For instance, a randomized time series has higher entropy than the original time series, although the process of generating surrogate data destroys correlations and degrades the information content of the original signal.

Diseased systems, when associated with the emergence of more regular behavior, show reduced entropy values compared to the dynamics of free-running healthy systems [3]. However, certain pathologies, including cardiac arrhythmias like atrial fibrillation, are associated with highly erratic fluctuations with statistical properties resembling uncorrelated noise [5–7]. Traditional algorithms will yield an increase in entropy values for such noisy, pathologic signals when compared to healthy dynamics showing correlated ($1/f$-type) properties, even though the latter represent more physiologically complex, adaptive states. This inconsistency may be related to the fact that widely used entropy measures are based on single-scale analysis and do not take into account the complex temporal fluctuations inherent in healthy physiologic control systems.

The entropy $H(X)$ of a single discrete random variable X is a measure of its average uncertainty. Entropy is calculated by the equation:

$$H(X) = -\sum_{x_i \in \Theta} p(x_i) \log p(x_i). \tag{1}$$

where X represents a random variable with set of values Θ and probability mass function $p(x_i)$.

For a time series representing the output of a stochastic process, that is, an indexed sequence of n random variables, $\{X_i\} = \{X_1, \ldots, X_n\}$, with set of values $\Theta_1, \ldots, \Theta_n$, respectively, the joint entropy is defined as

$$H_n = -\sum_{x_1 \in \Theta_1} \cdots \sum_{x_n \in \Theta_n} p(x_1, \ldots, x_n) \log p(x_1, \ldots, x_n), \tag{2}$$

where $p(x_1, \ldots, x_n)$ is the joint probability for the n variables $X_1, \ldots, X_n$.

The state of a system at a certain instant, X_n, is partially determined by its history, $X_1, X_2, \ldots, X_{n-1}$. However, each new state carries a certain amount of new information. The mean rate of creation of information, also known as the Kolmogorov-Sinai (KS) entropy, is a useful parameter to characterize the system dynamics [8]. Considering that the phase space of a system with $\mathcal{D}$ degrees of freedom is partitioned into hypercubes of content $\varepsilon^{\mathcal{D}}$ and the state of the system is measured at intervals of time τ, the KS entropy is defined as

$$H_{\mathrm{KS}} = \lim_{\tau \to 0} \lim_{\varepsilon \to 0} \lim_{n \to \infty} (H_{n+1} - H_n). \tag{3}$$

Numerically, only entropies of finite order n can be computed. As soon as n becomes large with respect to the length of a given time series, the entropy H_n is underestimated and decays towards zero. Therefore, the KS entropy for "real-world" time series of finite length cannot usually be estimated with reasonable precision.

For the analysis of such typically short, noisy time series, Pincus [9] introduced the approximate entropy (ApEn) family of parameters, which have been widely used in physiology and medicine [1]. Recently, a modified algorithm, sample entropy (SampEn) [4], has been proposed which has the advantage of being less dependent on the time series length. Such algorithms, however, assign a higher value of entropy to certain pathologic time series that are presumed to represent less complex dynamics than to time series derived from healthy function [3]. One possible reason for obtaining these results may be the fact that these measures are based on a single scale. Both the KS entropy and the related ApEn parameters depend on a function's one step difference (e.g., $H_{n+1} - H_n$) and reflect the uncertainty of the next new point, given the past history of the series. Therefore, such measures do not account for features related to structure on scales other than the shortest one.

Zhang [10,11] proposed a general approach to take into account the multiple time scales in physical systems. His measure, based on a weighted sum of scale dependent entropies, does, in fact, yield higher values for correlated noises compared to uncorrelated ones. However, since it is based on Shannon's definition of entropy, Zhang's method requires a large amount of almost noise-free data, in order to map a signal to a discrete symbolic sequence with sufficient statistical accuracy. Therefore, it presents obvious limitations when applied to typical physiologic signals that vary continuously and have finite length.

Here we introduce a multiscale entropy technique applicable to the analysis of the biologic time series. We study simulated noises as well as human cardiac interbeat interval time series, the latter representing the output of a major physiologic control system.

Given a one-dimensional discrete time series, $\{x_1, \ldots, x_i, \ldots, x_N\}$, we construct consecutive coarse-grained time series, $\{y^{(\tau)}\}$, determined by the scale factor, τ, according to the equation: $y_j^{(\tau)} = 1/\tau \sum_{i=(j-1)\tau+1}^{j\tau} x_i$, $1 \leq j \leq N/\tau$. For scale one, the time series $\{y^{(1)}\}$ is simply the original time series. The length of each coarse-grained time series is equal to the length of the original time series divided by the scale factor, τ. Here we consider time series with 3×10^4 points and coarse-grain them up to scale 20, so that the shortest time series has 1500 points. We then calculate an entropy measure (SampEn) for each coarse-grained time series plotted as a function of the scale factor τ [12]. We call this procedure multiscale entropy (MSE) analysis [13].

We first test the MSE method on simulated white and $1/f$ noises [14]. We find that for scale one, a higher value of entropy is assigned to white noise time series in comparison with $1/f$ time series. However, while the value of entropy for the coarse-grained $1/f$ series remains almost constant for all scales, the value of entropy for the coarse-grained white noise time series monotonically decreases, such that for scales > 5, it becomes smaller than the corresponding values for $1/f$ noise (Fig. 1). This result is consistent with the fact that, unlike white noise, $1/f$ noise contains complex structures across multiple time scales [10,11].

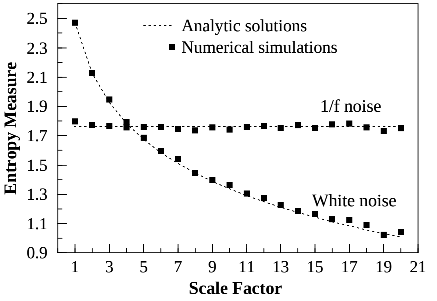

FIG. 1. MSE analysis of Gaussian distributed white noise (mean zero, variance one) and $1/f$ noise. On the y axis, the value of entropy (SampEn) for the coarse-grained time series is plotted. The scale factor specifies the number of data points averaged to obtain each element of the coarse-grained time series. Symbols represent results of simulations for time series of 3×10^4 points [12], and dotted lines indicate analytic results. SampEn for coarse-grained white noise time series, is analytically calculated by the expression $-\ln \int_{-\infty}^{+\infty} \frac{1}{2} \sqrt{(\frac{\tau}{2\pi})} [\mathrm{erf}(\frac{x+r}{\sqrt{(2/\tau)}}) - \mathrm{erf}(\frac{x-r}{\sqrt{(2/\tau)}})] e^{-(1/2)x^2\tau} dx$. τ and erf refer to the scale factor and to the error f nction, respecti el . r is defined in Refs. [4,9,12]. For $1/f$ noise time series, the anal tic al e of SampEn is a constant.

Next, we apply the MSE method to the analysis of selected physiologic time series (Fig. 2). We compare the time series of consecutive heartbeat intervals derived from healthy subjects, patients with severe congestive heart failure [15], and patients with the cardiac arrhythmia, atrial fibrillation. In Fig. 3, we observe three different types of behaviors: (1) The entropy measure for time series derived from healthy subjects increases on small time scales and then stabilizes to a constant value. (2) The entropy measure for time series derived from subjects with congestive heart failure, a life-threatening condition, markedly decreases on small time scales and then gradually increases. (3) The entropy measure for time series derived from subjects with atrial fibrillation monotonically decreases, similar to white noise. Of note, for scale one, atrial fibrillation time series are assigned the highest value of entropy [17], and healthy heartbeat time series are not distinguishable from those of heart failure patients. The largest separation between heart failure patients and healthy subjects is obtained for time scale 5. At the highest scales, the entropy values for the healthy heartbeat fluctuations are significantly higher than those of both pathologic groups.

We also find that the asymptotic value of entropy may not be sufficient to separate time series that represent the output of different dynamical processes. As seen in Fig. 3, for time scale 20, the value of the entropy measure for the

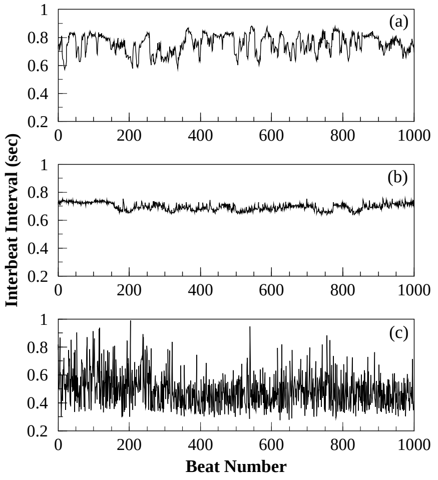

FIG. 2. Representative heartbeat intervals time series from (a) healthy individual (sinus rhythm), (b) subject with congestive heart failure (sinus rhythm), and (c) subject with the cardiac arrhythmia, atrial fibrillation.

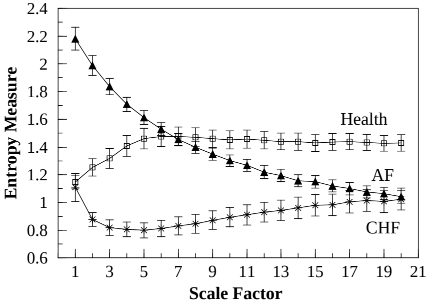

FIG. 3. MSE analysis of interbeat interval time series derived from healthy subjects, subjects with congestive heart failure (CHF), and subjects with atrial fibrillation (AF), as shown in Fig. 2. Values are given as means ± standard error [16]. Time series were filtered to remove outlier points due to artifacts and ventricular ectopic beats. The values of entropy depend on the scale factor. For scale one, AF time series are assigned the highest value of entropy, and the values corresponding to healthy and CHF groups completely overlap. For larger scales, e.g., 20, the entropy value for the coarse-grained time series derived from healthy subjects is significantly higher than those for AF and CHF. At this scale, AF and CHF groups become indistinguishable.

heart failure and atrial fibrillation time series is the same. However, these time series represent the output of a very different type of cardiac dynamics (Fig. 2). Therefore, not only the specific values of the entropy measure but also their dependence on resolution need to be taken into account to better characterize a physiologic process.

We further test the MSE algorithm by comparing the heartbeat time series from 20 healthy elderly subjects, 10 males and 10 females (mean age ±SD, 69 ± 3 yr), and 20 healthy young subjects, 10 males and 10 females (mean age ±SD, 32 ± 6 yr) (Fig. 4). We find that for all time scales, a higher value of entropy is assigned to time series from young subjects, consistent with the hypothesis of loss of complexity with age [3]. Of note, the weakest separation between the two groups occurs for scale one, the only scale studied by traditional entropy metrics. The strongest separation is obtained for time scale 5.

Finally, the MSE algorithm was tested on a set of surrogate data obtained from the heart rate time series of a healthy subject by simple randomization of its data points. The MSE algorithm discriminated the two time series and revealed that the randomized surrogate data was less complex than the original physiologic data. Furthermore, it assigned to the surrogate data set a behavior qualitatively very similar to the one already described for white noise time series.

Our findings are of interest from the following perspectives. The long-standing problem of deriving useful measures of time series complexity is germane to analyzing both the output of physical and biologic systems. In this respect, the MSE method appears to yield a more meaningful approach than conventional entropy measurements. MSE is based on the simple observation that complex physical and biologic systems generally exhibit dynamics that are far from the extrema of perfect regularity and complete randomness. Instead, complex dynamics typically reveal structure on multiple spatial and temporal scales. These multiscale features, ignored by conventional entropy calculations, are explicitly addressed in the MSE algorithm.

The MSE algorithm yields consistent findings when applied to assessing the complexity of both (a) simulated correlated and uncorrelated noises and (b) the integrated output of a major physiologic control system (cardiac interbeat intervals) under healthy and pathologic conditions. In particular, we find, in accord with Zhang [10], that correlated ($1/f$) noise has a higher complexity level than uncorrelated (white) noise when multiple time scales are taken into account (Fig. 1). We also find that pathologic dynamics associated with either increased regularity/decreased variability [Fig. 2(b)] or with increased variability due to loss of correlation properties [Fig. 2(c)] are both characterized by a reduction in complexity. This finding is compatible with the unifying concept that physiologic complexity is fundamentally related to the adaptive capacity of the organism, which requires integrative,

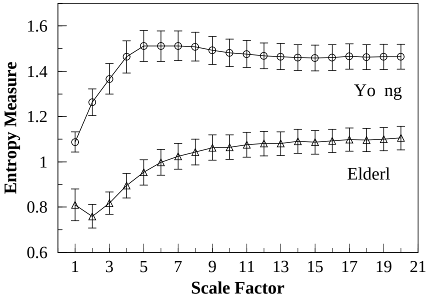

FIG. 4. MSE analysis of the cardiac interbeat time series derived from healthy young subjects and healthy elderly subjects. Values are given as means ± standard error [16]. For all time scales, the values of entropy for coarse-grained time series obtained from elderly subjects are significantly ($p < 0.005$; Student's t-test) lower than those from young subjects. The poorest separation between groups is obtained for scale one, indicating the importance of calculating entropy over different scales.

multiscale functionality. In contrast, disease states (Fig. 3), as well as aging (Fig. 4), may be defined by a sustained breakdown of long-range correlations and loss of information [18]. Finally, we note that the MSE method has potential applications to studying a wide variety of other physiologic and physical time series data.

We thank L. Glass, V. Schulte-Frohlinde, J. Mietus, and I. Henry for valuable discussions and assistance. We gratefully acknowledge support from the National Institutes of Health/National Center for Research Resources (P41-RR13622), NIH/NIA (P60-AG08812), the G. Harold and Leila Y. Mathers Charitable Foundation, the Fetzer Institute, the Centers for Disease Control and Prevention (H75-CCH119124), the Fulbright/FLAD, the Calouste Gulbenkian Foundation, and the Portuguese Foundation for Science and Technology (Praxis XXI/BD/13167).

[1] S. M. Pincus, Ann. N.Y. Acad. Sci. **954**, 245 (2001), and references therein.

[2] A. Porta, S. Guzzetti, N. Montano, R. Furlan, M. Pagani, A. Malliani, and S. Cerutti, IEEE Trans. Biomed. Eng. **48**, 1282 (2001); M. Palus, Physica (Amsterdam) **93D**, 64 (1996); N. Wessel, A. Schumann, A. Schirdewan, A. Voss, and J. Kurths, Lect. Notes Comput. Sci. **1933**, 78 (2000).

[3] A. L. Goldberger, C.-K. Peng, and L. A. Lipsitz, Neurobiol. Aging **23**, 23 (2002).

[4] J. S. Richman and J. R. Moorman, Am. J. Physiol. **278**, H2039 (2000).

[5] J. Hayano, F. Yamasaki, S. Sakata, A. Okada, S. Mukai, and T. Fujinami, Am. J. Physiol. **273**, H2811 (1997).

[6] W. Zeng and L. Glass, Phys. Rev. E **54**, 1779 (1996).

[7] R. Balocchi, C. Carpeggiani, L. Fronzoni, C.-K. Peng, C. Michelassi, J. Mietus, and A. L. Goldberger, in *Methodology and Clinical Applications of Blood Pressure and Heart Rate Analysis*, edited by M. Di Rienzo, G. Mancia, G. Parati, A. Pedotti, and A. Zanchetti (IOS Press, Amsterdam, 1999), pp. 91–96.

[8] For deterministic periodic systems, the KS entropy is zero because any state depends only on the initial conditions. In contrast, this entropy measure is maximum for uncorrelated random processes, since each state is totally independent of the previous ones. J.-P. Eckmann and D. Ruelle, Rev. Mod. Phys. **57**, 617 (1985).

[9] S. M. Pincus, Proc. Natl. Acad. Sci. U.S.A. **88**, 2297 (1991). Let $\{X_i\} = \{x_1, \ldots, x_i, \ldots, x_N\}$ represent a time series of length N. Consider the m-length vectors: $u_m(i) = \{x_i, x_{i+1}, \ldots, x_{i+m-1}\}$ and the following definition for the distance between two vectors: $d[u_m(i), u_m(j)] = \max\{|x(i+k) - x(j+k)| : 0 \leq k \leq m-1\}$. Let $n_{im}(r)$ represent the number of vectors $u_m(j)$ within r of $u_m(i)$. Therefore, $C_i^m(r) = n_{im}(r)/(N-m+1)$ represents the probability that any vector $u_m(j)$ is within r of $u_m(i)$. Define $\Phi^m(r) = 1/(N-m+1)\sum_{i=1}^{N-m+1}\ln C_i^m(r)$. ApEn is a parameter defined as follows: $\mathrm{ApEn}(m,r) = \lim_{N\to\infty}\Phi^m(r) - \Phi^{m+1}(r)$. For finite N, it is estimated by the statistics $\mathrm{ApEn}(m,r,N) = \Phi^m(r) - \Phi^{m+1}(r)$. Lower values of ApEn reflect more regular time series while higher values are associated with less predictable (more complex) time series.

[10] Y.-C. Zhang, J. Phys. I (France) **1**, 971 (1991).

[11] H. C. Fogedby, J. Stat. Phys. **69**, 411 (1992).

[12] SampEn was calculated for all time series with the following parameters: $m = 2$, $r = 0.15 \times \mathrm{SD}$. (SD is the standard deviation of the original time series.) We obtain the same qualitative results using either SampEn or ApEn algorithms.

[13] The term "multiscale entropy" has been employed in a different context in the image processing literature. See, for example, J.-L. Starck, F. Murtagh, and A. Bijaoui, *Image Processing and Data Analysis* (Cambridge University Press, Cambridge, 1998).

[14] The $1/f$ noise is generated as follows: we start with uniformly distributed white noise, calculate the fast Fourier transform (FFT), and after imposing a $1/f$ distribution on the power spectrum, we calculate the inverse FFT.

[15] MIT-BIH Normal Sinus Rhythm Database and BIDMC Congestive Heart Failure Database available at http://www.physionet.org/physiobank/database/#ecg.

[16] The error due to finite size of the data is substantially smaller (about $1/10$) than the intersubject variability.

[17] Time series derived from patients with atrial fibrillation have statistical properties similar to those of white noise on shorter time scales ($\lesssim 200$ s). For more details see [5–7].

[18] A. L. Goldberger, L. A. N. Amaral, J. M. Hausdorff, P. Ch. Ivanov, C.-K. Peng, and H. E. Stanley, Proc. Natl. Acad. Sci. U.S.A. **99** (suppl. 1), 2466 (2002).

Multiscale entropy analysis of biological signals

Madalena Costa,[1,2] Ary L. Goldberger,[1] and C.-K. Peng[1]
[1]*Margret and H. A. Rey Institute for Nonlinear Dynamics in Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts 02215, USA*
[2]*Institute of Biophysics and Biomedical Engineering, Faculty of Sciences of the University of Lisbon, Campo Grande, 1749-016 Lisbon, Portugal*

(Received 1 July 2004; published 18 February 2005)

Traditional approaches to measuring the complexity of biological signals fail to account for the multiple time scales inherent in such time series. These algorithms have yielded contradictory findings when applied to real-world datasets obtained in health and disease states. We describe in detail the basis and implementation of the multiscale entropy (MSE) method. We extend and elaborate previous findings showing its applicability to the fluctuations of the human heartbeat under physiologic and pathologic conditions. The method consistently indicates a loss of complexity with aging, with an erratic cardiac arrhythmia (atrial fibrillation), and with a life-threatening syndrome (congestive heart failure). Further, these different conditions have distinct MSE curve profiles, suggesting diagnostic uses. The results support a general "complexity-loss" theory of aging and disease. We also apply the method to the analysis of coding and noncoding DNA sequences and find that the latter have higher multiscale entropy, consistent with the emerging view that so-called "junk DNA" sequences contain important biological information.

DOI: 10.1103/PhysRevE.71.021906 PACS number(s): 87.19.Hh, 05.40.Ca, 05.45.Tp

I. INTRODUCTION

Physiologic systems are regulated by interacting mechanisms that operate across multiple spatial and temporal scales. The output variables of these systems often exhibit complex fluctuations that are not simply due to contaminative noise but contain information about the underlying dynamics.

Two classical approaches to time series analysis are related to deterministic and stochastic mechanisms. A fundamental underpinning of the former approach is Takens theorem [1,2], which states that it is possible to reach full knowledge of a high dimensional deterministic system by observing a single output variable. However, since experimental time series, even when generated by deterministic mechanisms, are most likely affected by dynamical noise, the purely deterministic approach may be of limited use. Nevertheless, for some practical applications, a low dimensional dynamics may be assumed and then the results tested for internal consistency [3].

The stochastic approach is aimed at quantifying the statistical properties of the output variables and developing tractable models that account for those properties. The diffusion process is a classic example of how a stochastic approach may contribute to the understanding of a dynamical system. At a macroscopic level, the diffusion equation can be derived from Fick's law and the principle of conservation of mass. Alternatively, at a microscopic level it is possible to derive the diffusion equation assuming that each particle can be modeled as a random walker, taking steps of length l in a given direction with probability p. The theory of Brownian motion, which is based on random walk models, together with experimental results, contributed to the understanding of the atomic nature of matter.

Time series generated by biological systems most likely contain deterministic and stochastic components. Therefore, both approaches may provide complementary information about the underlying dynamics. The method we use in this paper for the analysis of physiologic time series does not assume any particular mechanism. Instead, our method is aimed at comparing the degree of complexity of different time series. Such complexity-related metrics [4] have potentially important applications to discriminate time series generated either by different systems or by the same system under different conditions.

Traditional methods quantify the degree of regularity of a time series by evaluating the appearance of repetitive patterns. However, there is no straightforward correspondence between regularity, which can be measured by entropy-based algorithms, and complexity. Intuitively, complexity is associated with "meaningful structural richness" [5], which, in contrast to the outputs of random phenomena, exhibits relatively higher regularity. Entropy-based measures, such as the entropy rate and the Kolmogorov complexity, grow monotonically with the degree of randomness. Therefore, these measures assign the highest values to uncorrelated random signals (white noise), which are highly unpredictable but not structurally "complex," and, at a global level, admit a very simple description.

Thus, when applied to physiologic time series, traditional entropy-based algorithms may lead to misleading results. For example, they assign higher entropy values to certain pathologic cardiac rhythms that generate erratic outputs than to healthy cardiac rhythms that are exquisitely regulated by multiple interacting control mechanisms. Substantial attention, therefore, has been focused on defining a quantitative measurement of complexity that assigns minimum values to both deterministic/predictable and uncorrelated random/unpredictable signals [6]. However, no consensus has been reached on this issue.

Our approach to addressing this long-standing problem has been motivated by three basic hypotheses: (i) the complexity of a biological system reflects its ability to adapt and function in an ever-changing environment; (ii) biological systems need to operate across multiple spatial and temporal scales, and hence their complexity is also multiscaled; and (iii) a wide class of disease states, as well as aging, which reduce the adaptive capacity of the individual, appear to degrade the information carried by output variables. Thus, loss of complexity may be a generic feature of pathologic dynamics. Accordingly, our approach to defining a complexity measurement focuses on quantifying the information expressed by the physiologic dynamics over multiple scales.

Recently, we introduced a new method, termed multiscale entropy (MSE) [7–11]. Due to the interrelationship of entropy and scale, which is incorporated in the MSE analysis, the results are consistent with the consideration that both completely ordered and completely random signals are not really complex. In particular, the MSE method shows that correlated random signals (colored noise) are more complex than uncorrelated random signals (white noise). Compared to traditional complexity measures, MSE has the advantage of being applicable to both physiologic and physical signals of finite length.

In this paper, we apply the MSE method to the study of (i) the cardiac interbeat interval time series, the output of a major physiologic system regulated by the involuntary autonomic nervous system; and (ii) biological codes. First, we seek to characterize changes in the complexity of cardiac dynamics due to aging and disease, during both awake and sleeping periods. This analysis is a major extension of our previous work [7] that focused on application of MSE to a more limited database. In addition, we address the question of applying the MSE method to binary sequences in order to study the complexity of coding versus noncoding human DNA sequences.

The structure of the paper is as follows. In Sec. II we provide the mathematical background for calculating the entropy rate and discuss its physical meaning. We also present a short description of the approximate entropy (A_E) and the sample entropy (S_E) algorithms, which have been widely used in the analysis of short, noisy physiologic time series. In Sec. III, we review the MSE method, which incorporates the S_E statistics, and discuss the results of applying the MSE method to white and $1/f$ noises. The analytical calculations of S_E for both types of noises are presented in Appendix A. In Sec. IV, we apply the MSE method to a cardiac interbeat interval database comprising recordings of healthy subjects, subjects with atrial fibrillation, an erratic cardiac arrhythmia, and subjects with congestive heart failure. We address the question of quantifying the information in MSE curves for possible clinical use. We further discuss the effects of outliers, white noise superimposed on a physiologic time series, and finite sample frequency values in Appendix B. In Sec. V, we apply the MSE method to binary sequences of artificial and biological codes, aimed at quantifying the complexity of coding and noncoding DNA sequences. Technical aspects of applying the MSE method to such discrete sequences are described in Appendix C. Section VI presents conclusions.

II. BACKGROUND

The entropy $H(X)$ of a single discrete random variable X is a measure of its average uncertainty. Shannon's entropy [12] is calculated by the equation

$$H(X) = -\sum_{x_i\in\Theta} p(x_i)\log p(x_i) = -E[\log p(x_i)], \qquad (1)$$

where X represents a random variable with a set of values Θ and probability mass function $p(x_i)=P_r\{X=x_i\}$, $x_i\in\Theta$, and E represents the expectation operator. Note that $p\log p=0$ if $p=0$.

For a time series representing the output of a stochastic process, that is, an indexed sequence of n random variables, $\{X_i\}=\{X_1,\ldots,X_n\}$, with a set of values $\Theta_1,\ldots,\Theta_n$, respectively, and $X_i\in\Theta_i$, the joint entropy is defined as

$$\begin{aligned}H_n &= H(X_1,X_2,\ldots,X_n)\\ &= -\sum_{x_1\in\Theta_1}\cdots\sum_{x_n\in\Theta_n} p(x_1,\ldots,x_n)\log p(x_1,\ldots,x_n),\end{aligned} \qquad (2)$$

where $p(x_1,\ldots,x_n)=\text{P}_r\{X_1=x_1,\ldots,X_n=x_n\}$ is the joint probability for the n variables $X_1,\ldots,X_n$.

By applying the chain rule to Eq. (2), the joint entropy can be written as a summation of conditional entropies, each of which is a non-negative quantity,

$$H_n = \sum_{i=1}^{n} H(X_i|X_{i-1},\ldots,X_1). \qquad (3)$$

Therefore, one concludes that the joint entropy is an increasing function of n.

The rate at which the joint entropy grows with n, i.e., the entropy rate h, is defined as

$$h = \lim_{n\to\infty}\frac{H_n}{n}. \qquad (4)$$

For stationary ergodic processes, the evaluation of the rate of entropy has proved to be a very useful parameter [2,5,13–17].

Let us consider a $\mathcal{D}$-dimensional dynamical system. Suppose that the phase space of the system is partitioned into hypercubes of content $\varepsilon^{\mathcal{D}}$ and that the state of the system is measured at intervals of time δ. Let $p(k_1,k_2,\ldots,k_n)$ denote the joint probability that the state of the system is in the hypercube k_1 at $t=\delta$, in the k_2 at $t=2\delta$, and in the hypercube k_n at $t=n\delta$. The Kolmogorov-Sinai (KS) entropy is defined as

$$\begin{aligned}H_{KS} &= -\lim_{\delta\to 0}\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n\delta}\sum_{k_1,\ldots,k_n} p(k_1,\ldots,k_n)\log p(k_1,\ldots,k_n)\\ &= \lim_{\delta\to 0}\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n\delta}H_n.\end{aligned} \qquad (5)$$

For stationary processes [18], it can be shown that

$$\lim_{n\to\infty}\frac{H_n}{n} = \lim_{n\to\infty} H(X_n|X_{n-1},\ldots,X_1). \qquad (6)$$

Then, by the chain rule, it is straightforward to show that

$$H_{KS} = \lim_{\delta\to 0}\lim_{\varepsilon\to 0}\lim_{n\to\infty}(H_{n+1} - H_n). \tag{7}$$

The state of a system at a certain instant t_i is partially determined by its history, $t_1, t_2, \ldots, t_{i-1}$. However, each new state carries an additional amount of new information. The KS entropy measures the mean rate of creation of information, in other words, the decrease of uncertainty at a receiver by knowing the current state of the system given the past history.

Numerically, only entropies of finite order n can be computed. As soon as n becomes large with respect to the length of a given time series, the entropy H_n is underestimated and decays toward zero. Therefore, Eq. (7) is of limited use to estimate the entropy of finite length real-world time series. However, several formulas have been proposed in an attempt to estimate the KS entropy with reasonable precision. Grassberger and Procaccia [15] suggested characterizing chaotic signals by calculating the K_2 entropy, which is a lower bound of the KS entropy.

Let $\{X_i\}=\{x_1,\ldots,x_i,\ldots,x_N\}$ represent a time series of length N. Consider the m-length vectors: $\mathbf{u}_m(i)=\{x_i, x_{i+1},\ldots,x_{i+m-1}\}$, $1\leqslant i\leqslant N-m+1$. Let $n_i^m(r)$ represent the number of vectors $\mathbf{u}_m(j)$ that are close to the vector $\mathbf{u}_m(i)$, i.e., the number of vectors that satisfy $d[\mathbf{u}_m(i),\mathbf{u}_m(j)]\leqslant r$, where d is the Euclidean distance. $C_i^m(r)=n_i^m(r)/(N-m+1)$ represents the probability that any vector $\mathbf{u}_m(j)$ is close to the vector $\mathbf{u}_m(i)$. The average of the C_i^m, $C^m(r)=1/(N-m+1)\sum_{i=1}^{(N-m+1)}C_i^m(r)$, represents the probability that any two vectors are within r of each other. K_2 is defined as

$$K_2 = \lim_{N\to\infty}\lim_{m\to\infty}\lim_{r\to 0} -\ln[C^{m+1}(r) - C^m(r)]. \tag{8}$$

Following the same nomenclature, Eckmann and Ruelle (ER) [2] defined the function $\Phi^m(r)=1/(N-m+1)\sum_{i=1}^{N-m+1}\ln C_i^m(r)$, considering the distance between two vectors as the maximum absolute difference between their components: $d[\mathbf{u}_m(i),\mathbf{u}_m(j)]=\max\{|x(i+k)-x(j+k)|:0\leqslant k\leqslant m-1\}$. Note that $\Phi^{m+1}(r)-\Phi^m(r)\approx\sum_{i=1}^{N-m+1}\ln[C_i^m(r)/C_i^{m+1}(r)]$, represents the average of the natural logarithm of the conditional probability that sequences that are close to each other for m consecutive data points will still be close to each other when one more point is known. Therefore, Eckmann and Ruelle suggested calculating the KS entropy as

$$H_{\mathrm{ER}} = \lim_{N\to\infty}\lim_{m\to\infty}\lim_{r\to 0}[\Phi^m(r) - \Phi^{m+1}(r)]. \tag{9}$$

Although this formula has been useful in classifying low-dimensional chaotic systems, it does not apply to experimental data since the result is infinity for a process with superimposed noise of any magnitude [19]. For the analysis of short and noisy time series, Pincus [17] introduced a family of measures termed approximate entropy, $A_E(m,r)$, defined as

$$A_E(m,r) = \lim_{N\to\infty}[\Phi^m(r) - \Phi^{m+1}(r)]. \tag{10}$$

A_E is estimated by the statistics,

$$A_E(m,r,N) = \Phi^m(r) - \Phi^{m+1}(r). \tag{11}$$

A_E was not intended as an approximate value of ER entropy. Rather, A_E is a regularity statistic. It applies to real-world time series and, therefore, has been widely used in physiology and medicine [4]. Lower A_E values are assigned to more regular time series while higher A_E values are assigned to more irregular, less predictable, time series.

Recently, a modification of the A_E algorithm, sample entropy (S_E) [20], has been proposed. S_E has the advantage of being less dependent on time series length, and showing relative consistency over a broader range of possible r,m, and N values. Starting from the definition of the K_2 entropy, Richman and Moorman [20] defined the parameter

$$S_E(m,r) = \lim_{N\to\infty} -\ln\frac{U^{m+1}(r)}{U^m(r)}, \tag{12}$$

which is estimated by the statistic

$$S_E(m,r,N) = -\ln\frac{U^{m+1}(r)}{U^m(r)}. \tag{13}$$

The differences between $U^{m+1}(r)$ and $C^{m+1}(r)$, $U^m(r)$ and $C^m(r)$ result from (1) defining the distance between two vectors as the maximum absolute difference between their components; (2) excluding self-matches, i.e., vectors are not compared to themselves; and (3) given a time series with N data points, only the first $N-m$ vectors of length m, $\mathbf{u}_m(i)$, are considered, ensuring that, for $1\leqslant i\leqslant N-m$, the vector $\mathbf{u}_{m+1}(i)$ of length $m+1$ is also defined. S_E is precisely equal to the negative of the natural logarithm of the conditional probability that sequences close to each other for m consecutive data points will also be close to each other when one more point is added to each sequence. Figure 1 illustrates how S_E values are calculated.

Note that

$$A_E(m,r,N) \approx \frac{1}{N-m}\sum_{i=1}^{N-m}\ln\frac{n_i^m}{n_i^{m+1}} \tag{14}$$

and

$$S_E(m,r,N) = \ln\frac{\sum_{i=1}^{N-m} n_i'^{m}}{\sum_{i=1}^{N-m} n_i'^{m+1}}, \tag{15}$$

where $n_i'^m$ differs from n_i^m to the extent that for S_E self-matches are not counted ($i\neq j$) and $1\leqslant i\leqslant N-m$.

The difference between A_E and S_E can be related to the Renyi entropies, $S_R(q)$, which are defined by $S_R(q)=\ln(\sum_i p_i^q)/(1-q)$. A_E approximates the Renyi entropy of order $q=1$ (the usual Shannon entropy) and S_E the Renyi entropy of order $q=2$. The advantage of the latter is that the estimator [Eq. (15)] is unbiased [21].

Both S_E and A_E measure the degree of randomness (or inversely, the degree of orderliness) of a time series. However, as noted, there is no straightforward relationship between regularity, measured by entropy-based metrics, and complexity [22]. An increase in entropy is usually but not always associated with an increase in complexity. For example, higher entropy values are assigned to randomized surrogate time series than to the original time series even when the original time series represent the output of complex dynamics with correlational structures on multiple spatiotemporal scales. However, the process of generating surrogate data is designed to destroy correlations and, consequently, degrades the information content of the original signal. In fact, entropy-based metrics are maximized for random sequences, although it is generally accepted that both perfectly ordered and maximally disordered systems possess no complex structures [23]. A meaningful physiologic complexity measure, therefore, should vanish for these two extreme states.

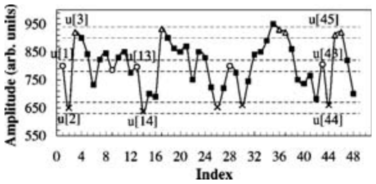

FIG. 1. A simulated time series $u[1],\ldots,u[N]$ is shown to illustrate the procedure for calculating sample entropy (S_E) for the case $m=2$ and a given positive real value r. Dotted horizontal lines around data points $u[1]$, $u[2]$, and $u[3]$ represent $u[1]\pm r$, $u[2]\pm r$, and $u[3]\pm r$, respectively. Two data points match each other, that is, they are indistinguishable, if the absolute difference between them is $\leqslant r$. Typically, r varies between 10% and 20% of the time series SD. The symbol ○ is used to represent data points that match the data point $u[1]$. Similarly, the symbols × and △ are used to represent data points that match the data points $u[2]$ and $u[3]$, respectively. Consider the two-component ○-× template sequence ($u[1]$, $u[2]$) and the three-component ○-×-Δ template sequence ($u[1]$, $u[2]$, $u[3]$). For the segment shown, there are two ○-× sequences, ($u[13]$, $u[14]$) and ($u[43]$, $u[44]$), that match the template sequence ($u[1]$, $u[2]$), but only one ○-×-△ sequence that matches the template sequence ($u[1]$, $u[2]$, $u[3]$). Therefore, in this case, the number of sequences matching the two-component template sequences is two and the number of sequences matching the three-component template sequence is 1. These calculations are repeated for the next two-component and three-component template sequence, which are ($u[2]$, $u[3]$) and ($u[2]$, $u[3]$, $u[4]$), respectively. The numbers of sequences that match each of the two- and three-component template sequences are again summed and added to the previous values. This procedure is then repeated for all other possible template sequences, ($u[3]$, $u[4]$, $u[5]$),…,($u[N$-2], $u[N$-1], $u[N]$), to determine the ratio between the total number of two-component template matches and the total number of three-component template matches. S_E is the natural logarithm of this ratio and reflects the probability that sequences that match each other for the first two data points will also match for the next point.

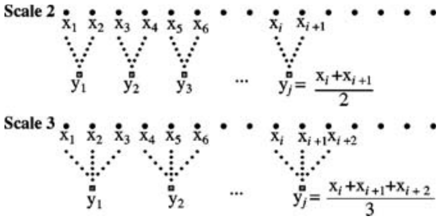
Scale 2
x1 x2 x3 x4 x5 x6 ... xi xi+1
y1 y2 y3 ... yj = (xi+xi+1)/2
Scale 3
x1 x2 x3 x4 x5 x6 ... xi xi+1 xi+2
y1 y2 ... yj = (xi+xi+1+xi+2)/3

FIG. 2. Schematic illustration of the coarse-graining procedure. Adapted from Ref. [8].

Of related note, when applied to physiologic data, both A_E and S_E algorithms assign higher entropy values to certain pathologic time series than to time series derived from free-running physiologic systems under healthy conditions [24]. However, pathologic time series represent the output of less adaptive (i.e., more impaired), and therefore, presumably, less complex systems [25,26]. One reason for obtaining these nonphysiologic results is the fact that A_E and S_E are based on a single scale. We note that both the KS entropy and the related A_E parameters depend on a function's one-step difference (e.g., $H_{n+1}-H_n$) and reflect the uncertainty of the next new point given the past history of the series. Therefore, these measures do not account for features related to structure and organization on scales other than the shortest one.

For physical systems, Zhang [23,27] proposed a general approach to take into account the information contained in multiple scales. Zhang's complexity measure is a sum of scale-dependent entropies. It has the desirable property of vanishing in the extreme ordered and disordered limits, and is an extensive quantity. However, since it is based on Shannon's definition of entropy, Zhang's method requires a large amount of almost noise-free data, in order to map the data to a discrete symbolic sequence with sufficient statistical accuracy. Therefore, it presents obvious limitations when applied to free-running physiologic signals that typically vary continuously and have finite length.

To overcome these limitations, we [7] recently introduced the multiscale entropy (MSE) method, applicable both to physical and physiologic time series. Our method is based on Zhang's and Pincus's approach.

III. MULTISCALE ENTROPY (MSE) METHOD

Given a one-dimensional discrete time series, $\{x_1,\ldots,x_i,\ldots,x_N\}$, we construct consecutive coarse-grained time series, $\{y^{(\tau)}\}$, corresponding to the scale factor, τ. First, we divide the original times series into nonoverlapping windows of length τ; second, we average the data points inside each window (Fig. 2). In general, each element of a coarse-grained time series is calculated according to the equation

$$y_j^{(\tau)} = \frac{1}{\tau}\sum_{i=(j-1)\tau+1}^{j\tau} x_i, \quad 1\leqslant j\leqslant N/\tau. \tag{16}$$

For scale one, the time series $\{y^{(1)}\}$ is simply the original time series. The length of each coarse-grained time series is equal to the length of the original time series divided by the scale factor, τ.

Finally, we calculate an entropy measure (S_E) for each coarse-grained time series plotted as a function of the scale factor τ. We call this proced re m ltiscale entrop (MSE) anal sis.

The MSE c r es are sed to compare the relati e comple it of normali ed time series (same ariance for scale one) based on the follo ing g idelines: (1) if for the majorit of the scales the entrop al es are higher for one time series than for another, the former is considered more comple than the latter; (2) a monotonic decrease of the entrop al es indicates the original signal contains information onl in the smallest scale.

Zhang defined comple it as the integral of all the scale-dependent entropies: $K=\int_1^N d\tau H(\tau)$, hich for a discrete signal co ld be estimated b $K=\sum_{i=1}^N H(i)(N\to\infty)$. D e to the finite length of real- orld time series, entrop can onl be calc lated for a finite range of scales. The s m to infinit is not feasible. Since different sets of entrop al es can ield the same K al e, e foc s on the anal sis of the MSE c r es instead of assigning a single comple it al e to each time series. F rther, for application to biological s stems, the MSE c r e ma pro ide sef l insights into the control mechanisms nderl ing ph siologic d namics o er different scales. We note, ho e er, that an appro imation of K for scales bet een one and t ent f rther s pports the concl sions e present in this paper.

Unless other ise specified, the al es of the parameters sed to calc late S_E are $N=2\times10^4$, $m=2$, and $r=0.15$.

The al e of the parameter r is a percentage of the time series SD. This implementation corresponds to normali ing the time series. As a conseq ence, S_E res lts do not depend on the ariance of the original time series, i.e., the absol te al e of the data points, b t onl on their seq ential ordering.

In general, ho e er, the entrop meas res reflect both the ariance of a time series and its correlation properties. To ill strate this point, e e amine t o special cases here these t o effects can be isolated. Case (1): Consider t o ncorrelated random ariables, X and Y, ith set of al es $\{\ _1,\ _2,\ldots,\ _N\}$ and $\{\ _1,\ _2,\ldots,\ _M\}$, respecti el . Ass ming that all al es are eq all probable, $p(\ _i)=1/N$, the entrop of the random ariables X is $H(X)=\ \sum_{i=1}^N 1/N\log 1/N=\log N$. Similarl , $H(Y)=\log M$. If $N>M$, then $H(X)>H(Y)$. Therefore, for the same le el of resol tion, the larger the set of alphabet of a random ariable, the larger its ariance and the entrop al e. Case (2): Consider a periodic signal ith ariance $|a|$ and a random signal ith ariance $|b|$, s ch that $|a|\gg|b|$. The entrop of a periodic signal is ero, since each data point occ rs ith probabilit 1. Therefore, the entrop of a periodic signal is ne er larger than the entrop of a random signal regardless of the ariance of the signals.

With the e ception of s ch er simple cases, it is not possible to eight separatel the contrib tions of the SD and the correlation properties to the entrop al e. Signals ith higher ariabilit and those that are more random tend to be more entropic. Ne ertheless, the act al entrop al e res lts from a comple combination of these t o factors.

In the MSE method, r is set at a certain percentage (s all 15%) of the original time series SD, and remains constant for all scales [10,28]. We do not recalc late r for each coarse-grained time series. After the initial normali ation, s bseq ent changes of the ariance d e to the coarse-graining proced re are related to the temporal str ct re of the original time series, and sho ld be acco nted for b the entrop meas re. The initial normali ation, ho e er, ins res that the MSE al es assigned to t o different time series are not a tri ial conseq ence of possible differences bet een their ariances b t res lt from different organi ational str ct res.

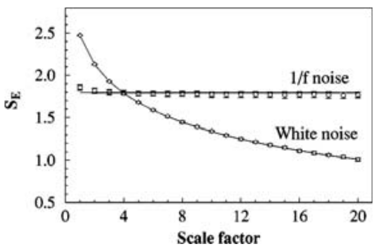

FIG. 3. MSE anal sis of 30 sim lated Ga ssian distrib ted (mean ero, ariance one) hite and $1/f$ noise time series, each ith 3×10^4 data points. S mbols represent mean al es of entrop for the 30 time series and error bars the SD, hich in a erage is 0.05 for hite noise and 0.02 for $1/f$ noise. Lines represent n merical e al ation of anal tic S_E calc lation. Note that the differences bet een the mean al es of S_E and the corresponding n merical al es are less than 1%. SD is larger for $1/f$ noise time series beca se of nonstationarit . Adapted from Ref. [7]. (See Appendi A.)

We first applied the MSE method to sim lated hite and $1/f$ noises and compared the n merical res lts ith the entrop al es calc lated anal ticall (Appendi A). Fig re 3 presents the res lts. For scale one, a higher al e of entrop is assigned to hite noise time series in comparison ith $1/f$ time series. Ho e er, hile the al e of entrop for the coarse-grained $1/f$ series remains almost constant for all scales, the al e of entrop for the coarse-grained hite noise time series monotonicall decreases, s ch that for scales >4 it becomes smaller than the corresponding al es for $1/f$ noise. This res lt is consistent ith the fact that, nlike hite noise, $1/f$ noise contains comple str ct res across m ltiple scales [23,27]. Note that in the case of hite noise, as the length of the indo sed for coarse-graining the time series increases (i.e., the resol tion decreases), the a erage al e inside each indo con erges to a fi ed al e since no ne str ct res are re ealed on larger scales. Conseq entl , coarse-grained time series are progressi el smoothed o t and the standard de iation monotonicall decreases ith the scale factor. Therefore, the monotonic decrease of entrop ith scale, hich mathematicall res lts from the decrease of standard de iation, reflects the fact that hite noise has information onl on the shortest scale. In contrast, for $1/f$ noise signals the a erage al es of the fl ct ations inside each indo do not con erge to a gi en al e. In other ords, the statistical properties of fl ct ations ithin a indo (e.g., 10 data points) are not the same as

those of the ne t indo beca se ne information is re ealed at all scales. The MSE ses the a erage al e of the fl ct ations as the representati e statistical propert for each block and meas res the irreg larit of the block-to-block d namics.

The discrepanc bet een the sim lation and the anal tical res lts is less than 0.5%. In Appendi B, e disc ss ho the time series length, N, and the al es of parameters r and m affect S_E res lts for both hite and $1/f$ noise time series. We f rther disc ss the effects of ncorrelated noise and o tliers on MSE res lts of cardiac interbeat inter al time series.

IV. MSE ANALYSIS OF CARDIAC INTERBEAT INTERVAL TIME SERIES

We ne t appl the MSE method to the cardiac interbeat (RR) inter al time series deri ed from 24 ho r contin o s electrocardiographic (ECG) Holter monitor recordings of health s bjects, s bjects ith congesti e heart fail re, a life-threatening condition, and s bjects ith atrial fibrillation, a major cardiac arrh thmia.[1] We test the h pothesis that nder free-r nning conditions, health interbeat inter al d namics are more comple than those ith patholog d ring both da time and nighttime ho rs.

The data for the normal control gro p ere obtained from 24 ho r Holter monitor recordings of 72 health s bjects, 35 men and 37 omen, aged 54.6 16.2 ears (mean SD), range 22-78 ears. ECG data ere sampled at 128 H . The data for the congesti e heart fail re gro p ere obtained from 24 ho r Holter recordings of 43 s bjects (28 men and 15 omen) aged 55.5 11.4 ears (mean SD), range 22-78 ears. Ne York Heart Association (NYHA) f nctional classification [30] is pro ided for each s bject: 4 s bjects ere assigned to class I, 8 to class II, 17 to class III, and 14 to class III-IV. Fo rteen recordings ere sampled at 250 H and 29 recordings ere sampled at 128 H . The data for the atrial fibrillation gro p ere obtained from 10 ho r Holter recordings sampled at 250 H of nine s bjects. Datasets ere filtered to e cl de artifacts, premat re entric lar comple es, and missed beat detections (see Appendi B). Of note, the incl sion of the premat re entric lar comple es does not q alitati el change o r anal sis.

Representati e time series of health , congesti e heart fail re, and atrial fibrillation gro p s bjects are presented in Fig. 4.

When disc ssing the MSE res lts of cardiac interbeat inter al time series, e refer to large and small time scales hen the scales are larger or smaller than one t pical respirator c cle length, that is, appro imatel fi e cardiac beats.

In Fig. 5, e present the res lts of the MSE anal sis of the RR inter al time series for the three gro ps of s bjects. We obser e three different t pes of beha iors: (i) The entrop meas re for time series deri ed from health s bjects increases on small time scales and then stabili es to a relati el constant al e. (ii) The entrop meas re for time series deri ed from s bjects ith congesti e heart fail re markedl decreases on small time scales and then grad all increases. (iii) The entrop meas re for time series deri ed from s bjects ith atrial fibrillation [31] monotonicall decreases, similar to hite noise (Fig. 3).

For scale one, hich is the onl scale considered b traditional single-scale based comple it methods, the entrop assigned to the heartbeat time series of s bjects ith atrial fibrillation and those ith congesti e heart fail re is higher than the entrop assigned to the time series of health

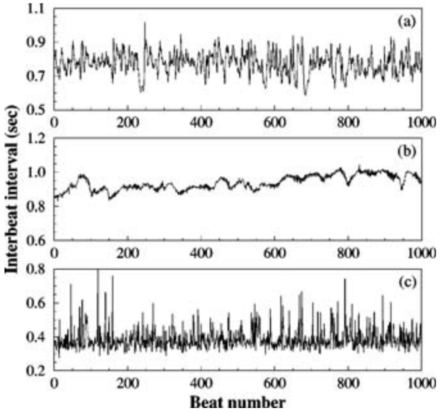

FIG. 4. Representati e interbeat inter al time series from (a) health indi id al (sin s rh thm), (b) s bject ith congesti e heart fail re, and (c) s bject ith atrial fibrillation, a highl erratic cardiac arrh thmia.

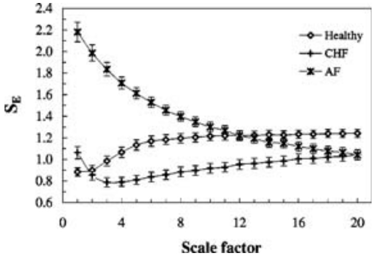

FIG. 5. MSE anal sis of RR time series deri ed from long-term ECG recordings of health s bjects in normal sin s rh thm, those ith congesti e heart fail re (CHF) in sin s rh thm, and those ith atrial fibrillation (AF). S mbols represent the mean al es of entrop for each gro p and bars represent the standard error ($SE=\mathrm{SD}/\sqrt{n}$), here n is the n mber of s bjects). Parameters to calc late S_E are $m=2$ and $r=0.15$. Time series length is 2×10^4 beats. The S_E al es from health s bjects are significantl (t-test, $p<0.05$) higher than from CHF and AF s bjects for scales larger than scale 2 and scale 20, respecti el .

[1] All data anal ed here are a ailable at http://ph sionet.org and ha e been described in Ref. [29].

subjects. In contrast, for sufficiently large scales, the time series of healthy subjects are assigned the highest entropy values. *Thus, the MSE method indicates that healthy dynamics are the most complex, contradicting the results obtained using the traditional S_E and A_E algorithms*.

The time series of subjects with AF exhibit substantial variability in beat-to-beat fluctuations. However, the monotonic decrease of the entropy with scale reflects the degradation of the control mechanisms regulating heart rate on larger time scales in this pathologic state.

The largest difference between the entropy values of coarse-grained time series from congestive heart failure and healthy subjects is obtained for time scale 5. On small time scales, the difference between the profiles of the MSE curves for these two groups may be due to the fact that the respiratory modulation of heart rate (respiratory sinus arrhythmia) has higher amplitude in healthy subjects than in subjects with congestive heart failure. Since entropy is a measure of regularity (orderliness), the higher the amplitude of the respiratory modulation, the lower the entropy values tend to be. However, the coarse-graining procedure filters out the periodic respiratory-related heart rate oscillations. Therefore, coarse-grained time series from healthy subjects on large time scales are likely more irregular (and are assigned higher entropy values) than the original time series.

For congestive heart failure subjects, the entropy of coarse-grained time series decreases from scales 1–3 and then progressively increases. This result suggests that for these subjects, the control mechanisms regulating heart rate on relatively short time scales are the most affected. However, this finding could also result from the measurement uncertainty of the interbeat intervals due to the finite sample frequency. Since time series from subjects with congestive heart failure have, in general, lower variance than time series from healthy subjects, the signal-to-noise ratio tends to be lower for datasets from heart failure subjects. We note that the MSE coarse-graining procedure progressively eliminates the uncorrelated random components such that the entropy of white noise coarse-grained time series monotonically decreases with scale (Fig. 3). Therefore, the monotonic decrease of the entropy values with heart failure over short time scales may be related to the relatively low signal-to-noise ratio.

We also found that the asymptotic value of entropy may not be sufficient to differentiate time series that represent the output of different dynamical processes. As seen in Fig. 5, for time scale 20, the value of the entropy measure for the heart failure (sinus rhythm) and atrial fibrillation time series is the same. However, these time series represent the output of very different types of cardiac dynamics. *Therefore, not only the specific values of the entropy measure but also their dependence on time scale need to be taken into account to better characterize the physiologic process*.

Next, to assess the effects of activity level, we compare the complexity of the RR interval time series during sleep and wake periods for the different subject groups. Using the 24 h heartbeat interval time series of healthy and congestive heart failure subjects, the sleep and wake datasets were then obtained by extracting the segments of 2×10^4 consecutive data points ($\sim$5 h) with highest and lowest heart rate, respectively. Figures 6(a) and 6(b) show that during both the waking and sleeping periods, the highest entropy values on most time scales are assigned, in descending order, to the coarse-grained time series derived from healthy young subjects, healthy elderly subjects, and congestive heart failure subjects. These results further support the concept that under free-running conditions, the cardiac dynamics of healthy young subjects are the most complex and are consistent with the hypothesized loss of complexity with aging and disease [24].

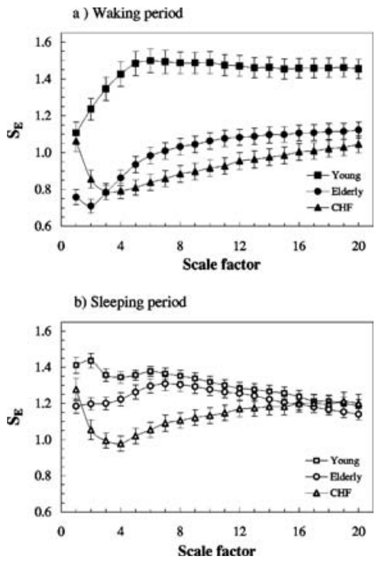

FIG. 6. MSE analysis of RR time series derived from 24 h ECG recordings of 27 healthy young subjects, aged 34.5 ± 7.3 years (mean ± SD), range 20 - 50 years, 45 healthy elderly subjects, aged 70 ± 3.97 years, range 66 - 75 years, and 43 congestive heart failure (CHF) subjects, aged 55 ± 11.6 years, range 22 - 78 years. (a) Waking period. For all scales the S_E values from healthy young subjects are significantly (t-test, $p<0.05$) higher than from CHF subjects. The S_E values from healthy young subjects are significantly higher than from healthy elderly subjects for scales larger than scale 1. The S_E values from healthy elderly subjects are significantly (t-test, $p<0.05$) higher than from CHF subjects for scales between scales 5 and 13, inclusively. (b) Sleeping period. Both the S_E values from healthy elderly and healthy young subjects are significantly (t-test, $p<0.05$) higher than from CHF subjects for scales between scales 2 and 11, inclusively. The S_E values from healthy young subjects are significantly higher than from healthy elderly subjects for scales shorter than scale 5. Symbols represent the mean values of entropy for each group and the bars represent the standard error. Parameters of S_E calculation are $m=2$ and $r=0.15$. Time series length is 2×10^4 beats.

Despite the fact that the entropy values for healthy elderly subjects are lower than those for healthy young subjects, the profiles of MSE curves for both groups are similar, in particular over large time scales. Indeed, during sleep, a period of minimal activity, the difference between the entropy values of both groups is significant over only small time scales. These results are consistent with the known loss of high-frequency modulation of the cardiac rhythm with age [32], and suggest that the control mechanisms operating over small time scales, including the parasympathetic branch of the autonomic nervous system, are the most affected with aging. The monotonic decrease in entropy on large time scales for both young and elderly groups indicates that the coarse-grained time series become progressively more regular (less complex) than those corresponding to shorter time scales, which is compatible with a previous study [33] reporting a reduction in long-range correlations in healthy subjects during the sleeping period.

The MSE results for the waking and sleeping periods of each group of subjects are shown in Fig. 7. For both young and elderly healthy subjects, the profiles of the MSE curves corresponding to the waking and sleeping periods are qualitatively different from each other [Figs. 7(a) and 7(b)]. For subjects with congestive heart failure, however, there is only a shift of the entropy values but not a significant change in the profile of the MSE curves [Fig. 7(c)]. Thus, differences between the day-night dynamics of subjects with a severe cardiac pathology are less marked than for healthy subjects. This loss of differentiation in the complexity of sleep/wake dynamics may be a useful new index of reduced adaptive capacity.

Further, we found that, contrary to the results obtained for healthy young subjects, in healthy elderly and congestive heart failure subjects, the coarse-grained time series obtained from the waking period have lower entropy than those obtained from the sleeping period. To the extent that aging and disease degrade adaptive capacity, environmental stimuli may exceed the system's reserve. This situation would be equivalent to what might occur if a young individual were subject to prolonged physical or other stress throughout the daytime hours.

Finally, to assist in clinical classification, we extracted two simple features of MSE curves, the slopes for small and large time scales, i.e., the slopes of the curves defined by S_E values between scale factors 1 and 5, and scale factors 6 and 20, respectively. Results for the healthy and congestive heart failure groups corresponding to the sleeping period are presented in Fig. 8. There is a good separation between the two groups. Considering other features of the MSE curves, in addition to these slopes, may further improve the separation. Alternatively, methods derived from pattern recognition techniques, e.g., Fisher's discriminant, may also be useful for clinical discrimination [9].

V. MSE ANALYSIS OF ARTIFICIAL AND BIOLOGICAL CODES

In all cells, from microbes to mammals, proteins are responsible for most structural, catalytic, and regulatory functions. Therefore, the number of protein-coding genes that an organism makes use of could be an indicator of its degree of complexity. However, several observations contradict this reasoning [34,35].

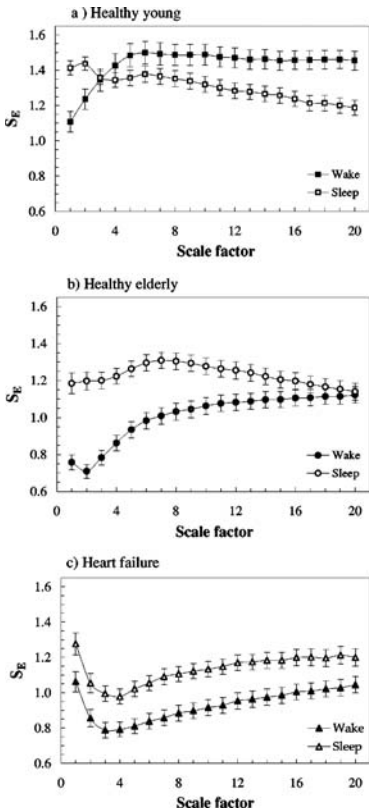

FIG. 7. MSE analysis of RR time series derived from 24 h ECG recordings during waking and sleeping periods. (a) Young healthy subjects. The S_E values for the waking period are significantly (t-test) higher ($p<0.05$) than for the sleeping period on scales larger than scale 7. (b) Elderly healthy subjects. The S_E values for the sleeping period are significantly (t-test) higher ($p<0.05$) than for the waking period on scales shorter than scale 16. (c) Congestive heart failure subjects. The S_E values for the sleeping period are significantly (t-test) higher ($p<0.05$) than for the waking period on all scales but scale 1. Symbols represent mean values of entropy for each group and the bars represent the standard error. Parameters of S_E calculation are $m=2$ and $r=0.15$. Time series length is 2×10^4 beats.

Large regions of DNA, which in humans account for about 97% of the total genome, do not code for proteins and were previously thought to have no relevant purpose. These regions have been referred to as “junk” DNA or gene

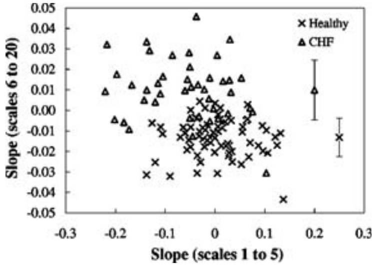

FIG. 8. Scatter plot of the slope of the MSE curves between scale factors 6 and 20 vs the slope of the MSE curves between scale factors 1 and 5, for healthy and congestive heart failure (CHF) groups during the sleeping period. For both groups, symbols with error bars represent the mean of y-axis values, and the error bars the corresponding SD. The groups are well separated ($p<0.005$).

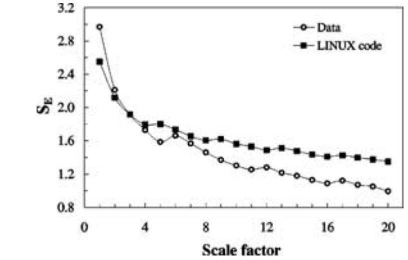

FIG. 9. MSE results for binary files of a computer executable program (LINUX kernel) and a compressed data file. The original binary file has only two symbols, 0 and 1. However, the number of symbols in coarse-grained sequences increases with the scale factor, which introduces a characteristic artifact on the MSE curves. In order to avoid this artifact, instead of the original sequences, we analyze a derived sequence, which is constructed as follows: we divide the original sequence into consecutive nonoverlapping segments, each with 128 data points, and then calculate the number of 1's (0's) within each segment. Some structural information is lost since the procedure is not a one-to-one mapping. The derived sequences are expected to be more regular than the original ones. However, this procedure does not alter the conclusions drawn from our analysis.

deserts. However, these noncoding sequences are starting to attract increasing attention as more recent studies suggest that they may have an important role in regulation of transcription, DNA replication and chromosomal structure, pairing, and condensation.

Detrended fluctuation analysis [37–39] revealed that noncoding sequences contained long-range correlations and possessed structural similarities to natural languages, suggesting that these sequences could in fact carry important biological information. In contrast, coding sequences were found to be more like a computer data file than a natural language.

The biological implications of the presence of long-range correlations in noncoding sequences, their origin, and their nature are still being debated. Audit *et al.* [40,41] have investigated the relation between long-range correlations and the structure and dynamics of nucleosomes. Their results suggest that long-range correlations extending from 10 to 200 bp are related to the mechanisms underlying the wrapping of DNA in the nucleosomal structure.

Gene regulatory elements or enhancers are types of functional sequences that reside in noncoding regions. Until recently, enhancers were thought to be located near the genes that they regulate. However, subsequent *in vivo* studies [42,43] have demonstrated that enhancers and the genes to which they are functionally linked may be separated by more than thousands of bases. These results reinforce earlier evidence that the noncoding sequences contain biological information and further support the notion that there are several layers of information in genomic DNA.

In this section, we apply the MSE method to the analysis of the complexity of both coding and noncoding DNA sequences of human chromosomes.

Because of possible parallelisms between artificial and biological codes, we first considered two examples of artificial language sequences: the compiled version of the LINUX Operating System, an executable computer program, and a compressed nonexecutable computer data file, which can both be analyzed as binary sequences. Although both files contain useful information, the structure of that information is very different. The sequence derived from the executable program exhibits long-range correlations [38], while the sequence derived from the data file does not. These results indicate that the computer program, which executes a series of instructions and likely contains several loops running inside each other, possesses a hierarchical structure, in contrast to the computer data file. Therefore, the former is expected to be more complex than the latter.

When applied to discrete sequences (binary codes), the MSE results present a typical artifact due to the dependence of the entropy values on the size of the sequence alphabet, which we discuss in Appendix C.

MSE analysis of the nonbiological codes reveals (Fig. 9) the following. (i) For scale one, the sequence derived from the data file is assigned a higher entropy value than the sequence derived from the executable program. (ii) Between scales 2 and 6, the S_E measure does not separate the coarse-grained sequences of the two files. (iii) For scales larger than scale 6, the highest entropy values are assigned to coarse-grained sequences derived from the executable program file. Furthermore, the difference between S_E values assigned to coarsegrained sequences of the executable file and the computer data file increases with scale factor. These results indicate, as hypothesized, that the structure of the executable file is more complex than the structure of the data file. Of note, conventional (single scale) S_E and A_E algorithms applied to sequences of artificial languages fail to meaningfully quantify their overall complexity.

Finally, we apply the MSE method to the analysis of DNA sequences, likely one of the most complex natural information databases.

The DNA building units are four nucleotides. Two of them contain a purine base, adenine (A) or guanine (G), and

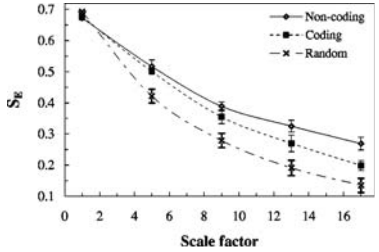

FIG. 10. MSE results for four coding, nine noncoding DNA sequences from human chromosome 22 and 30 binary random time series. All coding sequences with more than identified 4×10^3 bp were selected. The longest coding sequences has 6762 bp. All noncoding sequences with more than 6000 and fewer than 6050 bp were selected. The length of the random sequences is 6000 data points. The symbols and the error bars represent the S_E mean values and SD, respectively. Due to a typical artifact that affects the MSE results of discrete sequences (Appendix C), only the entropy values for scales 1, 5, 9, 13, and 17 are plotted. Note the higher complexity of the noncoding vs coding sequences ($p=0.006$ for scale 9). The lowest entropy values are assigned to the random (white noise: mean zero, variance 1) time series mapped to a binary sequence: 1 if $x_i>0$ and 0 if $x_i<0$.

the other two contain a pyrimidine base, cytosine (C) or thymine (T). There are many ways of mapping the DNA sequences to a numerical sequence that take into consideration different properties of the DNA sequences. For this application, we consider the purine-pyrimidine rule [37–39]. Given the original DNA sequence, bases A and G are mapped to number 1, and bases C and T are mapped to number -1.

In Fig. 10, we present the MSE results for selected coding and noncoding human DNA sequences. For scales larger than scale 5, S_E values for noncoding sequences are higher than for coding sequences. Consistently, for all scales but the first one, the lowest S_E values are assigned to coarse-grained time series derived from uncorrelated white noise mapped to a binary sequences. Comparable results were obtained from the analysis of coding versus noncoding sequences ($\geqslant 4\times10^3$ bp) of all human chromosomes. These results show that the structure of noncoding sequences is more complex than the structure of coding sequences analyzed here.

These findings support previous studies [37–39] suggesting a parallelism between executable computer programs and noncoding sequences, and data storing files and coding sequences. They also support the view that noncoding sequences contain important biological information. As pointed out by others [35,36,40,41], biological complexity and phenotype variations should relate not only to proteins, which are the main effectors of cellular activity, but also to the organizational structure of the control mechanisms responsible for the networking and integration of gene activity.

VI. LIMITATIONS AND FUTURE DIRECTIONS

The MSE method requires an adequate length of data to provide reliable statistics for the entropy measure on each scale. As discussed in Appendix B, for simulated white and $1/f$ noises, both the mean value of S_E and the SD increase as the length of the time series decreases. However, for all time series tested, the consistency of the results was preserved, i.e., given two time series, *a* and *b*, each with 3×10^4 data points, whenever S_E was higher (lower) for time series *a* than for time series *b*, the same result held if only 1×10^3 data points were considered.

The minimum number of data points required to apply the MSE method depends on the level of accepted uncertainty. Typically, we use time series with 2×10^4 data points for analyses extending up to scale 20, in which case the shortest coarse-grained time series has 1×10^3 data points.

Another important consideration is related to nonstationarity. To calculate S_E, one has to fix the value of a parameter that depends on the time series SD. Therefore, the results may be significantly affected by nonstationarities, outliers, and artifacts. As we discuss in Appendix C, removing local artifacts and a small percentage of outliers ($<2\%$) does not substantially modify the structure of the time series and its related statistical properties. In contrast, attempts to remove nonlocal nonstationarities, e.g., trends, will most likely modify the structure of the time series over multiple time scales.

Further studies are needed to construct clinically useful indices for monitoring the complexity of biological systems, and for developing and testing the utility of complexity measures designed to quantify the degree of synchronization of two time series over multiple scales [20].

We note that the cardiac analyses reported here pertain to interbeat interval dynamics under free-running conditions. The high capability of healthy systems to adapt to a wide range of perturbations requires functioning in a multidimensional state space. However, under stress, the system is forced to work in a tighter regime. For example, during physical exercise, there is a sustained increase in heart rate and a decrease in the amplitude of the interbeat interval fluctuations in response to an increased demand for oxygen and nutrients. The dynamics is, therefore, limited to a subset of the state space. We anticipate that under a variety of stressed conditions, healthy systems will generate less complex outputs than under free-running conditions [11].

Finally, the potential applications of the MSE method to the study of artificial and biological codes, with attention to the effects of evolution on the complexity of genomic sequences, require systematic analysis.

VII. CONCLUSIONS

The long-standing problem of deriving useful measures of time series complexity is important for the analysis of both physical and biological systems. MSE is based on the observation that the outputs of complex systems is far from the extrema of perfect regularity and complete randomness. Instead, they generally reveal structures with long-range correlations on multiple spatial and temporal scales. These multiscale features, ignored by conventional entropy calculations, are explicitly addressed by the MSE method.

When applied to simulated time series, the MSE method shows that $1/f$ noise time series are more complex than hite noise time series. These res lts are consistent ith the presence of long-range correlations in $1/f$ noise time series b t not in hite noise time series.

Ph siologic comple it is associated ith the abilit of li ing s stems to adj st to an e er-changing en ironment, hich req ires integrati e m ltiscale f nctionalit . In contrast, nder free-r nning conditions, a s stained decrease in comple it reflects a red ced abilit of the s stem to f nction in certain d namical regimes possibl d e to deco pling or degradation of control mechanisms.

When applied to the cardiac interbeat inter al time series of health s bjects, those ith congesti e heart fail re and those ith atrial fibrillation, the MSE method sho s that health d namics are the most comple . Under pathologic conditions, the str ct re of the time series ariabilit ma change in t o different a s. One d namical ro te to disease is associated ith loss of ariabilit and the emergence of more reg lar patterns (e.g., heart fail re). The other d namical ro te is associated ith more random t pes of o tp ts (e.g., atrial fibrillation). In both cases, MSE re eals a decrease in s stem comple it .

Finall , e emplo ed the MSE method to compare the comple it of an e ec table comp ter program ers s a compressed none ec table comp ter data file, and selected coding ers s noncoding DNA seq ences. We fo nd that the e ec table comp ter program has higher comple it than the none ec table comp ter data file, and similarl that the noncoding seq ences are more comple than the coding seq ences e amined. O r res lts s pport recent *in itro* and *in i o* st dies s ggesting, contrar to the j nk DNA theor , that noncoding seq ences contain important biological information [44].

ACKNOWLEDGMENTS

We thank J. Miet s, I. Henr , and J. Heale for al able disc ssions and assistance. We gratef ll ackno ledge s pport from the National Instit tes of Health/National Center for Research Reso rces (P41-RR13622), the G. Harold and Leila Y. Mathers Charitable Fo ndation, the Centers for Disease Control and Pre ention (H75-CCH119124), the NIH/NICHD (R01-HD39838), and the James S. McDonnell Fo ndation.

APPENDIX A: MSE RESULTS FOR WHITE AND 1/*f* NOISES

In this appendi , e pro ide detailed anal tical deri ations of MSE for t o special cases: correlated and ncorrelated noises ith Ga ssian distrib tions. Linear Ga ssian correlation is a necessar ass mption to make the deri ation possible. In general, it is diffic lt to deri e anal tical sol tions for MSE of stochastic processes ith nonlinear correlations.

First, e start ith the case of ncorrelated noise (hite noise). For the case $m=1$, S_E is the negati e nat ral logarithm of the conditional probabilit that the distance bet een t o data points is less than or eq al to r (i.e., $|{}_i \quad {}_j| \leqslant r$) gi en that the distance bet een the t o preceding data points is also less than or eq al to r (i.e., $|{}_{i\ 1} \quad {}_{j\ 1}| \leqslant r$). Since there is no correlation bet een an data point and the preceding data points in hite noise, S_E red ces to the negati e nat ral logarithm of the probabilit that the distance bet een an t o data points is less than or eq al to r.

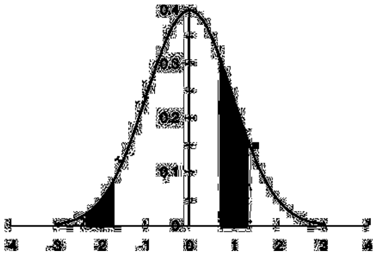

FIG. 11. Ga ssian distrib tion. Shado ed areas centered at points 2 and 1 represent the probabilit that the distances bet een each of these points and an other point chosen randoml from the time series are less than or eq al to r.

To be specific, the joint probabilit of a finite seq ence of independent random ariables is simpl

$$p(\ {}_1,\ {}_2,\ldots,\ {}_n) = \prod_{i=1}^{N} p(\ {}_i). \tag{A1}$$

One can sho that

$$\begin{aligned} P_r(|{}_i \quad {}_j| \leqslant r||{}_{i\ 1} \quad {}_{j\ 1}| \leqslant r) \\ &= \frac{P_r(|{}_i \quad {}_j| \leqslant r \wedge |{}_{i\ 1} \quad {}_{j\ 1}| \leqslant r)}{P_r(|{}_{i\ 1} \quad {}_{j\ 1}| \leqslant r)} \\ &= \frac{P_r(|{}_i \quad {}_j| \leqslant r) \times P_r(|{}_{i\ 1} \quad {}_{j\ 1}| \leqslant r)}{P_r(|{}_{i\ 1} \quad {}_{j\ 1}| \leqslant r)} \\ &= P_r(|{}_i \quad {}_j| \leqslant r). \end{aligned}$$

Using this approach rec rsi el , it can be pro ed that this res lt is alid for an m al e, hene er the ariables are independent. In this appendi , e adhere to the standard notations of sing $P_r()$ for probabilit distrib tions and $p()$ for probabilit densit f nctions.

To s mmari e, hite noise is a random process s ch that all ariables are independent. Therefore,

$$S_E = \quad \ln P_r(|{}_j \quad {}_i| \leqslant r). \tag{A2}$$

Ne t, e calc late the probabilit distrib tion $P_r(|{}_j \quad {}_i| \leqslant r)$.

For a gi en al e of , the probabilit of finding other data points ithin the distance r from is

$$P_r(| \quad | \leqslant r) = \int_{\ r}^{\ +r} p(\ )d\ . \tag{A3}$$

For e ample, if ${}_i=1$ and $r=0.3$, (Fig. 11), $P_r(|1 \quad {}_j| \leqslant 0.3)$ is the area nder the Ga ssian c r e bet een the ertical lines $=0.7$ and $=1.3$. Similarl , for ${}_i= \ 2$ and the same r al e, $P_r(|2 \quad {}_j| \leqslant 0.3)$ is the area nder the Ga ssian c r e bet een the ertical lines $= \ 2.3$ and $= \ 1.7$. Since ${}_i$ can ass me an al e bet een ∞ and $+\infty$, $P_r(|{}_i \quad {}_j| \leqslant r)$ is the a erage area centered at all possible ${}_i$ al es. In other ords,

$$\begin{aligned} P_r(|{}_j \quad {}_i| \leqslant r) &= \int_{\ \infty}^{+\infty} \left\{ \int_{{}_i\ r}^{{}_i+r} p(\ {}_j)d\ {}_j \right\} p(\ {}_i)d\ {}_i \\ &= \frac{1}{2\pi\sigma^2} \int_{\ \infty}^{+\infty} \left\{ \int_{{}_i\ r}^{{}_i+r} e^{\ {}_j^2/2\sigma^2} d\ {}_j \right\} e^{\ {}_i^2/2\sigma^2} d\ {}_i \\ &= \frac{1}{2\sigma\sqrt{2\pi}} \int_{\ \infty}^{+\infty} \left\{ \mathrm{erf}\left(\frac{{}_i+r}{\sigma\sqrt{2}}\right) \quad \mathrm{erf}\left(\frac{{}_i\quad r}{\sigma\sqrt{2}}\right) \right\} \\ &\quad \times e^{\ {}_i^2/2\sigma^2} d\ {}_i, \end{aligned}$$

here erf refers to the error f nction.

Witho t loss of generalit , e considered a ero mean ($\mu=0$) Ga ssian distrib tion. Coarse-grained hite noise time series still ha e a ero mean Ga ssian densit beca se the are the o tp t of a linear combination of Ga ssian random ariables. Ho e er, the ariance decreases as the scale factor increases,

$$\sigma_\tau = \frac{\sigma}{\sqrt{\tau}}, \tag{A4}$$

here τ refers to the scale factor, σ_τ to the ariance of the coarse-grained time series corresponding to scale τ, and σ to the ariance of the original time series (scale 1). Conseq entl , the probabilit that the distance bet een t o data points of the coarse-grained time series corresponding to scale τ is less than or eq al to r is

$$\begin{aligned} P_r(|{}_j^{\tau} \quad {}_i^{\tau}| \leqslant r) &= \frac{1}{2\sigma}\sqrt{\frac{1}{2\pi}} \int_{\ \infty}^{+\infty} \left\{ \mathrm{erf}\left(\frac{{}_i+r}{\sigma\sqrt{2/\tau}}\right) \right. \\ &\quad \left. \mathrm{erf}\left(\frac{{}_i\quad r}{\sigma\sqrt{2/\tau}}\right) \right\} e^{\ {}_i^2 r/2\sigma^2} d\ {}_i. \end{aligned}$$

The abo e e pression can be appro imated n mericall . We set the follo ing conditions for o r n merical calc lation: (1) $d\ \to \Delta\ =1/5000$; (2) the range of the integration is $[\ 3,3]=[\ (N/2)\Delta\ ,(N/2)\Delta\ ]$, ith $N=30\,000$. Th s, e ha e

$$\begin{aligned} &\frac{1}{2}\sqrt{\frac{\tau}{2\pi}} \sum_{k=\ N}^{N} \left\{ \mathrm{erf}\left(\frac{k\Delta\ +r}{\sqrt{2/\tau}}\right) \quad \mathrm{erf}\left(\frac{k\Delta\ \quad r}{\sqrt{2/\tau}}\right) \right\} \\ &\quad \times e^{[\ (k\Delta\ )^2\tau]/2}\Delta\ , \end{aligned}$$

The al es obtained ith the abo e form la are plotted in Fig. 3. These n merical al es are in good agreement ith those obtained b the MSE algorithm on sim lated hite noise time series.

Ne t, e sho the MSE deri ation for $1/f$ noise. Note that a random process ith a po er spectr m that deca s as $1/f$ is correlated. In order to n mericall calc late S_E for $1/f$ noise, e ill sho that there e ists an orthogonal transformation that maps the correlated ariables into a basis in hich the are independent. The dimension of this basis reflects the e tension of the s stem memor .

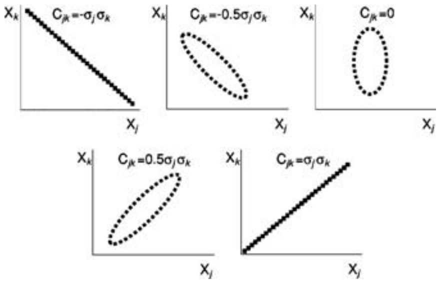

FIG. 12. Correspondence bet een the co ariance and the shape of the conto rs of a bi ariate Ga ssian densit f nction. If t o random ariables, X_j and X_k, are independent [$C_{jk}=C(X_j,X_k)=0$], the shapes of the conto rs are ellipses ith major and minor a es parallel to X_j and X_k a es, respecti el . If the ariables ha e eq al ariance ($\sigma_j=\sigma_k$), the shape of the conto r is a circle. In contrast, if t o ariables are not independent, the shapes of the conto rs are ellipses ith major and minor a es that are not aligned ith the a es X_j and X_k.

Let s consider N random ariables, $X_1,X_2,\ldots,X_N$, ith mean al es $\overline{X_j}$ for $j=1,\ldots,N$. Elements of the co ariance matri are defined b

$$C(X_j,X_k) = E[(X_j \quad \overline{X_j})(X_k \quad \overline{X_k})]. \tag{A5}$$

The diagonal elements are the ariance of each random ariable X_j, i.e., $C(X_j,X_j)=\sigma_j^2$ (see Fig. 12).

The co ariance matri is Hermitian since it is s mmetric and all of its elements are real. Therefore, it has real eigen al es hose eigen ectors form a nitar basis. Each of the eigen ectors, U_i, and the corresponding eigen al es, λ_i, satisf the eq ation

$$CU_j = \lambda_j U_j. \tag{A6}$$

Hence,

$$U_j^T C U_k = \lambda_k U_j^T U_k = \begin{cases} \lambda_k & \text{if } j=k \\ 0 & \text{if } j \neq k \end{cases}. \tag{A7}$$

Let U represent the matri hose col mns are the eigen ectors of the co ariance matri . Then,

$$U^T C U = \begin{bmatrix} \lambda_1 & 0 & \cdots & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & \cdots & \ddots & \cdots & 0 \\ 0 & \cdots & 0 & \lambda_{N\ 1} & 0 \\ 0 & \cdots & \cdots & 0 & \lambda_N \end{bmatrix} = \Lambda. \tag{A8}$$

We sho ne t that U^TCU is also the co ariance matri of the transformed ectors $Y=U^TX$, here $X=[X_1,X_2,\ldots,X_N]^T$,

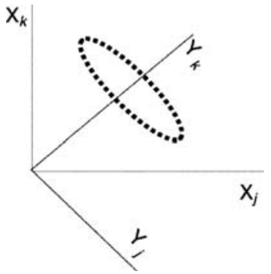

FIG. 13. The ellipse represents the contour of a bivariate Gaussian density function. The major and minor axes of the ellipse are not parallel to the axes X_j and X_k, meaning that the random variables are correlated in this frame. However, there exists a rotation that transforms the original frame into one defined by the axes Y_j and Y_k, which are aligned with the major and minor axes of the ellipse. Therefore, in this frame the original variables are uncorrelated.

$$\begin{aligned}U^TCU &= U^TE[(X-\bar{X})(X-\bar{X})^T]U = E[U^T(X-\bar{X})(X-\bar{X})^TU]\\ &= E[(U^TX-U^T\bar{X})(X^TU-\bar{X}^TU)]\\ &= E[(U^TX-U^T\bar{X})(U^TX-U^T\bar{X})^T]\\ &= E[(Y-\bar{Y})(Y-\bar{Y})^T].\end{aligned}$$

Combining this result with Eq. (A8), we prove that all transformed variables are uncorrelated in the basis formed by the eigenvectors of the covariance matrix C. Furthermore, the variances, σ_j', of the transformed variables, Y_j, are $\sqrt{\lambda_j}$.

The physical meaning of the transformation U^T is illustrated in Fig. 13. U^T is an orthogonal transformation that amounts to a rotation of the original coordinate system into one defined by the eigenvectors of the covariance matrix, in which the transformed variables are independent.

The probability density function for an n-dimensional Gaussian random vector, X, is

$$p(X) = \frac{1}{\sqrt{(2\pi)^n|C|}}e^{[-(1/2)(X-\bar{X})^TC^{-1}(X-\bar{X})]}, \qquad \text{(A9)}$$

where $|C|$ is the determinant of the covariance matrix.

For the transformed vector, $Y=U^TX$, the probability density function is

$$\begin{aligned}p(Y) &= \frac{1}{\sqrt{(2\pi)^n|\Lambda|}}e^{[-(1/2)(Y-\bar{Y})^T\Lambda^{-1}(Y-\bar{Y})]}\\ &= \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\lambda_i}}\exp\left[-\frac{(Y_i-\bar{Y}_i)^2}{2\lambda_i}\right] = \prod_{i=1}^{N}p(Y_i), \qquad \text{(A10)}\end{aligned}$$

where

$$p(Y_i) = \frac{1}{\sigma_i'\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{Y_i-\bar{Y}_i}{\sigma_i'}\right)^2\right\}. \qquad \text{(A11)}$$

In order to calculate the covariance matrix numerically, we limit the frequency range of the power spectral density, denoted as $S(\omega)$, of the $1/f$ noise signal to

$$S(\omega) = \begin{cases}K/\omega & \text{for } \omega_1 \leqslant \omega \leqslant \omega_2\\ 0 & \text{otherwise},\end{cases} \qquad \text{(A12)}$$

where K is a constant. The upper and lower limits on frequency range are useful constraints for numerical calculation and also realistic in real-world applications where the resolution (sampling frequency of signal) and length of data are bounded.

The autocorrelation function, Φ, is obtained using the Wiener-Khintchine theorem,

$$\Phi(\tau) = \frac{K}{2\pi}\int_{\omega_1}^{\omega_2}\frac{\cos\omega\tau}{|\omega|}d\omega = \frac{K}{2\pi}\{\mathrm{Ci}(\omega_2\tau)-\mathrm{Ci}(\omega_1\tau)\}, \qquad \text{(A13)}$$

where τ represents the time lag and Ci is the cosine integral. The series expansion of the Ci is

$$\mathrm{Ci}(\tau) = \gamma + \ln(\tau) + \sum_{k=1}^{+\infty}\frac{(-1)^k\tau^{2k}}{(2k)!\,2k}, \qquad \text{(A14)}$$

where $\gamma=0.5772\ldots$ is Euler's constant.

Therefore,

$$\Phi(\tau) = \frac{K}{2\pi}\left\{\ln\left(\frac{\omega_2}{\omega_1}\right) + \sum_{k=1}^{+\infty}\frac{(-1)^k}{(2k)!\,2k}\times[(\omega_2\tau)^{2k}-(\omega_1\tau)^{2k}]\right\}. \qquad \text{(A15)}$$

The autocorrelation function is the autocovariance divided by the variance. For an ergodic process, as is the case of $1/f$ noise, the relation between the autocovariance function and the covariance matrix is

$$C = \begin{bmatrix}\Phi(0) & \Phi(\tau) & \Phi(2\tau) & \cdots & \Phi(N\tau)\\ \Phi(\tau) & \Phi(0) & \Phi(\tau) & \cdots & \Phi((N-1)\tau)\\ \Phi(2\tau) & \Phi(\tau) & \Phi(0) & \cdots & \Phi((N-2)\tau)\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \Phi(N\tau) & \cdots & \cdots & \Phi(\tau) & \Phi(0)\end{bmatrix}. \qquad \text{(A16)}$$

The eigenvalues of the covariance matrix are the variances of the transformed variables. Since the variables Y_i are independent, S_E is calculated using

$$p(Y_1) = \frac{1}{\sqrt{2\pi\lambda_1}}\exp\left(-\frac{[Y_1-\overline{Y_1}]^2}{2\lambda_1}\right). \qquad \text{(A17)}$$

We consider $k=\ln(\omega_1/\omega_2)$ for numerical calculation, which corresponds to normalizing the power spectrum. We also set $\omega_1=1/(2\Delta)$ and $\omega_2=N$. The numerical calculation yields the value $S_E=1.8$. We note that coarse-graining $1/f$ noise does not alter the correlation and the variance of the signal. Therefore, the S_E value calculated is valid for any scale.

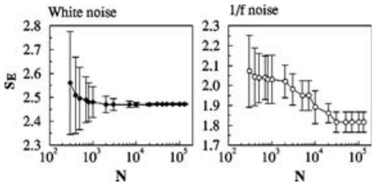

FIG. 14. S_E as a function of time series number of data points N. $r=0.15$ and $m=2$ for all time series. Symbols represent the mean values of S_E for 30 simulated white and $1/f$ noise time series, and the error bars represent the SD.

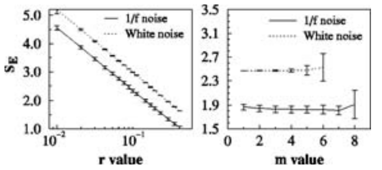

FIG. 15. S_E as a function of the parameter r (left plot) and m (right plot). $N=3\times10^4$ and $r=0.15$ for all time series. Symbols represent the mean values of S_E for 30 simulated $1/f$ and white noise time series, and error bars represent the SD.

APPENDIX B: TECHNICAL ASPECTS OF MSE CALCULATIONS

1. Dependence on time series length and the values of parameters *m* and *r*

The MSE method uses the S_E family of statistics. Therefore, in this appendix we use simulated Gaussian distributed (mean zero, variance 1) white and $1/f$ noise time series to illustrate the effects on S_E of (i) the time series finite length and (ii) the choice of parameters m and r.

Figure 14 shows that the mean value of S_E diverges as the number of data points decreases for both white and $1/f$ noise. However, since $1/f$ noise time series are not stationary, as the number of data points decreases, the discrepancy between the S_E value calculated numerically and the mean value for 30 simulated time series increases faster for $1/f$ noise than for white noise time series. For both types of noise, for $N=1\times10^5$, the discrepancy between the numerical and the mean value of S_E for simulated time series is less than 0.5%. However, for $N=1\times10^3$ the discrepancy between these values is approximately 12% in the case of $1/f$ noise but still less than 1% in the case of white noise. Furthermore, even for very large time series, the SD of S_E values for $1/f$ noise is never as small as for white noise. These results are due to the fact that stationarity is a basic requirement of S_E. The MSE method presents the same limitation. One possible solution to this problem is to decompose the original time signal into multiple well-behaved signals, each corresponding to different time scales.

We also note that as the number of data points decreases, the consistency of S_E results is progressively lost. Therefore, there is no guarantee that if S_E is higher for time series *a* than for time series *b*, both with N data points, the same result will hold if only N' data points are used to calculate S_E, in particular if $N \gg N'$ or $N' \gg N$.

We note that the coarse-graining procedure generates times series with a decreasing number of data points. However, coarse-grained time series are not a subset of the original time series. Instead, they contain information about the entire original time series. Therefore, the error due to the decrease of coarse-grained time series length is likely lower than that resulting from selecting a subset of the original time series.

As stated in Sec. II, the r value defines the similarity criterion used to compare vectors. If the absolute difference between any two matched vector components is larger than $r\times$SD, then the vectors are different; otherwise, the are considered equal. Theoretically, for continuous processes, r varies between 0 and 1; but for experimental time series, the recording resolution level determines the lowest possible r value. In any case, the actual r value determines the level of accepted noise, since for larger r values, fewer vectors are distinguishable. Figure 15 (left plot) shows that as the r value increases, the S_E value for both simulated $1/f$ and white noise time series decreases. Of note, the consistency of S_E values is preserved. Therefore, the SD of S_E values (error bars) reflects the scattering of values corresponding to different time series (intersubject variability).

Figure 15 (right plot) shows the variation of S_E with m value, i.e., the vector length. Between $m=1$ and $m=5$, the mean values of S_E vary less than 2% and the coefficient of variation (CV=SD/mean) is less than 3% for both types of noise. For larger m, both the S_E and the CV increase dramatically due to the finite number of data points, since longer and longer time series are required in order to calculate the frequency of the m and $(m+1)$-component vectors with sufficient statistical accuracy.

For a discussion of the optimal selection of m and r parameters, and the confidence intervals of S_E estimates, see [49]. We note that for $m=2$ and $r=0.15$, the discrepancies between the mean values of S_E for simulated time series and the numerically calculated values are less than 1% for both $1/f$ and white noises. This result suggests that for most practical applications, the error bars associated with computation of S_E values are likely smaller than the error bars related to experimental sources and also to inter- and intrasubject variability.

2. Effect of noise, outliers, and sample frequency

The output of an experiment may be contaminated by different types of noise. Here, we discuss the effects of MSE analysis of superimposing uncorrelated (white) noise on a physiologic time series. Common sources of uncorrelated noise for interbeat interval time series are the analog-digital conversion devices, whose accuracy depends both on the sample frequency and the number of bits used, and computer rounding errors. Figure 16 shows that (i) superimposing un-

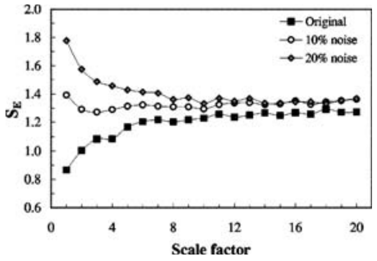

FIG. 16. Effects of different amounts of Gaussian white noise on MSE curves. The MSE curve labeled original corresponds to the MSE results for the RR intervals time series from a healthy subject.

correlated noise on a time series affects mainly the entropy values on small scales; (ii) the discrepancy between the entropy values assigned to the original time series and those assigned to time series with superimposed uncorrelated noise increases as the signal-to-noise ratio decreases; (iii) for small scales, S_E values monotonically decrease with scale factor similar to white noise time series. This effect becomes more prominent as the signal-to-noise ratio decreases.

Outliers may also affect S_E values because they change the time series SD and, therefore, the value of parameter r that defines the similarity criterion.

In the interbeat interval time series, two types of outliers are commonly found resulting from (i) missed beat detections by automated or visual electrocardiographic analysis, and (ii) recording artifacts [Fig. 18(a)]. These outliers do not have physiologic meaning. However, they may dramatically affect the entropy calculation if their amplitude is a few orders of magnitude higher than the mean value of the time series.

For the analysis of physiologic rhythm dynamics, cardiac beats not originating in the sinus node may be treated as outliers [Fig. 18(b)]. Of note, the amplitude of all cardiac (sinus and nonsinus) interbeat intervals is of the same order of magnitude. Therefore, the inclusion of a relatively low percentage of nonsinus beats should not significantly change the entropy values.

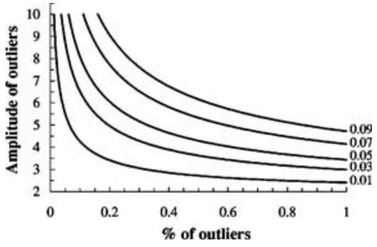

FIG. 17. Contour plot showing how the percentage of outliers and their amplitude (relative to the mean value of the time series) affects the variance of the time series. Lines connect pairs of values that change the variance by the same amount.

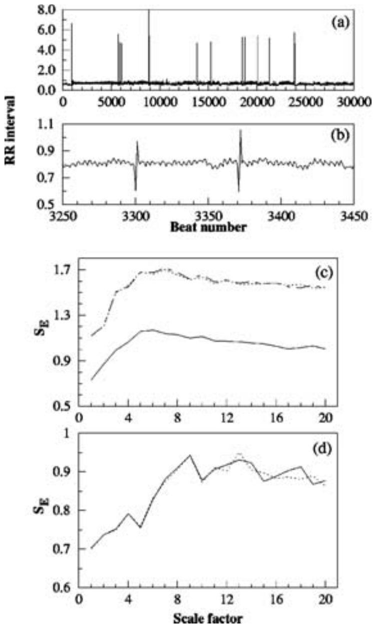

FIG. 18. (a) The interbeat interval time series of a young healthy subject with 15 outliers that represent artifacts or missed beat detections. Note that the absolute value of the outliers is much larger than the mean RR interval. (b) The interbeat interval time series of an elderly healthy subject with frequent premature ventricular complexes (PVCs) (two are represented in the figure). (c) MSE results for the time series shown in plot (a): the solid line is the MSE result for the unfiltered time series; the dotted line is the MSE results for the same time series excluding outliers; and the dashed line is the MSE result for the original time series but using an r value that is calculated by excluding the outliers. (d) MSE results for time series shown in plot (b): solid and dotted lines are the MSE results for unfiltered and filtered (PVCs removed) time series.

Consider a time series, X, with N data points, M of which are outliers with amplitude Δ. Let X' represent the time series that is obtained from the time series X by excluding the outliers. Assume that $M \ll N$ and that $\Delta = a\overline{X'}$, where $\overline{X'}$ is the time series mean value. It can be shown that $\sigma^2(X) - \sigma^2(X') = (a^2\epsilon - \epsilon^2 a^2 - 2\epsilon a)\mu(X')^2$, where $\epsilon = M/N$, and σ and μ are the time series SD and mean value, respectively.

Figure 17 shows that a small number of outliers with high amplitude has similar effects on the variance as a higher percentage of outliers with lower amplitude.

Figure 18(a) presents a time series with 0.05% outliers which account for an increase in the time series SD of about 44%. Figure 18(b) presents a time series with approximately ten times more outliers than in Fig. 18(a). Since the amplitude of the outliers is of the same order of magnitude as the remaining data points, the difference between the SD of the time series which includes these outliers and that which excludes them is only 1%.

Changes of the time series SD proportionally affect the value of parameter r. Higher r values mean that fewer vectors will be distinguishable and that the time series will appear more regular. Figure 18(c) presents the MSE results for the unfiltered time series (a) (solid line) and the corresponding time series obtained by excluding the outliers (dotted line). As expected, the MSE curve corresponding to the unfiltered time series is lower than the MSE curve corresponding to the filtered time series.

The presence of a small percentage of outliers may significantly alter the SD but should not substantially modify the temporal structure of the time series. In Fig. 18(c), the dashed line represents the MSE results for the unfiltered time series obtained using the r value derived from the filtered time series. Note that when using the correct r value, the MSE curves for the unfiltered and the filtered time series overlap.

Figure 18(d) compares the MSE results for time series (b) and for the time series that results from excluding the outliers. The two MSE curves almost overlap, showing that the entropy measure is robust to the presence of a relatively small percentage of low-amplitude outliers.

For a time series sampled at frequency f, the temporal location of the actual heartbeat can be identified only up to an accuracy of $\Delta = 1/f$. Each data point of a coarse-grained heartbeat interval time series is an average of consecutive differences. For example, $y_1^{\tau} = (RR_1 + \cdots + RR_{\tau-1})/\tau = [(t_2 - t_1) + \cdots + (t_\tau - t_{\tau-1})] = (t_\tau - t_1)/\tau$. Therefore, the accuracy of averaged heartbeat intervals of coarse-grained time series is Δ/τ, i.e., the accuracy increases with scale.

S_E is underestimated for finite sample frequency values [48]. However, the discrepancy between the value of S_E calculated for a time series sampled at a finite frequency and the value of S_E corresponding to the limit $\lim_{\Delta \to 0} S_E$ decreases with scale. For analysis on small time scales, it may be important to consider a correction of this effect [48]. We note that the conclusions that we present in this paper are not altered by the value of sample frequency.

APPENDIX C: MSE ANALYSIS OF DISCRETE TIME SERIES

Here we discuss an important artifact that affects the MSE analysis of discrete time series, such as DNA sequences.

Let us consider an uncorrelated random variable, X, with alphabet $\Theta = \{0,1\}$. Both symbols occur with probability 1/2.

All possible different two-component sequences built from the binary series are 00, 01, 10, and 11. Therefore, the alphabet of the coarse-grained time series corresponding to scale 2 is $\Theta_2 = \{0, 1/2, 1\}$. The probabilities associated with the occurrence of the different values are 1/4, 1/2, and 1/4, respectively. Let us consider that the r value used to calculate S_E is 0.5. In this case, only the distance between the coarse-grained values 0 and 1 (and not between values 0 and 1/2, and between 1/2 and 1) is higher that r. Therefore, the probability of distinguishing two data points randomly chosen from the coarse-grained time series, $P_r(|x_a - x_b| > r)$, is $p(0) \times p(1) = 1/4 \times 1/4 = 1/16 = 0.0625$.

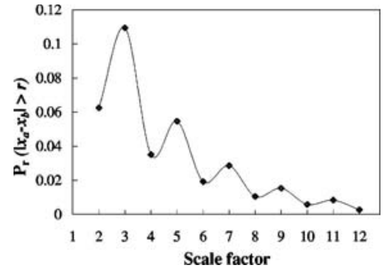

FIG. 19. Probability of distinguishing any two data points randomly chosen from the coarse-grained time series of binary discrete time series ($r = 0.5$).

Similarly, there are eight different three-component sequences that can be built from the original binary series: 000, 001, 010, 100, 110, 011, 101, and 111. Consequently, the alphabet of the coarse-grained time series corresponding to scale 3 is $\Theta_2 = \{0, 1/3, 2/3, 1\}$ and the probabilities associated with the occurrence of each value are 1/8, 3/8, 3/8, and 1/8, respectively. For $r = 0.5$, only the distances between the coarse-grained data points 0 and 2/3, 1/3 and 1, and 0 and 1 are higher than r. Therefore, $P_r(|x_a - x_b| > r) = p(0) \times p(2/3) + p(1/3) \times p(1) + p(0) \times p(1) = 0.1094$.

Note that the probability of distinguishing two data points of the coarse-grained time series increases from scale 2 to scale 3 (Fig. 19). As a consequence, S_E also increases, contrary to both analytic and numerical results presented in Fig. 3. This artifact, which affects discrete time series, is due to the fact that the size of the alphabet of the coarse-grained time series increases with scale.

In general, for scale n, the alphabet set is $\Theta_n = \{i/n\}$ with $0 \leqslant i \leqslant n$, and the corresponding probability set $\{p(i/n)\}$ is generated by the expression $n!/[2^n \times i!(n-i)!]$, $0 \leqslant i \leqslant n$. The value of $P_r(|x_a - x_b| > r)$ is calculated by the equation

$$P_r(|x_a - x_b| > r) = \sum_{j=0}^{N-1} p(j/n) \sum_{i=i'}^{n} p(i/n), \tag{C1}$$

where $i' = N + j + 1$ if $n = 2N$ (even scales) and $i' = N + j$ if $n = 2N - 1$ (odd scales).

Figure 19 shows how the probability varies with the scale factor. We note an attenuated oscillation, which as a consequence also shows up on the MSE output curve for the same time series. The period of this oscillation depends only on the r value.

To overcome this artifact, one approach is to select the scales for which the entropy values are either local minima or maxima of the MSE curve. We adopted this procedure in calculating the complexity of coding versus noncoding DNA sequences (Fig. 10). Note that for uncorrelated random binary time series (Fig. 19), and for $r=0.5$, the sequence of entropy values at odd or even scales monotonically decreases with scale factor, similar to the MSE curve for white noise time series, as described in Sec. III (Fig. 3).

An alternative approach is to map the original discrete time series to a continuous time series, for example by counting the number of symbols (1's or 0's) in nonoverlapping windows of length 2^n. Since this procedure is not a one-to-one mapping, some information encoded on the original time series is lost. Therefore, relatively long time series are required. We adopted this procedure in calculating the complexity of binary time series derived from a computer executable file and a computer data file (Fig. 9).

[1] F. Takens, in *Dynamical Systems and Turbulence*, edited by D. A. Rand and L. S. Young, Lecture Notes in Mathematics Vol. 898 (Springer, Berlin, 1981), p. 366.

[2] J.-P. Eckmann and D. Ruelle, Rev. Mod. Phys. **57**, 617 (1985).

[3] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, Physica D **58**, 77 (1992).

[4] S. M. Pincus, Ann. N.Y. Acad. Sci. **954**, 245 (2001), and references therein.

[5] P. Grassberger in *Information Dynamics*, edited by H. Atmanspacher and H. Scheingraber (Plenum, New York, 1991), p. 15.

[6] B.-Y. Yaneer, *Dynamics of Complex Systems* (Addison-Wesley, Reading, Massachusetts, 1997).

[7] M. Costa, A. L. Goldberger, and C.-K. Peng, Phys. Rev. Lett. **89**, 068102 (2002).

[8] M. Costa, A. L. Goldberger, and C.-K. Peng, Comput. Cardiol. **29**, 137 (2002).

[9] M. Costa and J. A. Healey, Comput. Cardiol. **30**, 705 (2003).

[10] M. Costa, A. L. Goldberger, and C.-K. Peng, Phys. Rev. Lett. **92**, 089804 (2004).

[11] M. Costa, C.-K. Peng, A. L. Goldberger, and J. M. Hausdorff, Physica A **330**, 53 (2003).

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

[13] R. Shaw, Z. Naturforsch. A **36**, 80 (1981).

[14] P. Grassberger and I. Procaccia, Physica D **56**, 189 (1983).

[15] P. Grassberger and I. Procaccia, Phys. Rev. A **28**, 2591 (1983).

[16] F. Takens, in *Proceedings of the 13th Colóquio Brasileiro de Matemática* (Instituto de Matemática Pura e Aplicada, Rio de Janeiro, 1983).

[17] S. M. Pincus, Proc. Natl. Acad. Sci. U.S.A. **88**, 2297 (1991).

[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991), p. 64.

[19] S. M. Pincus, I. M. Gladstone, and R. A. Ehrenkranz, J. Clin. Monit. **7**, 335 (1991).

[20] J. S. Richman and J. R. Moorman, Am. J. Physiol. **278**, H2039 (2000).

[21] P. Grassberger, T. Schreiber, and C. Schaffrath, Int. J. Bifurcation Chaos Appl. Sci. Eng. **1**, 521 (1991).

[22] D. P. Feldman and J. P. Crutchfield, Phys. Lett. A **238**, 244 (1998).

[23] Y.-C. Zhang, J. Phys. I **1**, 971 (1991).

[24] A. L. Goldberger, C.-K. Peng, and L. A. Lipsitz, Neurobiol. Aging **23**, 23 (2002).

[25] A. J. Mandell and M. F. Shlesinger in *The Ubiquity of Chaos*, edited by S. Krasner (American Association for the Advancement of Science, Washington, D.C., 1990), p. 35.

[26] M. P. Paulus, M. A. Geyer, L. H. Gold, and A. J. Mandell, Proc. Natl. Acad. Sci. U.S.A. **87**, 723 (1990).

[27] H. C. Fogedby, J. Stat. Phys. **69**, 411 (1992).

[28] V. V. Nikulin and T. Brismar, Phys. Rev. Lett. **92**, 089803 (2004).

[29] J. E. Mietus, C.-K. Peng, I. Henry, R. L. Goldsmith, and A. L. Goldberger, Heart **88**, 378 (2002).

[30] The New York Heart Association functional classification is used to characterize patients' limitations from left ventricular failure. Subjects assigned to class I can perform ordinary physical exercise with no limitations. Subjects assigned to class II are comfortable at rest but experience fatigue or shortness of breath when performing ordinary physical exercise. Class III subjects are also comfortable at rest but their ability to exercise is markedly reduced. Class IV comprises those subjects who have symptoms at rest.

[31] Time series derived from subjects with atrial fibrillation have statistical properties similar to those of white noise on shorter time scales ($\lesssim 200$ s). For more details, see [45–47].

[32] K. K. L. Ho, G. B. Moody, C.-K. Peng, J. E. Mietus, M. G. Larson, D. Levy, and A. L. Goldberger, Circulation **96**, 842 (1997).

[33] A. Bunde, S. Havlin, J. W. Kantelhardt, T. Penzel, J.-H. Peter, and K. Voigt, Phys. Rev. Lett. **85**, 3736 (2000).

[34] T. Cavalier-Smith, in *The Evolution of Genome Size*, edited by T. Cavalier-Smith (Wiley, Chichester, U.K., 1985).

[35] J. S. Mattick, BioEssays **25**, 930 (2003).

[36] J. S. Mattick, EMBO Rep. **2**, 986 (2001).

[37] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992).

[38] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. Simons, and H. E. Stanley, Physica A **221**, 180 (1995).

[39] S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **51**, 5084 (1995).

[40] B. Audit, C. Thermes, C. Vaillant, Y. d'Aubenton-Carafa, J. F. Muzy, and A. Arneodo, Phys. Rev. Lett. **86**, 2471 (2001).

[41] B. Audit, C. Vaillant, A. Arneodo, Y. d'Aubenton-Carafa, and C. Thermes, J. Mol. Biol. **316**, 903 (2002).

[42] L. A. Lettice *et al.*, Proc. Natl. Acad. Sci. U.S.A. **99**, 7548 (2002).

[43] M. A. Nobrega, I. Ovcharenko, V. Afzal, and E. M. Rubin, Science **302**, 413 (2003).

[44] Consider a time series with only two symbols: 0 and 1. The coarse-grained time series corresponding to scale τ contains the symbols $0, 1/\tau, \ldots, i/\tau, \ldots 1 (0 \leqslant i \leqslant \tau)$. If the time series is the output of a stochastic process without correlations and all values are equally probable, then the entropy of the process is $S=-\sum_{i=1}^{N} p_i \log p_i = \log N$, where N is the total number of data points. Therefore, entropy monotonically increases as the number of symbols increases.

[45] J. Hayano, F. Yamasaki, S. Sakata, A. Okada, S. Mukai, and T. Fujinami, Am. J. Physiol. **273**, H2811 (1997).

[46] W. Zeng and L. Glass, Phys. Rev. E **54**, 1779 (1996).

[47] R. Balocchi, C. Carpeggiani, L. Fronzoni, C.-K. Peng, C. Michelassi, J. Mietus, and A. L. Goldberger, in *Methodology and Clinical Applications of Blood Pressure and Heart Rate Analysis*, edited by M. Rienzo, G. Mancia, G. Parati, A. Pedotti, and A. Zanchetti (Ios Press Inc., Amsterdam, 1999), p. 91.

[48] D. E. Lake and R. J. Moorman (private communication).

[49] D. E. Lake, J. S. Richman, M. P. Griffin, and J. R. Moorman, Am. J. Physiol. **283**, R789 (2002).