

# Entropy, Complexity, Predictability and Data Analysis of Time Series and Letter Sequences

Werner Ebeling<sup>1</sup>, Lutz Molgedey<sup>1</sup>, Jürgen Kurths<sup>2</sup> and Udo Schwarz<sup>2</sup>

<sup>1</sup> Institute of Physics, Humboldt–University, D-10115 Berlin

<sup>2</sup> Institute of Physics, Potsdam University, D-14469 Potsdam

December 23, 1999

**Abstract.** The structure of time series and letter sequences is investigated using the concepts of entropy and complexity. First conditional entropy and transinformation are introduced and several generalizations are discussed. Further several measures of complexity are introduced and discussed. The capability of these concepts to describe the structure of time series and letter sequences generated by nonlinear maps, data series from meteorology, astrophysics, cardiology, cognitive psychology and finance is investigated. The relation between the complexity and the predictability of informational strings is discussed. The relation between local order and the predictability of time series is investigated.

## 1 Introduction

The category of entropy was introduced in 1864 by Rudolf Clausius into physics and in a different context in 1949 by Claude Shannon into information theory. Applications to the structure of sequences were already given by Shannon [66], who published in 1951 the seminal paper on "Predictions and Entropy of Printed English". Later Shannon's approach was applied also to other languages [81, 18], to biosequences and to many other information carriers [25, 28, 11, 12, 82, 15, 16, 37, 30, 31, 48, 32, 18, 3]. The extension of Shannon's concept to the investigation of dynamic processes is due to Kolmogorov and Sinai [41, 69].

The concept of the Kolmogorov-Sinai entropy belongs to the key concepts of the modern theory of dynamical systems [64].

A few years later Kolmogorov developed a concept for the characterization of the complexity of sequences [42]. Several related concepts were developed later [47, 39, 21, 1, 77, 29, 65, 43, 44, 22].

Entropy and complexity concepts provided new tools for the investigation of irregular time series which play a great role in many branches of science [38].

Our investigation of time series is restricted on the concepts of conditional entropies, mutual information and complexity, as well as on certain generalizations.

Our working hypothesis is that many time series and most information carriers as texts, pieces of music and biosequences are not first order Markov processes both have higher order correlations. Further we expect some structural analogies to strings generated by nonlinear processes.

In some cases we expect the existence of long-range correlations. This hypothesis was checked by the analysis of the behavior of the dynamic entropies, the transinvariances and other correlation measures [16, 31, 10, 83]. Here this line of investigations will be continued. In comparison to our earlier work special attention is paid here to

## 2 Conditional entropies and predictability

In physics the entropy concept is connected with the names of Boltzmann, Gibbs, Einstein, Onsager, Prigogine and others. The relation between physical and information-theoretical concepts has been discussed by Maxwell, Szilard, Brillouin and other workers [18, 3].

Here we are mainly concerned with the applications of the entropy concept to time series and to information carriers. In order to proceed let us assume that the processes or structures to be studied are modeled by trajectories on discrete state spaces having the total length  $L$ . Let  $\lambda$  be the length of the alphabet. Further let  $A_1 A_2 \dots A_n$  be the letters of a given subtrajectory of length  $n \leq L$ . Let further  $p^{(n)}(A_1 \dots A_n)$  be the probability to find in the total trajectory a block (subtrajectory) with the letters  $A_1 \dots A_n$ . Then we may introduce the entropy per block of length  $n$ :

$$H_n = - \sum p^{(n)}(A_1 \dots A_n) \log p^{(n)}(A_1 \dots A_n) \quad (1)$$

From the block entropies we derive the conditional (dynamic) entropies by the definition

$$h_n = H_{n+1} - H_n \quad (2)$$

Further we define  $r_n = 1 - h_n$  as the average predictability of the state following immediately after a measured  $n$ -trajectory.

These quantities are called by Shannon  $n$ -gram entropies. The limit of the conditional entropy for large  $n$  is the entropy of the source (Kolmogorov - Sinai entropy). We have seen, that the predictability of processes is closely connected with the conditional (dynamic) entropies. Let us consider now certain section of length  $n$  of the trajectory, a time series or another sequence of symbols  $A_1 \dots A_n$ , which often is denoted as a subcylinder. We are interested in the uncertainty of the predictions of the state following after this subtrajectory of length  $n$ . We define now the expression

$$h_n^{(1)}(A_1 \dots A_n) = \sum p(A_{n+1}|A_1 \dots A_n) \log p(A_{n+1}|A_1 \dots A_n)^{-1} \quad (3)$$

as the uncertainty of the next state (1 step into the future) of the state following behind the measured trajectory  $A_1 \dots A_n (A_i \in \text{alphabet})$ . Here and in the following all logs are measured in  $\lambda$ -units. We note that in these units the inequality holds:

$$0 \leq h_n^{(1)}(A_1 \dots A_n) \leq 1 \tag{4}$$

Further we define

$$r_n^{(1)}(A_1 \dots A_n) = 1 - h_n^{(1)}(A_1 \dots A_n) \tag{5}$$

as the predictability of the next state following after a measured subtrajectory, which is a quantity between zero and one. [19, 20, 63].

We note that the average of the local uncertainty

$$h_n = h_n^{(1)} = \langle h_n^{(1)}(A_1 \dots A_n) \rangle \tag{6}$$

$$= \sum p(A_1 \dots A_n) h_n^{(1)}(A_1 \dots A_n) \tag{7}$$

leads us back to Shannon's n-gram conditional entropy. Let us consider other possible generalizations. If we want to predict the state which follows not immediately after the observed n-string, but only after k steps into the future we may define the quantity

$$h_n^{(k)}(A_1 \dots A_n) = \sum p(A_{n+k} | A_1 \dots A_n) \log p(A_{n+k} | A_1 \dots A_n)^{-1} \tag{8}$$

This is the uncertainty of the state which occurs  $k$  steps into the future after the observation of an  $n$ -block, or symbolically

$$[A_1 \dots A_n](k - 1 \text{ states})[?] \tag{9}$$

Further we may define accordingly the local predictabilities [19]

$$r_n^{(k)}(A_1 \dots A_n) = 1 - h_n^{(k)}(A_1 \dots A_n) \tag{10}$$

$$\tag{11}$$

For  $n = 1$  the average predictability is closely related to the transinformation (mutual information) [33, 34] and the information flow [63]. The mutual information can be expressed by our predictabilities by

$$I(k) = r_1^{(k)} + r_0$$

where  $r_0 = 1 - H_1$  is the predictability of a letter, if no preknowledge is available. For systems with long memory it makes sense to study the a whole series of predictabilities with increasing n-values (where n is denoted by the lower index)

$$r_1^{(k)}, r_2^{(k)}, r_3^{(k)}, \dots, r_m^{(k)}$$

where  $m$  is an estimate for the length of the memory. Since

$$r_{n+1}^{(k)} \geq r_n^{(k)}$$

the average predictability may be improved by taking into account longer blocks. In other words, one can gain advantage for prediction by basing the predictions not only on actual states but on whole trajectory blocks which represent the actual state and its history. Let us mention that the conditional entropies may be exactly calculated for several model systems [28, 12, 17]. In our empirical investigations described below we considered only the transinformation, also called mutual information, which is a special case of our concepts. The transinformation is explicitly defined as

$$I(n) = \sum_{A_i A_j} p^{(n)}(A_i, A_j) \log \left[ \frac{p^{(n)}(A_i, A_j)}{p^{(1)}(A_i) \cdot p^{(1)}(A_j)} \right] \quad (12)$$

For completeness let us further define  $I(0) = H_1$  and  $I(-n) = I(n)$ . The transinformation is a special measure of correlations [37, 30, 31, 33, 34] which is closely related to the autocorrelation function [57, 71].

For our analysis the following relations between the entropies and predictabilities defined above and the transinformation are of special importance.

$$h_1 = H_1 - I(1) \quad (13)$$

$$r_1^{(n)} = I(n) + r_0 \quad (14)$$

In other words, the predictability of a letter  $n$  steps ahead is the sum of the mutual information and the overall predictability of letters  $r_0 = 1 - H_1$ . As shown by several authors [25, 37, 33, 76], the transinformation is also a reliable measure for the correlations of letters in the distance  $n$ . Every peak at  $n$  corresponds to a strong positive correlation. [33, 34]

### 3 Concepts of complexity

Observational data from astrophysical, geophysical or physiological experiments are typically quite different from those obtained in laboratories. Often we have rather short, noisy, irregularly sampled time series. More important, however, is that non-stationary and very complicated behavior in time is usually observed. In such cases, well-known global characteristics of the underlying processes, such as periodicities or fractal dimension, do not provide a sufficient description. With respect to modeling often the question arises

whether the data have any structure at all, for example correlations, and of what kind the structure is. The concept of complexity is an appropriate approach to analyze such data.

During the past decade numerous definitions of complexity have been proposed (e.g. [42, 77, 28, 54, 7, 4, 51, 2, 24, 68, 50]) and successfully used in various fields, ranging from information processing (e.g. [42, 6], and theory of dynamical systems (e.g. [77, 28, 80]) to thermodynamics (e.g. [7, 8, 68]), astrophysics [29, 65], geophysics [79], evolution theory (e.g. [46, 67]) and medicine diagnostics (e.g. [44, 60]). Recently, some generalizations of this approach to analyze two-dimensional objects have been proposed [26, 60, 27]. Many of these definitions rely on the intuitive impression that complexity should reflect some hidden order of a phenomenon, which nevertheless possesses a certain degree of randomness. Neither well-ordered nor completely disordered objects are seemingly complex; thus complexity appears somewhere at the borderline between disorder and order. Formally, this implies that complexity is a convex function of the disorder, provided the latter is appropriately defined (see [68] for a discussion).

Quite commonly, the definition of disorder is based on the comparison of the Boltzmann – Gibbs – Shannon entropy  $H_1$  (Eq. 1) with the maximal possible entropy of the system  $H_1^{max}$ . The value of the maximal entropy  $H^{max}$  depends on the nature of the system, but for the simplest case when  $N$  states are available, the maximal entropy is achieved for the equiprobable distribution:

$$H^{max} = \log_2 N. \quad (15)$$

It is important to note that Shannon entropy is a measure of randomness, i.e. it assigns highest complexity for white noise-like behavior, where past and future are uncorrelated. Other popular measures, especially the algorithmic complexity [39] or the approximative entropy [58] have the same property; we call this class traditional measures of complexity.

Such a characterization is, however, not sufficient for many systems, especially in nonlinear dynamics. We, therefore, present another kind: non-traditional or alternative measures of complexity which relate highest complexity at phase transitions, e.g. the onset of chaos.

A straightforward notion of such an alternative measure has been recently proposed by combining disorder and order. The disorder is defined as  $H/H^{max}$  and correspondingly the disorder-based complexity  $\Gamma_{\alpha,\beta}$  [68]:

$$\Gamma_{\alpha,\beta} = (1 - H/H^{max})^\alpha (H/H^{max})^\beta \quad (16)$$

For  $\alpha > 0, \beta > 0$ ,  $\Gamma_{\alpha,\beta}$  is a convex function of the disorder. Other values of these parameters may correspond to alternative definitions of complexity [68].

The relevance of the complexity measure  $\Gamma_{\alpha,\beta}$  as introduced by Eq. (16) has been demonstrated in application to the logistic map and to one-dimensional spin-systems [68]. We however wish to stress two features of the definition

(16): (i) it exploits a concept of “maximal possible entropy”, which for some systems may not be easily computed and even unambiguously defined, and (ii) it lacks in accounting for inherent correlations in the system, which are certainly an important component of order, and thus of complexity. By construction this measure relates zero complexity to most random behavior, e.g. equiprobable distribution as well as to simple ordered states.

Grassberger [28] introduced another approach to complexity which is based on the differences of block entropies  $h_n$  (Eq. 2). This Effective Measure Complexity (EMC) is defined as

$$EMC = \sum_{n=1}^{\infty} n (h_{n-1} - h_n) \quad (17)$$

$EMC$  describes the behavior of the local difference  $h_n$  as it converges toward the dynamical entropy of the dynamical system. It can also be written as an average Kullback information, for instance in terms of conditional probabilities, as demonstrated in [49].

It is easy to see that EMC vanishes for both, for most random case, such as white noise, and for constant symbolic strings. It goes to infinity in period doubling sequences, i.e. it goes to infinity along this typical route to chaos.

One of the most interesting complexity measures is the renormalized entropy, originally introduced by Klimontovich [40] in thermodynamics. It takes into account that the energy of an open system changes with its control parameter, which makes a direct comparison of Shannon entropies impossible. The main idea is that the Shannon entropy for different system states is normalized to a fixed value of mean effective energy. This approach, loosely speaking, renormalizes the entropy obtained from a time series  $x(t)$  of a certain system state in such a manner that the mean effective energy coincides with that of a reference state  $x_r(t)$ .

Starting from these two time series, we can easily estimate the corresponding probability distributions  $f(x)$  and  $f_r(x)$ . By using formal arguments from thermodynamics the effective energy is defined as:

$$h_{eff}(x) = -\log f_r(x) \quad (18)$$

The renormalization of  $f_r$  into  $\tilde{f}_r$  is constructed such that the mean effective energies  $\langle h_{eff} \rangle$  of  $f$  and  $\tilde{f}_r$  are equal. To make this idea operational, we first represent the distribution in terms of the canonical Gibbs distribution

$$\tilde{f}_r(x) = \exp\left(\frac{\Phi(T_{eff}) - h_{eff}(x)}{T_{eff}}\right) \quad (19)$$

which can be rewritten as

$$\tilde{f}_r(x) = C(T_{eff}) \cdot \exp\left(-\frac{h_{eff}(x)}{T_{eff}}\right), \quad (20)$$

where  $T_{eff}$  and  $\Phi(T_{eff})$  are the effective temperature resp. the free effective energy. Because  $h_{eff}$  can be calculated from Eq. (18), there are two unknowns

in Eq. (20):  $C(T_{eff})$  and  $T_{eff}$ . They are determined from the following two conditions.

a) Normalization:

$$\int \tilde{f}_r(x) dx = 1$$

b) Equality of mean effective energy:

$$\int h_{eff}(x) \tilde{f}_r(x) dx = \int h_{eff}(x) f(x) dx.$$

Hence,  $\tilde{f}_r$  fulfills the properties wanted. Consequently, we can compare the Shannon entropies of  $f$  and  $\tilde{f}_r$

$$H = - \int f(x) \log f(x) dx \quad \text{and} \quad \tilde{H}_r = - \int \tilde{f}_r(x) \log \tilde{f}_r(x) dx \quad (21)$$

For that the renormalized entropy difference

$$\Delta \tilde{H} = H - \tilde{H}_r \quad (22)$$

is introduced. It is important to note that  $\Delta \tilde{H}$  is a relative measure that depends on the reference state chosen.

If this reference state is chosen suitable, then  $\Delta \tilde{H}$  relates smaller values to periodic and random behavior than to chaotic dynamics [59].

There are some other alternative measures of complexity, such as the epsilon-complexity [7] or the fluctuation complexity. We compared the properties of these measures in detail for the logistic map [77] and found that there is till now no outstanding alternative measure of complexity; each of them is sensitive against certain structural changes. The proper choice of these measures is context-dependent. Therefore we recommend to compare the proper choice of these measures in each special application. We will demonstrate in Chap. 6 how efficient these measures can be used in several applications.

## 4 Applications to biosequences and other information carriers

In genetic data banks one can find nowadays a very large number of DNA-sequences and there is an urgent need for the development of tools which formalize their analysis. Formally, DNA-sequences are linear string written on an alphabet consisting on four letters

$$X = A, C, G, T \quad (23)$$

A genom contains about 1 - 100 billion nucleotides and corresponds therefore to a very long string, which however consists in general of several pieces.

The strings which are available for a statistical analysis comprise in general  $10^4 - 10^6$  letters. Since Gatlins pioneering work "Information theory and the living system" [25], the calculation of entropic measures for genetic sequences as conditional entropies and transinormations has found many fruitful applications [37, 30, 31, 13]. In particular we mention the development of criteria to differ between coding and noncoding regions [33, 34] For genetic sequences several authors have pointed out the existence of long range correlations [37, 31, 76, 71, 57, 10].

However an analysis of the average uncertainties (the dynamic entropies) yields rather high values. Measured in bits the limit uncertainty is in most cases larger than 1.8 bit, i.e. larger than 0.9 in  $\lambda$  units [25, 30]. For this reason the average dynamic entropies do not seem to be the appropriate instrument to analyze DNA-strings. However as we can show, local investigations of the entropy and the transinormation might be very powerful for the analysis of long correlations. In [14] we presented the lowest uncertainties of predictions for the interval 9800 - 10300 of the DNA of the virus HIV2BEN. In spite of the fact that the average uncertainty is rather high  $h_3 = 0.94$ , we may find special positions where the uncertainty is lower than 0.83 [14]. For example, if we observe the triple gtc, then with large probability an a or a g is expected to follow. If however the triple aac is fold then the most probable continuation is either a or c. These simple rule increase locally the predictability. Since for DNA a huge material on the mutual information is available [37, 30] we will not go here into further detail. Let us only mention that DNA-strings show some (formal) analogies to texts.

Let us discuss now in brief several results which are available for texts. We have studied for example MELVILLE's Moby Dick ( $L \approx 1,170,200$ ) and GRIMM's Tales ( $L \approx 1,435,800$ ). Our methods for the analysis of the entropy of sequences were in detail explained elsewhere [14]. We have shown that at least in a reasonable approximation the scaling of the entropy against the word length is at large  $n$  given by a root law.

For example a reasonable fit of the data obtained for texts on the 32-alphabet (measured in  $\log(32)$  units) reads

$$h_n \approx (0.25/\sqrt{n}) + 0.07. \quad (24)$$

The dominating term is given by a root law corresponding to a rather long memory tail. We mention that a scaling law of the root type was first found by Hilberg who made a new fit for Shannon's original data [35]. We used our own data for  $n = 1 \dots 26$  but included also Shannon's result for  $n = 100$ . The slow decay of the conditional entropies may be interpreted by the existence of long correlations in texts.

Let us now briefly summarize results obtained from using other measures of correlations [10]. At first we have calculated the algorithmic entropy according to Lempel and Ziv which is introduced as the relation of the length

of the compressed sequence (with respect to a Lempel–Ziv compression algorithm) to the original length. The results obtained for the Lempel–Ziv complexities (entropies) of several DNA-sequences and for texts were compared with the exponents of the mean square fluctuation of the composition and with diffusion exponents [10]. Further we studied also the power spectrum which is defined as the Fourier transform of the correlation function. The results of spectra calculations for the original file of the Bible, for Moby Dick and for the same files shuffled on the word level or on the letter level correspondingly were presented in a foregoing work [10].

We have shown that the spectra of the original texts have a characteristic shape with a well-expressed low frequency part. This shows again the existence of long-range correlations in texts [10]. Similar results were obtained for DNA-sequences [71, 57]. We see from this analysis that DNA-sequences also show some type of long correlations. However as shown in [10] the long-range correlations are only due to slow changes in the composition. As a matter of fact, the composition (the letter content) of DNA (and also of texts) fluctuates slowly with a wave length of  $10^2 - 10^3$ . This fluctuation of the composition is evidently the main reason for the observed characteristic exponents. Shuffling destroys these properties [10].

Let us go now to the investigation of protein sequences [14]. As well known protein structures play a fundamental role in all living processes [14]. The building blocks of the proteins are the 20 amino acids which we denote by the letters of the alphabet

$$X = A, C, E, G, \dots, Y, W \quad (25)$$

In this way the primary structure of a protein can be mapped to a linear string on an alphabet with 20 letters. Typically the protein strings have a length between  $10^2$  and  $10^4$ . In other words, protein strings are much shorter than DNA-strings. Further, protein string show a very high degree of randomness [21] but nevertheless they contain also many intriguing informations connected with their function and their history [5]. All this together leads to serious difficulties in the statistical analysis of the primary structure of proteins. A possible way to reduce these difficulties is to use reduced alphabets [36].

Here we shall use nevertheless the full alphabet; this restricts our investigation to a rather small statistical significance. As a prototype we considered in an earlier work [14] the sequences pacvspen and paphuman which have a length  $L \simeq 5000$ . We calculated the transinformation for the protein pacvspen and the transinformation for the protein paphuman. The predictability was obtained by adding  $r_0 = 0.05$ . We obtained several well expressed peaks which show structural regularities where the predictability is a little bit better than on other places.

Looking at the length and at the size of the alphabet, protein sequences show at least a formal similarity to musical strings. Therefore we have made a comparison of the transformation of these both types of information carriers [14]. In the mentioned work [14] we calculated the transformation and the predictabilities for the Beethoven Sonatas 10 no 2 and no 3., for the Beethoven Sonatas 28 and 48 and for Mozart's pieces KV 311 and 330. The peaks show that there exist strong correlations between two notes at certain distances. In this way several -from far analogies between pieces of music and protein sequences have been shown.

## 5 Applications of entropy concepts to data analysis

Prediction of strong noisy data using classical linear methods usually fails to give accurate and reliable confidence level of the prediction. Moreover the linear methods are dominated by the most frequent events. However predictability may not be constant in time and even higher for rare events. The concept of entropy and local predictability in combination with classical methods is good candidate to give reliable results. Applications of these concepts to meteorological strings were given in [55, 78] and to nerve signals in [13].

In the following our concept will be demonstrated on daily stock index data  $S_t$ : Dow Jones 1900-1999 (27044 trading days). Since the stock index itself has an exponentially growing trend one uses daily logarithmic price changes

$$x_t = \ln(S_t) - \ln(S_{t-1}) \quad . \quad (26)$$

A direct application of the entropy concept requires a partitioning of the real value data  $x_t$  into symbols  $A_t$  of an alphabet having the length  $\lambda$ . To find an optimal partition and alphabet is a process of maximizing the Kolmogorov-Sinai entropy. However for strong noisy signals with short memory an equal frequency of the letters is near to optimal. To be concrete  $\lambda = 3$  and  $A_t = 0$ ;  $x_t < -0.0025$  (strong decrease in the stock value),  $A_t = 2$ ;  $x_t > 0.0034$  (strong increase),  $A_t = 1$  (intermediate) were chosen.

The result of the calculation of the local uncertainty  $h_n(A_1, \dots, A_n)$  for the next trading day following behind an observation of  $n$  trading days  $A_1, \dots, A_n$  according (3) for  $n = 5$  is plotted in Fig. 1. The local uncertainty is almost near one, i.e. the local predictability is almost very small. However behind certain patterns of stock movements  $A_1, \dots, A_n$  the local predictability reaches 8% – a notable value for the stock market, which is usually pure random. The mean predictability over the full data set is less than 2% (see Fig. 2).

The question of the significance of the prediction is treated by calculating a distribution of local uncertainty  $h_n^S(A_1, \dots, A_n)$  by help of surrogates. The

surrogate sequences have the same two point probabilities  $p^{(2)}(A_2|A_1)$  as the original sequence [73, 61, 62, 56, 9]. The level of significance  $K$  is calculated as

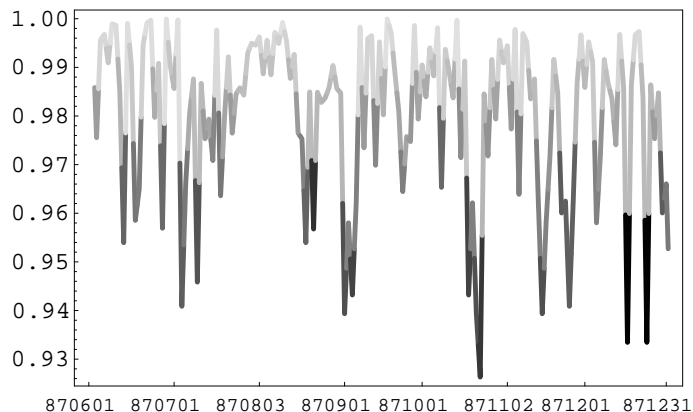
$$K_n(A_1, \dots, A_n) = \frac{h_n(A_1, \dots, A_n) - \langle h_n^S(A_1, \dots, A_n) \rangle}{\sigma}, \quad (27)$$

where  $\langle h_n^S(A_1, \dots, A_n) \rangle$  is the mean and  $\sigma$  is the standard deviation of the local uncertainty distribution for the word  $A_1, \dots, A_n$ .

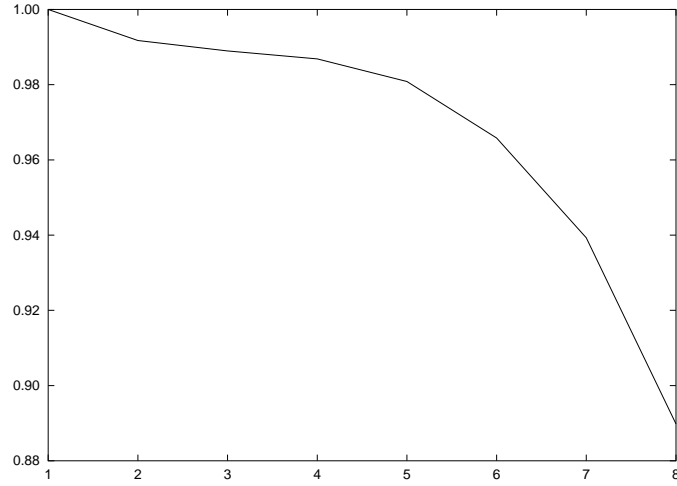
Assuming Gaussian statistics  $|K| \leq 2$  represents confidence greater than 95%. However the local uncertainty distribution is more exponential distribution like. Therefore larger  $K$ -values are required to guarantee significance. For the analyzed data set a word length up to 6 seems to give reliable results.

Fortunately higher local predictabilities coincides with larger levels of significance as seen in Fig. 1 and from Table 1.

Since we used a timeseries over a very long period we have to address the problem of non-stationary by dividing the original timeseries into smaller pieces. Furthermore instead of producing surrogates on the level of Symbols one can discuss surrogates obtained by models of a stockmarkets like ARCH/GARCH-models[52]. This has been done in [53].



**Fig. 1.** Local uncertainty of the 6th symbol when seen 5 symbols for the second half of 1987. The grey value codes the level of significance calculated from a surrogate with memory of 2. Dark represents a large deviation from the noise level (good significance). Note the higher predictability following the October-Crash.



**Fig. 2.** Conditional entropy  $h_n = H_{n+1} - H_n$  as a function of word length  $n$

word	uncert.	K	word	uncert.	K	word	uncert.	K
020	0.971	-27.9	1112	0.954	-14.3	11110	0.919	-9.5
112	0.971	-30.5	0000	0.957	-12.0	11120	0.926	-9.1
110	0.977	-23.4	1110	0.958	-15.0	20000	0.926	-6.9
120	0.981	-29.4	0110	0.960	-14.9	11112	0.931	-8.7
000	0.982	-18.5	0020	0.961	-12.5	10120	0.933	-5.2
212	0.983	-19.5	1102	0.962	-12.4	22202	0.933	-9.4
202	0.984	-22.9	2020	0.966	-10.6	00000	0.934	-6.4
111	0.985	-20.1	0200	0.968	-10.6	11011	0.937	-9.7
121	0.985	-12.4	0202	0.969	-14.6	02000	0.939	-5.7
012	0.987	-14.2	0120	0.971	-12.3	02020	0.941	-5.1
102	0.988	-10.5	2112	0.971	-9.0	00020	0.943	-6.4

**Table 1.** Words with the highest predictability

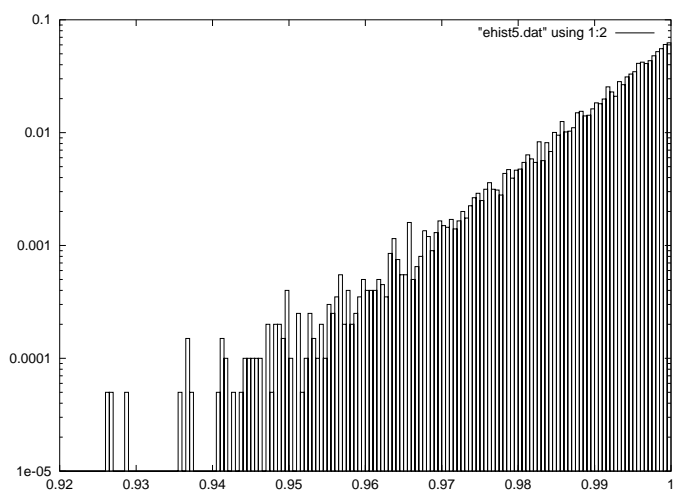


Fig. 3. Local uncertainty distribution of the surrogate sequence for the word 11110

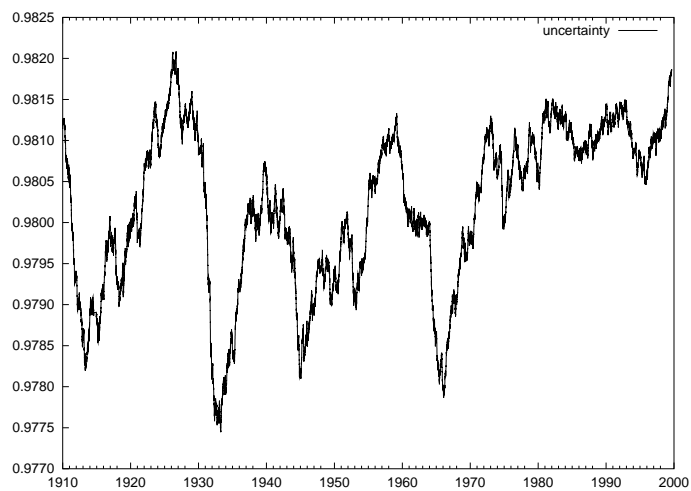
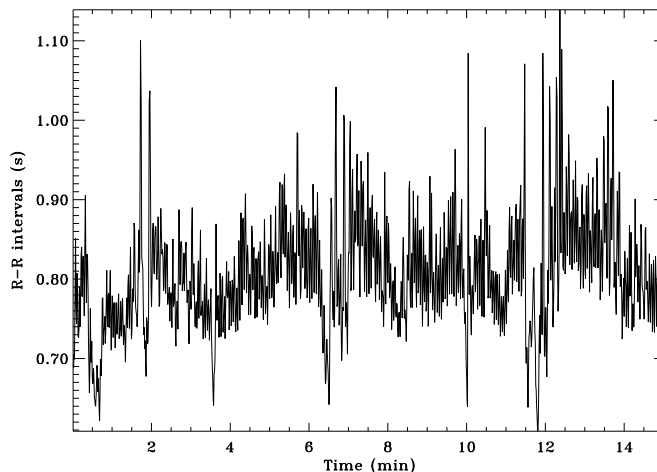


Fig. 4. Moving exponential average of the local uncertainty  $h_5$  with an halfife period of 5 years for the full dataset.

## 6 Applications of complexity concepts

We have applied the concept of measures of complexity to experimental data where other tools of linear as well as nonlinear data analysis fail. The areas of application range from astro- and geophysics, via physiology to cognitive psychology [29, 43, 44, 59, 65, 79, 23, 26, 60, 27].



**Fig. 5.** Tachogram of a healthy person.

Such experimental data consists usually of real numbers. Therefore, the first necessary step is to transform them into a symbolic string, i.e. the data are transformed into a series of the same length but the elements are only a few symbols. In doing so one loses a certain amount of detailed information, but some of the invariant, robust properties of the dynamics are retained. In the best case, such a transformation generates a Markov partition. However, in most examples of natural systems we know neither the existence of such a partition nor their construction. Therefore, more pragmatic transformations have to be used which may be not Markovian ones [77, 75].

We distinguish static transformations, where the transformation is based on a few thresholds (see application to cardiology) and dynamic ones, where we consider the step-to-step difference of data points adjacent in time (see application to cognitive complexity). The choice of the kind of transformation is of particular importance. In case of rather small data records, as it typically occur in applications, a transformation into only few symbols is to recommend. This manner, the transformation is context-dependent. If possible,

such a coarse-graining should, hence, be based on some physical motivation. Otherwise, we highly recommend to compare different transformations into symbols.

### **Cardiology: Detecting High-Risk-Patients for the Sudden Cardiac Death**

Every year several 100,000 persons die due to the sudden cardiac death that is caused by cardiac fibrillation. Without any warning it even can occur by subjects who are up to this time apparently healthy or medically inconspicuous. By use of conventional methods only 30% of all affected persons are diagnosed as high-risk-patients. This is an important challenge to nonlinear dynamics.

The aim of our investigation is to get by means of nonlinear dynamics a clear improvement of the detection rate of persons with a high risk for the sudden cardiac death. An essential point is the finding of new parameters, which describe the complex processes and their interactions for detecting those high-risk-patients, who could not recognized by traditional - mostly linear methods.

The basis of our analysis is the heart-rate-variability (HRV) which is yielded from 30 minutes and 24-hour ECG-measurements, i.e. from non-invasive methods. It is important to emphasize even in case of healthy volunteers, the variability shows a broad variety of structures. This is essentially caused by the fact that the system which generates the HRV has to be considered as an open one whose energy changes temporally as well as from person to person.

The following application of the concept of complexity leads to an improved risk stratification: From various tests it comes out that for our purpose at least 4 different symbols are necessary. The most appropriate transformation is a static one:

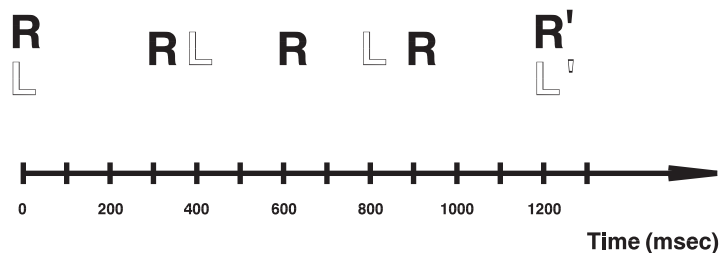
$$s_i = \begin{cases} 0 & \text{if } t_i > (1+a)\mu \\ 1 & \text{if } \mu < t_i \leq (1+a)\mu \\ 2 & \text{if } (1-a)\mu < t_i \leq \mu \\ 3 & \text{if } t_i \leq (1-a)\mu \end{cases} \quad (28)$$

where  $t_i$  are the RR-intervals,  $\mu$  is their mean value and  $a = 0.1$ .

To analyze the symbolic strings, Shannon and Renyi entropies of length-3 words are calculated. As expected, the Shannon entropy is not so useful as the generalized Renyi entropies. We use in particular the  $H_k^{(q)}$  for  $q = 0.25$  and for  $q = 4$  to describe the complexity. It is interesting to note that already the distribution of length-3 words yields a criterion for a distinction of both groups: For persons with high cardiac risk, this distribution is mainly concentrated on about 10 words (of 64 possible ones), whereas healthy persons are characterized by a more uniform distribution.

The renormalized entropy is especially related to compare different states of one system. An important problem of its application is to choose a suitable reference state, i.e. special person in our case. We choose that healthy

proband as reference person who has the largest renormalized entropy. Note that this choice does not sensitively influence the results. From this, we indeed get an indication for high cardiac risk in two directions. If  $\Delta\tilde{H}$  (Eq. 22) is very low, a strongly reduced variability is expressed and, on the other side, if  $\Delta\tilde{H}$  is rather large, an exceptional variability is indicated.



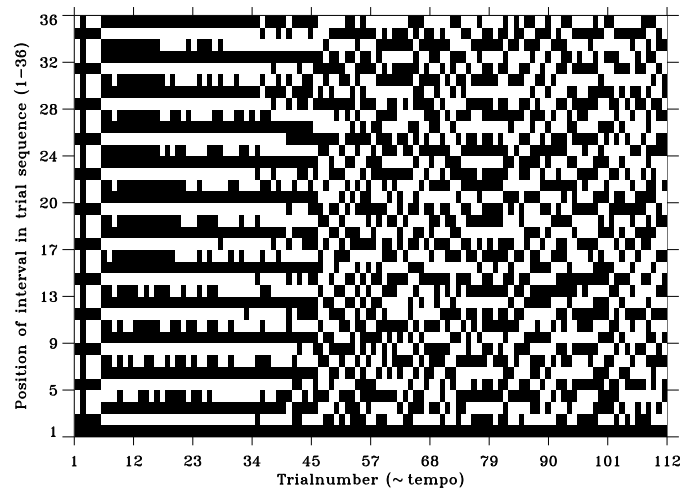
**Fig. 6.** Bimanual rhythm production: The figure illustrates the sequence of finger movements and their ideal timing for a polyrhythm (3 beats per cycle in the left hand versus 4 beats per cycle in the right hand) at a cycle duration of 1200 ms. R refers to the right, L to the left index finger. A cycle starts with simultaneous strokes of the two fingers.

It is important to note that none of these measures of complexity alone is sufficient to describe the risk. Therefore, we combine several measures of complexity with traditional methods from time and frequency domain. Applying this combination method to ECG measurements from 572 survivors of acute myocardial infarction, we get a significant better prediction of high arrhythmia risk than the standard measurement of global heart rate variability [74].

### **Cognitive Psychology: Synchronization and Coordination of Movements**

The abilities to perform precise movements, coordinate movements between different limbs, or adjust them to external performance constraints, constitute general, but highly complex human capacities. Biologists and psychologists have for a long time been interested in these capacities in order to gain insights into the functionality of the central nervous system. More recently, the dynamic systems perspective has been applied to a number of phenomena related to motor control [72] and human development of cognition and action [70].

Krampe et al. [45] investigated the production of bimanual rhythms in a large number of subjects differing in musical (pianist) skills, and also age. One task in these experiments required polyrhythmic performance, that is, the combination of different rhythms in the two hands (see Fig. 6).



**Fig. 7.** Phase shift in the performance of a 3 against 4 polyrhythm as a function of tempo in one subject. Trials have been sorted by tempo for this illustration. The trial number is provided by the x-axis and corresponds to the external control parameter (tempo); trial number 1 refers to a cycle duration of 800 ms, trial number 112 to a tempo of 8200 ms per cycle. The position of a single interval within a given trial is provided by the y-axis. Only intervals from the left hand (1-36) are shown. Black pixels indicate intervals which are longer than their immediate predecessor, otherwise a white pixel is set. A change of the pattern can be observed in a region around trial number 47, which corresponds to a performance tempo of 1400 ms per cycle.

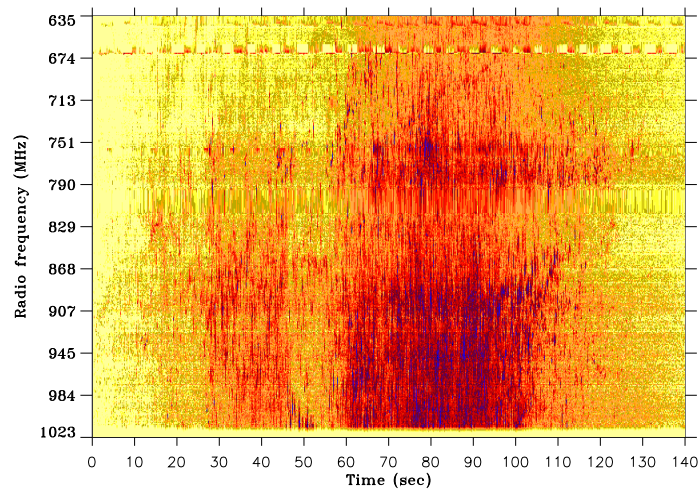
The task was performed on two keys of an electronic piano connected to a computer which measured the time intervals between successive keystrokes. The external control parameter, that is the prescribed tempo, was experimentally varied across trials for each subject between 800 ms and 8200 ms cycle duration; for highly skilled subjects, performance was assessed at speeds as fast as 500 ms per cycle. After a short synchronization phase during which subjects could play along with the rhythm generated by the computer and adjust to the prescribed tempo, participants had to continue their performance without external support for another 12 cycles. A complete continuation trial consists of 36 intervals between left, and 48 intervals between right hand keystrokes, which should ideally be of equal duration for a given hand.

Methods from symbolic dynamics permit to investigate whether the empirically observed variation in the duration of these intervals can be described as an orderly sequence of violations of the prescribed duration within trials, and whether these systematic patterns emerge or dissolve as a function of the

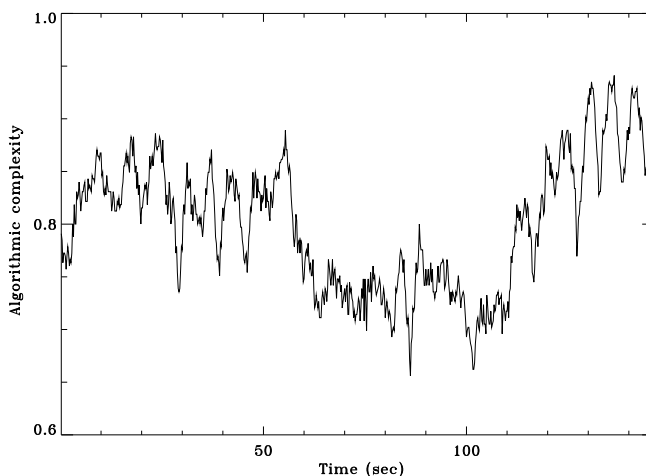
external control parameter performance tempo. We have chosen one particular coding scheme to get symbolic strings (Fig. 7). The coding scheme used here transforms continuous measures of interval duration to a dichotomous variable based on local comparisons between successive keystrokes performed with the same hand. A given interval terminated by a left hand keystroke (only these data are considered here for illustration) receives a value of '1', if its duration exceeds the duration of the preceding interval between left-hand keystrokes, otherwise the value is '0'. Values of '1' are indicated by black pixels in Fig. 7.

Already this simple transformation clearly suggests a qualitative shift in strategies for the realization of the polyrhythmic task as a function of performance tempo. The calculation of different measures of complexity exhibits that this transition is significant and that it is typical for young and old amateurs but also for concert pianists [23]. This qualitative transition has been modeled using a delayed feedback control. We conclude that the complexity of coordinated bimanual movements results from interactions between nonlinear control mechanisms with delayed feedback and stochastic timing components.

#### Astrophysics: Organization of Solar Spikes



**Fig. 8.** Dynamic spectrograms of millisecond spikes. The data was recorded by the frequency-agile solar radio spectrometer in ETH Zürich on 6 June, 1983. Low flux: white. High flux: black. The resolution in frequency is 1 MHz and in time 0.2 sec.



**Fig. 9.** Algorithmic complexity calculated from the irregular spike pattern presented in Fig. 8.

The variation of solar electromagnetic radiation and the particle emission is mainly caused by solar activity. Flares are the most violent manifestation of solar activity. They are caused by a rapid release of energy stored in the coronal magnetic field. Understanding the flare phenomenon requires to identify and model a large variety of physical processes involved. From observations of solar radio emission we know that the impulsive phase of this primary energy release in flares is fragmented into a multitude of substructures, called spikes (Fig. 8) which are triggered almost simultaneously. Therefore, properties of spikes give some detailed insight into the nature of this impulsive phase.

Depending on the assumed energy release and emission processes, two types of fragmentation are now under discussion: a scenario of global organization (spikes are emitted in a succession of similar events by the same system) or a scenario of local organization (many systems are triggered by an initial event).

We have searched for interrelations of spikes emitted simultaneously at different frequencies during the impulsive phase of a flare event [65]. To characterize such complex spatio-temporal patterns, such as dynamic spectra (Fig. 8) measured in solar radio astronomy, we use quantities of symbolic dynamics, such as Shannon information and algorithmic complexity. This approach is appropriate to characterize these patterns, whereas the popular estimate of fractal dimensions and related techniques fail here.

In the case of the analyzed dynamic spectra the length of the symbolic strings is about 400. To improve the statistics, we concatenate up to ten

symbol strings of successive scans of the dynamic spectrum. We observe in all cases that the only effect of such concatenation is a smoothing in Shannon information and algorithmic complexity.

The Shannon information and algorithmic complexity of the symbol sequence are used for comparing the considered observation with others, or with models (surrogates) and for characterizing the observation. These measures concern the whole frequency region, i.e. the global source region. This way, we find out that global organization is also apparent in quasi-periodic changes of these Shannon information and algorithmic complexity in the range of 2 – 8 seconds (Fig. 9).

Our analysis of spike events suggests that the structure in frequency is not stochastic but a process in which spikes at nearby locations are simultaneously triggered by a common exciter.

## 7 Conclusions

Let us summarize the main results obtained here.

- The dynamic entropy (uncertainty of next state) is connected with predictability.
- The Shannon entropy and algorithmic complexity are measures of randomness, but alternative measures, such as EMC or renormalized entropy, relate highest complexity to phase transitions and are, therefore, more appropriate to describe complex systems.
- Typical time series and information carrying sequences (DNA, texts, proteins, music) show correlations on many scales (including those of very long range).
- Measures of complexity are useful tools in various fields (e.g. physiology, astro- and geophysics) where techniques of linear and nonlinear data analysis fail.

Our results show that the dynamic entropies and other complexity measures are an appropriate measure for studying the predictability of evolutionary processes. Of particular interest are local studies of the predictabilities of certain local histories. Long correlations are of specific interest since they improve the predictability. This means, one can in principle improve the predictions by basing the predictions at longer observations. Further we can conclude that there are specific substrings, which are relatively seldom, where the local uncertainty is much smaller than the average i.e. the predictability is much better than in average. In other words, there are specific situations where the predictability is much better than the average predictability. It may be of practical importance to find out all substrings which belong to this particular class.

It was no space here to discuss the relation to other measures of long-range relations based on methods of statistical physics, as e.g. algorithmic entropy,

correlations functions, meansquare deviations [10],  $1/f^\delta$  noise [76, 10], and scaling exponents [57, 71, 10].

In conclusion we would like to express the hope that the analysis of entropies, predictabilities and other complexity measures could be developed to useful instruments for studies of the large-scale structure of a rather broad class of time series from various fields and information-carrying sequences.

## 8 Acknowledgments

The authors thank A.O. Benz, R. Engbert, F.-W. Gerstengarbe, H. Herzel, M.A. Jimenez-Montano, R. Kliegl, R. Krampe, F. Moss, A. Neiman, C. Nicolis, G. Nicolis, T. Pohl, P. Saparin, R. Steuer, A. Voss, P.C. Werner, A. Witt for many fruitful discussions and a collaboration on special topics of the problems discussed here.

## References

1. H. Atmanspacher, J. Kurths, H. Scheingraber, R. Wackerbauer, A. Witt, *Open Systems & Information Dynamics* **1**, 269 (1992).
2. H. Atmanspacher, C. Rath, G. Wiedemann, *Physica A* **234**, 819 (1997).
3. Badii, R., A. Politi, *Complexity : Hierarchical Structures and Scaling in Physics*, Cambridge University Press, 1997.
4. C.H. Bennett, in *Complexity, Entropy and the Physics of Information*, ed. W.H. Zurek, Addison-Wesley, Reading, MA, 1990.
5. A. Berman et al., *Proc. Natl. Acad. Sci. USA* **91**, 4044 (1994); E. Kolker, E.N. Trifonov, *Proc. Natl. Acad. Sci. USA* **92**, 557 (1995).
6. G.J. Chaitin, *J. ACM* **13**, 547 (1996).
7. J.P. Crutchfield, K. Young, *Phys. Rev. Lett.* **63**, 105 (1989).
8. J.P. Crutchfield, D.P. Feldman, *Phys. Rev.* **E 55**, R1239 (1997); S. Lloyd, H. Pagels, *Ann. Phys. (NY)* **188**, 186 (1988).
9. K. Dolan, A. Witt, M. L. Spano, A. Neiman, F. Moss *Phys. Rev. E* **59**, 5235 (1999).
10. W. Ebeling, A. Neiman, T. Pöschel, Dynamic Entropies, Long-Range Correlations and Fluctuations in Complex linear Structures. in *Coherent Approach to Fluctuations* (Proc. Hayashibara Forum 1995), World Scientific, Singapore 1995
11. W. Ebeling, G. Nicolis, *Europhys. Lett.* **14**, 191 (1991).
12. W. Ebeling, G. Nicolis, *Chaos, Solitons & Fractals* **2**, 635 (1992).
13. W. Ebeling, M.A. Jimenez-Montano, T. Pohl, *Entropy and Complexity of Sequences*, Festschrift devoted to Jagat N. Kapur, New Delhi 1999.
14. W. Ebeling, C. Frömmel, *BioSystems* **46**, 47 (1998).
15. W. Ebeling, T. Pöschel, *Europhys. Lett.* **26**, 241 (1994).
16. W. Ebeling, T. Pöschel, K. F. Albrecht, *Int. J. Bifurcation & Chaos*, **5**, 51 (1995).
17. W. Ebeling, J. Freund, K. Rateitschak, *J. Bif. & Chaos* **6**, 611 (1996).

18. W. Ebeling, J. Freund, F. Schweitzer, *Entropie, Struktur, Komplexität*, Teubner-Verlag, 1998
19. W. Ebeling, in *Proc. Conf. Complex Systems and Chaos*, Zakopane 1995.
20. W. Ebeling, *Physica D* (1997).
21. W. Ebeling, Jimenez-Montano, *Math. Biosci.* **52**, 53 (1980).
22. W. Ebeling, K. Rateitschak, *Discrete Dynamics in Nature and Society* **2**, 187 (1998).
23. R. Engbert, C. Scheffczyk, R.T. Krampe, M. Rosenblum, J. Kurths, R. Kliegl, *Phys. Rev. E* **56**, 5823 (1997).
24. D.P. Feldman, J.P. Crutchfield, *Phys. Lett. A* **238**, 244 (1998).
25. L. Gatlin, *Information Theory and the Living System*. Columbia University Press, New York 1972.
26. W. Gowin, P.I. Saporin, J. Kurths, Felsenberg, *Radiology* **205**, 428 (1997).
27. W. Gowin, P.I. Saporin, J. Kurths, Felsenberg, *Technology and Health Care* **6**, 373 (1998).
28. P. Grassberger, *Int. J. Theor. Phys.* **25**, 907 (1986); *Physica* **140 A**, 319-325 (1986).
29. A. Hempelmann, J. Kurths, *Astron. & Astrophys.* **232**, 356 (1990).
30. H. Herzel, A.O. Schmitt, W. Ebeling, *Phys. Rev. E* **50**, 5061 (1994).
31. H. Herzel, W. Ebeling, A. Schmitt, *Chaos, Solitons, Fractals* **4**, 97 (1994).
32. H. Herzel, W. Ebeling, A.O. Schmitt, M.A. Jimenez-Montano, in *From Simplicity to Complexity in Chemistry* (eds. A. Müller et al.) Vieweg Braunschweig 1996; H. Herzel, W. Ebeling, I. Grosse, *Proc. Conf. Bioinformatics*, GBF Monographs Vol. 18, Braunschweig 1995.
33. H. Herzel, I. Grosse, *Physica A* **216**, 518 (1995).
34. H. Herzel, I. Grosse, *Phys. Rev. E* **55**, 1 (1997).
35. W. Hilberg, *Frequenz* **44**, 243 (1990).
36. M.A. Jimenez Montano, *Bull. Math. Biol.* **64**, 641 (1984); *BioSystems* (1996).
37. W. Li, K. Kaneko, *Europhys. Lett.* **17**, 655 (1992).
38. H. Kantz, T. Schreiber, *Nonlinear Time Series Analysis*, Cambridge University Press, Cambridge 1997.
39. F. Kaspar, H.G. Schuster, *Phys. Rev. A* **36**, 842 (1987).
40. Yu. L. Klimontovich, *Turbulent Motion and the Structure of Chaos*, Kluwer Academic Publishers, Dordrecht, 1991.
41. A.N. Kolmogorov, *Dokl. Akad. Nauk USSR* **124**, 754 (1959).
42. A.N. Kolmogorov, *Probl. of Inform. Theory* **1**, 3 (1965).
43. J. Kurths, U. Schwarz, *Space Science Reviews* **68**, 171 (1994).
44. J. Kurths, A. Voß, P. Saporin, A. Witt, H.J. Kleiner, N. Wessel, *Chaos* **5**, 88 (1995).
45. R. Krampe, R. Kliegl, U. Mayr, *The Fast and the Slow of Bimanual Movement Timing*. Research Report. Max-Planck-Institute for Human Development, Berlin, Germany, 1993.
46. P.T. Landsberg, *Phys. Lett. A* **102**, 107 (1984); P.T. Landsberg, in *On Self-Organization*, ed. R.K. Misra, D. Maas, E. Zwierlein, Springer-Verlag, Berlin, 1994.
47. A. Lempel, J. Ziv, *IEEE Trans. Inf. Theory* **IT-22**, 75 (1976).
48. L. Levitin, Z. Reingold, *Chaos, Solitons, Fractals* **4**, 709 (1994).
49. K. Lindgren, M. Nordahl, *Complex Systems* **2**, 409 (1988).

50. R. Lopez-Ruiz, H.L. Mancini, X. Calbet, *Phys. Lett. A* **209**, 321 (1995).
51. D.W. McShea, *Biol. Physiol.* **6**, 303 (1991).
52. L. Molgedey, *IJTAF* **3**, in print (2000).
53. L. Molgedey, W. Ebeling *Eur. Phys. J.* **B**, submitted (2000).
54. A. Neiman, B. Shulgin, V. Anishchenko, W. Ebeling, L. Schimansky-Geier, J. Freund, *Phys. Rev. Lett.* **76**, 4299 (1996).
55. C. Nicolis, W. Ebeling, C. Baraldi, *Tellus* **49A**, 10-18 (1997).
56. X. Pel, F. Moss *Nature* **379**, 618 (1996).
57. C.-K. Peng et al. *Phys. Rev. E* **49**, 1685 (1994).
58. S.M. Pincus, *Proc. Natl. Acad. Sci. USA* **88**, 2297 (1991).
59. P. Saparin, A. Witt, J. Kurths, V. Anishenko, *Chaos, Solitons and Fractals* **4**, 1907 (1994).
60. P.I. Saparin, W. Gowin, J. Kurths, Felsenberg, *Phys. Rev. E* **58**, 6449 (1998).
61. A. O. Schmitt, H. Herzel, W. Ebeling *Europhys. Lett.* **23**, 303 (1993).
62. A. O. Schmitt, W. Ebeling, H. Herzel *Biosystems* **37**, 199 (1996).
63. C. Schittenkopf, A. Deco, *Physica D*, (1998).
64. H.G. Schuster, *Deterministic Chaos: An Introduction*, VCH, Weinheim, 1988.
65. U. Schwarz, J. Kurths, A. Witt, A.O. Benz, *Astron. & Astrophys.* **277**, 215 (1993).
66. C. Shannon, *Bell Systems Tech.* **30**, 50 (1951).
67. J.S. Shiner, in *Self-Organization of Complex Structures: From Individual to Collective Dynamics*, ed. F. Schweitzer, Gordon and Breach, London, 1996.
68. J.S. Shiner, M. Davison, P.T. Landsberg, *Phys. Rev. E* **59**, 1459 (1999), P.T. Landsberg, J.S. Shiner, *Phys. Lett. A* **245**, 228 (1998).
69. Ya. B. Sinai, *Dokl. Akad. Nauk USSR* **124**, 768 (1959); **125**, 1200 (1959).
70. L.B. Smith, E. Thelen (Eds.), *A Dynamic Systems Approach to the Development of Cognition and Action*, MIT Press, Cambridge, MA, 1994.
71. H.E. Stanley et al., *Physica A* **205**, 214 (1994).
72. G.E. Stelmach, J. Requin (Eds.), *Tutorials in Motor Behavior II*, North Holland, Amsterdam, 1992.
73. E.N. Trifonov, V. Brendel, *Gnomic - A Dictionary of Genetic Codes* VCH Weinheim 1987.
74. A. Voss, K. Hnatkova, N. Wessel, J. Kurths, A. Sander, A. Schirdewan, A.J. Camm, M. Malik, *Pace* **21**, 186 (1998).
75. H. Voss, J. Kurths, *Phys. Rev. E* **58**, 1155 (1998).
76. R.F. Voss, *Phys. Rev. Lett.* **68**, 3805 (1992); R.F. Voss, *Fractals* **2**, 1(1994).
77. R. Wackerbauer, A. Witt, H. Atmanspacher, J. Kurths, H. Scheingraber, *Chaos, Solitons & Fractals* **4**, 133 (1994).
78. P.C. Werner, F.-W. Gerstengarbe, W. Ebeling, *Theor. Appl. Climatol.* **62**, 125 (1999).
79. A. Witt, J. Kurths, F. Krause, K. Fischer, *Geophysical and Astrophysical Fluid Dynamics* **77**, 79 (1994).
80. A. Witt, A. Neiman, J. Kurths, *Phys. Rev. E* **55**, 5050 (1997).
81. A.M. Yaglom, I.M. Yaglom, *Probability and Information*, Reidel, 1983.
82. H.P. Yockey, *Information Theory and Molecular Biology*, Cambridge University Press, Cambridge 1992.
83. Zaks, M., Pikovsky, A., Kurths, J., *Physica D* **117**, 77 (1998).