Kolmogorov's legacy about entropy, chaos and complexity

Massimo Falcioni[1], Vittorio Loreto[2], Angelo Vulpiani[3]

[1] University of Rome "La Sapienza", Physics Department and INFM Center for Statistical Mechanics and Complexity, P.zzle Aldo Moro, 2 Rome, Italy, I-00185, massimo.falcioni@phys.uniroma1.it
[2] University of Rome "La Sapienza", Physics Department and INFM Center for Statistical Mechanics and Complexity, P.zzle Aldo Moro, 2 Rome, Italy, I-00185, loreto@pil.phys.uniroma1.it
[3] University of Rome "La Sapienza", Physics Department and INFM Center for Statistical Mechanics and Complexity, P.zzle Aldo Moro, 2 Rome, Italy, I-00185, angelo.vulpiani@roma1.infn.it

1 Introduction

The concept of entropy was initially introduced in thermodynamics in a phenomenological context. Later, mainly by the contribution of Boltzmann, a probabilistic interpretation of the entropy was developed in order to clarify its deep relation with the microscopic structure underlying the macroscopic bodies. In the unrelated field of communication systems, Shannon showed that, with a suitable generalization of the entropy concept, one can establish a mathematically self-consistent information theory. Kolmogorov realized, just after Shannon's work, the conceptual relevance of information theory and the importance of its ideas for the characterization of irregular behaviors in dynamical systems. Going beyond a probabilistic point of view, the Algorithmic Complexity (introduced by Chaitin, Kolmogorov and Solomonoff) allows a formalization of the intuitive notion of randomness of a sequence. In this chapter we discuss the connections among entropy, chaos and algorithmic complexity; in addition we briefly discuss how these concepts and methods can be successfully applied to other unrelated fields: linguistics, bioinformatics, finance.

2 Entropy in thermodynamics and statistical physics

The concept of entropy made its appearance in physics to describe in a mathematical way the irreversible behaviors of the macroscopic bodies, those we experience in everyday life. The goal is achieved by the 2^{nd} principle of the thermodynamics, which can be formulated as follows [16].

There exists a function, the *entropy* S, defined for the equilibrium states of macroscopic systems, such that its difference between (equilibrium) states A and B of a given system satisfies the relation:

$$\Delta S_{AB} \equiv S(B) - S(A) \geq \int_A^B \frac{\delta Q}{T_{\mathrm{source}}}\,, \tag{1}$$

where the integral is computed along an arbitrary transformation leading from state A to state B, in which the system exchanges heat δQ with external sources at thermodynamic temperatures T_{source}. When the transformation is chosen to be reversible, relation (1) has the equality sign and may be used to compute ΔS_{AB}. In reality, irreversibility transforms relation (1) to an inequality which sets limits to what can happen. In particular, in isolated systems, where $\delta Q = 0$, a transformation $A \to B$ is allowed only if $S(B) > S(A)$.

In the case of n moles of a classical, perfect and mono-atomic gas, contained in a vessel of volume V at temperature T, one is able to write:

$$S(n,U,V) = n\,R\,\ln\left[\gamma_0 \left(\frac{U}{n}\right)^{3/2} \left(\frac{V}{n}\right)\right], \tag{2}$$

where γ_0 is a suitable constant, R is the universal gas constant and $U = \frac{3}{2}\,n\,R\,T$ is the internal energy of the gas.

If we accept the atomistic point of view [5], we put $R = N_A k$, where $N_A = 6.02\ 10^{23}$ is Avogadro's number and $k = 1.38\ 10^{-23} J/K$ is the Boltzmann constant, and we rewrite eq. (2), for $N = n\,N_A$ molecules, as

$$S(N,U,V) = N\,k\,\ln\left[\gamma_1 \left(\frac{U}{N}\right)^{3/2} \left(\frac{V}{N}\right)\right]\,, \tag{3}$$

where γ_1 is another constant. In this formula, the specific volume, V/N, is the mean volume available for the motion of a single molecule, where we assume that molecules are indistinguishable. If we define $V/N = (\Delta x)^3$, Δx may be thought of as the typical variability of the space coordinates for each molecule. For the mean (kinetic) energy U/N one can write:

$$\frac{U}{N} = \frac{1}{2m}\left(\langle p_x^2\rangle + \langle p_y^2\rangle + \langle p_z^2\rangle\right) = \frac{3}{2m}\langle p_x^2\rangle \equiv \frac{3}{2m}(\Delta p)^2\,, \tag{4}$$

where p_x, p_y, p_z are the components of the momentum of one molecule, so that Δp, the root mean-square of one component, gives the order of variability of the single molecule momentum. Then, denoting by γ_2 and γ two suitable constants, we have:

$$S = N\,k\,\ln\left[\gamma_2\,(\Delta x)^3\,(\Delta p)^3\right] = k\,\ln\left[\frac{(\Delta x)\,(\Delta p)}{\gamma}\right]^{3N}\,. \tag{5}$$

Here, $[(\Delta x)(\Delta p)/\gamma]^{3N}$ represents the volume, in units of $(\gamma)^{3N}$, of the $6N$-dimensional phase space region where the motion of the system takes place.

Equation (5) is a special case of the relation between thermodynamic entropy and the extention of the motion in the phase space, the "Boltzmann principle", that one can write as:

$$S = k \ln \Gamma(N, U, V) \ , \tag{6}$$

with

$$\Gamma(N, U, V) = \int\limits_{U-\delta U/2 \le H \le U+\delta U/2} \frac{\mathrm{d}^{3N}x \ \mathrm{d}^{3N}p}{N! \, \mathrm{h}^{3N}} \equiv \int\limits_{U-\delta U/2 \le H \le U+\delta U/2} \mathrm{d}\Gamma_N \ . \tag{7}$$

In (7) $H = \sum_{i=1}^{N} \mathbf{p}_i^2/2m$ is the Hamiltonian function (the energy) of the system, U is its actual value, known within an uncertainty δU, h is a constant that in classical mechanics is indeterminate but that, with the help of quantum mechanics, may be taken to be the Planck's constant. The term $N!$ takes into account the indistinguishability of the molecules of the gas. So $\Gamma(N, U, V)$ measures the amount of freedom the system has in its phase space. This is the starting point for a statistical mechanics description of the thermodynamics. According to the interpretative rules of statistical mechanics,

$$\frac{\mathrm{d}\Gamma_N}{\Gamma(N, U, V)} \tag{8}$$

is the probability to find the system, in equilibrium with energy around U, in a volume $\mathrm{d}^{3N}x \ \mathrm{d}^{3N}p$ of its phase space.

It is interesting to observe that eq. (8) defines $1/\Gamma(N, U, V)$ as the (uniform) probability density of finding a system in equilibrium in anyone of its allowed microscopic states, the micro-canonical distribution:

$$\rho(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_N, \mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \ldots, \mathbf{p}_N) = 1/\Gamma(N, U, V) \ , \tag{9}$$

and we can write

$$\frac{S}{k} = \ln \Gamma = \int \mathrm{d}\Gamma_N \ \rho \ (-\ln \rho) \equiv \langle -\ln \rho \rangle \ . \tag{10}$$

This is a possible and useful definition of the entropy in statistical mechanics, even when the equilibrium states are described by a density ρ, not necessarily the micro-canonical one (Gibbs entropy),

$$S = -k \ \langle \ln \rho \rangle . \tag{11}$$

3 Entropy in information theory

The concept of entropy enjoyed a renewed consideration in the scientific community when, at the end of 1940s, in a very different context, C. E. Shannon [22] adressed the problem of an efficient transmission of messages, thus laying the foundations of a mathematical theory of communication. Among the various results obtained by Shannon, the following is of interest for us.

Each particular message to be transmitted is thought to belong to the ensemble of all the messages one source can generate. The source is able to emit M different symbols, constituting the so-called *alphabet*, and every message is a sequence of N letters that we indicate with $(s_1, s_2, s_3, \ldots, s_N)$ or s^N, where each s_i is one of the M symbols. One supposes that every message may be produced with a probability $p(s^N)$. Usually the signal sent through a transmission channel is a sequence of $m\,(\neq M)$ possible physical states; we assume a binary channel, i.e. a channel operating with two states that can be named 0 and 1. At this point one has the coding problem: one needs to transform every M-ary sequence into a binary sequence, in the most economical way, i.e. such that the average length of the coding sequences be minimum (one also wants to recover the original message so the code must be invertible). Shannon was able to show that it is possible to associate each message s^N with a sequence of $\ell(s^N)$ binary symbols, i.e. of $\ell(s^N)$ bits, where

$$\ell(s^N) \approx \log_2 \frac{1}{p(s^N)} \,, \tag{12}$$

thus obtaining a (quasi) optimal coding. So one has that $\log_2 1/p(s^N)$ is a measure of the resources needed to send a particular message. An interesting result is obtained when considering long messages, the $N \to \infty$ limit. In this case one introduces the quantity, called Shannon entropy:

$$h = \lim_{N\to\infty} \frac{1}{N} \, \langle \ell \rangle_N = \lim_{N\to\infty} \frac{1}{N} \sum_{s^N} p(s^N) \; \log_2 \frac{1}{p(s^N)} \tag{13}$$

that represents the average number of bits needed to code a source symbol. The importance of h is clarified hereafter. According to the Shannon-McMillan theorem [12], for $N \to \infty$ the ensemble of sequences s^N divides up into two sets:

- the set of sequences with $p(s^N) \approx 2^{-N\,h}$, that are said the typical sequences;
- the set containing all the other sequences.

The set of typical sequences, the ones that are "reasonably probable", has $\mathcal{N}(N) \approx 2^{N\,h}$ elements. Moreover the probability of this set is ≈ 1; while the probability of the remaining set is ≈ 0. This means that the typical sequences almost exhaust the total probability and constitute a set of $\mathcal{N}(N)$ practically equiprobable elements, with $p(s^N) \simeq 1/\mathcal{N}(N)$. The Shannon entropy gives, at the same time, both the number of bits per symbol that are necessary to code each one of these (long) sequences and the rate of growth of their number with the sequence length. From this point of view, writing $h = (1/N)\log \mathcal{N}(N)$ and comparing Eqs. (13) and (11), one realizes that h is the analogue of the entropy per particle in statistical mechanics: $S/N = (1/N)k \ln \Gamma$, hence the name. Therefore, for an efficient transmission of long messages, one must be able to forward h bits per emitted source symbol. This is the same as identifying, with $N \cdot h$ bits, one among the $\mathcal{N}(N)$ substantially equiprobable messages that one can reasonably expect [22].

The statistical mechanics entropy (Boltzmann entropy) $S = k \ln \Gamma$ and the analog quantity in information theory $h\,N = \log_2 \mathcal{N}(N)$ may be seen as two measures of the amount of uncertainty in the associated system: this uncertainty grows as either Γ (the number of accessible microstates) or $\mathcal{N}(N)$, (the number of messages that can potentially be transmitted) grows. From this point of view, entropy may also be seen as a measure of the quantity of information one obtains after the state of the system has been determined or the message has been transmitted (and received). Given a probabilistic scheme, assigning probabilities p_i $(i = 1, \ldots, M)$ to the M occurrences of an event x, the expression:

$$\sum_{i=1}^{M} p_i \; \log_2 \frac{1}{p_i} = \langle -\log_2 p \rangle \tag{14}$$

defines the entropy or uncertainty of the event, that we indicate with $H(x)$. This quantity indeed can be thought of as the average uncertainty on the event before it happens, or the average amount of information one gets from a single actual realization. It satisfied $0 \leq H \leq \log_2 M$; one has $H = 0$ if all p_i but one are zero, since in this case there is no uncertainty at all (and no new information from observing the certain event). On the other hand the maximum value both of uncertainty and of information, $H = \log_2 M$, is reached when all occurrences are equiprobable: $p_i = 1/M$. In the case when x is a symbol of a source of messages, $x = s^1$,

$$H(x) = \sum_{s^1} p(s^1) \; \log_2 \frac{1}{p(s^1)} \equiv H_1 \tag{15}$$

is the average uncertainty on the emission of one symbol. It is possible to show that the quantity

$$\frac{1}{N} \sum_{s^N} p(s^N) \; \log_2 \frac{1}{p(s^N)}$$

may be read as the mean uncertainty on the emission of the N^{th} symbol of a sequence, given that the preceding $N-1$ are known. Therefore we may interpret h as the mean residual uncertainty on the appearance of an additional symbol in a sequence, after the information contained in the (infinitely long) past history has been acquired. If time correlations exist in the sequences, so that some constraint on the future is given by the past, the mean uncertainty becomes lower. So, in general, we have $h \leq H_1 \leq \log_2 M$.

In the coding problem of information theory, one wants to find an (economical) invertible relation between different sequences that represent the original message and the transmitted one. If one thinks about the sequences as describing evolutions of dynamical systems (see next subsection), one easily realizes that information theory concepts may play a role in problems of dynamical systems. For instance in the isomorphism problem, where one wants to characterize the conditions under which one is sure that two formally different systems are actually the same. This implies that an invertible relation exists which preserves the probabilistic and dynamical relations among their respective elements. In facing this problem Kolmogorov defined a quantity that is known as the *Kolmogorov-Sinai entropy* of a dynamical system [14, 24].

The above results have been obtained by supposing that the message source has well defined probabilistic rules, such that each message has a certain probability to appear. The effect of this probability distribution is to generate a selection among all the M^N *a priori* possible messages, reducing them to the set of the typical sequences that are "reasonable" from a probabilistic point of view. In cases where a message is generated by unknown rules, one can search for the most economical, i.e. the shortest, (binary) code to transmit it. Thus one poses the problem of a "universal coding", depending only on "intrinsic" properties of the symbol sequence. To find the answer, Kolmogorov introduced the idea of *Algorithmic Complexity* (AC) (also known as *Kolmogorov complexity*) [15, 6, 25].

4 Entropy in dynamical systems

Let us now discuss the relevance of information theory in the context of the deterministic systems. The idea is due to A.N.Kolmogorov who, just after the Shannon work, realized the conceptual importance of the information theory beyond its practical use in communications [13, 14].

For the sake of self-consistency, we introduce some basic elements on chaotic dynamical systems. A dynamical system is defined as a deterministic rule for the time evolution of the state (described by a d-dimensional vector $\mathbf{x}$). Well-known examples are the ordinary differential equations:

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}(t)) \tag{16}$$

and the discrete time maps:

$$\mathbf{x}(n+1) = \mathbf{g}(\mathbf{x}(n)) \, , \tag{17}$$

where $\mathbf{x}, \mathbf{f}, \mathbf{g} \in R^d$. Examples of regular behavior of the solutions of ordinary differential equations are stable fixed points, periodic or quasi-periodic motions.

After the seminal work of Poincaré, and later Lorenz, Hénon and Chirikov (to cite only some of the most eminent ones) the existence of visibly irregular non periodic behaviors, that are called chaotic, is now well established [20]. In Fig. 1 one can see the time evolution of the celebrated Lorenz's model [19]:

$$\begin{cases} \dot{x} = -\sigma(x-y) \\ \\ \dot{y} = -xy + rx - y \\ \\ \dot{z} = xy - bz \end{cases} \tag{18}$$

where the dot means the time derivative.

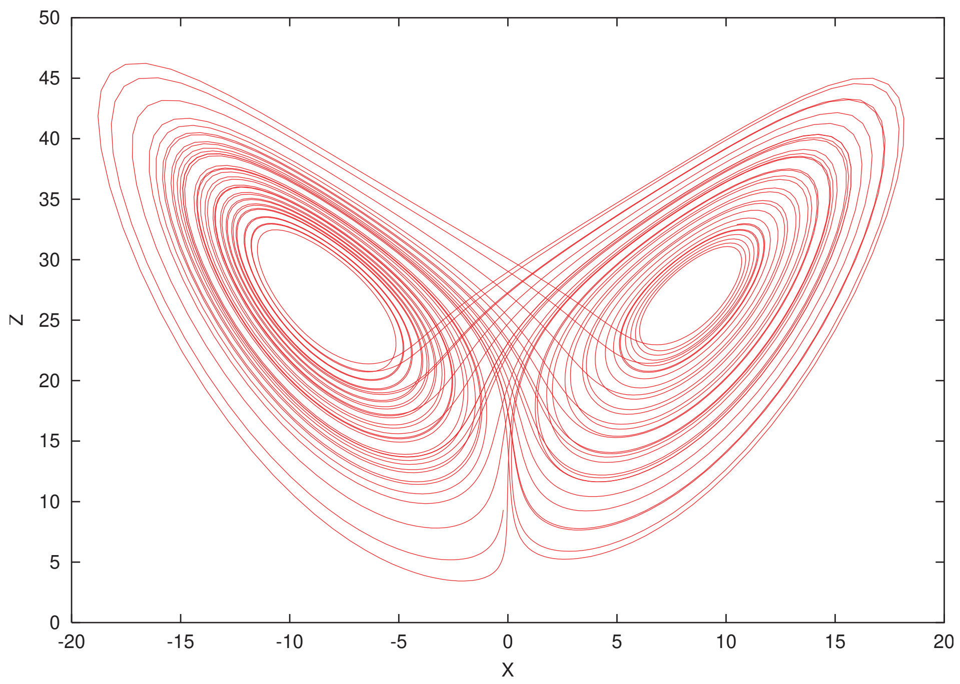

Fig. 1. Lorenz's model: Projection of a trajectory on the $x\,z$ plane for $r = 28$, $\sigma = 10$ and $b = 8/3$.

For the purposes of this chapter, the most relevant feature of chaotic dynamics is the so-called sensitive dependence on initial conditions. Consider two trajectories, $\mathbf{x}(t)$ and $\mathbf{x}'(t)$, initially (say, at time $t = 0$) very close; in chaotic systems we find that the distance between them, $|\delta \mathbf{x}(t)| = |\mathbf{x}(t) - \mathbf{x}'(t)|$, increases exponentially in time as $|\delta \mathbf{x}(0)| \to 0$ and $t \to \infty$:

$$|\delta \mathbf{x}(t)| \sim |\delta \mathbf{x}(0)| e^{\lambda_1 t} , \tag{19}$$

λ_1 is called first (or maximal) Lyapunov exponent, see Fig. 2. The exponent λ_1 characterizes the typical trajectories of the system, just like the Shannon entropy characterizes the typical sequences of an ensemble of messages. The sensitive dependence on the initial conditions (i.e. $\lambda_1 > 0$) is a practical definition of deterministic chaos, since it implies that, in presence of any uncertainty of the knowledge on the initial state, the system is unpredictable for long times. If we know the initial conditions with a finite, but small, precision $\delta_0 = |\delta \mathbf{x}(0)|$ and we want to predict the state with a certain tolerance δ_M then, from eq. (19), we find that our forecast cannot be pushed to times larger than T_P, called predictability time:

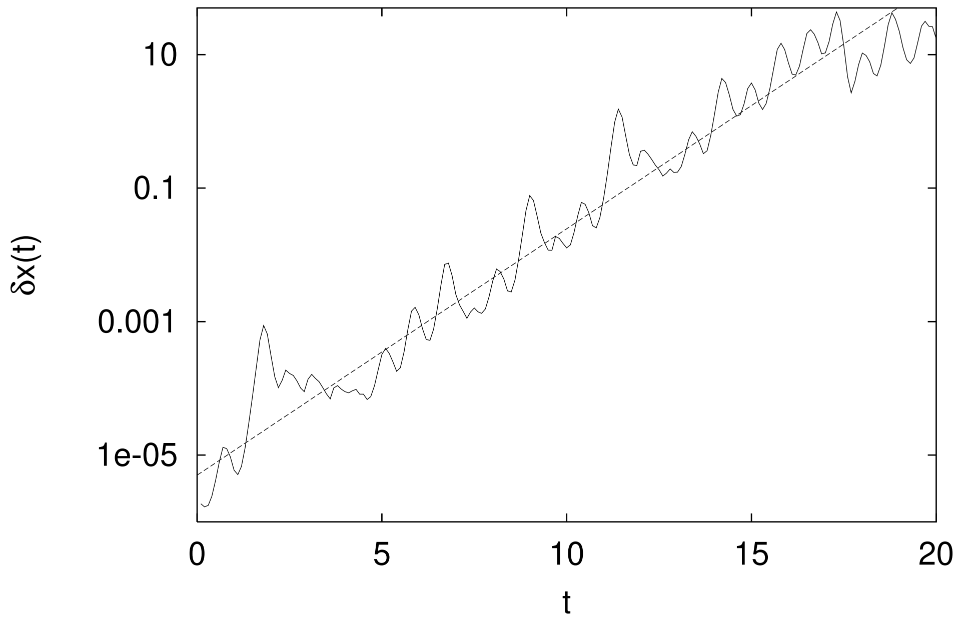

Fig. 2. Lorenz's model: $|\delta \mathbf{x}(t)|$ vs t for $r = 28$, $\sigma = 10$ and $b = 8/3$.

$$T_P \sim \frac{1}{\lambda_1} \ln \frac{\delta_M}{\delta_0} . \tag{20}$$

Because the logarithm varies in an extremely slow way, T_P is essentially determined by λ_1.

A more complete characterization of the instability properties of the trajectories of a chaotic system is given by the set of Lyapunov exponents

$\lambda_1, \lambda_2, ...\lambda_d$ (where conventionally the labeling is such that $\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_d$) defined as follows. Consider a small d-dimensional ball of radius ϵ with center $\mathbf{x}(0)$. Under the evolution (16) or (17), at time t the ball will be deformed into a d-dimensional ellipsoid with semi-axes $l_1(t) \geq l_2(t) \geq ... \geq l_d(t)$. The Lyapunov exponents are defined as

$$\lambda_i = \lim_{t\to\infty} \lim_{\epsilon\to 0} \ln \frac{l_i(t)}{\epsilon} \,. \tag{21}$$

An important consequence of the instability with respect to initial conditions is that the time sequences generated by a chaotic system, from the point of view of information theory, have the same properties of genuine stochastic processes. This is another way to view unpredicatability. In order to discuss this idea, let us consider the example of a one-dimensional map, eq. (17) with $d = 1$, the celebrated logistic map at Ulam's point:

$$x(t+1) = 4x(t)(1 - x(t)) \,. \tag{22}$$

It is possible to show that such a system, according to our definition, is chaotic since $\lambda_1 = \ln 2$ [20]. Consider a trajectory $\{x(0), x(1), ..., x(T-1)\}$ and the symbolic sequence $\{i_0, i_1, ..., i_{T-1}\}$ obtained by taking $i_k = 0$ if $x(k) \leq 1/2$, otherwise $i_k = 1$. Let us now study the Shannon entropy h of the ensemble of sequences so generated and its relation with the maximal Lyapunov exponent λ_1. Under rather general conditions, almost always satisfied by the commonly investigated dynamical systems, h can be numerically estimated by analyzing one very long (infinitely long, in principle) sequence as follows. Denote with W_m a m-long subsequence $\{i_j, i_{j+1}, ..., i_{j+m-1}\}$; from its frequency of apparition in $\{i_0, i_1, ..., i_{T-1}\}$, determine its probability $P(W_m)$ and finally compute [4]

$$\lim_{m\to\infty} \frac{1}{m} \sum_{W_m} P(W_m) \, \ln \frac{1}{P(W_m)} = h \,.$$

Note that in dynamical systems theory, entropy is usually defined with the natural logarithm, while in information theory the binary log is preferred. In the above exemple (22), one obtains $h > 0$ and indeed, in this case, one can exactly determine $h = \ln 2$. This coincidence of values for λ_1 and h is not accidental. The results of Section 3 show that the number of bits N_T necessary to specify one of these sequences of length T is

$$N_T \simeq T \, \frac{h}{\ln 2} \,. \tag{23}$$

[4] Strictly, the estimated probability P depends on T and one would have to let T tend to infinity before letting m tend to infinity. in practise, T is given (finite) and $\lim_{m\to\infty}$ only means taking m large (but still small with respect to T to get a good statistical estimate of $P(W_m)$).

On the other hand, we know that in the presence of chaos, see Eq. (19), $|\delta x(t)| \sim |\delta x(0)| \exp(\lambda_1 t)$. Therefore, in order to be able to specify i_{T-1}, one needs $|\delta x(0)|$ small enough such that $|\delta x(T-1)| \leq 1/2$. Namely $|\delta x(0)| \leq (1/2)\exp(-\lambda_1 T)$, which means that one has to use a number of bits $N_T \simeq \ln(1/|\delta x(0)|)/\ln 2 \simeq T\lambda_1/\ln 2$. From these considerations, comparing the above result with Eq. (23), one can conclude that, in a one-dimensional map, the maximal Lyapunov exponent has also an entropic meaning, i.e. it gives (to within a factor $\ln 2$) the number of bits per unit time necessary to specify a trajectory.

Consider again the Lorenz model (which has a unique positive Lyapunov exponent) shown in Fig. 1. The chaotic trajectory evolves turning around two unstable fixed points, C_1 and C_2, with an average period $\langle \tau \rangle$. Let us introduce a binary sequence $\{i_1, i_2, \dots\}$ as follows. For each turn around C_1, put $i = 1$, otherwise $i = 0$. If one computes the Shannon entropy h of such a sequence, one finds the nice relation

$$\frac{h}{\langle \tau \rangle} = \lambda_1 \ , \tag{24}$$

again providing a link between the Lyapunov exponent of chaotic trajectories and the Shannon entropy of symbolic sequences generated by the system.

At first glance one might believe that the above procedure, in which one associates a trajectory (given by real numbers) of a chaotic system to a sequence of integer numbers, is too crude and some relevant features are suppressed. An important result is that, in presence of (unavoidable) finite precision, this reduction may be realized with no suppression of information, in the case of deterministic systems. Previously we saw that, because of the sensitive dependence on the initial conditions, the uncertainty on the initial state implies the practical impossibility of a detailed description of the trajectory, in chaotic systems. Therefore a coarse-grained description appears rather natural.

For the sake of simplicity, we consider a discrete time system (i.e. a map). To model finite precision measurements, we introduce an ϵ-partition of the phase space with $\mathcal{N}$ disjoint sets (hypercubes of edge ϵ): $A_1, A_2, ..., A_{\mathcal{N}}$. From any initial condition $\mathbf{x}(0)$, one has a trajectory $\{\mathbf{x}(1), \mathbf{x}(2), ..., \mathbf{x}(T)\}$. This trajectory identifies a symbolic sequence $\{i_1, i_1, ..., i_T\}$, where i_j $(1 \leq i_j \leq \mathcal{N})$ means that $\mathbf{x}(j)$ is in the cell A_{i_j}.

In such a way, for any given ϵ-partition, one can compute the Shannon entropy $h(\epsilon)$, either following the approach discussed in Sect. 3, if one considers the ensemble of the trajectories, or as described above, if one analyzes one single trajectory. The meaning of $h(\epsilon)$ is clear: the average number of bits per unit time necessary to specify the time evolution with a precision ϵ is nothing but $h(\epsilon)/\ln 2$. In the limit $\epsilon \to 0$, we have the so called Kolmogorov-Sinai entropy:

$$h_{KS} = \lim_{\epsilon \to 0} h(\epsilon) \, . \tag{25}$$

For deterministic systems with finite dimensional phase space, $h_{KS} < \infty$. This quantity was introduced by Kolmogorov as a way to distinguish among non isomorphic dynamical systems: h_{KS} is isomorphism invariant, so two systems with different h_{KS} are surely non isomorphic. The hope was that identity of the Kolmogorov-Sinai entropy would entail isomorphism. In general, however, this is not the case.

Just as the Shannon entropy in information theory, the Kolmogorov-Sinai entropy gives the information per unit time necessary to completely specify one typical time evolution of the system. Moreover, as remarked at the end of Sect. 3, it is also a measure of the unavoidable residual randomness in the prediction of the deterministic chaotic evolution, due to the impossibility of perfect measurements. From both points of view, one important result is that a discrete time deterministic system with finite entropy, satisfying fairly general properties, may be represented by (is equivalent to) a random process with a finite number of states; the minimum number of required states, m, depends on the Kolmogorov-Sinai entropy: $\exp(h_{KS}) < m < 1 + \exp(h_{KS})$ (*Krieger's theorem*).

In the above discussion we used a partition with hyper-cubic cells of edge ϵ, and then we took the limit $\epsilon \to 0$, so that the number of elements of the partition went to ∞. It is possible to introduce h_{KS} using more general partitions with a finite number of elements and then to take the supremum over all the partitions.

As suggested by the examples above, there is a close relation between the Kolmogorov-Sinai entropy and the positive Lyapunov exponents [20], which, in many interesting cases, can be written as

$$h_{KS} = \sum_{\lambda_i > 0} \lambda_i \ . \qquad (26)$$

If one regards chaos as random behavior, this relation gives a definite and clear meaning to the above definition that a system with $\lambda_1 > 0$ is a chaotic one.

5 Algorithmic complexity

In Sect. 3 we saw how the Shannon entropy gives a (statistical) measure of complexity by putting a limit on the possibility of compressing an ensemble of sequences. In the Algorithmic Complexity theory (introduced independently by G.Chaitin, A.N.Kolmogorov and R.J.Solomonoff) one goes further, i.e. one treats a single sequence, avoiding to use an ensemble. One can address the problem of compressing a given sequence. However, one soon realizes that a big compression is associated with a “simple structure” (little information is required to reproduce the sequence). On the other hand, a “complicated structure" requires much information to be reproduced, resulting in little compression. So, the posed problem inevitably meets with that of defining the randomness of the sequence.

Consider 3 different sequences of 24 coins flips (0 for heads and 1 for tails):

1) 000000000000000000000000 ;

2) 010101010101010101010101 ;

3) 100001001010011000111010 .

The first and the second sequences appear regular and somehow atypical. In contrast, the last one seems irregular and typical. On the other hand, probability theory does not seem able to express the notion of randomness of an individual sequence. In our case, each sequence has the same occurrence probability, i.e. 2^{-24}.

The notion of Algorithmic Complexity (or Kolmogorov complexity) is a way to formalize the intuitive notion of randomness (and regularity) of a sequence. Consider a binary digit sequence of length T generated by a computer code on some machine $\mathcal{M}$: one defines Algorithmic Complexity $K_{\mathcal{M}}(T)$ of this sequence as the bit length of the shortest program able to generate the T-sequence and to stop afterwards. Kolmogorov was able to show the existence of universal computers $\mathcal{U}$ such that

$$K_{\mathcal{U}}(T) \leq K_{\mathcal{M}}(T) + c_{\mathcal{M}} \tag{27}$$

where $c_{\mathcal{M}}$ depends only on the computer $\mathcal{M}$. At this point, in the limit $T \to \infty$, we can define the Algorithmic Complexity per symbol with respect to a universal computer (we omit the $\mathcal{U}$-label)

$$\mathcal{C} = \lim_{T \to \infty} \frac{K(T)}{T} . \tag{28}$$

Let us note that as consequence of (27), $\mathcal{C}$ does not depend on the machine and it is an intrinsic quantity.

Using the above definition, one easily understands that a T-sequence of all 0 (as in the first series) or a periodic repetition of 01 (as in the second series) can be obtained with a program whose size is $O(\ln T)$ and therefore in the limit $T \to \infty$, the complexity is zero. On the other hand in an irregular sequence one expects that $K(T) \sim T$ and therefore $\mathcal{C}$ is positive and one calls the sequence complex.

Unfortunately Algorithmic Complexity cannot be computed; this impossibility is related to the Gödel incompleteness theorem [6]. Beyond this impossibility, the concept of Algorithmic Complexity has an important role to clarify the vague notion of randomness.

We can wonder whether the sequences generated by a chaotic dynamics are complex (in the sense of the Algorithmic Complexity) or not. In this respect, there exists an important relation between the Shannon entropy and the Algorithmic Complexity K_{W_T} of the T-sequences W_T of an ensemble [3, 18]:

$$\lim_{T\to\infty} \frac{1}{T}\langle K_{W_T}\rangle = \lim_{T\to\infty} \frac{1}{T}\sum K_{W_T} P(W_T) = \frac{h}{\ln 2} ,$$

where $P(W_T)$ is the occurrence probability of W_T. Putting this together with the results of the preceding section, we have that

chaotic systems produce algorithmically complex sequences.

Of course, one has that for periodic sequences $\mathcal{C} = 0$, while for a series of random variables $\mathcal{C} > 0$.

At first glance the above sentence may appear questionable, as one could (erroneously) argue as follows. Since chaos is present also in very simple systems, whose temporal evolution can be obtained with numerical codes of few instructions, a sequence of arbitrary length can be obtained with a finite number of bits, i.e. those necessary to specify the evolution rules, therefore the Algorithmic Complexity is zero. The missing point in the above naive argument is the quantity of information necessary (i.e. how many bits N_T) to specify the initial condition in such a way the code produce a certain T-sequence. Previously, we saw that in chaotic systems for a typical initial condition one has $N_T/T \simeq \sum_i^+ \lambda_i/\ln 2$ (where $\sum_i^+ \lambda_i = h_{KS}$ is the sum of the positive Lyapunov exponents). Therefore, the Algorithmic Complexity per symbol in chaotic systems is positive and strictly related to the Lyapunov exponents, therefore to the Kolmogorov-Sinai entropy.

Let us clarify the above points with the following simple example [9]. Consider the one-dimensional map (called Bernoulli shift):

$$x(t+1) = 2x(t) \bmod 1 ,$$

with Lyapunov exponent $\ln 2$. Writing the initial condition in binary notation, $x(0) = \sum_{j=1}^{\infty} a_j 2^{-j}$, where the a_j can be 1 or 0, one immediately sees that the action of the Bernoulli shift is nothing but a shift of the binary coordinates: $x(1) = \sum_{j=1}^{\infty} a_{j+1} 2^{-j}$, $x(2) = \sum_{j=1}^{\infty} a_{j+2} 2^{-j}$, and so on. Therefore the Algorithmic Complexity of the sequence $\{i_1, i_2, ..., i_T\}$, obtained by taking $i_k = 0$ if $x(k) \leq 1/2$ otherwise $i_k = 1$, is reduced to the study of the sequence $\{a_1, a_2, ..., a_T\}$ related to the initial condition. Of course there are initial conditions $x(0)$ with zero complexity. On the other hand, a remarkable result, due to Martin-Löf, shows that almost all (in the sense of the Lebesgue measure) binary sequences $\{a_1, a_2, ..., a_T\}$ have maximum complexity, i.e. $K \simeq T$, in agreement with the fact that $\lambda_1 = \ln 2$ [18].

The above results can be summarized with a sort of slogan condensating the deep link between the unpredictability of chaotic systems (i.e. at least one positive Lyapunov exponent) with the impossibility to compress the chaotic trajectories beyond the limit given by the fact that the Kolmogorov-Sinai entropy is positive [3]:

Complex = **Incompressible** = **Unpredictable**.

6 Complexity and Information in Linguistics, Genomics and Finance

In this section we briefly describe some examples of disciplines for which the concepts discussed so far could represent important tools of investigation.

Let us first study the sequences of characters representing texts written in a given language. This is the most familiar example of sequences of characters which bring and transmit information. It is interesting to see how the previously stated slogan linking complexity, compressibility and predictability applies to language. Consider, without loss of generality, the English language as our case example. The entropy of the English language can be defined as the minimum number of bits per character necessary to encode an ideally infinite message written in English. In order to estimate this quantity one should be able to subtract the unavoidable redundancy which always comes along with any linguistic message. The redundancy can also be seen as the number of constraints (for instance lexical or grammatical rules) imposed on the English text. For example the fact that a **q** must always be followed by a **u** or the impossibility to have two subsequent **h** are dependencies that make the English language more redundant. Rules of grammar, parts of speech, and the fact that we cannot invent words all make English redundant as well. Redundancy is actually benefical in order to make the message transmission efficient in *noisy* conditions or when only part of a message comes across. For example if one hears "Turn fo th lef!", one can make a fairly good guess as to what the speaker meant.

Redundancy makes language more predictable. Imagine watching a sequence of symbols emitted on a ticker-tape. The question one could ask is how much information will be added by the next symbol s_i once one already knows the sequence $s_1 ... s_{i-1}$. How much information will be gained upon seeing s_i also fixes the amount of surprise we experience. An extreme surprise will convey a large amount of information, while if one can reliably predict the next symbol from context, there will be no surprise and the information gain will be low. The entropy will be highest when you know least about the next symbol, and lowest when you know most. Shannon, with an ingenious experiment [23], showed that the entropy of the English text is something between 0.6 and 1.3 bits per character [21]. He devised an experiment where he was asking subjects to guess letters in a phrase one by one using only the knowledge of the letters previously guessed. At each step, he recorded the total number of guesses taken to reach the correct letter for that position. Shannon looked at the sequence of numbers of guesses as an encoding of the original sentence and used the entropy of the resulting random variable as an estimate for the entropy of an English letter.

More generally, estimating the complexity of a generic symbol sequence is a formidably difficult task which involves the identification and extraction of non-trivial long-range correlations inside the sequences. Kolmogorov proposed the following strategy [28]. Suppose that a subject knows the probability p_i that the next symbol in the sequence is the i-th character of the alphabet, conditioned by the appearance of the previous characters. The subject scans through the text and calculates the running averages of $-\log p_k$ where k indicates the actual character observed at each time. Kolmogorov argued that, if the probabilities p_i were correct, this average would converge to the entropy of the sequence. This approach can be seen as a natural by-product of another very interesting approach to estimate the complexity of a string which was proposed by Kelly [11], shortly after Shannon's seminal paper. This approach is related to the concept of *gambling* and it deserves a short discussion.

6.1 From gambling to entropy estimate

Let us now discuss the deep relation existing between information theory and the optimal strategy in building a portfolio. In particular, compare the growth rate of wealth and the entropy rate of a generic sequence of characters, for instance one describing the market (i.e. of the prices evolving according probabilistic rules). This apparently surprising result was originally introduced by Kelly [11] who was looking for an interpretation of the information theory outside the context of communication. Later, many authors reconsidered and generalized this approach for the growth-optimal investment strategy.

In order to illustrate the basic idea, we consider the simplest case of a discrete time price of a unique stock:

$$S_{t+1} = u_t S_t \, ,$$

where S_t is the price at time t and the gain factors u_t are independent identically distributed random variables, which can assume m possible values. Let us consider an investor, with wealth S_0 at the initial time, who decides to gamble on such a stock market many times. At each time t the agent invests a fraction l_t of the capital in stock, and the rest in risk-less security, i.e. a bank account. For simplicity, we set the case of the risk-less rate to zero. It is easy to realize that at time t, the wealth S_t is given by the following multiplicative random process:

$$S_{t+1} = (1 - l_t)S_t + u_t l_t S_t = [1 + (u_t - 1)l_t]S_t \ .$$

The problem is to find the optimal strategy, i.e. the optimal values of $l_1, l_2, \ldots$, which maximize the capital at the final time T.

In the financial literature, the standard way to face the problem is to introduce the so-called utility function $U(S_T)$, where U is a monotonic convex function that depends on the preferences of the investor, and to maximize the average of $U(S_T)$ [8]. If T is large, using the large numbers law, one has that the exponential growth of S_T is constant with probability one. That is:

$$\lambda = \lim_{T\to\infty} \frac{1}{T} \ln \frac{S_T}{S_0} = \langle \ln[1 + (u-1)l] \rangle = \sum_{i=1}^{m} \ln[1 + (u^i - 1)l] p_i \,,$$

where p_i is the probability to have $u = u^i$. The optimal strategy is specified by the value l^* which maximizes $\lambda(l)$. Kelly's strategy corresponds, in terms of the utility function approach, to use $U(S_T) = \ln(S_T)$.

From the maximization procedure, one obtains a new set of probabilities:

$$q_i = \frac{p_i}{1 + (u^i - 1)l^*}$$

and

$$\lambda(l^*) = \sum_{i=1}^{m} p_i \ln \frac{p_i}{q_i} \,.$$

The quantity $\mathcal{I} = \sum_{i=1}^{m} p_i \ln p_i/q_i$ is the relative entropy (or Kullback-Leibler divergence), which is a measure of the statistical distance between two distributions $\{p_i\}$ and $\{q_i\}$ (we shall come back to the definition of relative entropy in the coming sections). In the "symmetric" case, i.e. $m = 2$ and $u^1 - 1 = 1 - u^2$ one has the very transparent formula

$$\lambda(l^*) = \ln 2 - h$$

where $h = -\sum_i p_i \ln p_i$ is the Shannon entropy of the random process describing the stock market.

The above results show the connection between the information available from the past data and the growth rate of wealth in an optimal strategy: in the case of a process with the maximal entropy, i.e. $h = \ln 2$ the growth rate is zero. The simple case treated by Kelly can be generalized to more realistic cases with many stocks and time dependent stochastic processes, e.g. Markovian process. However, the connection between the growth rate of wealth and the rate entropy of the market is still valid (see [8] Chap.s 6 and 15).

The concept of gambling has also been applied in a linguistic context. Cover and King [7] have proposed a strategy for estimating the entropy of English language which extends Shannon's technique of guessing the next symbol. In particular, they devised a gambling strategy where the gambler bets on the outcome of the next symbol.

By generalising Kelly's strategy, they imagine a gambler with an initial capital $S_0 = 1$. At the generic time step $k+1$, i.e. after k characters $x_1, ..., x_k$ of the string are known, the gambler bets the whole capital over the d possible symbols ($d = 27$ in [7]). On every symbol, he puts a fraction of his capital, given by the probability of occurrence of this symbol conditioned to the knowledge of the k previous symbols:

$$q(x_{k+1}|x_k, ..., x_1), \tag{29}$$

with the condition $\sum_{x_{k+1}} q(x_{k+1}|x_k, ..., x_1) = 1$. This scheme is called *proportional gambling* because the whole capital is bet proportionally to the probabilities of occurrence of the possible symbols, conditioned to the knowledge of the past symbols. For each bet, the maximum possible gain is equal to d, i.e. to the cardinality of the alphabet. The capital changes recursively according to the expression:

$$S_{k+1}(x_1, ..., x_{k+1}) = dq(x_{k+1}|x_k, ..., x_1)S_k(x_1, ..., x_k), \tag{30}$$

where $k = 1, 2, ...$ It can be shown [7] that for any sequential gambling scheme $q(x_{k+1}|x_k, ..., x_1)$, one has

$$(k - E\log_d S_k(x_1, ..., x_k)) \log_2 d \geq K(x_1, ..., x_k), \tag{31}$$

where $E\log_d S_k(x_1, ..., x_k)$ indicates the expectation value of $\log_d S_k(x_1, ..., x_k)$ along the sequential scheme $q(x_{k+1}|x_k, ..., x_1)$ and $K(x_1, ..., x_k)$ is the algorithmic complexity of the string $x_1, ..., x_k$. The equality holds if and only if $q(x_{k+1}|x_k, ..., x_1) = p(x_{k+1}|x_k, ..., x_1)$, $p(x_{k+1}|x_k, ..., x_1)$ being the true conditional probability of the process. This result (which generalizes Kelly's [11]) states that an optimal gambling scheme allows for an estimate of the Algorithmic Complexity of a finite string. In the limit of an infinite sequence emitted by an ergodic source, this optimal scheme would allow for an optimal estimation of the Shannon entropy of the source. Using this scheme, implemented on humans, Cover and King estimated the entropy of English to a value on the order of 1.3 bits/symbol, in agreement with the earlier estimates by Shannon, also based on humans.

As a by-product of the previous result, it turns out that one can use the gambling scheme as a compressor for finite strings. In particular, using an alphabet of d symbols, it is possible to construct a deterministic compression scheme that allows for saving $\log_2 S_n$ bits in the coding of a sequence of n symbols. This results in a length of the compressed sequence equal to $(1 - 1/n \log_d S_n) \log_2 d$ which, in the limit of optimal gambling scheme, converges to the Algorithmic Complexity of the string. The deterministic compression scheme uses a gambler and an identical twin to the gambler who shares the same thoughts processes of the gambler. The gambler and the twin play the role of encoder and decoder respectively. We refer to [7], which contains an extensive bibliography, for the details of the compression scheme.

6.2 Looking for relevant information

One of the most challenging issues is represented by the huge amount of data available nowadays. While this abundance of information and the extreme accessibility to it represents an important cultural advance, one is also faced with the problem of retrieving relevant information. Imagine entering the largest library in the worlds, in search of all the relevant documents on your favorite topic. Without the help of an efficient librarian, this would be a difficult, perhaps hopeless, task [4]. The desired references would likely remain buried under tons of irrelevant volumes. Clearly, the need for effective tools for information retrieval and analysis is becoming more urgent as the databases continue to grow.

To accomplish such an ambitious task, we must determine what is useful or relevant information and where and how it is coded–say, in a written document. This is a non-trivial problem since "information" means different things in different contexts [1, 30]. That is, information has no absolute value, but depends on some specific "filters" each observer imposes on his data. Consider a simple coin-toss experiment: The gambler will likely be interested only in the outcome (heads/tails) of the toss. The physicist on the other hand, might be interested in whether the outcomes reveal anything on the nature of the coin—such as whether it is honest or dishonest.

Information extraction takes place *via* a two-step process. First comes the so-called *syntactic* step, where one identifies the structures present in the messages without associating any specific meaning to them. It is only in the second (or *semantic*) step that comprehension of meaning takes place, by connecting the syntactic information to previous experience and knowledge.

As an example of this two-step process, consider how we might identify the language in which a given text is written. In the first step we scan through the text and identify the syntactic structures: articles, verbs, adjectives, etc. But only one who "knows" the language can carry out the second phase, where the incoherent jumble of syntactic data is summarized in the specific meaning of the sentences. Other examples of this two-step process are how we recognize the subject of a given text, its historical background, and possibly its author.

The very same problem of retrieving relevant information turns out to be crucial for another field that has experienced impressive growth in the last few years: the bioinformatics.

Bioinformatics represents an emblematic example of a situation where too much information comes into play. Since it is impossible to learn everything about all living organisms, biologists *solve* the dilemma by focusing on some model organisms and trying to find out as much as they can about them. On the other hand, all available evidence indicates that the complete information necessary for making organisms is encoded in the genome, i.e. in one or several DNA molecules composed by strings of nucleotides. Having a DNA sequence is like having a text written in an unknown language, coded with an alphabet of four letters (A,C,G,T). The big challenge in this field is that the typical sizes of the few genomes completely sequenced up to now range from 10^6 to 10^{11} bases, numbers which pose two main problems: how and where to store all these data and, more importantly, how to extract relevant information from them.

When analyzing a string of characters, for instance a text, a DNA or a protein sequence, the main question is to extract the information it contains. This is known as sequence analysis or information retrieval. At present, most of bioinformatics and of the activities in the area of the so-called computational linguistics is concerned with sequence analysis.

Here are some of the questions studied in computational linguistics: automatic recognition of the language in which a given text is written, authorship attribution, automatic indexing and classification of large corpora of documents, retrieval of specific documents according to their content, reconstruction of phylogenetic trees for languages, etc..

On the other hand, current trends in bioinformatics include: (i) *gene finding.* One of the most important steps in the analysis of a new DNA sequence is finding out whether or not it contains any genes, and if so, determining exactly where they are. This process is also called segmentation; several algorithms have been devised for this problem and implemented as systems that are widely used today. (ii) *gene function prediction.* Suppose you have identified a gene. What is its role in the biochemistry of its organism? Sequence databases can help us in formulating reasonable hypotheses based for instance on homology considerations. (iii) *protein 3D structure prediction.* The structure of the protein is directly related to the protein's functionality, probably even determining it. Moreover, 3D structure is more highly conserved than the primary structure (the sequence). (iv) *reconstruction of phylogenetic trees.* A phylogenetic tree gives a graphical representation of the evolution of contemporary species from their possible common ancestor. Typically this can be done by defining and measuring some sort of evolutionary distance and giving an estimate of the time needed for this divergence to occur.

6.3 Relative entropy and remoteness of sequences

One of the most important tools in sequence analysis is the definition of suitable measures of remoteness between pairs of sequences which could be used for information retrieval. This can be done in general by looking for regularities in the sequences. The question is then how to detect such similarities? The example from bioinformatics is again enlightening. In this case a measure of similarities is obtained by means of an *alignment* procedure: roughly speaking, different sequences are made to coincide as much as they can by inserting gaps (corresponding to insertions or deletions with respect to the possible common ancestor sequence) and paying a corresponding *price.* The price to pay to align two sequences is proportional to their remoteness.

Another possible way to define the concept of remoteness between two sequences of characters (used in Linguistics or Bioinformatics) comes from the definition of Kolmogorov complexity. This concept is particularly useful for information retrieval purposes since, as mentioned above, Kolmogorov complexity is a property of a single sequence and does not refer to the existence of a source of messages. Using Algorithmic Complexity one can hope to be able to extract information from individual sequences instead of referring to the sources which are typically not known (think of a text or a DNA sequence).

In this framework an important concept to recall is that of relative entropy or Kullback-Leibler divergence [8] which is a measure of the statistical remoteness between two distributions. Its essence can be easily grasped with the following example. Let us consider two ergodic sources A and B emitting sequences of independent 0 and 1: A emits a 0 with probability p_A and 1 with probability $1-p_A$, while B emits 0 with probability p_B and 1 with probability $1-p_B$. As already described, the compression algorithm applied to a sequence emitted by A will be able to encode the sequence almost optimally, i.e. with an average number of bits per character equal to $-p_A \ln p_A - (1-p_A)\ln(1-p_A)$. This optimal coding will not be the optimal one for the sequence emitted by B. In particular, the entropy per character of the sequence emitted by B in the coding optimal for A will be the cross entropy per character:

$$h(B\|A) = -p_B \ln p_A - (1-p_B) \ln(1-p_A). \tag{32}$$

Similarly, the entropy per character of the sequence emitted by B in its optimal coding is $-p_B \ln p_B - (1-p_B) \ln(1-p_B)$. The number of bits per character wasted to encode the sequence emitted by B with the coding optimal for A is the relative entropy per character of A and B:

$$d(B||A) = -p_B \ln \frac{p_A}{p_B} - (1-p_B) \ln \frac{1-p_A}{1-p_B}. \tag{33}$$

A linguistic example will help clarify the situation: transmitting an Italian text with a Morse code optimized for English will result in the need for transmitting more bits than another coding optimized for Italian: the difference is a measure of the relative entropy.

The concept of relative entropy is crucial in the so-called language modeling. We have already discussed the problem of defining the entropy of a language. In the simplest approximation, one can define a language with a probability distribution $p(x)$ where x indicates a generic sequence of characters (a word). Since the distribution $p(x)$ is typically unknown, one should construct a suitable modeling by making a guess $q(x)$ about what the actual distribution could look like. Now the question is how good an estimate of $p(x)$ is $q(x)$. An answer to this question can be given in terms of the relative entropy. The relative entropy between the $q(x)$ distribution and the actual distribution $p(x)$ decreases as the language modeling improves.

More generally, relative entropy can be considered as a measure of the remoteness between two individual sequences; it represents an important tool for classification and information retrieval purposes. Suppose one wants to estimate the "distance" (or similarity) between texts A and B in terms of their information content. For instance, for two texts written in different languages (e.g. English and Italian), their "distance" is a measure of the difficulty experienced by a typical speaker of tongue A in understanding the text written in language B. The remoteness of one text from the other can be measured by estimating the relative entropy. We should remark that the relative entropy is not a distance (metric) in the mathematical sense: it is neither symmetric, nor does it satisfy the triangle inequality. As we shall see below, in many applications such as phylogenesis, it is vital to define a true metric that measures the actual distance between sequences. In this perspective a very important contribution has been made by the group of Li [17] where a rigorous definition of distance between unaligned sequences was proposed, by using the information theoretical concepts of Kolmogorov complexity [18].

6.4 Data compression and measures of complexity

One last important point is how to obtain approximate measurements of quantities such as the Kolmogorov complexity or the relative entropy. We have already noted the impossibility of computing the Kolmogorov complexity related to Turing's theorem on the halting problem and to Gödel's theorem [18].

Despite the impossibility to compute the Algorithmic Complexity of a sequence, one has to recall that there are algorithms explicitly conceived to provide a good approximation [18]. Since the Algorithmic Complexity of a string fixes the minimum number of bits one need to reproduce it (optimal coding), it is intuitive that a typical compression software (zipper), besides trying to reduce the space occupied on a memory storage device, can be considered as an entropy meter. The better the compression algorithm, the closer the length of the zipped file to the optimal coding limit and the better the estimate of the Algorithmic Complexity provided by the zipper. It should be remarked that the zippers can provide a reliable approximation (always an upper bound) to the Algorithmic Complexity of a typical string produced by a stationary stochastic process with finite memory (finite correlation length). By compressing the sequence of digits of π a compressor would never realize that there is a simple rule to generate it.

We have already discussed how gambling techniques can be used for compression purposes. Now we focus on specific zipping algorithms. A great improvement in this field is represented by the so-called Lempel and Ziv 77 algorithm [29]. This algorithm zips a file by exploiting the existence of repeated sub-sequences of characters in it. Its compression efficiency becomes optimal as the length of the file goes to infinity. This algorithm has been extensively studied and many variants and applications have been drawn from it.

In the field of the so-called computational linguistics, there have been several contributions showing how data compression techniques [27] can be useful in solving different problems: language recognition, authorship recognition or attribution, language classification, classification of large corpora by subject, etc. A detailed description of the state of the art is outside our present scope and we refer the reader to the specialized literature. Just to give an example, we quote one recent work [2] where data compression techniques were used to

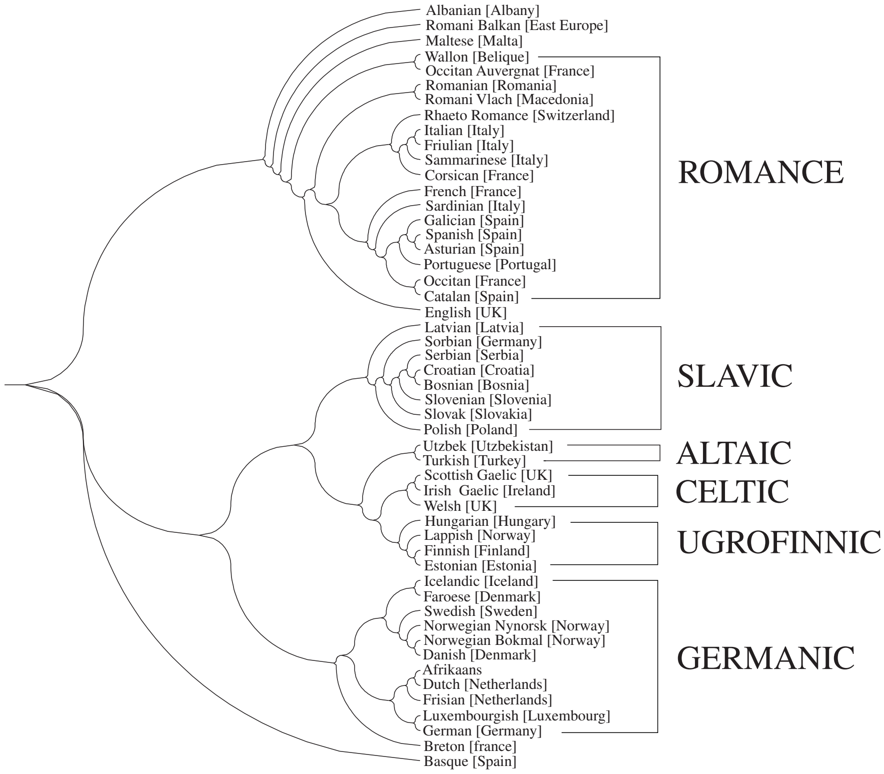

Fig. 3. This figure illustrates the phylogenetic-like tree constructed on the basis of more than 50 different versions of the "The Universal Declaration of Human Rights". The tree is obtained using the Fitch-Margoliash method applied to a symmetrical distance matrix whose elements are computed in terms of the relative entropy between pairs of texts. The tree features essentially all the main linguistic groups of the Euro-Asian continent (Romance, Celtic, Germanic, Ugro-Finnic, Slavic, Baltic, Altaic), as well as few isolated languages such as Maltese, typically considered an Afro-Asian language, and Basque, classified as a non-Indo-European language and whose origins and relationships with other languages are uncertain. The tree is unrooted, i.e. it does not require any hypothesis about common ancestors for the languages. What is important is the relative positions between pairs of languages. The branch lengths do not correspond to the actual distances in the distance matrix. It is important to stress how this tree is not intended to reproduce the current trends in the reconstruction of genetic relations among languages. It is clearly biased by the use of modern texts for its construction. In the reconstruction of genetic relationships among languages, one is typically faced with the problem of distinguishing *vertical* (i.e. the passage of information from parent languages to child languages) from *horizontal* transmission (i.e. which includes all the other pathways in which two languages interact). Horizontal borrowings should be expunged if one is interested in reconstructing the actual genetic relationships between languages.

define a suitable measure of distance between two sequences, to be used for authorship attribution and language phylogenesis (see Fig. 3).

Of course the possibilities of data-compression-based methods go beyond computational linguistics. Another important example is that of genetic problems. Here also there have been important contributions and we refer to [17, 10] for a recent overview.

It is evident how the specific features of data compression techniques in measuring entropic quantities make them potentially very important also for fields where human intuition can fail: DNA and protein sequences, as already mentioned, but also geological time series, stock market data or medical monitoring.

Acknowledgments The authors are grateful to R. Baviera, D. Benedetto, L. Biferale, G. Boffetta, A. Celani, E. Caglioti, M.Cencini, G. Mantica, A. Puglisi, M. Serva and D. Vergni for the collaboration in the course of the years and for many enlightening discussions; and A. Baronchelli for the preparation of figure 3, extracted from his master thesis.

References

1. R. Badii and A. Politi, *Complexity, Hierarchical Structures and Scaling in Physics*, (Cambridge University Press, Cambridge, 1997).
2. D. Benedetto, E. Caglioti and V. Loreto, "Language trees and zipping", *Phys. Rev. Lett.* **88**, 048702 (2002).
3. G. Boffetta, M. Cencini, M. Falcioni and A. Vulpiani, "Predictability, a way to characterize complexity", *Phys. Rep.* **356**, 367 (2002).
4. J.L. Borges, "La biblioteca de Babel", in *Ficciones*, Primera edición, Buenos Aires, Editorial Sur, (1944).
5. C. Cercignani, *Ludwig Boltzmann. The man who trusted atoms*, (Oxford University Press, Oxford 1998).
6. G.J. Chaitin, *Information, randomness and incompleteness* (2nd edition, World Scientific, Singapore 1990).
7. T. Cover and R.C. King, "A convergent gambling scheme of the entropy of english", *IEEE Trans. Inf. Th.*, **24**, 413 (1978).
8. T. Cover and J. Thomas, *Elements of information theory*, (Wiley, New York 1991).
9. J. Ford, "How random is a coin tossing ?" *Physics Today*, **36**, 40 (1983).
10. Proceedings of the *IEEE Computer Society Bioinformatics Conference (CSB'02)*, (IEEE publisher, Palo Alto, California 2002).
11. J. Kelly, "A new interpretation of information rate", *Bell. Sys. Tech. Journal* **35**, 917-926 (1956).
12. A.I. Khinchin, *Mathematical Foundations of Information Theory*, (Dover, New York, 1957).
13. A.N. Kolmogorov, "On the Shannon theory of information transmission in the case of continuous signals", *IRE Trans. Inf. Theory.* **1**, 102 (1956).
14. A.N. Kolmogorov, "New metric invariant of transitive dynamical systems and auto-morphism of Lebesgue spaces", *Dokl. Akad. Nauk. SSSR* **119**, 861 (1958).

15. A.N. Kolmogorov, "Three Approaches to the Concept of the Amount of Information", *Prob. Info. Trans.* **1**, 1 (1965).
16. K. Huang, *Statistical Mechanics*, (Wiley, London, 1967).
17. M. Li, J. Badger, X. Chen, S. Kwong, P. Kearney and H. Zhang, "An information-based sequence distance and its application to whole mitochondrial genome phylogeny", *Bioinformatics* **17**, 149 (2001).
18. M. Li and P. Vitányi, *An introduction to Kolmogorov complexity and its applications*, (Springer, New York, 1997).
19. E. Lorenz, "Deterministic Nonperiodic flow", *J. Atmos. Sci* **20**, 130 (1963).
20. E. Ott, *Chaos in Dynamical Systems*, (Cambridge University Press, New York), Chap. 5, pp. 151-183.
21. J.R. Pierce, *Introduction to information theory: Symbols, signals and noise* (Dover Publications, Inc., New York, 2nd edition, 1980).
22. C.E. Shannon, "A mathematical theory of communication", *The Bell System Technical J.* **27**, 379 and 623 (1948).
23. C. Shannon, "Prediction and Entropy of Printed English", *Bell Systems Technical Journal* **30**, 50 (1951).
24. Ya.G. Sinai, "On the concept of entropy for a dynamical system", *Dokl. Akad. Nauk. SSSR* **124**, 768 (1959).
25. R.J. Solomonoff, "A formal theory of inductive inference (part 1 and 2)", *Inform. Contr.* **7**, 1 (part 1) and 224 (part 2) (1964).
26. A. D. Wyner, 1994 Shannon Lecture, "Typical sequences and all that: Entropy, Pattern Matching and Data Compression", *IEEE Information Theory Society Newsletter*, July 1995.
27. I.H. Witten, A. Moffat and T.C. Bell, *Managing Megabytes*, (Morgan Kaufmann Publishers, 1999).
28. A.M. Yaglom and J.M. Yaglom, *Probability and Information*, (Science House, Leningrad 1973) chapter IV, Part.3, pp.236-329.
29. J. Ziv and A. Lempel, "A universal algorithm for sequential data compression", *IEEE Trans. Inf. Th.*, 337 (1977).
30. W.H. Zurek (editor), *Complexity, Entropy and Physics of Information* (Addison-Wesley, Redwood City, 1990).