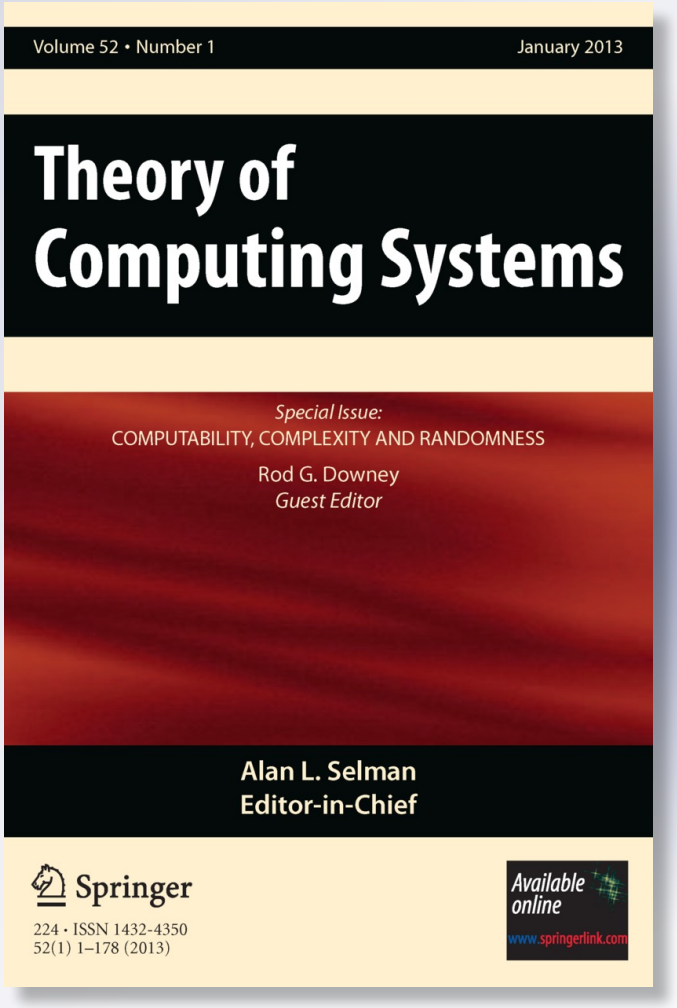
Using Ideas of Kolmogorov Complexity for Studying Biological Texts

Boris Ryabko, Zhanna Reznikova, Alexey Druzyaka & Sofia Panteleeva

Theory of Computing Systems

ISSN 1432-4350
Volume 52
Number 1

Theory Comput Syst (2013) 52:133-147
DOI 10.1007/s00224-012-9403-6



Springer

Theory Comput Syst (2013) 52:133–147
DOI 10.1007/s00224-012-9403-6

Using Ideas of Kolmogorov Complexity for Studying Biological Texts

Boris Ryabko · Zhanna Reznikova · Alexey Druzyaka · Sofia Panteleeva

Published online: 3 May 2012

Abstract Kolmogorov complexity furnishes many useful tools for studying different natural processes that can be expressed using sequences of symbols from a finite alphabet (texts), such as genetic texts, literary and music texts, animal communications, etc. Although Kolmogorov complexity is not algorithmically computable, in a certain sense it can be estimated by means of data compressors. Here we suggest a method of analysis of sequences based on ideas of Kolmogorov complexity and mathematical statistics, and apply this method to biological (ethological) "texts." A distinction of the suggested method from other approaches to the analysis of sequential data by means of Kolmogorov complexity is that it belongs to the framework of mathematical statistics, more specifically, that of hypothesis testing. This makes it a promising candidate for being included in the toolbox of standard biological methods of analysis of different natural texts, from DNA sequences to animal behavioural patterns (ethological "texts"). Two examples of analysis of ethological texts are considered in this paper. Theses examples show that the proposed method is a useful tool for

Research was supported by Russian Foundation for Basic Research (grants 12-07-00125 and 11-04-00536), by the Integrated Project of Siberian Branch RAS (grant N 21), the Program "Living Nature" of the Presidium of Russian Academy of Science (grant No. 30.6), and by the Program of cooperative investigations of SB RAS and third parties (grant No. 63).

B. Ryabko (✉)
Siberian State University of Telecommunications and Information Sciences, Novosibirsk, Russia
e-mail: boris@ryabko.net

B. Ryabko
Institute of Computational Technology of Siberian Branch of Russian Academy of Science, Novosibirsk, Russia

Z. Reznikova · A. Druzyaka · S. Panteleeva
Institute of Systematics and Ecology of Animals, Russian Academy of Science, Novosibirsk, Russia

Z. Reznikova · A. Druzyaka · S. Panteleeva
Novosibirsk State University, Novosibirsk, Russia

distinguishing between stereotyped and flexible behaviours, which is important for behavioural and evolutionary studies.

Keywords Kolmogorov complexity · Hypothesis testing

1 Introduction

Sequences of symbols from a finite alphabet (or "texts") appear as basic objects of study in many scientific fields, including molecular biology and genetics (genetic texts), linguistics (literary and musical texts), zoosemiotics (animal communications), ethology (behavioural sequences) and others. The main problem that researchers face in these domains is finding an adequate model which would allow for assessment of certain characteristics of a text, while using a relatively small number of parameters. One of the most popular approaches is based on the description of sequences by stochastic processes. Modeling of DNA sequences by Markov processes of finite depth, or connectivity, can serve as a good example here. This way of looking at the problem presents some methodological limitations. Indeed, it is difficult to imagine that a text by Shakespeare or a frog's genome could be described adequately by a stochastic process with a relatively small number of parameters. In order to obtain an approximately adequate model of a text of a certain type, it is necessary to increase the number of parameters which, in turn, should be estimated statistically, based on real data. For example, if the frequency of each letter in a genetic text depends on $n, n > 0$ previous letters, then the number of parameters equals 4^n. When $n = 10$, the numbers of parameters is 2^{20}, that is, about one million, and one needs several times more data to estimate this quantity exactly. Moreover, the finite-memory assumption (with memory say, 5 or 10) is even less realistic for texts in natural human languages. As it was noted in [31], if we meet the word "lemma" in a book, we hardly meet the word "love" in it. (This means that, in fact, the memory of human-written texts is unlimited.)

One can approach the methodological problems listed above with methods that are close in spirit to the ideas of Kolmogorov complexity. More precisely, the degree of complexity of a "text" could be estimated by its Kolmogorov complexity. Although Kolmogorov complexity is not algorithmically computable, it can be, in a certain sense, estimated by means of data compressors [26, 27]. This approach does not contradict the probabilistic one, because if one looks at a sequence as generated by stochastic process, the length of the compressed sequence can be considered an estimate of the Shannon entropy, which, in turn, equals Kolmogorov complexity. It is worth noting that a quantitative estimation of complexity of sequences in natural "texts" is of interest in its own right. There is a huge body of literature that analyses symbolic sequences by means of Kolmogorov complexity, including diagnostic of the authorship of literary and musical texts [1, 3, 5, 14, 15, 24, 25, 30]. The importance of the use of ideas and methods of Kolmogorov complexity in biological sciences can hardly be overestimated. Such methods were used for estimation of closeness of DNA sequences [1, 3, 14, 15], construction of phylogenetic trees [3, 5, 14, 15], and distinguishing between innate and acquired behaviour in ants [19]. Unfortunately, these

approaches do not give a possibility to use methods of mathematical statistics and, in particular, hypothesis testing. This limits the applicability of ideas of Kolmogorov complexity to biological studies, because, since Fisher's classic works [6], statistical testing of biological hypotheses became the main method of quantitative analysis of biological data.

In this paper we suggest an approach which allows us to combine the advantages of methods based on Kolmogorov complexity with classic methods of testing statistical hypotheses. As distinct from other approaches to the analysis of sequences by means of Kolmogorov complexity, we stay within the framework of mathematical statistics. As examples, we consider ethological "texts" of two kinds, namely, territorial behaviour of gulls and hunting behaviour of ants. In the first case we found that the complexity of territorial behaviour in gulls differ in two situations: when a trespasser approaches a resident's nest, and when it escapes from a resident's defended territory. In the second case we found that complete (successful) hunting stereotypes in members of a natural ant colony are characterized by smaller complexity than incomplete hunting stereotypes in naive laboratory-reared ants.

In sum, being applied to members of different branches of Animal Kingdom, the proposed method appears to be a promising tool to distinguish between "basic" stereotypical behavioural patterns and flexible behaviour. We believe that this method is applicable to the analysis of biological texts of different kinds.

2 Description of the Tests

The goal here is to estimate the complexity of sequences of different kinds and to use these estimates to test hypotheses. This gives the possibility to make decisions on the basis of standard statistical tests. For example, two sets of sequences could be DNA of two different species (say, viruses or bacteria), and the problem is to compare complexities of these sequences. Similar problems arise in many fields of biology and other sciences.

In order to describe the suggested approach precisely, in the next two subsections we formally define the notation "complexity" and describe possible hypotheses and statistical tests.

2.1 General Scheme of the Suggested Method and Its Applicability

First we describe the scheme of the suggested approach. Let there be a sequence $x = x_1 \dots x_t$, $t > 0$, of letters from a finite alphabet A and let φ be a data compressor. We denote the compressed sequence by $\varphi(x)$, its length by $|\varphi(x)|$ and define the complexity (per letter) as follows:

$$K_\varphi(x_1 \dots x_t) = \big|\varphi(x_1 \dots x_t)\big|/t. \tag{1}$$

Generally speaking, we suggest to use $K_\varphi(x)$ for hypothesis testing.

It turns out that, under certain conditions, the proposed approach can be used to evaluate hypotheses about the Kolmogorov complexity of the considered sequences; see Theorem 1 below (even though it is impossible to calculate Kolmogorov complexity). These conditions are as follows:

(i) the considered sequences are generated by stationary ergodic sources, and
(ii) the data compressor φ is a universal code.

The definition of a stationary ergodic source is given in Appendix 1 as well as the definitions of universal codes, Kolmogorov complexity, Shannon entropy and some auxiliary notions. Let us discuss the conditions (i) and (ii) from a practical point of view. Informally, a universal code can "compress" a sequence $x_1 \dots x_t$ up to its Shannon entropy (per letter) if the sequence is generated by a stationary ergodic source; see Appendix 1. Moreover, if a universal code φ is applied to the sequence $x_1 \dots x_t$, the ratio $|\varphi(x_1 \dots x_t)|/t$ goes to the Kolmogorov complexity (with probability 1). Hence, according to (1), the limit of $K_\varphi(x_1 \dots x_t)$ is equal to the limit Kolmogorov complexity per letter, see Theorem 1. Nowadays there are many efficient universal codes, which are described in numerous papers, see, for ex., [12, 22, 23]. It is important for practical applications that the modern data compressors (or archivers) are based on the universal codes and, hence, the main properties of universal codes are valid for them (as far as asymptotic properties can be valid for a real computer program). So, the condition (ii) is not restrictive.

The condition (i) that the considered sequences are generated by stationary ergodic sources (SES) is potentially too restrictive. For example, if sequences are (digitized) bird songs, it seems to be natural to apply the model of SES, because the number of possible songs is very large and they can be created under conditions that remain unchanged. If the sequences are genetic texts of several species of birds, seemingly, the applicability of the SES model is not so obvious, because, say, the number of species (and, hence, corresponding DNA sequences) is limited. And, at last, if someone investigates the complexity of Tolstoy's novels (or Bach's fugues), the SES model seems to be even less natural. So, the feasibility of the SES model should be assessed beforehand by a researcher basing on biological (linguistics, musical, etc.) considerations.

Besides, it is worth noting that stationarity and ergodicity is one of the most general assumptions in statistics: it is much more general than, say, the finite-memory assumption commonly used for such problems. Testing this assumption directly is not possible, since it would require formulating an even more general alternative. Tests for stationarity exist within specific parametric models, but they are not applicable in the general setting such as ours.

2.2 The Tests

In this part we describe the suggested approach for testing certain hypotheses about Kolmogorov complexity of sequences. The main idea of the suggested approach is very simple and can be formulated as follows: Apply some universal code φ to estimate the Kolmogorov complexity of a word $x_1 \dots x_t$. Then, use (consistent) statistical tests to study $\varphi(x_1 \dots x_t)/t$ $(= K_\varphi(x_1 \dots x_t))$ in the same manner as one would study any other natural parameter, such as the weight, the length, the speed, etc. If the assumptions (i) and (ii) hold and if the length of sequences (t) grows, the following Theorem 1 guarantees that the obtained statistical inference can be interpreted as a result about Kolmogorov complexity (per letter) of the sequences.

Theorem 1 *Let there be a stationary ergodic source generated letters from a finite alphabet and a universal code φ. Then, with the probability* 1

$$\lim_{t\to\infty} t^{-1}\big|\varphi(x_1\dots x_t)\big| = \lim_{t\to\infty} t^{-1}K(x_1\dots x_t), \tag{2}$$

where $x_1\dots x_t$ is generated by the source and $K(x_1\dots x_t)$ is the Kolmogorov complexity.

Proof is given in Appendix 2. □

Here we analyse theoretically one example of the use of this approach; numerical examples will be given in the next section.

First, we briefly describe some notions of mathematical statistics that we use. We will consider only tests for assessing whether two independent samples of observations have equally large values, because this example can be easily extended to other statistical problems. Let there be two sets of samples S_1, S_2 and the following statistical hypotheses be considered: The null hypothesis H_0 is that the elements of both sets obey the same probability distribution, whereas the alternative hypothesis (H_1) is that elements of the sets S_1, S_2 obey different distributions $F_1()$ and $F_2()$ and one distribution is stochastically greater than the other, i.e. either for every x $F_1(x) < F_2(x)$ or for every x $F_2(x) < F_1(x)$. (As it is typical of the mathematical statistics [10], this test does not consider other possibilities.)

There exist several consistent tests for testing H_0 against H_1 [10, 16].

A test is called *consistent* if its Type II error goes to 0 when $\min(|S_1|, |S_2|)$ goes to ∞, while the Type I error is not greater than a required level of significance α, $\alpha \in (0, 1)$. Note that the Type I error occurs if the test rejects H_0 when it is true and Type II error occurs if the test accepts H_0 when it is not true.

Among the known consistent tests we mention Mann–Whitney–Wilcoxon test (U-test) [10, 16], which is often used in biological research, and which will be used in our biological examples later. Informally, the main idea of this test (and many others) can be described as follows: First, order the joint sample $S_1 \cup S_2$ (say, in ascending order). Then tag each element of the resulting ordered sample with s_i if the element is coming from the sample S_i. For example, the resulting sequence of tags may look like the following

$$s_1s_2s_1s_1s_2s_1s_2s_2s_2s_1s_1s_2s_2s_1. \tag{3}$$

(Here the first symbol s_1 means that the smallest element of the two samples belongs to S_1, second largest value belongs to S_2, etc.) The hypothesis testing procedure can be informally described as follows: if the data from two sets are uniformly mixed, like in (3), the hypothesis H_0 is accepted. On the other hand, if the data looks like two separate subintervals, for example

$$s_1s_1s_1s_1s_1s_1s_1s_2s_2s_2s_2s_2 \quad \text{or} \quad s_2s_2s_2s_2s_1s_1s_1s_1s_1, \tag{4}$$

then H_0 should be rejected. (In a certain sense the examples (4) represent limiting situations, and they will be used later in the proof.)

Let us come back to the problem of estimating the complexity of sequences. Now the sets S_1, S_2 contain sequences of the length t, $t \geq 1$, generated by stationary ergodic sources. Our goal is to find a statistical test that can distinguish between the two following hypotheses: $H_0 =$ {the sequences from both sets are generated by one source} and $H_1 =$ {the sequences from the different sets are generated by stationary and ergodic sources with different Kolmogorov complexities (per letter of generated sequences)}. First of all, let us note that for sequences $x_1 x_2 \ldots$ generated by a stationary ergodic source Ω there exists a constant k_Ω such that

$$\lim_{t\to\infty} t^{-1} K(x_1 \ldots x_t) = k_\Omega \tag{5}$$

with probability 1; see [32] and Appendix 1. It is natural to call the constant k_Ω the per-letter Kolmogorov complexity. (It is worth noting that k_Ω equals the limit Shannon entropy; see Appendix 1.) So, the definition of H_1 is correct, because, with probability 1, sequences from S_i have the same Kolmogorov complexity per letter.

In order to construct the test for H_0 against H_1 we consider an auxiliary hypothesis $H_1^* =$ {the estimation of the average complexity $K_\varphi()$ is not equally large for sequences from different sets S_1, S_2}. Let T be a consistent test for distinguishing between H_0 and H_1^* (say, the Mann–Whitney–Wilcoxon test mentioned above).

The suggested test T_φ' for H_0 against H_1 (not H_1^*!) uses a universal code φ and a consistent test T for assessing whether two independent samples of observations have equally large values (say, the Mann–Whitney–Wilcoxon test). Let there be two sets S_1, S_2 of sequences of the length t, $t > 1$, which are generated by stationary ergodic sources. The test T_φ' is as follows: First, calculate $K_\varphi()$ for all sequences from S_1, S_2 and then apply T for testing H_0 against H_1^* based on the new sets $\{K_\varphi(x), x \in S_1\}$ and $\{K_\varphi(x), x \in S_2\}$. The following theorem describes properties of T_φ':

Theorem 2 *The Type I error of the test T_φ' is not greater than α and, if* $\min(|S_1|, |S_2|) \to \infty$ *and* $t \to \infty$, *the probability to accept H_0 instead of $H_1 =$* {the sequences from the different sets are generated by stationary and ergodic sources with different Kolmogorov complexities (per letter of the generated sequences)} *goes to* 0.

Proof is given in Appendix 2. □

Comment 1 The suggested approach can be applied in the case where the lengths of sequences (t) are not constant, but obey a certain probability distribution. It can be shown that the theorem is true in this case if the distribution of the lengths is the same for both sets S_1, S_2 and the average of this distribution goes to infinity when $|S_i| \to \infty$.

Comment 2 Having taken into account Theorem 1, we can see that for any t the "common" statistical test T is used to test "common" statistical hypotheses H_0 and H_1^*, but, if the length of sequences t goes to infinity, T_φ' can be considered the test for comparison of the Kolmogorov complexity of the sequences, i.e. the test for distinguishing between H_0 and H_1. As we mentioned above, this construction can be used in many situations when statistical tests are applied. For example, one can investigate

whether the complexity of a bird song depends on the age of the bird. In this case one can estimate the complexity by (1) and calculate the correlation coefficient between K_φ and the age.

3 Biological Examples: Analysis of Ethological Experimental Data

3.1 Why Study the Complexity of Animal Behavioural Patterns?

One of the main problems in studying animal behaviour at different levels of organisation, from individual to collective behaviour of organisms, is searching for a reliable criterion for evaluating the variability and complexity of behavioural patterns. In evolutionary biology variability is known as an important mechanism of speciation in animals, and differences in behavioural patterns have high diagnostic value for species identification. Within populations behavioural variability serves as a basis for behavioural, cognitive and social types of specialization which facilitate tuning of integrative reactions of the whole animal community to unpredictable influences of its changing environment (see details in [20]). The concept of complexity of animal behaviour is still mainly intuitive. First of all, one has to distinguish between the complexity of flexible and stereotypic behaviour. In the first case we mean levels of complexity of problems to solve and decisions to make, whereas in the second case we mean the inner coordination and regularity of species-specific repertoire. Surprisingly, despite many attempts to examine the organizational complexity of signal repertoires (see, for example, [17, 18]), there are no reliable tools for studying the complexity of animal behavioural patterns.

The prevalent method of ethological studies is based on the analysis of the so-called ethograms, that is, recordings of behavioural sequences as alphabets consisting, in average, of 10–15 symbols or letters each, corresponding to a certain behavioural unit (an act). For example, hunting attacks in many species, both vertebrates and invertebrates, are organized as more or less constant sequences of acts, and can be presented roughly as a recording like this: R (running)–A (approaching)–J (jumping)–F (fight)–C (capture)–H (handling)–K (killing). Attempts to apply the probabilistic approach to describe and compare animal behaviours meet methodological difficulties, and among them the problem of the large number of parameters mentioned above. We suggest that the proposed method of evaluation of complexity of behaviours based on the concept of Kolmogorov complexity and approaches of mathematical statistics is more adequate.

3.2 First Ethological Example: Territorial Behaviour in Gulls

One of the most well-known examples of ethological "texts" is the description of behavioural sequences in gulls. In particular, since Tinbergen's classic works [28, 29], gulls are well known by their expressive territorial demonstrations towards intruders that pretend to enter their nesting territories (see, for example, [8]). Using the so called "resident–intruder" experimental paradigm, we compared territorial behaviours in gulls *Larus ridibundus* displayed in two situations: (1) an intruder is approaching a nest in which a gull is clutching (i.e. the gull is sitting on the eggs), and (2) an intruder is moving away from the nest.

We hypothesised that reactions of a resident towards an approaching intruder are more variable and "chaotic" than its reactions towards an escaping one, because, as we could observe in nature, in the first situation a resident hectically tries various ways to drive a trespasser away, whereas in the second situation it simply repeats successful combinations of behaviours. We tested this hypothesis by means of the method described above. Experiments were conducted in 2011, on a lake in South Siberia (in latitude 53.751° North, longitude 77.975° East). *Larus ridibundus* gulls hatch on nests which float in the lake. During the hatching period, we selected 24 gulls for the field experiments. In order to provoke defensive behaviour in birds, we used a taxidermically prepared gull (a "model gull") as a mock intruder. This model gull, being distantly operated by an experimenter, moved to the border of the territory of a resident bird clutching on her nest (i.e. sittings on the eggs), stayed there for 10–15 sec., and then moved away from the territory. A whole session took 60–90 seconds. All reactions of the resident were video-recorded. In order to compose a "dictionary" of gulls' territorial behaviour, we used the following protocol. Typical states of a bird, in combination with current movements were selected: the position of the bird (such as sitting on the eggs, flying, floating, etc.), its demonstrative body posture (such as "upright", "oblique", etc.), its position of wings (such as stretching, folding, etc.), and its vocalization (such as aggressive calls "cah-cah!," long calls, that is, a series of notes during which the head is dipped then raised, etc.) (see Appendix 3). From 315 combinations, 60 behavioural units were selected. As a result, we represented behavioural sequences as "texts," in which behavioural units (60 in total), singled out from video records and denoted by symbols (see Appendix 3) served as an alphabet. Pooling individual ethograms, we obtained 72 behavioural sequences composed of combinations of 60 units. Then two samples of text files were composed where each file contained from 2 to 6 sequences of symbols separated by a semicolon. In sum, we obtained two samples: the first one from 24 files that corresponded to birds' reactions towards an intruder approaching the nest (the average file size a file is 177.2 ± 2.9 bytes), and the second one from 10 files that corresponded to birds' reaction towards an escaping intruder (an average is 182.8 ± 9.5 bytes). We compressed text files with the use of the so-called KGB archiver [11] and compared the compression ratio of different behavioural sequences (Table 1). We tested the Hypothesis H_0 (the sequences from two sets are generated by one source) against H_1 (the complexity of sequences from one set is, in average, larger than the complexity of sequences from the other) by the Mann–Whitney–Wilcoxon test, as described in Sect. 3. It turned out that the files corresponding to the reactions of resident gulls towards the escaping intruder compress better than those corresponding to birds' reactions towards the approaching and staying intruder ($U = 1.97$; $p < 0.05$). So, H_0 is rejected and we can conclude that, in average, the complexity of sequences in the first set is larger than in the second. Thus, these data support our initial suggestion that reactions of a resident gull towards an approaching trespasser are more variable and "chaotic" than its reactions towards the escaping one.

3.3 Second Ethological Example: Hunting Behaviour in Ants

We analysed the hunting stereotype of ants *Myrmica rubra* towards jumping springtails. As this was revealed earlier [21], this stereotype includes determining the vic-

Table 1 The compression ratios of the resulting files containing reactions of resident gulls toward an approaching and escaping intruder

Reactions of resident gulls towards	Approaching intruder	Escaping intruder
size of the file before compression (in bytes)	177.2 ± 2.9	182.8 ± 9.5
size of the file after compression (in bytes)	73.8 ± 2.1	85.9 ± 3.31
compression ratio, %	41.83 ± 1.34	47.71 ± 2.78

tim, approaching it, and then performing the so-called fixed action pattern that we call "tip-and-run attack:" the ant attacks the prey, bends the abdomen and head to the thorax, jumps towards the springtail, falls on it abruptly, and stings. Then the ant takes the victim and transports it to the nest. We compared two groups of highly genetically variable ants: members of a natural colony ("wild" for brevity) and naive (laboratory–reared from pupae) ants of age from 3 to 12 days. Ants were housed in transparent laboratory nests on arenas. The wild colony included about 3000 completely matured workers, and three naive colonies included 300 workers each. Examined ants were placed one by one into glass containers with 30 live springtails, and each individual was tested once. To analyse ethograms from video records, we used the Observer *XT* 7.0 (version: 7.0.214, Noldus Information Technology). In total, we analysed 6.5 hours of video recordings by the second, for 26 ants. In order to select behavioural units, we used the following protocol. For the abdomen, legs, head, antennae and mandibles, in combination with current movements, typical states were designated by symbols. With the use of these symbols, we described behavioural units that included "blocks" of locomotions and postures. We represented behavioural sequences as "texts" in which behavioural units (10 in total), singled out from video records and denoted by symbols (letters), served as an alphabet: W (waiting), S (slow walking), R (running), T (turning), U (turning around), B (belligerent posture), A (attack), C (capturing the prey), K (kicking the victim with the sting), T (transporting the prey).

Using the "alphabet" of these 10 units, we expressed the hunting stereotypes as text files. Every sequence (file) was constructed manually (by the researchers) from the corresponding video fragment. As the starting point of a hunting stereotype we took the ant's approach to the victim and the display of purposive movements; transportation of the killed victim was considered the end of the complete stereotype. All cases of loss of a victim and switching to another one were considered ends of incomplete stereotypes. Pooling individual ethograms of members of four groups, we obtained 4 files which included: 19 complete and 20 incomplete hunting stereotypes in "wild" ants and, correspondingly, 20 and 31 stereotypes in "naive" ants. We reduced these files to equal initial length of 147 units, compressed them with the use of the same archiver as mentioned in Sect. 3.2 and compared the ratios of compression in different stereotypes. The length of the complete stereotypes varied from 6 to 22 units (13.42 ± 1.08 on average) in the wild ants and from 5 to 18 in naive ants (8.75 ± 0.71 on average). The length of the incomplete stereotypes varied from 4 to 14 units (6.55 ± 0.51 on average) in the wild ants and from 3 to 17 in naive ants (6.03 ± 0.57 on average).

We tested the Hypothesis H_0 (the sequences from two sets are generated by one source) against H_1 (the complexity of sequences from one set is, on average, larger

Table 2 The compression ratios of the resulting files containing complete and incomplete stereotypes

Parameters	Wild ants Stereotypes		Naive ants Stereotypes	
	Complete	Incomplete	Complete	Incomplete
size of the file before compression (in bytes)	147	147	147	147
compression ratio, %	63.27	70.07	56.46	68.03

than the complexity of sequences from the other) by the Mann–Whitney–Wilcoxon test, as described in Sect. 2.1. It turned out that files corresponding to the successful hunting stereotypes compress better than those corresponding to incomplete hunting stereotypes both in wild and in naive ants (Table 2). Moreover, H_0 is rejected (with $\alpha = 0.05$), and we can conclude that, on average, the complexity of sequences from the first set is larger than in the second. In sum, these data support our initial suggestion that complete successful hunting stereotypes in ants are less complex.

In general, the use of the suggested method for studying animal behavioural patterns is a promising tool to be used in different areas of behavioural and evolutionary research. In particular, this method can help to extract "basic" (completely innate) behavioural patterns by comparing behavioural sequences of different levels of complexity and flexibility. It becomes possible for ethologists to extract innate behavioral patterns by comparing behavioural sequences of different levels of complexity without resorting to rearing naive animals. Analyzing complexity of behavioural patterns in naive and experienced animals, we gain an additional possibility to link the experience with structure and function. This is particularly important for evolutionary studies including behavioural mechanisms of speciation.

Appendix 1: Universal Codes, Shannon Entropy and Kolmogorov Complexity

First we briefly describe stochastic processes (or sources of information). Consider a finite alphabet A, and denote by A^t and A^* the set of all words of length t over A and the set of all finite words over A correspondingly ($A^* = \bigcup_{i=1}^{\infty} A^i$).

A process P is called stationary if

$$P(x_1, \ldots, x_k = a_1, \ldots, a_k) = P(x_{t+1}, \ldots, x_{t+k} = a_1, \ldots, a_k)$$

for all $t, k \in N$ and all $(a_1, \ldots, a_k) \in A^k$. A stationary process is called stationary ergodic if the frequency of occurrence of every word $a_1, \ldots, a_k$ converges (a.s.) to $P(a_1, \ldots, a_k)$. For more details see [2, 4, 7].

Let τ be a stationary ergodic source generating letters from a finite alphabet A. The m-order (conditional) Shannon entropy and the limit Shannon entropy are defined as follows:

$$h_m(\tau) = -\sum_{v \in A^m} \tau(v) \sum_{a \in A} \tau(a|v) \log \tau(a|v), \qquad h_\infty(\tau) = \lim_{m \to \infty} h_m(\tau), \quad (6)$$

[2, 7]. The well known Shannon-MacMillan-Breiman theorem states that

$$\lim_{t\to\infty} -\log \tau(x_1 \dots x_t)/t = h_\infty(\tau) \tag{7}$$

with probability 1, see [2, 4, 7].

Now we define codes and Kolmogorov complexity. Let A^∞ be the set of all infinite words $x_1x_2\dots$ over the alphabet A. A data compression method (or code) φ is defined as a set of mappings φ_n such that $\varphi_n : A^n \to \{0,1\}^*, n = 1, 2, \dots$ and for each pair of different words $x, y \in A^n$ $\varphi_n(x) \neq \varphi_n(y)$. Informally, it means that the code φ can be applied for compression of each message of any length n over the alphabet A and the message can be decoded if its code is known. It is also required that each sequence $\varphi_n(u_1)\varphi_n(u_2)\dots\varphi_n(u_r), r \geq 1$, of encoded words from the set $A^n, n \geq 1$, can be uniquely decoded into $u_1u_2\dots u_r$. Such codes are called uniquely decodable. For example, let $A = \{a, b\}$, the code $\psi_1(a) = 0$, $\psi_1(b) = 00$, obviously, is not uniquely decodable. It is well known that if a code φ is uniquely decodable then the lengths of the codewords satisfy the following inequality (Kraft inequality): $\Sigma_{u\in A^n} 2^{-|\varphi_n(u)|} \leq 1$, see, for ex., [4, 7].

In this paper we will use the so-called prefix Kolmogorov complexity, whose precise definition can be found in [9, 13]. Its main properties can be described as follows. There exists a uniquely decodable code κ such that (i) there is an algorithm for decoding (i.e. there is a Turing machine, which maps $\kappa(u)$ to u for every $u \in A^*$) and (ii) for any uniquely decodable code ψ, whose decoding is algorithmically realizable, there exists a non-negative constant C_ψ that

$$\big|\kappa(u)\big| - \big|\psi(u)\big| < C_\psi \tag{8}$$

for every $u \in A^*$; see Theorem 3.1.1 in [13]. The prefix Kolmogorov complexity $K(u)$ is defined as the length of $\kappa(u)$: $K(u) = |\kappa(u)|$. The code κ is not unique, but the second property means that codelengths of two codes κ_1 and κ_2, for which (i) and (ii) are true, are equal up to a constant: $||\kappa_1(u)| - |\kappa_2(u)|| < C_{1,2}$ for any word u (and the constant $C_{1,2}$ does not depend on u, see (8)). So, $K(u)$ is defined up to a constant. In what follows we call this value "Kolmogorov complexity".

We can see from (ii) that the code κ is asymptotically (up to a constant) the best method of data compression, but it turns out that there is no algorithm that can calculate the codeword $\kappa(u)$ (and even $K(u)$). That is why the code κ (and Kolmogorov complexity) cannot be used for practical data compression directly.

The following Claim is by Levin [32, Proposition 5.1]:

Claim *For any stationary ergodic source* τ

$$\lim_{t\to\infty} t^{-1} K(x_1 \dots x_t) = h_\infty(\tau) \tag{9}$$

with probability 1.

Comment In [32] this claim is formulated for "common" Kolmogorov complexity, but it is also valid for the prefix Kolmogorov complexity, because for any word $x_1 \dots x_t$ the difference between both complexities equals $O(\log t)$, see [13].

Let us describe universal codes, or data compressors. For their description we recall that (as it is known in Information Theory) sequences $x_1 \dots x_t$, generated by a source p, can be "compressed" till the length $-\log p(x_1 \dots x_t)$ bits and, on the other hand, there is no code ψ for which the expected codeword length $(\Sigma_{x_1 \dots x_t \in A^t} p(x_1 \dots x_t)|\psi(x_1 \dots x_t)|)$ is less than $-\Sigma_{x_1 \dots x_t \in A^t} p(x_1 \dots x_t) \log p(x_1 \dots x_t)$. The universal codes can reach the lower bound $-\log p(x_1 \dots x_t)$ asymptotically for any stationary ergodic source p with probability 1. The formal definition is as follows: A code φ is universal if for any stationary ergodic source p

$$\lim_{t \to \infty} t^{-1}\big(-\log p(x_1 \dots x_t) - \big|\varphi(x_1 \dots x_t)\big|\big) = 0 \tag{10}$$

with probability 1. So, informally speaking, universal codes estimate the probability characteristics of the source p and use them for efficient "compression."

Appendix 2: Proofs of Theorems

Proof of Theorem 1 For any universal code from the Shannon-McMillan-Breiman theorem (7) and the definition (10) we obtain the following equation

$$\lim_{t \to \infty} t^{-1}\big|\varphi(x_1 \dots x_t)\big| = h_\infty(x_1 \dots x_t), \tag{11}$$

with probability 1. Having taken into account this equality and (9), we obtain the statement of the Theorem 1. □

Proof of Theorem 2 For any t the level of significance of the test T equals α, so, by definition, the Type I error of the test T'_φ equals α, too. In order to prove the second statement of the theorem we suppose that the hypothesis H_1 is true. From (8) we can see that there exist constants k_1 and k_2 such that with probability 1

$$\lim_{t \to \infty} t^{-1} K(x_1 \dots x_t) = k_i, \tag{12}$$

where $x_1 \dots x_t \in S_i$, $i = 1, 2$, and $k_1 \neq k_2$. Let us suppose that $k_1 > k_2$ and define $\Delta = k_1 - k_2$. From (1), (12) and Theorem 2 we can see that

$$\lim_{t \to \infty} K_\varphi(x_1 \dots x_t) = k_i \tag{13}$$

with probability 1 (here $x_1 \dots x_t \in S_i$, $i = 1, 2$). By definition, it means that for any $\epsilon > 0$ and $\delta > 0$ there exists such t' that

$$P\big\{\big|K_\varphi(x_1 \dots x_t) - k_i\big| < \epsilon\big\} \geq 1 - \delta$$

for $x_1 \dots x_t \in S_i$, $i = 1, 2$, when $t > t'$. Hence, if $\epsilon = \Delta/4$, then with probability at least $1 - \delta$ all values $K_\varphi(x_1 \dots x_t)$, $x_1 \dots x_t \in S_1$ are less than all values $K_\varphi(x_1 \dots x_t)$, $x_1 \dots x_t \in S_2$. So, with probability at least $1 - \delta$ a set of the ranked values will look like the right part of (4) and, hence, the hypothesis H_1 will be rejected (for large enough $|S_i|$, $i = 1, 2$). Having taken into account that $\min(|S_1|, |S_2|) \to \infty$ and $t \to \infty$, we can see that the last statement is valid for any δ. The theorem is proved. □

Appendix 3: Dictionary of Gull's Behaviours

Symbol	Gull's position	Demonstrative postures and actions	Position of wings	Vocalisation
0	sitting on eggs	upright	folding	aggressive call
1	sitting on eggs	upright	folding	no
2	sitting on eggs	oblique	folding	aggressive call
3	sitting on eggs	oblique	folding	long call
4	sitting on eggs	oblique	folding	no
5	sitting on eggs	biting with a beak	folding	no
6	sitting on eggs	snapping with a beak	folding	no
7	sitting on eggs	no	folding	aggressive call
7	sitting on eggs	no	folding	long call
8	sitting on eggs	no	folding	no
9	standing on a nest	upright	folding	aggressive call
a	standing on a nest	upright	folding	no
b	standing on a nest	oblique	stretching	aggressive call
c	standing on a nest	oblique	stretching	long call
d	standing on a nest	oblique	stretching	no
e	standing on a nest	oblique	folding	aggressive call
f	standing on a nest	oblique	folding	long call
g	standing on a nest	oblique	folding	no
h	standing on a nest	oblique	flapping	aggressive call
i	standing on a nest	oblique	flapping	long call
j	standing on a nest	oblique	flapping	no
k	standing on a nest	biting with a beak	stretching	no
l	standing on a nest	biting with a beak	folding	no
m	standing on a nest	biting with a beak	flapping	no
n	standing on a nest	snapping with a beak	stretching	no
o	standing on a nest	snapping with a beak	folding	no
p	standing on a nest	snapping with a beak	flapping	no
q	standing on a nest	slashing with a wing	flapping	aggressive call
r	standing on a nest	slashing with a wing	flapping	long call
s	standing on a nest	slashing with a wing	flapping	no
t	standing on a nest	no	stretching	aggressive call
u	standing on a nest	no	stretching	no
v	standing on a nest	no	folding	aggressive call
w	standing on a nest	no	folding	no
x	standing on a nest	no	flapping	aggressive call
y	standing on a nest	no	flapping	no
z	flying	no	flapping	aggressive call
A	flying	no	flapping	no
B	flying	swooping to attack	flapping	long call
C	flying	swooping to attack	flapping	aggressive call
D	flying	swooping to attack	flapping	no
E	flying	biting with a beak	flapping	no
F	flying	snapping with a beak	flapping	no
G	sitting on a perch	oblique	stretching	aggressive call
H	sitting on a perch	oblique	stretching	long call
I	sitting on a perch	oblique	stretching	no
J	sitting on a perch	oblique	folding	aggressive call
K	sitting on a perch	oblique	folding	long call

References

1. Anel, C., Sanderson, M.J.: Missing the forest for the trees: phylogenetic compression and its implications for inferring complex evolutionary histories. Syst. Biol. **54**(1), 146–157 (2005)
2. Billingsley, P.: Ergodic Theory and Information. Wiley, New York (1965)
3. Cilibrasi, R., Vitanyi, P.: Clustering by compression. IEEE Trans. Inf. Theory **51**(4), 1523–1545 (2005)
4. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley-Interscience, New York (2006)
5. Ferragina, P., Giancarlo, R., Greco, V., Manzini, G., Valiente, G.: Compression-based classification of biological sequences and structures via the Universal Similarity Metric: experimental assessment. BMC Bioinf. **8** (2007)
6. Fisher, R.A.: Statistical Methods, Experimental Design, and Scientific Inference. Oliver & Boyd, Edinburgh (1956)
7. Gallager, R.G.: Information Theory and Reliable Communication. Wiley, New York (1968)
8. Groothuis, T.: The influence of social experience on the development and fixation of the form of displays in the black-headed gull. Anim. Behav. **43**(1), 1–14 (1992)
9. Hutter, M.: Universal Artificial Intelligence. Sequential Decisions Based on Algorithmic Probability. Springer, Berlin (2005)
10. Kendall, M.G., Stuart, A.: The Advanced Theory of Statistics, Inference and Relationship, vol. 2. Griffin, London (1961)
11. KGB archiver (v. 1.2). http://www.softpedia.com/get/Compression-tools/KGB-Archiver.shtml
12. Kieffer, J., Yang, E.: Grammar-based codes: a new class of universal lossless source codes. IEEE Trans. Inf. Theory **46**, 737–754 (2000)
13. Li, M., Vitanyi, P.: An Introduction to Kolmogorov Complexity and Its Applications, 2nd edn. Springer, New York (1997)
14. Li, M., Badger, J., Chen, X., Kwong, S., Kearney, P., Zhang, H.Y.: An information based distance and its application to whole mitochondrial genome phylogeny. Bioinformatics (Oxford) **17**, 149–154 (2001)
15. Li, M., Chen, X., Li, X., Ma, B., Vitanyi, P.: The similarity metric. IEEE Trans. Inf. Theory **50**(12), 3250–3264 (2004)
16. Lloyd, E. (ed.): Handbook of Applied Statistics, vol. 2. Wiley-Interscience, New York (1984)
17. McCowan, B., Doyle, L.R., Hanser, S.F.: Using information theory to assess the diversity, complexity, and development of communicative repertoires. J. Comp. Psychol. **116**(2), 166–172 (2002)
18. Oller, D.K., Griebel, U. (eds.): Evolution of Communicative Flexibility: Complexity, Creativity, and Adaptability in Human and Animal Communication. MIT Press, Cambridge (2008)
19. Panteleeva, S., Danzanov, Zh., Reznikova, Zh.: Estimate of complexity of behavioral patterns in ants: analysis of hunting behavior in myrmica rubra (hymenoptera, formicidae) as an example. Entomol. Rev. **91**(2), 221–230 (2011)
20. Reznikova, Z.: Animal Intelligence: From Individual to Social Cognition. Cambridge University Press, Cambridge (2007)
21. Reznikova, Zh., Panteleeva, S.: An ant's eye view of culture: propagation of new traditions through triggering dormant behavioural patterns. Acta Ethol. **11**, 73–80 (2008)
22. Rissanen, J.: Universal coding, information, prediction, and estimation. IEEE Trans. Inf. Theory **30**(4), 629–636 (1984)
23. Ryabko, B.: Prediction of random sequences and universal coding. Probl. Inf. Transm. **24**(2), 87–96 (1988)
24. Ryabko, B., Reznikova, Zh.: Using Shannon entropy and Kolmogorov complexity to study the communicative system and cognitive capacities in ants. Complexity **2**, 37–42 (1996)
25. Ryabko, B., Reznikova, Z.: The use of ideas of information theory for studying "language" and intelligence in ants. Entropy **11**, 836–853 (2009)
26. Ryabko, D., Schmidhuber, J.: Using data compressors to construct order tests for homogeneity and component independence. Appl. Math. Lett. **22**(7), 1029–1032 (2009)
27. Ryabko, B., Astola, J., Gammerman, A.: Application of Kolmogorov complexity and universal codes to identity testing and nonparametric testing of serial independence for time series. Theor. Comput. Sci. **359**, 440–448 (2006)
28. Tinbergen, N.: An objective study of the innate behaviour of animals. Bibl. Biotheor. **1**, 39–98 (1942)

29. Tinbergen, N.: The Study of Instinct. Oxford University Press, London (1951)
30. Vitanyi, P.M.B.: Information distance in multiples. IEEE Trans. Inf. Theory **57**(4), 2451–2456 (2011)
31. Yaglom, A.M., Yaglom, I.M.: Probability and Information. Theory and Decision Library. Springer, Berlin (1983)
32. Zvonkin, A.K., Levin, L.A.: The complexity of finite objects and concepts of information and randomness through the algorithm theory. Russ. Math. Surv. **25**(6), 83–124 (1970)