INTEGRATIONS OF IDEAS AND METHODS OF KOLMOGOROV COMPLEXITY AND CLASSICAL MATHEMATICAL STATISTICS WITH APPLICATION TO ANIMAL BEHAVIOUR

Boris Ryabko
Siberian State University of Telecommunication and Informatics; www.boris.ryabko.net

Zhanna Reznikova
www.reznikova.net
Institute of Systematics and Ecology of Animals, Novosibirsk, Russian Academy
Abstract

- A new approach is suggested which allows to combine the advantages of methods based on Kolmogorov complexity with classical methods of testing statistical hypotheses.

- As distinct from other approaches to analysis of sequences by means of Kolmogorov complexity, we stay within the framework of mathematical statistics.

- As examples, we consider behavioural sequences of animals (ethological “texts”): territorial behaviour of gulls and hunting behaviour in ants.
Introduction

The idea of quantitative evaluation of complexity of sequences of symbols from a finite alphabet (or “texts”) is important for many scientific fields, including molecular biology, linguistics, zoosemiotics, and ethology.

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.

This simple observation has been a basis of many applications of Kolmogorov complexity to comparative analysis of many literary, musical and biological “texts” (see, for example, Cilibrasi and Vitanyi: 2005).
Sir Ronald Aylmer Fisher (1890 – 1962)
A famous English statistician, evolutionary biologists, eugenecist, and geneticist;
Alma mater: Cambridge University

Introduction: The problem

However, this approach does not give a possibility to use methods of mathematical statistics and, in particular, hypothesis testing.
This is a severe limitation for the applicability of ideas of Kolmogorov complexity to natural sciences, in particular, to biological studies.
Indeed, since Fisher's classical works, mathematical statistics and, in particular, hypothesis testing became the main method of quantitative analysis of biological data.
Biologists need statistical tests to argue about complexity of sequences.

The standard “Data compressors estimate Kolmogorov complexity” argument does not lead to statistical tests.
Our approach

We suggest an approach which allows us to combine the advantages of methods based on Kolmogorov complexity with classical methods of testing statistical hypotheses.

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 describe the behaviour of an animal in different situations, or behaviours of different animals. Alternatively, the sequences can be DNAs of two different species. The problem is to compare complexities of such sequences. Similar problems arise in many fields of biology and other sciences.
The scheme of the suggested approach

First we describe the scheme of the suggested approach. Let there be a sequence \( x = x_1 \ldots x_t, \ t > 0, \) of letters from a finite alphabet \( A \) and let \( \varphi \) 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:

\[
K_{\varphi}(x_1 \ldots x_t) = \frac{|\varphi(x_1 \ldots x_t)|}{t}.
\]

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 (even though it is impossible to calculate Kolmogorov complexity).
The conditions

- The considered sequences are generated by stationary ergodic sources.
- The data compressor \( \varphi \) is a universal code.

Informally, a universal code can “compress” a sequence \( x_1 \ldots x_t \) up to its Shannon entropy (per letter) if the sequence is generated by a stationary ergodic source.

Nowadays there are many efficient universal codes, which are described in numerous papers. It is important for practical applications that modern data compressors are based on universal codes.
The main theorem

The following theorem gives a theoretical basis for the suggested approach:

**Theorem**

Let there be a stationary ergodic source generated letters from a finite alphabet and a universal code $\varphi$. Then with probability 1 we have

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

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

The main idea of the suggested approach is very simple: Apply a universal code $\varphi$ to estimate the Kolmogorov complexity of a word $x_1...x_t$. Then, use (consistent) statistical tests to study $\varphi(x_1...x_t)/t$ ($= K_\varphi(x_1...x_t)$) in the same manner as one would study any other natural parameter, such as the weight, length, speed, etc.
Given two sets of sequences \( V \) and \( B \), our goal is to find a statistical test that can distinguish between the following two hypotheses: \( H_0 = \{ \text{the sequences from both sets are generated by one source} \} \) and \( H_1 = \{ \text{the sequences from the different sets are generated by stationary and ergodic sources with different Kolmogorov complexities (per letter of generated sequences)} \} \).

In order to construct the test for \( H_0 \) against \( H_1 \) we consider an auxiliary hypothesis \( H_1^* = \{ \text{the estimation of the average complexity } K_\varphi() \text{ is not equally large for sequences from different sets } V, B \} \). The suggested test \( T'_\varphi \) for \( H_0 \) against \( H_1 \) (not \( H_1^* \) !) uses a universal code \( \varphi \) and a consistent test \( T \) for assessing whether two independent samples of observations have equally large values (say, the Mann–Whitney–Wilcoxon \( U \) test).
The test $T'_\varphi$ is as follows: First, calculate $K_\varphi()$ for all sequences from $V$, $B$ and then apply $T$ for testing $H_0$ against $H_1^*$ based on the new sets $\{K_\varphi(x), x \in V\}$ and $\{K_\varphi(x), x \in B\}$. The following Theorem describes properties of $T'_\varphi$:

**Theorem**

*The Type I error of the test $T'_\varphi$ is not greater than $\alpha$ and, if $\min(|B|, |V|) \to \infty$ and $\min_{v \in V, b \in B}(|v|, |b|) \to \infty$, the probability to accept $H_0$ instead of $H_1 = \{\text{the sequences from the different sets are generated by stationary and ergodic sources with different Kolmogorov complexities (per letter)}\}$ goes to 0.*
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.
An ant approaches a victim with open mandibles.

The ant attacks the prey, bend the abdomen and head to the thorax, jumps to the springtail, falls on it abruptly, and sting.
Behaviours of two groups of ants were compared:

1) Members of a natural colony («wild»)

2) Naive (laboratory reared) ants of age from 3 to 12 days
Observations

30 alive springtails

D = 5 cm; h = 6 cm

We carried out 104 sessions with “wild” (n=33) and naïve, laboratory reared, (n=61) ants. In total, 26 hours.
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).
Analysis of video records

Video → The Observer XT 7.0 → Sequence of elements

«alphabet»

... RRTURRTUURRWRWWSTRR ATWRATRRTTRTURWWWWR ARACCKCKPRTSRRTRRRAT RURRRACKPRWRR...

Stereotypes:
- **Complete (successful)**
  - N=110 in «wild» ants
  - N=151 in naïve ants
- **Incomplete**
  - N=119 in «wild» ants
  - N=165 in naïve ants

4 Summarized text files (.txt)

Complete stereotypes
- in «wild» ants

Incomplete stereotypes
- in «wild» ants

Complete stereotypes
- in naïve ants

Incomplete stereotypes
- in naïve ants
Analysis of video records

4 Summarized text files (.txt)

Complete stereotypes in «wild» ants
Complete stereotypes in naive ants
Incomplete stereotypes in «wild» ants
Incomplete stereotypes in naive ants

KGB Archiver v.1.2.

The compression ratio = \frac{\text{file size after compression}}{\text{file size before compression}}
We tested the Hypothesis H0 (the sequences from two sets are generated by one source) against H1 (the complexity of sequences from one set is, on average, larger than the complexity of sequences from the other) by the Mann–Whitney–Wilcoxon test.

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. Moreover, H0 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 suggestion that complete successful hunting stereotypes in ants are less complex.
The “nature/nurture” problem

To distinguish between innate and learned behaviour
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.
Territorial behaviour in gulls. “Resident-intruder” experiments
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.)

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 served as an alphabet. Pooling individual ethograms, we obtained 72 behavioural sequences composed of combinations of 60 units.

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.

These data suggest that reactions of a resident gull towards an approaching trespasser are more variable and “chaotic” than its reactions towards the escaping one.
Applied to different animals, the proposed method is a promising tool for distinguishing between “basic” stereotypical behavioural patterns and flexible behaviour.

More generally, the method provides a statistically sound tool for reasoning about Kolmogorov complexity of biological texts. This is potentially of great importance in many branches of biology.