Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Objects and processes: Two notions for understanding biological information

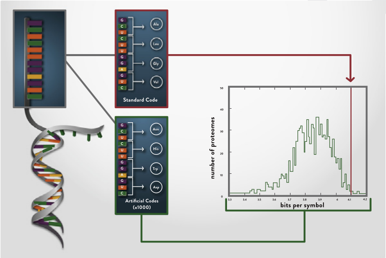
Agustín Mercado-Reyes [a], Pablo Longoria Padilla [b], Alfonso Arroyo-Santos [c,d,*]

[a] Posgrado en Filosofía de la Ciencia, Instituto de Investigaciones Filosóficas UNAM, Ciudad Universitaria, Circuito Mario de la Cueva s/n, 04510 Mexico City, Mexico
[b] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM, Ciudad Universitaria, 04510 Mexico City, Mexico
[c] Centro de Información Geoprospectiva, Berlín 209B, Del Carmen Coyoacán, 04100 Mexico City, Mexico
[d] Facultad de Filosofía y Letras, Ciudad Universitaria, UNAM Circuito Interior s/n, 04510 Mexico City, Mexico

HIGHLIGHTS

- Information in biology has been linked almost exclusively to Shannon's theory.
- We show experimentally the limitations of such quantitative analyses.
- Our results suggest the need to complement formal analyses with semantic approaches.
- We propose two separate theoretical frameworks called object- and process-information.
- Processual terms help describe biological semiosis and meaning.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 20 November 2014
Received in revised form
7 May 2015
Accepted 13 May 2015

Keywords:
Information theory
Complexity
Semantics
Object-information
Process-information

ABSTRACT

In spite of being ubiquitous in life sciences, the concept of information is harshly criticized. Uses of the concept other than those derived from Shannon's theory are denounced as metaphoric. We perform a computational experiment to explore whether Shannon's information is adequate to describe the uses of said concept in commonplace scientific practice. Our results show that semantic sequences do not have unique complexity values different from the value of meaningless sequences. This result suggests that quantitative theoretical frameworks do not account fully for the complex phenomenon that the term "information" refers to. We propose a restructuring of the concept into two related, but independent notions, and conclude that a complete theory of biological information must account completely not only for both notions, but also for the relationship between them.

1. Introduction

The concept of information has a central role in contemporary biology. For example, information is at the core of molecular biology, one of the most important theoretic structures to emerge in the 20th century life sciences, and the one that currently informs our way of understanding the process of life. Despite its central role in contemporary biology, the notion of information remains controversial. Some scientists and philosophers believe that the only legitimate use of the notion of information in biology is that coming from quantitative approaches such as Shannon's information theory (Shannon, 1948; Weaver and Shannon, 1963) or Kolmogorov–Chaitin's complexity (Kolmogorov, 1965; Chaitin, 1969). In the view of these authors, all other uses of information are metaphoric, terms without a proper referent, and even

* Corresponding author at: Centro de Información Geoprospectiva, Berlín 209B, Del Carmen Coyoacán, 04100 Mexico City, Mexico.
E-mail address: agripas@gmail.com (A. Arroyo-Santos).

http://dx.doi.org/10.1016/j.jtbi.2015.05.021
0022-5193/© 2015 Elsevier Ltd. All rights reserved.

detrimental to the proper understanding of biological systems (i.e., Sarkar, 2001; Griffiths, 2001; Godfrey-Smith and Sterelny, 2008; Moss, 2003).

In the present paper, we argue that informational terms are far from metaphoric but the conceptual structure that underlies them does need clarification. In general, we believe that minimally, a theory of biological information should explain how certain data are used to transmit a message. In our opinion, most popular accounts on information have paid a lot of attention on data (i.e. on their attributes, on how they are encoded and transmitted), and little on how such data becomes meaningful information.

To defend our point, we designed an experiment to determine whether quantitative approaches can account for the broad, albeit fuzzy understanding of the concept of information. In our experiment, we measure information as understood in Shannon's information theory, where “measuring information” amounts to calculating the *complexity* of a given structure, meaning the minimum amount of information that would be required to reconstruct completely the original structure, in this case, a given DNA sequence. Our results show that functional biological sequences have high complexity but, more importantly, it shows that there are *alternative, meaningless sequences with similar complexity measures*. This means that no particular value of algorithmic complexity is inherently bound to meaningful content and in consequence, *quantitative accounts on information can explain a part, but not everything we want to convey when talking about biological information in terms of coding, transmission and content.* Our results give support to those authors who believe that such quantitative approaches should be complemented with semantic theories.

From the results of our experiment, we argue that there are at least two notions of biological information: the first involves a notion where information is generally understood as a set of attributes pertaining to an object, typically the genetic sequence, which can be analyzed by means of information theory. The second notion deals with the ways in which certain attributes acquire meaning. We have called these kinds *object-information* and *process-information*, respectively. We suggest that the controversy surrounding the notion of information is in part the result of conflating two related but independent notions of information. We believe that our distinction provides a basis for the construction of a theory of biological information that can be used to better understand the problems and possible solutions to current controversies of information.

We proceed as follows: in Section 2, we present the computational experiment; in Section 3, we discuss our results, placing them in context of other authors and proposing a separation of the concept of information into two notions, pointing out possible ways to articulate them; and we offer brief concluding remarks and possible directions for further inquiry in Section 4.

2. A computational experiment

2.1. Aims of the experiment

Our experiment aims to answer the following question: what is the relationship between the values obtained when measuring genetic sequences using quantitative approaches, and what we usually want to convey in biological discourse when talking of information? To keep the discussion as simple as possible, in this experiment information is limited to the processes of transcription and translation, that is, to the whole process that goes from “reading” the genetic sequence to synthesizing a given protein. Even though information permeates an enormous diversity of biological processes at different levels of description, the so-called *genetic information* serves our purpose well for a host of reasons: it stands at the center of the information controversy, data is readily available and the mechanisms of gene expression have been thoroughly researched. Furthermore, any biological information theory should explain how a code is transmitted and transformed into meaningful data (or at least, how to tell what's meaningful from what is not).

The basic premise of our experiment is: if information was a univocal notion, quantifiable and dependent on the structure of the sequence, it could be represented wholly in internal structural measures, such as Shannon's entropy or complexity. Under this scenario, structural measures would function as a kind of diagnosis to predict semantic content and nothing else would be needed. However, if semantic content and structural measures were different in any ways – that is, if the complexity features of a sequence were independent of semantics – it would mean that there are aspects of the notion of information that are not touched upon by sequence-structure analysis. It would not mean that information-theoretic approaches are incorrect, but that they are incomplete.

2.2. Methods

In our experiment we use the total translatable DNA sequences of four organisms. The organisms chosen were *Nanoarchaeum equitans* (Waters, 2003), *Mycoplasma genitalium* (Fraser et al., 1995), *Schizosaccharomyces pombe* (Wood et al., 2002), and the Mimivirus from *Entamoeba* (Raoult, 2004). The first three model organisms were chosen as representative of the three separate domains of life (Archaea, Eubacteria and Eukarya, respectively), to encompass phylogenetically distant organisms. The inclusion of Mimivirus, a complex and large virus that infects amoebas, presented a decision point for us. Viruses have long been problematic in terms of classification and under some definitions of life may even be considered to be non-living, but we decided to include them to further increase the diversity of the analysis.

We used the complementary DNA (cDNA) of all four organisms selected and obtained their proteome. We then measured the information content of all four proteomes (see Fig. 1). As a method of measuring the information of each proteome we turned to string compression, a common method used to estimate the value of algorithmic complexity. Briefly, the general idea is to calculate the minimum algorithm that would be necessary to reconstruct a given sequence. If the sequence is random, then the amount of information necessary to reconstruct the sequence is the same as the sequence itself as there would be no way of telling what symbol comes next. This is called maximum complexity, or maximum value. However, if the sequence is not random, then it is possible to obtain an algorithm that has less information than the original sequence (and hence is “compressed” in relation to the original source), because there would be a way of calculating, probabilistically, what symbol comes next in the sequence (for a review see Li and Vitányi, 2008).

In this paper we used the algorithm described in Cao et al. (2007), as it was especially developed to deal with biological sequences, both nucleic and peptidic. The measurements yielded, expressed in bits per symbol (bps), indicate more complexity as they approach the maximum value. The maximum value is calculated by the formula $V_{\max}=\log 2A$, where A is the number of symbols in the alphabet. Thus, for nucleic acids, which can be constituted by 4 different bases, $V_{\max}=\log 2(4)=2$, and for amino acid chains, formed by 20 different possible amino acids, $V_{\max}=\log 2(20)=4.322$.

Once the calculations were performed, we asked ourselves whether the values obtained were enough to account for our minimal understanding of information, that is, if the values