Objects and processes: two notions for understanding biological information

Agustín Mercado-Reyes [1]

Pablo Longoria Padilla [2]

Alfonso Arroyo-Santos [3,4,5]

1. Posgrado en Filosofía de la Ciencia, Instituto de Investigaciones Filosóficas, UNAM
2. Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM
3. Centro de Información Geoprospectiva.
4. Facultad de Filosofía y Letras, UNAM
5. Corresponding author: agripas@gmail.com

DRAFT. FINAL VERSION TO APPEAR IN THE JOURNAL OF THEORETICAL BIOLOGY.

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 pernicious metaphors. 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 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 II, we present the computational experiment; in section III, 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 IV.

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 were 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 way – 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 et al, 2003), *Mycoplasma genitalium* (Fraser et al, 1995), *Schizosaccharomyces pombe* (Wood et al., 2002), and the Mimivirus from *Entamoeaba* (Raoult et al, 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 figure 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 Vmax = log2A, where A is the number of symbols in the alphabet. Thus, for nucleic acids, which can be constituted by 4 different bases, Vmax = log2(4) = 2, and for amino acid chains, formed by 20 different possible amino acids, Vmax = log2(20) = 4.322.

Figure 1. The experiment. We obtained the proteome (the entire set of proteins expressed by the genome) of four different organisms using their cDNA (1). Then, we produced a set of artificial proteomes for each organism by assigning to each naturally occurring codon a random amino acid (figure 2a illustrates the natural code and figure 2b the artificial assignment). Finally, we measured the information content of all proteomes, both natural (vertical line) and artificial (histogram), using string compression (3).

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 obtained were enough to account for meaningful data in the context of the organisms under study. To answer this question, we produced a set of *1000 artificial genetic codes* by randomly assigning a specific amino acid to each DNA codon from each of our four organisms. After that, we used each of the artificial genetic codes to produce *artificial proteomes* (Figure 1 central panel). Note that in this experimental set-up, artificial genetic codes are meaningless in the context of the organisms under study because in nature codons code univocally for specific amino acids. For this reason, we will say that natural genomes carry *meaningful information* whereas artificial genomes do not and the question will be to know if complexity measures can tell apart the difference.

2.3 Results

Figure 2 shows the results of the experiment for each of the four organism (figures 2A to 2D). Bars represent the distribution of the values of algorithmic complexity for 1000 alternative proteomes produced with artificial genetic codes, and the vertical line shows the value for the proteomes produced by the standard genetic code.

Figure 2. Comparison between the compression ratio of the standard proteomes (vertical line) and the distribution of the alternative proteomes derived from the genomes of a) *Entamoeba histolytica* mimivirus; b) *Nanoarchaeum equitans*; c) *Mycoplasma genitalium*; and d) *Saccharomyces pombe*. Histograms illustrate compression values for 1000 artificial proteomes and the single vertical line shows compression values obtained for the naturally-occurring proteomes for each organism under study. Note that even though the standard proteomes have an unusually high informational complexity, they are not unique, as we obtained alternative codes with similar complexity measures. See text for further details.

The measurements for the standard proteomes are very high in all four graphs, approaching the maximum complexity value, which means that, structure-wise, biological sequences with semantic content are not easily compressed and approach complete randomness. For example, the standard genome compression value of *E. hystolytica* mimivirus deviates 1.1430 standard deviations (σ) from the mean value of the artificial proteomes distribution, which means that it is higher than 87.35% of the proteomes analyzed (see Table 1). Values for artificial genomes were distributed along a wide array of values. However, the important result is that we obtained a subset of artificial proteomes with similar complexity values to those obtained for standard codes (Figure 2).

Table 1. Statistical comparison between the standard genetic code and the mean of the artificially generated proteomes. Columns 1 and 2 compare the mean compression values of the alternative proteomes we constructed and the compression value of the actual, standard proteome respectively. Columns 3-5 summarize a statistical analysis of the difference of the values in 1 and 2.

The codes that yielded the compression values most similar to that of the standard code were completely different, meaning that few (if any) of the translation rules were coincident with the standard ones (figure 3). That is, the codons were assigned to different amino acids in almost all of the alternative codes. In actual biological systems, the change in the assignment of a single codon brings catastrophic consequences in the functionality of the translated proteins. We did not constrain the assignments of codons to families of redundancy, further increasing the difference of each alternative proteome with the original. It is thus safe to assume that the alternative proteomes possess no functional information in the context of the original organism, even if they are derived from the same cDNA genome.

Our experimental results show that no particular value of algorithmic complexity is inherently bound to meaningful content. Meaningful content is intuitively associated to a high level of complexity because it is very unlikely that a periodically repetitive sequence can carry a meaningful message and the syntactic cues necessary to interpret it. In addition, the complexity of a semantic sequence can't be maximal, because maximal complexity would entail that the sequence is random and, once again, meaning and syntax would be impossible without an internal structure of some kind. As remarked by Abel and Trevors (2005), it is likely that most sequences containing some sort of semantic information will be located in a subset of high, but not maximum complexity. We partially agree with their suggestion, but we show that semantic sequences have no unique complexity value, different from the value of meaningless sequences. We show that there are meaningless sequences that have compression values remarkably similar to the standard proteome. In other words, *it is impossible to diagnose the presence of semantic information from the value of a measure that refers to sequence structure, though it is possible to predict that the structure of semantic sequences will have high complexity*.

Figure 3. Comparison between two possible "genetic codes". Briefly, the genetic code is constructed by the combination of the four nucleotides Uracyl, Cysteine, Adenine and Guanine shown in the table with the letters U, C, A and G respectively. Combination formed by three nucleotides are called codons, and particular codons code for the 20 aminoacids plus three STOP codons. In the figure, Code a) is the standard genetic code; Code b) is an alternative code, which was constructed artificially by randomly reassigning aminoacids to naturally-occurring codons, and differs from code a) in its general structure and the great majority of individual assignments. In spite of their different origins and structures, the translations effected by the artificial code were consistently similar to those of the standard genetic code in terms of informational complexity. See text for further details.

3. Discussion

3.1 The quantitative notion of information

Our computational experiment stresses different manifestations of information that are not reducible to one another. We propose that quantitative conceptualizations of information, centered on the description and analysis of the structure of a given sequence, be called "object-information", whereas the second kind to be explained in section 3.2, be called "process-information".

Object-information is associated to a given sequence, and the sequence is an object with defined, quantifiable characteristics such as Shannon's entropy (Shannon, 1948). These quantifiable characteristics are exclusively derived from the internal structure of the object. Under the scope of Information Theory, each element of a sequence has a frequency which can be interpreted as the probability of its appearance. The information entropy (*H*) of a sequence is expressed in the formula

$$H = -\sum_{i=1}^{N} P(i)\log P(i)$$

where P(i) stands for the frequency of each symbol *i*. The value of *H* ranges from 0 to 1; the greater the value, the more information a sequence contains. In the context of Shannon's theory, information is not meant to be semantic or functional. A higher content of information means that each symbol has a similar frequency; thus, a maximum informational entropy value means that the next symbol emitted by the source is completely unpredictable.

The limitations of information theory, of which Shannon himself was well aware, emerge when we consider other kinds of internal order. For example, a sequence can have maximum informational entropy and still be highly structured. Imagine, for example, a sequence that consists only of the symbol 0 repeated a hundred times, followed by the symbol 1 repeated another hundred times. This sequence is far from random but it is maximally entropic. Informational entropy only tallies the appearance of each symbol and yields a probability distribution for the next; as our imaginary sequence has exactly the same number of each of two possible symbols, the probability for the next symbol produced by the source is 0.5 (zero or one). In other words, the next symbol is completely uncertain and our sequence is unpredictable. This means that a sequence can be unbiased, have maximum informational entropy and a clear internal structure – it can even be a periodic repetition of signs, which obviously is far from complex. To solve this problem, Kolmogorov (1965), Solomonoff (1960) and Chaitin (1969) proposed independently the concept of algorithmic complexity: a sequence of length n is maximally complex if it needs an algorithm of length $\geq n$ to be specified, and this in turn means that the sequence is random as it has no discernible patterns to exploit. Without any sort of pattern, the algorithm or computer program to produce such sequence as output would necessarily be a program that specifies, symbol by symbol, the whole sequence.

More recently, other authors have attempted to circumvent the shortcomings of object-information. For example, Hazen et al. (2007) propose a measure of “functional information”. In general terms, a

sequence of signs has higher functional information if it raises the probability of bringing about some predetermined outcome. For example, the authors state that the sequence MAPLENMAIN has higher functional information than DANGERFIRE as it is more likely to summon the fire department to a particular address (the corner of Maple and Main streets). This work is particularly relevant as the authors explicitly try to deal with biological information as one of the uses of their quantification. However, once again the intention is to make meaning a characteristic internal to the sequence by introducing the assumption that systems remain fixed (its behavior and structure is always the same), and that somehow, there are universal cues that not only carry information but also help the system decide what is meaningful from what is not. Hazen et.al. call such cues *prescriptive information*. For example, in the sequence MAPLENMAIN the middle N must be universally understood as being short for AND and that whenever we say Maple and Main we all mean the corner formed by the streets Maple and Main. Note that these cues also help decide that DANGERFIRE is uninformative even if the system understands that such string of symbols tells of a fire happening somewhere. Note that Hazen et.al's analysis makes sense only in the context of a fire department. In any other context (say, a civilian that happens to tune in the frequency transmitting the message, or a passerby that stumbles upon a piece of paper with the message written on it) "DANGERFIRE" would certainly be more informative as it would bring to mind the catastrophe, while "MAPLENMAIN" would sound at most like a cryptic reference to an intersection. Hazen et al., however, argue that functional information can be measured regardless of the receiver, the context, etc. While they do not make this assertion explicitly (the example is well constrained inside the narrative of a fire department), as long as they try to measure information referring only to the sequence, they make information an attribute of the sequence itself. Thus, they are conflating an object (the sequence) with a process (with in this case includes but is not limited to the context of transmission).

The attempt to quantify prescriptive information relies on constructing a theory that ignores an indeterminate number of contextual events: in the example above, relaying the information to a fire department, the competence of firefighters in deciphering a mangled telegraphic message or their proficiency in the English language to name a few. Because of these problems, functional information faces a dilemma: either the measure is relative (which defeats the purpose of the theory) or the theory is riddled with a series of philosophical problems such as intentionality, determinism, mentalism or teleology.

Neither algorithmic complexity nor quantifications like Hazen et al.'s are absolute measurements like Shannon's. They are not computable and in any case, they can only approximate to a concrete value, as it is dependent on the method for determining it, but unlike informational entropy, they have other virtues. Algorithmic complexity has the advantage of exploring certain structural features of a sequence, and functional information tries to account for the response to a sequence. In any case, quantitative-centered analyses have in common the presupposition that information resides completely in the structural features of a sequence. Our experimental results show that biological information is not only a function of the structure of a sequence, so another theoretical framework is needed.

3.2 Expanding the notion of information

Both Shannon's and Kolmogorov's measures are just one aspect of information. They refer exclusively to a string of symbols, even though the string itself can be thought of as the result of a process – a continuous emission of signs in the case of Shannon, or a repetition of a specific sequence executed by a computer program in the case of Kolmogorov. The measures can be evaluated in the sequence itself, and no consideration of its context or its temporal changes is needed. Thus, the meaning of information is lost in the quantitative analyses. To deal with this shortcoming, we propose another kind of information which will need a new theoretical approach, and which we will name "process-information"[1].

Process information depends on an interpretative activity. This interpretation is the event that makes any given sequence *meaningful.* It can be immediately noted that process information deals with that facet of information that is sometimes deemed metaphoric by certain authors (Sarkar, 2001; Griffiths, 2001; Kay, 2000), as it can be extremely variable in its nature and results. Process information doesn't refer exclusively to genetic information: a meaningful message can be actualized through DNA and amino acid chains, but also through cellular receptors and external molecules or the sensorial data used by an organism to construct a representation of its surroundings, among many other examples. What matters is the relationship between the elements, and the dynamic change in the system caused by interpretation. In this respect, this notion of information is close to Gregory Bateson's (2000) formulation of information as "a difference that makes a difference".

---

[1] Both the division of the two notions of information and the description of the different domains they refer to were inspired by Anthony Eagle's excellent treatment of product- and process- randomness. See Eagle, 2014.

The complete separation of what we call object- and process-information can seem difficult, given that object-information can serve as a sort of raw material for informational processes. For example, a stretch of DNA (an object-information) can instruct the synthesis of a specific protein, via the conventional relationships between codons and amino acids that constitute the genetic code (process information). However, it should be kept in mind that the distinction between object-information and process-information resides at different levels of description. The concept of object-information is concerned with internal relationships; in the case of DNA, the ratio of each nucleotide, the possible patterns in which they appear and so on. Process information deals with external relationships; that is, the possible semiotic relations that can arise between nucleic acids and their amino acid translations via an interpretant element that cannot be part of the sequence. But what can this interpretant element be? In sections 3.3 and 3.4 we explore this issue by discussing previous attempts to understand biological information. However, it is important to say at this point that we are a long way from understanding what this interpretant element may be. Our experiment gives empirical support to an intuition held by many experts in the field that quantitative measure are not enough to account for physiological processes related for example, to genetics or epigenetics, fields where a clear understanding of information is crucial. For this reason, we believe that a clear conceptual separation of the two notions of information presented here can dispel the frequent misgivings about the appropriateness of the terminology. Thus, we strive to distance the concepts of information from the information-as-blueprint view and the determinism it entails. We also try to show that information, understood as two separate notions, does not entail a special kind of causation that is context-independent. Rather the opposite: contextualization, along with the physicochemical characteristics of the system in which information arises, are necessary to understand biological processes.

3.3 Comparison to other approaches to information

It is undeniable that many of the terms in molecular biology derive from the contingent synchronic rise of the DNA sciences, information technology and cybernetics; but it is not our goal to analyze the social or historical context of the scientific community and its linguistic conventions. Our discussion is closer to everyday scientific practice, as we try to evaluate the validity of the concept of biological information by stressing a separation between different theoretical frameworks. Commonplace scientific discourse usually makes no distinction between a sequence of DNA (object-information) and the meaningful information it supposedly carries (process-information). It is true that most of the time,

this can be considered a conceptual shorthand: “a gene for a trait”, be it anatomical, physiological or developmental. The shorthand might be even more precise in some cases, as the gene will be associated only with a specific post-editing mRNA molecule and the corresponding amino acid chain. We must be quite insistent that this is not the issue at hand; we are not discussing the nature of the concept “gene”, nor we are interested in localizing an arbitrary point in time in which the “informational molecule” expresses its message. The central issue of this paper is the same one that worries some authors like Lewontin (1983), Oyama (2000) or Sarkar (2001): that genetic information is conceptualized as an entity containing the instructions that will cause some trait to emerge, as if it were a scripted program interpretable in only one way. This attitude towards information can hardly be blamed. Minuscule changes in a sequence are readily associated with specific and often disastrous changes in a trait. These effects are in some cases so predictable that they are powerful tools for research in molecular biology.

Our approach to information differs from certain attempts to understand it systematically. For example, the framework we propose is not compatible with Floridi's treatment of information. Floridi (2005) supports an extended DOS (“declarative, objective and semantic”) definition of information, in which the communication is considered semantic if and only if it consists of data, is well-formed, meaningful and truthful. We have no place for truth-value in this account of semantics, as any concept of a “true” communication would assume the possibility of a correspondence between object-information and a state of the world prior the occurrence of the process. This assumption takes for granted, then, that some kind of symbolic information is inherently encoded in the DNA and it will either speak about the world (if it is not misinforming) or it will indicate a precise sequence of states that will take place within the living being. Neither of those presuppositions are considered in our informational account. In our account, meaning cannot be characterized as being “true” or “false; meaning can be evaluated only in terms of the changes in the dynamics of the system. Our distinction is similar to Sterelny and Griffiths' (1999) *causal* and *intentional* separation; indeed, they speak about an “aboutness” that can be misinterpreted as a trademark of intentional information. However, they completely reject an intentional element in biological systems. In fact, they place a divide between information derived from physical causation and an utterance with no necessary connection with a factual state of the world. The present paper argues that this intentional information can and must be included in a complete description of a biological system. This is done by carefully avoiding any attribution of process-information as a characteristic of an object. With this conceptual shift, we can escape the otherwise inevitable pitfall of affirming that *every possible* informational phenomenon is contained, previous to

the process of interpretation, in the informational sequence (as would seem necessary for accounts such as Hazen et al.).

The proposed division is closer, in any case, to Bergstrom and Rosvall (2009). Trying to keep Shannon's framework but still being true to what biologist understand when they talk of information, Bergstrom and Rosvall propose the "transmission sense of information", where "[A]n object X conveys information if the function of X is to reduce, by virtue of its sequence properties, uncertainty on the part of an agent who observes X (Bersgstrom and Rosvall 2009, p.165)." In their proposal, the marriage between Shannon and a decision-theoretic approach conveyed through the explicit introduction of an agent is meant to incorporate the notion that information is not simply a correlation between variables (Godfrey-Smith 2007), but a message transmitted to someone for something, in this case, from parents to offspring to help them make a living. While the intuition is certainly correct, Bergstrom and Rosvall still try to pack the two general notions of information (to be fair, only genetic information) into one general framework but, as we show, Shannon or Kolmogorov can deal with some, but not with all the information conveyed in the genetic sequence and therefore, seems better to us to use two different notions to better clarify the phenomena.

3.4 Bridging the two notions

The separation of concepts proposed throughout this paper is a necessary clarification. It contributes to understanding the information terminology that has been denounced on various grounds, from the inception of several genetic terms taken from the communication sciences, to the preformationist misconceptions about the dominance of genes as the principal source of order in the processes of life – and particularly in development. Firstly, the separation helps to rule out the preformationistic ideas: if a sequence, which fits the object-information theoretical framework, is treated like process-information, it is automatically endowed with more attributes than it can possibly have. We can extract useful data from the sequences, such as informational entropy or algorithmic complexity; but this kind of quantitative analyses will never reveal the network of relationships that an information-carrying object has in a biological system. This is precisely the usefulness of process-information: it treats information as this network of relationships, which eventually elicit changes in the dynamics of the system. A particularly striking example comes from one of our model organisms, the Entamoeba Mimivirus. For all their recently found complexity (Raoult and Forterre, 2008), viruses still depend on the interpretation of information by their host organisms. In this case, the Entamoeba molecular

mechanism is able to interpret DNA that is foreign to it; if it were not for this interpretative possibility, the viral cycle would be interrupted inescapably. The freedom that our process-information notion imparts upon interpretation makes it easy to describe and explain the informational interaction that occurs between seemingly distinct organisms, because our notion permits a point of view in which context and temporality are not only important, but necessary to understand the informational event. In the case of the Mimivirus for example, the host cell becomes an integral part of explaining and describing information as a process.

In addition, the separation is the groundwork on which a coherent theory of biological information can be built. The concept of information seems simple enough, but defies the efforts to construct a univocal definition. As Floridi (2004) points out, it is inherently polisemic and on par with the most philosophically complex terms such as “truth” or “being”. In this paper we do not defend the idea of information as a physical object, comparable to matter and energy, as Battail (2009, 2012) (maybe metaphorically) suggests. In our account, information is not some *thing,* but rather some *relationship*. Both process- and object-information are powerful concepts because of the richness and diversity of the situations and objects that fit the network of relations they posit. Moreover, a complete theory of informational relations is useful, at least heuristically, as it would treat its objects of enquiry as part of a symbolic system. This would permit an analysis of certain characteristic traits of symbols, such as arbitrariness, which can't be captured by a mechanistic causal explanation; it would not only justify the commonplace use of informational terms, but it would give a way to relate diverse elements of biological systems that is not restricted to mere physicochemical causation. Trevors and Abel (2004) make a similar point when discussing the origins of biological information systems, saying that the dichotomic view of life as the product of either chance or necessity is insufficient to account for the emergence of the genetic coding system. In any case, these discussions mean that information terminology would rely on a set of theories that are complementary, rather than antagonistic, to other approaches.

For example, the separation we try to clarify echoes the duality found in von Neumann's (1966), H. H. Pattee's (1995) and L. M. Rocha's (1997) accounts on the symbol-matter problem, which points to the possibility of expanding the divide proposed in this work to a broader scope. A radical point of departure from these authors is that they assume that a purely physicalist explanation is sufficient to account for the symbolic mode of behavior. Pattee, for example, places a divide between physical laws and boundary conditions which ultimately is localized as a sort of artifact derived from the presence of

an observer (his famous concept of “epistemic cut”). We consider Rocha to be closer to the views we propose, but ultimately, he too resolves the difference between matter and symbols with mechanistic presuppositions: “material symbol systems... must be formed out of certain available material parts” (section 2.2 in Rocha 1997), and thus are limited to a finite, predefined set of configurations. Rocha and Pattee's efforts are to parallel ours, as their starting point is also the realization of some elements of the description that are not accounted for by the usual theoretical framework. They both try to include some sort of contextual cues from the environment to make sense of the notion of information processing in biological systems. However, in the end they still restrict their analyses to a combinatorial logic and remain within a mechanistic view of information.

There have been some theoretical proposals that aim to deal with what we call process-information in a way that tries to account for their open-endedness, such as the innovative views of biosemiotics (see El-Hani et al, 2006; Favareau, 2008; Abel 2009; Barbieri, 2013). This relatively recent field centers fundamental analyses on the relationship between of three elements: an object, a signifying element and an interpretant. This triad of elements was proposed by the semiotician Charles Sanders Peirce (1955) as a way to answer a fundamental question that riddles us too: how can it be that symbols have an effect on the world, but are not describable by physical laws? His answer, which we find appealing, was to propose a more basic idea, his general, triadic, “short list” of categories. In a semiotic system, the sign is recognized as meaning something (i.e., the object), and elicits a change (i.e., the interpretant) in some aspect of the system. Note that the interpretant is not meant to be a little genie hiding somewhere in our cells, but a process that elicits a change. For example, Arnellos et al (2012) provides a physiological example: an antigen (the sign) refers to an object external to the cell such as a pathogen and elicits an immune response that is the interpretant. The semiotic relationship is said to be dynamical, because the identities are not fixed upon the different parts of the semiotic process; for example, a sign produced by an interpretative action of a semiotic system can act as an interpretant in a further triad, and so on. Thus, the semiotic phenomenon forms a web of sign relations, and the overall effect of this web is a structured change of dynamics of the system, when the system is confronted by different signs. In the example provided by Arnellos et.al, think of allergies: for as much as we know about the immune response, we are still a long way from grasping how is it possible that all of a sudden, your daily aspirin triggers anaphylaxis. In this case, the sign comes to mean something that the interpretant (a set of elements (objects and their interactions) belonging to the immune system) takes to be one thing when it used to mean something else. This contextual sensitivity and the focus on

temporal development of the semiotic process are what we want to convey with our description of process-information.

4. Concluding remarks.

Throughout this paper we have argued for a clear separation of different informational domains as a means to preserve part of the theoretical framework of information, and to address some of the valid criticisms made by various authors. There are still many unanswered questions regarding the role of information in living beings. For example, and most clearly, we lack a structuring account of process-information that helps us understand the complex interwoven net of interpretation. There have been many efforts to construct a viable formalization drawing from Shannon's theory (for a classical example, Bar-Hillel and Carnap, 1953), but these theories of semantic content make idealizations that render them unfit for our purposes. Particularly, they try to pinpoint semantic content as a property of object-information, and we have insisted that we need a semantic theory that deals with dynamic relationships of the different elements *as a process*.

As necessary the separation may be, the next step for developing a complete theory of information must be bridging the two notions. We have tried to point to the direction of this bridging, which in our view will take the form of a theory of information which can't be reduced to a quantitative framework. The consequences of this theoretical irreducibility reach far and deep, as many metaphysical commitments that are taken for granted must be reassessed. With this paper, our goal has been to contribute to the discussion of bioinformation by providing tangible evidence in the form of a computer experiment. Our results show how the concept of information in biology needs to be restructured and we propose it can be done by considering two sides of information in terms of objects and processes. The present paper has endeavored to present both sides, object- and process-information, as complementary. However, the manifestations of information that are not accounted for by Shannon's theory (or other similar quantitative frameworks) are generally considered metaphorical and, consequently, not integrated in as a solution for the criticisms of informational terminology. As our two notions of information are independent but clearly related, the separation and subsequent articulation we propose can serve as the common ground that is needed to build a general, more satisfactory theory of biological information.

Acknowledgements.

This work was supported in part by CONACYT (PhD scholarship awarded to A. Mercado-Reyes no. 358381). The authors wish to thank Carlos Rodríguez Silva for designing the figures of the paper; Fabrizzio Guerrero McManus and Mark Olson for reading and commenting on earlier drafts of the manuscript, and two anonymous reviewers for providing insightful comments and suggesting relevant literature. We are especially grateful to Dr. Minh Duc Cao for providing advice via personal communication for working with his compression algorithm.

References

Abel, D.L. (2009). The biosemiosis of prescriptive information. *Semiotica*, 174, 1-19.

Abel, D. L., & Trevors, J. T. (2005). Three subsets of sequence complexity and their relevance to biopolymeric information. *Theoretical Biology & Medical Modelling*, 2, 29.

Arnellos, A., Bruni, L. E., El-Hani, C. N., & Collier, J. (2012). Anticipatory Functions, Digital-Analog Forms and Biosemiotics: Integrating the Tools to Model Information and Normativity in Autonomous Biological Agents. *Biosemiotics*, *5*(3), 331–367.

Barbieri, M. (2013). Organic Semiosis and Peircean Semiosis. *Biosemiotics*, 6(2), 273-289.

Bar-Hillel, Y., & Carnap, R. (1953). Semantic information. *The British Journal for the Philosophy of Science*, *4*(14), 147–157.

Bateson, G. (2000). *Steps to an ecology of mind*. Chicago: University of Chicago Press.

Battail, G. (2009). Applying Semiotics and Information Theory to Biology: A Critical Comparison. *Biosemiotics*, *2*(3), 303–320.

Battail, G. (2012). Biology Needs Information Theory. *Biosemiotics, 6*(1), 77–103.

Bergstrom, C. T., & Rosvall, M. (2009). The transmission sense of information. *Biology & Philosophy*, *26*(2), 159–176.

Cao, M. D., Dix, T. I., Allison, L., & Mears, C. (2007). A Simple Statistical Algorithm for Biological Sequence Compression (pp. 43–52). Snowbird, UT, USA.

Chaitin, G. (1969). On the simplicity and speed of programs for computing infinite sets of natural numbers. *Journal of the ACM*, *16*, 407–422.

Eagle, A. (2014) "Chance versus Randomness", *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), E.N. Zalta (ed.), <http://plato.stanford.edu/archives/spr2014/entries/chance-randomness/>.

El-Hani, C. N., Queiroz, J., & Emmeche, C. (2006). A semiotic analysis of the genetic information system. *Semiotica*, *2006*(160), 1–68.

Favareau, Donald (2008). The Biosemiotic Turn. *Biosemiotics* 1 (1):5-23.

Floridi, L. (2004). Information. *The Blackwell guide to the philosophy of computing and information*. Blackwell Publishing Ltd.

Floridi, L. (2005). Is Semantic Information Meaningful Data? *Philosophy and Phenomenological Research*, *70*(2), 351–370.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., et al. (1995). The Minimal Gene Complement of Mycoplasma genitalium. *Science*, *270*(5235), 397–404.

Godfrey-Smith, P. (2007) Information in biology. In D. Hull and M. Ruse (eds.), *The Cambridge Companion to the Philosophy of Biology*. Cambridge University Press, pp. 103-119

Godfrey-Smith, P. and Sterelny, K. "Biological Information", *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), E.N. Zalta (ed.), <http://plato.stanford.edu/archives/fall2008/entries/information-biological/>

Griffiths, P. E. (2001). Genetic information: A metaphor in search of a theory. *Philosophy of Science*, *68*(3), 394–412.

Hazen, R.M., Griffin, P.L., Carothers, J.M., Szostak J. W. (2007). Functional information and the emergence of biocomplexity. *Proceedings of the National Academy of Sciences, 104,* 8574-8581.

Kay, L. E. (2000). *Who wrote the book of life?: a history of the genetic code*. Stanford, Calif.: Stanford University Press.

Kolmogorov, A. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, *1*(1), 1–7.

Lewontin, R. C. (1983). Gene, organism and environment. *Evolution from molecules to men* (pp. 273–285). Cambridge University Press.

Li, M. & Vitányi, P. (2008) *An introduction to Kolmogorov complexity and its applications.* Springer Verlag.

Moss, L. (2003). *What genes can't do*. Cambridge: MIT Press.

Oyama, S. (2000). *The ontogeny of information : developmental systems and evolution*. Durham, N.C.: Duke University Press.

Pattee, H. H. (1995). Evolving Self-Reference: Matter, Symbols, And Semantic Closure. *Communication and Cognition - Artificial Intelligence*, *12*, 9–27.

Peirce, C. (1991). *Peirce on signs : writings on semiotic*. Chapel Hill: University of North Carolina Press.

Raoult, D. (2004). The 1.2-Megabase Genome Sequence of Mimivirus. *Science*, *306*(5700), 1344–1350.

Raoult, D. & Forterre, P. (2008). Redefining viruses: lessons from Mimivirus. *Nature Reviews Microbiology*, *6*(4), 315–319.

Rocha, L.M. (1997). Evidence sets and contextual genetic algorithms: Exploring uncertainty, context, and embodiment in cognitive and biological systems. Binghampton: Systems Science, http://informatics.indiana.edu/rocha/dissert.html [last accessed March 14,2015], State University of New York.

Sarkar, S. (2001). Biological information: A skeptical look at some central dogmas of molecular biology. *The philosophy and history of molecular biology: New perspectives* (pp. 187–232). Kluwer Academic Publishers.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423.

Solomonoff, R. J. (1960). *A preliminary report on a general theory of inductive inference*. Report V-131, Zator Company, Cambridge, Mass.

Sterelny, K. & Griffiths, P.E. (1999) *Sex and death: an introduction to philosophy of biology.* University of Chicago Press

Trevors, J. T., & Abel, D.L. (2004). Chance and Necessity Do Not Explain the Origin of Life. *Cell Biology International* 28 (11): 729–39.

Von Neumann, J. (1966). *Theory of Self-Reproducing Automata*. University of Illinois Press.

Waters, E. (2003). The genome of Nanoarchaeum equitans: Insights into early archaeal evolution and derived parasitism. *Proceedings of the National Academy of Sciences*, *100*(22), 12984–12988.

Weaver, W. & Shannon, C (1963) *A mathematical theory of communication.* University of Illinois Press.

Wood, V., Gwilliam R., Rajandream, M.-A. Et al. (2002) The genome sequence of Schizosaccharomyces pombe. *Nature* 415, 871-880

	(1) Mean compression of the alternative proteomes (bps)	(2) Standard genetic code proteome compression (bps)	σ value of (1)	Right tail P-value
A.polyphaga mimivirus	3.83	4.0410	0.2135	0.1265
M.genitalium	3.7742	4.0999	0.1801	0.0692
N.equitans	3.8377	4.0735	0.2131	0.1093
S.pombe	3.9440	4.1651	0.1112	0.0234

Table 1. Statistical comparison between the standard genetic code and the mean of the artificially generated proteomes. Columns 1 and 2 compare the mean compression values of the alternative proteomes we constructed and the compression value of the actual, standard proteome respectively. Columns 3-5 summarize a statistical analysis of the difference of the values in 1 and 2.

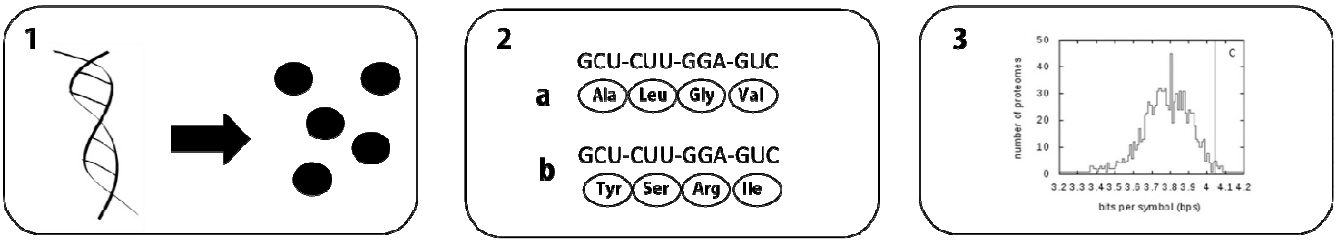

Figure 1. The experiment. We obtained the proteome (the entire set of proteins expressed by the genome) of four different organisms using their cDNA (1). Then, we produced a set of artificial proteomes for each organism by assigning to each naturally occurring codon a random amino acid (figure 2a illustrates the natural code and figure 2b the artificial assignment). Finally, we measured the information content of all proteomes, both natural (vertical line) and artificial (histogram), using string compression (3).

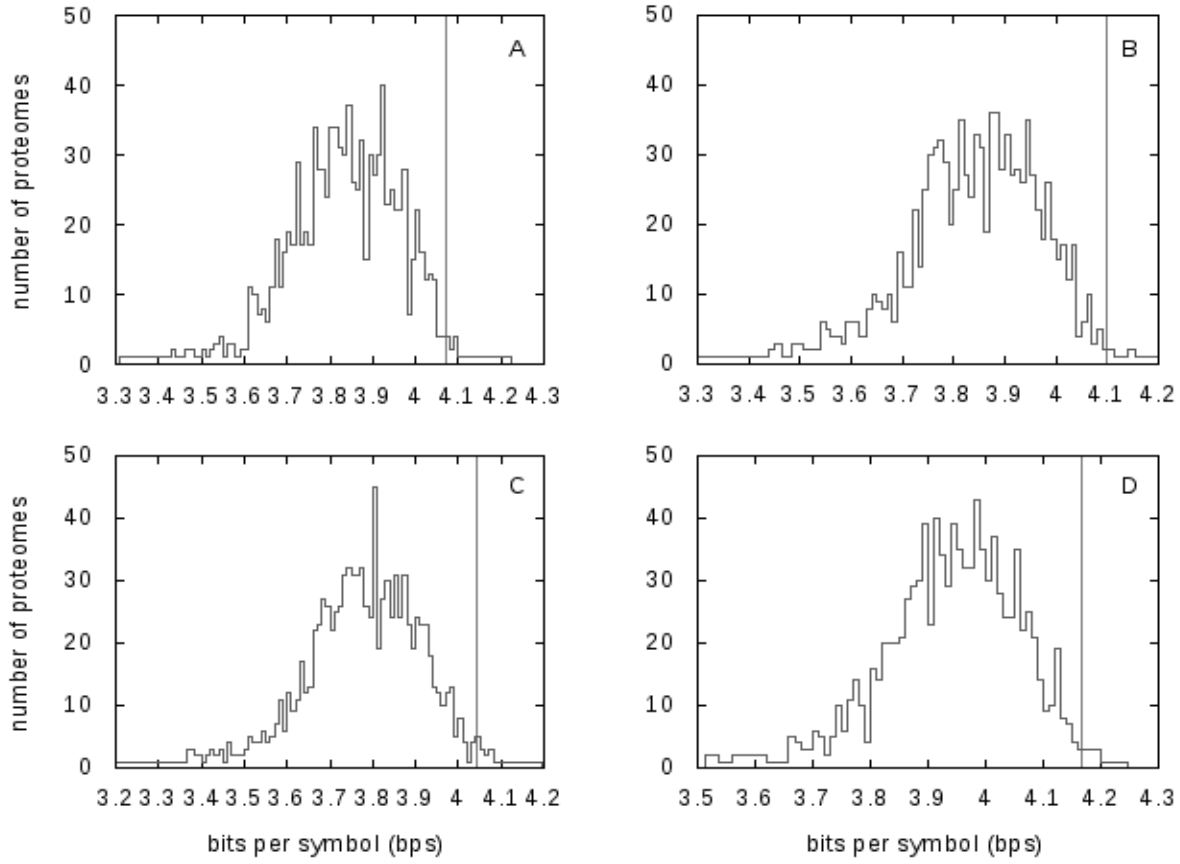

Figure 2. Comparison between the compression ratio of the standard proteomes (vertical line) and the distribution of the alternative proteomes derived from the genomes of a) *Entamoeba histolytica* mimivirus; b) *Nanoarchaeum equitans*; c) *Mycoplasma genitalium*; and d) *Saccharomyces pombe*. Histograms illustrate compression values for 1000 artificial proteomes and the single vertical line shows compression values obtained for the naturally-occurring proteomes for each organism under study. Note that even though the standard proteomes have an unusually high informational complexity, they are not unique, as we obtained alternative codes with similar complexity measures. See text for further details.

a)

	U	C	A	G	
U	UUU F UUC F UUA L UUG L	UCU S UCC S UCA S UCG S	UAU Y UAC Y UAA STOP UAG STOP	UGU C UGC C UGA STOP UGG W	U C A G
C	CUU L CUC L CUA L CUG L	CCU P CCC P CCA P CCG P	CAU H CAC H CAA Q CAG Q	CGU R CGC R CGA R CGG R	U C A G
A	AUU I AUC I AUA I AUG M	ACU T ACC T ACA T ACG T	AAU N AAC N AAA K AAG K	AGU S AGC S AGA R AGG R	U C A G
G	GUU V GUC V GUA V GUG V	GCU A GCC A GCA A GCG A	GAU D GAC D GAA E GAG E	GGU G GGC G GGA G GGG G	U C A G

b)

	U	C	A	G	
U	UUU V UUC Y UUA Y UUG P	UCU W UCC R UCA R UCG E	UAU S UAC F UAA D UAG W	UGU Y UGC F UGA D UGG W	U C A G
C	CUU H CUC M CUA I CUG T	CCU S CCC N CCA M CCG D	CAU E CAC W CAA C CAG M	CGU K CGC R CGA R CGG R	U C A G
A	AUU S AUC K AUA A AUG Q	ACU C ACC A ACA F ACG A	AAU P AAC K AAA T AAG V	AGU M AGC T AGA F AGG D	U C A G
G	GUU E GUC D GUA R GUG R	GCU N GCC W GCA H GCG T	GAU L GAC T GAA L GAG N	GGU P GGC I GGA W GGG R	U C A G

Fig. 3. Comparison between two possible genetic codes. Briefly, the genetic code is constructed by the combination of four nucleotides Uracyl, Cysteine, Adenine and Guanine shown in the table with the letters U, C, A and G respectively. Combinations formed by three nucleotides are called codons, and particular codons code for the 20 aminoacids (single capital letters) plus three STOP codons. In the

figure, Code a) is the standard genetic code; Code b) is an alternative code constructed artificially by randomly reassigning aminoacids to naturally-occurring codons, and differs from code a) in its general structure and the great majority of individual assignments. In spite of their different origins and structures, the translations effected by the artificial code were consistently similar to those of the standard genetic code in terms of informational complexity. See text for further details.