Language: A fresh concept to integrate syntactic and semantic information in life sciences

Rémi Tournebize [a], Cédric Gaucherel [a,b,*]

[a] CNRS, UMIFRE IFP, 605001, Pondicherry, India
[b] AMAP, INRA, CNRS, IRD, Université Montpellier, Montpellier, France

ABSTRACT

Several fields in biology tend to view the concept of information from one or the other of two extreme positions. Exclusionists base their stance of total rejection on gene-centrism and gene-determinism, typified by the recently-established endo-Darwinist school of life sciences. At the other end of the spectrum, there is total acceptance, as in the newly developed information-centred paradigms that populate biosemiotics. We propose in this paper to split the informational concepts into two irreducible (but linked) poles: the syntactic (concerned with the quantification of the information structure or complexity in a system), and the semantic (concerned with the organization rules and causality weights of interactions in a system). We claim that the past and present uses of the concept could then be classified as various degrees of oscillation between the two poles. The concept of *language* presents itself as a good tool with which to bridge the syntactic and the semantic poles, combining as it does the form-related and the meaning-related aspects of information, while methodologically supporting formal grammatical models in life sciences. We aim to show, at the same time, that neither of these poles alone can suffice to efficiently and holistically describe, model, and predict natural phenomena.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Since the 1920s, and through its exponential growth from the 1950s, the concept of information has occupied a significant amount of space in intellectual and scientific discourses in biology (Dretske, 1981; Griffiths, 2001; Godfrey-Smith, 2007). The concept entered the field of biology most conspicuously during the 1960s with the notion of *genetic information* and *genetic code* (Sarkar, 1996). Quite early on, other biological fields like physiology and developmental biology (through the concept of signalling pathways (Tkačik et al., 2008), ecology (Dall et al., 2005; Rivoire and Leibler, 2011) or evolutionary biology (Griffiths, 2001; Godfrey-Smith, 2007) recognized the heuristic power of the information concept. In each case, they adapted the concept to their own disciplinary needs and contexts, with very different senses: either as a measure of contingency between the states of a system (Godfrey-Smith, 2007), as a mean for organisms to reduce uncertainty associated to foraging, mating and other processes (van Baalen, 2013), or as a biological instruction (Monod, 1972; Barbieri, 2008). In evolution-

ary biology for instance, dynamics of allelic frequency changes have been interpreted as patterns bearing an *information* about the historical environments into which gene populations were embedded (Maynard-Smith, 2000; Angelier, 2008). A result of these multiple introductions and disciplinary re-definitions was the present-day semantic complexity associated with the notion of information in biology (Godfrey-Smith, 2007; Barbieri, 2008; Kupiec et al., 2016).

If information concept remains useful in life sciences, how to define it and how can we understand its high polysemy? In this paper, we claim that the concept of *language* presents itself as a useful tool with which to bridge the gap between the two irreducible but linked poles of the information continuum, the *syntactic* and the *semantic* poles. Besides, to use the concept of language as a basis for interpretation of biological phenomena is not far from other views originally proposed by Waddington (1972) or by Thom (1968). First, we aim at showing that neither pole by itself can adequately describe, model, or predict natural phenomena, and that the historical and present-day uses of the information concept can be classified in terms of degrees of oscillation between these two poles. Second, we propose to consider the conceptual advantage and mathematical modelling fecundity of formal language theory in biological sciences (Palacios et al., 1998; Witzany, 2016).

* Corresponding author at: INRA – EFPA, UMR AMAP, TA A.51/PS2, 34398, Montpellier Cedex 5, France.
E-mail address: gaucherel@cirad.fr (C. Gaucherel).

2. Biological information in question

The concept of information has been classically addressed by two opposed views in biology: either conceived as a solely cognitive construct used for the description and modelling of reality (a nominalist position, *e.g.* Thom, 1973), or as an agent possessing an objective ontological existence in the world akin to matter and energy, independent from sole human cognition (Bateson, 1979; Dall et al., 2005). While the contemporarily dominant view in biology may adopt the first position, this paper intends to suggest that many natural phenomena may actually rests on the agency of information-processing systems. Yet, before exploring this facet of information on the basis of non-random structures present in living systems, it is worth reminding that existing information-based approaches are not necessarily incompatible with noise or stochasticity.

Information proved to be a fashionable and inspiring metaphor in genetics (Crick, 1958 and even earlier in a more speculative form: Schrödinger, 1944), evolutionary biology (Maynard-Smith, 2000; Godfrey-Smith, 2007), developmental biology (Tkačik et al., 2008) or ecology (Stephens and Krebs, 1986), but many instances of its acceptance met with growing criticism. It has sometimes been mentioned the lack of a theoretical (or simply more rigorous) basis in some of these attempts (Griffiths, 2001). A good example of information rejecting views is given by the endo-Darwinist school that burgeoned with the work of French biologist Kupiec et al. (2016). In this section, we intend to show however that these information-free epistemological views are faced with many challenges.

Endo-darwinism contests the validity of models based on the concept of information in ontogenesis and genetics, largely based on the reject of a "programmatic representation of gene expression" conceived as a set of *biological rules of instruction* that are followed by the cell, like algorithmic instructions by a computer (Noble, 2010). Endo-Darwinism (also known as stochastic gene expression theory), is a recently proposed generalization of the synthetic theory of evolution when applied to infra-individual populations of some biological reproducing entities such as cells (Kupiec et al., 2016). For instance, it assumes that any organism development is a process analogically driven by the evolutionary forces (drift, mutation, migration and selection). Instead of attributing the determination of cell phenotypes to instructions natively encoded into the DNA molecule, they propose that cells divide with stochastically-expressed phenotypes later selected (both epigenetically and by selection during physicochemical interactions) into their cellular ecosystems (Heams, 2004). In contrast to the conventional gene-centred biology, endo-Darwinism therefore refutes a genetic determinism in the development of the individual phenotype. It suggests that every developing cell has the intrinsic ability to undergo a specific phenotypic pathway among a wide range of possibilities resulting from the stochastic expression of genes. Therefore, it also discards the role played by genetic information in the determination of cellular phenotypes. In this new context, the semio-chemicals (*i.e.* the chemicals bearing instructional content) of the program-centred theory are said to be a superstitious concept: they bear no "meaning", nor could they be involved in any form of symbolic coding: "The signal metaphor [becomes] entirely superfluous and misleading" (Kupiec et al., 2016).

Criticism of instructional information-based approaches, such as Turing machine models which were previously proposed to depict biological phenomena in a strict program-centred view, was already expressed by Longo (2006). Endo-Darwinism calls for a reappraisal of stochasticity in natural phenomena, dismissing *de facto* a large explanatory value previously attributed to information. Furthermore, one observes a range of conceptions between pure endo-Darwinists applying stochastic processes at all scales and at all organisation levels of life, and soft endo-Darwinists see-ing stochastic processes at some (large) scales only. We argue that endo-Darwinism, as a neutral theory regarding the role of information in biology, should help us to critique the shortcomings of what we call "semantic-based information theories". These theories focus on a representation of information in terms of rules, defined as single instructional information (*i.e.* a set of specific and stable discrete causalities). Hence, endo-Darwinism will also help us in pointing out the necessity of a fresh information-based theory, specifically driven by the wider, holistic concept of language.

Stochastic theories refute the specificity of the key-lock metaphor of signal induction and transduction, arguing for instance that "a majority of identified molecules are highly widespread and commonplace, and transcription regulations or transductions pathways are highly degenerate" (Edelman and Mountcastle, 1978). Therefore, they refute the existence of rules in life sciences, because of the polysemy of the supposed informing agents (*e.g.* signalling molecules in physiology, species in an ecological community). The polysemy argument is relatively weak to refute the whole semantic information structure, as the definition of "polysemy" already implies signs or meanings, however plural and ambiguous they appear. Besides, in a program-centred view, rather than being an insoluble obstacle, polysemy may reveal coding *in its historicity*, *i.e.* the variations and interplays of coding systems during life evolution (Atamas, 1996; Barbieri, 2008). Evolution can build up in hierarchical scales or reshape previous coding rules, thus generating a polysemy of codes (this property will be reviewed more precisely in the "semantic information" section). In addition, the forces involved in the stochastic "stabilization" of cell differentiation require that cells in the selected pool often be in a great amount. To be an emergent property of living phenomena, regulation requires a large spectrum of diversity. Some cases prove this requirement to be unrealistic, even if some models give positive results (Heams, 2004). It appears that most information-free models need some specific bounds in order to be realistic, these bounds being informative in the sense that they are defined by information-related constraints (Tkačik et al., 2008), which constitute the informational (structural) make-up of the organism.

No less significant is the fact that the endo-Darwinist theory advocates the extension of the theory of natural selection to cellular entities in order to overcome the lingering use of the information concept in genetics and developmental biology. Yet, does natural selection actually entail shrugging the information concept off? An important literature in evolutionary biology has in fact specifically addressed the question of thinking in terms of information in evolution (Maynard-Smith, 2000; Shea, 2007): "Evolution can be characterized as a process that shapes and maintains information across generations" (van Baalen, 2013). These views have in common to recognize natural selection as a process which encodes environmental dynamics and properties into the genetic information: adaptive traits are therefore viewed as *functions* which derived from the interaction between generation of entities and their environment (Godfrey-Smith, 2007). Recent developments in this field were concerned with the objective/subjective definition of information and its evaluation in terms of fitness (van Baalen, 2013).

In defence of an information-centred approach to understanding living phenomena, following the cybernetists (Blandin and Chapouthier, 1970; Bateson, 1979), we assume that codes, equivalent to an ordered set of rules in the first instance, are a category of mediating agencies that could perform just as well as, or perhaps better than, stochasticity-based theories. Indeed, codes can accommodate a stochastic scenario, which can be easily implemented into a more constrained structure, like a formal grammar algorithm. On the contrary, it is less parsimonious and feasible for a system based on mere stochastic interactions to generate stable coding and organized structures, apart as necessarily emerging properties of the dynamical system. *Prima facie*, endo-Darwinism is more