Biosemiotics (2009) 2:303–320
DOI 10.1007/s12304-009-9062-4

ORIGINAL PAPER

Applying Semiotics and Information Theory to Biology: A Critical Comparison

Gérard Battail

Received: 15 July 2008 / Accepted: 10 October 2008 /
Published online: 2 October 2009

Abstract Since the beginning of the XX-th century, it became increasingly evident that information, besides matter and energy, is a major actor in the life processes. Moreover, communication of information has been recognized as differentiating living things from inanimate ones, hence as specific to the life processes. Therefore the sciences of matter and energy, chemistry and physics, do not suffice to deal with life processes. Biology should also rely on sciences of information. A majority of biologists, however, did not change their mind and continued to describe life in terms of chemistry and physics. They merely borrowed some vocabulary from the information sciences. The first science of information available to biological applications, semiotics, appeared at the end of the XIX-th century. It is a qualitative and descriptive science which stemmed from efforts of linguists and philosophers to understand the human language and is thus mainly concerned with *semantics*. Applying semiotics to biology resulted in today's *Biosemiotics*. Independently, an explosive expansion of communication engineering began in the second half of the XX-th century. Besides tremendous progresses in hardware technology, it was made possible by the onset of a science of literal communication: *Information Theory* (Shannon, Bell Syst Tech J 27:379–457, 623–656, 1948). Literal communication consists of faithfully transporting a message from a place to another, or from an instant to another. Because the meaning of a message does not matter for its transportation, information theory ignores semantics. This restriction enables

Parts of this material have been presented at the 8-th Biosemiotics Gathering, 23–28 June 2008, Syros, Greece. Submitted to *Biosemiotics*.

The author has retired from ENST, Paris, France.

G. Battail (✉)
La Chanatte, le Guimand, 26120 Chabeuil, France
e-mail: gbattail@club-internet.fr

defining information as a measurable quantity on which a mathematical theory of communication is founded. Although lacking implementation means at its beginning, information theory became later very successful for designing communication means. Modern ones, like mobile phones, can be thought of as experimentally proving the relevance and accuracy of information theory since their design and operation heavily rely on it. Information theory is plainly relevant to biological functions which involve literal communication, especially heredity. This paper is intended to compare the two approaches. It shows that, besides obvious differences, they have some points in common: for instance, the quantitative measurement of information obeys Peirce's triadic paradigm. They also can mutually enlighten each other. Using information theory, which is closer to the basic communication mechanisms, may appear as a preliminary step prior to more elaborated investigations. Criticizing genetics from outside, information theory furthermore reveals that the ability of the template-replication paradigm to faithfully conserve genomes is but a prejudice. Heredity actually demands error-correcting means which impose severe constraints to the living world and must be recognized as biological facts.

Keywords Biosemiotics · Error-correcting codes · Heredity · Information theory · Semantics

Introduction

Accounting for the exchange of signals of all kind (chemical, electrical, optical, acoustical, ...) which occurs as an integral part of the life processes, at all scales from the molecules to the ecosystems, biosemiotics happily complements traditional biology. It is intended to fully take into account the communication of information in the living world, which is more and more recognized as differentiating it from the inanimate world, hence as fundamental in the life processes. Communication plays indeed in the living world an essential rôle which moreover is specific to life. It is not reducible to physics and chemistry but what is communicated—*information*—is an entity of its own. Biology *must* thus integrate the science of information. But there is no unified science of information. Two such sciences are available: *semiotics* and *information theory*. The former and oldest is at the origin of *biosemiotics* which can already gather a number of experts. The second one has been extraordinarily successful in engineering but very few people seriously try to apply it to life. The great pioneer here is Yockey (1992, 2005).

Biosemiotics borrows its paradigms from the communication between humans and thus incurs the risk of anthropocentrism. The founding fathers of semiotics lived well before the explosive expansion of communication engineering in the second half of the XX-th century. The idea that functions of abstract nature actually need a physical implementation was foreign to them.

The only instances of such an implementation they thought of were mental, i.e., involving the human brain, an organ of immense complexity which, a century later, is still poorly understood despite much works and progresses. The brain mechanisms were completely unknown when semiotics arised so it could only develop as an abstract, speculative discipline deprived of the support of experiments, except for mental ones.

By now, communication (and even communicating) machines have literally invaded our lives. The progress of communication technology was made possible not only by tremendous advances in the physical hardware, but also because information theory, which originated in a reflection about communication technology, provided a sound theoretical basis for the design and analysis of communication systems. Information theory lays emphasis on *literal communication* and quantitative methods. Its development was made possible thanks to a bold *a priori* position: *to discard semantics*. In the very first page of the paper which founds information theory (Shannon 1948), Shannon wrote:

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design. (Shannon's italics.)

Just like a messenger has not to know about the message he/her carries, a communication system ignores meaning and should operate regardless of the particular message it has to transmit. Replacing in the above quotation the word 'point' which refers to a location in space with the word 'instant' moreover extends its relevance to communication in time, hence to heredity.

Two Contrasting Sciences

It is hard to imagine that two sciences of communication could be based on such opposite postulates. For semiotics, semantics is the essence of communication and thus its main object; if not overlooked, literal communication is a trivial matter which deserves no special interest. For information theory, semantics must be discarded as irrelevant to the engineering problem of communication, which merely consists of making the transmitted message available to its destination. It is a necessary step in any communication however not the ultimate one. Information theory may thus be considered as the prerequisite to further refinements accounting for semantics. It may be thought of as an unavoidable intermediate in the path from conventional biology towards biosemiotics.

Table 1 Main features of semiotics and information theory

	Semiotics	Information theory
Beginning	End of XIX-th century de Saussure, Peirce	Shannon (1948)
Originated in	Philosophy and Linguistics	Engineering
Concerned with	Semantic communication	Literal communication
Difficulty	Intrinsically difficult as facing semantic problems	Basically easier as discarding semantics
Style	Descriptive and qualitative	Mathematical and quantitative
Impact on the physical world	Foreign to implementation	Useful in designing devices, systems and processes
Knowledge of borders		Found absolute limits for possible communication

The main features of semiotics and information theory have been gathered in Table 1. Each of its lines points out a feature about which they disagree and the remainder of this paper will elaborate on these contrasts, trying to show what benefits information theory can bring to biology.

What Engineering can Bring to Biology

Reviewing in Benner (2008) a book by Regis (2008), Steven Benner wrote:

> "Because building something requires a deep understanding of its parts [and of their mutual relationship], synthesis also stops scientists from fooling themselves. Data are rarely collected neutrally during analyses by researchers, who may discard some, believing the data to be wrong if they do not meet their expectations. Synthesis helps manage this problem. Failures in understanding mean that the synthesis fails, forcing discovery and paradigm change in ways that analysis does not." (The phrase in brackets has been added by me.)

Synthesis being the engineers' job, this remark is an excellent plea for a close collaboration of biologists and engineers. It is originally intended to genetic engineering but actually applies to any instance where nature and engineers are faced with the same problems. It puts in the forefront the necessary *implementation* of biological functions. Indeed, assuming the existence of some biological function without caring about how it is implemented pertains to wishful thinking. The engineering approach advocated in the above quotation should avoid it. Besides a renewed *understanding* of biological facts, another benefit of an engineering approach regards methodology: it makes possible *quantitative* assessments.

Communication engineering benefits from the theoretical framework of information theory. Literal communication of sequences of symbols ('literal' meaning that semantics is ignored) is actually a mathematical problem, and information theory is just that branch of mathematics which deals with it.

Information theory can bring to biology its concepts and methods as well as its results. Maybe its most important concept is that of *channel capacity*, proven to set an impassable limit to any communication. Information theory actually proves that communication without errors (more precisely, with an arbitrarily small error rate) is possible over a channel despite the errors which affect it, provided the information rate is less than the channel capacity, a quantity which decreases as the channel error rate increases. However, the very means which enable 'errorless' communication hinder any communication at all beyond the channel capacity. Both the possibility of 'errorless' communication below the capacity and its impossibility above it, although rather counterintuitive, are fully confirmed by the engineers' experience, besides being theoretically proven.

As an engineering discipline, information theory had a tremendous impact on the communication techniques. It established the limits of what is possible, and the engineers strove for approaching them. We briefly summarize the parallel development of information theory and communication engineering in the next section.

Information Theory and Communication Engineering

Two events of capital importance for the future of communication engineering occurred simultaneously at the same place: in 1948, at the Bell Telephone Laboratories, Claude Shannon published 'A mathematical theory of communication' (Shannon 1948); and John Bardeen, William Shockley and Walter Brattain invented the *transistor*. The technological developments based on the second event, i.e., the semi-conductor technology, provided means to implement solutions to communication problems having their origin in the first. It turns out that the information-theoretic solutions to communication problems are the more efficient, the more complex. The progress of semi-conductor technology resulted in devices becoming at the same time more and more tiny and more and more complex. By now, 60 years after the transistor invention, a silicon chip of a few square centimetres can bear about a billion transistors. The tremendous evolution of semi-conductor technology towards increasing complexity perfectly fitted the needs of communication engineering for implementing solutions inspired by information theory. Information theory especially led to the development of very sophisticated *error-correcting codes* which reliable and inexpensive devices can by now implement. They actually invaded our daily life: computer memories, mobile phones, CD, DVD, digital television … However, they remain invisible and very few people are aware of the enormous complexity which subtends electronic objects of daily use. With their trend towards complexity and small size, electronic devices tend to mimic biological devices. Just like we are unaware of the physiological processes which keep us alive, we are less and less conscious of the complexity of the electronic objects which we routinely use. Most of us completely ignore how they work. Moreover, explaining their operation often needs advanced concepts, mainly borrowed from information theory.

Information theory originated in Shannon's paper (Shannon 1948). It gave rise to a new science but, according to an approach almost unique in history, this paper also constitutes a complete treatise of the nascient science. Theoretical developments were nevertheless needed to confirm Shannon's statements, especially as regards the mathematical rigour of his proofs, but little was left to Shannon's successors for deepening and expanding information theory. One of the most important theoretical events in this respect has been the introduction of the *algorithmic information theory*, Kolmogorov and Chaitin around 1965, see Chaitin (2005), which, at variance with Shannon's, does not rely on probabilities. If Shannon left comparatively little to be done on the theoretical side, his papers prompted countless, entirely unexpected applications in the field of *source-* and *channel coding*. Source coding consists of replacing an initial message by a *shorter* but fully equivalent one. Channel coding aims at protecting an initial message against transmission errors; it necessarily introduces redundancy, i.e., replaces the original message by a *longer* one. Then within certain limits errors in the encoded message do not hinder recovering the initial one. As regards source coding, the Huffman algorithm asymptotically achieved the theoretical limit stated by information theory, namely, the source entropy, as early as 1952. Other efficient source coding algorithms were found later (arithmetic coding, Lempel-Ziv algorithm, ...). In sharp contrast, while information theory also stated the limit of what is possible as regards channel coding, namely, the *channel capacity*, no practical means to closely approach it were found during decades although it has been perceived as a challenge by thousands of mathematicians and engineers and thus prompted intense researches. The goal was not achieved earlier than 1993 when the invention of turbocodes by Berrou and Glavieux (Berrou et al. 1993; Berrou and Glavieux 1996; Guizzo 2004) provided practical means to communicate at information rates close to the channel capacity, hence experimentally proving the relevance of Shannon's channel coding theorem.

Is Literal Communication a Trivial Problem?

Is literal communication so trivial a problem? As defined in Shannon's quotation of "Introduction", it simply consists of making the transmitted message available to its destination. The message is generated at a distance (in space and/or time) from the destination so it needs be transported. *Transporting* the message generally needs its *transformation* by source- and/or channel coding. It can actually be transformed into an infinity of equivalent messages which possibly differ as regards their physical support, the size of the alphabet in use, and *coding* operated on the original message. Hence *an* information must be seen as an *equivalence class* with respect to all such transformations.

As an example, the sequence of Latin letters:

Information theory discards semantics (1)

and the binary sequence
1001001 1101110 1100110 1101111 1110010 1101101 1100001 1110100 1101001
1101111 1101110 0100000 1110100 1101000 1100101 1101111 1110010 1111001
0100000 1100100 1101001 1110011 1100011 1100001 1110010 1100100 1110011

0100000 1110011 1100101 1101101 1100001 1101110 1110100 1101001 1100011
1110011
share the same *information*, since the latter just resulted from transforming sequence (1) using the ASCII (American Standard Code for Information Interchange) 'code' which is currently used in computer memories. Each Latin letter is replaced by a 7-bit[1] word according to a one-to-one correspondence.

The binary sequence
10010011 11011101 11001100 11011110 11100100 11011011 11000011 11101000
11010010 11011110 11011101 01000001 11101000 11010001 11001010 11011110
11100100 11110011 01000001 11001001 11010010 11100111 11000110 11000011
11100100 11001001 11100111 01000001 11100111 11001010 11011011 11000011
11011101 11101000 11010010 11000110 11100111
also bears the same information as sentence (1) since an 8-th bit has been appended to each of the 7-bit words of the previous binary sequence, equal to the sum modulo 2 of its bits thus making the total number of '1's even. This may be thought of as a rudimentary means of error control: if an error affects a symbol in a 8-bit word, the number of '1's becomes odd so counting the '1's in each word enables detecting a single-symbol error. Of course, the first binary sequence could be transformed by sophisticated error-correcting codes into an equivalent one made resilient to errors (up to a limit) and bearing again the same information.

As a counterexample, the sentence:

La théorie de l'information exclut la sémantique (2)

is a French translation of the English sequence (1). Although it looks close to it, articles and a preposition have been appended to comply with the French grammar and the word order is different, so sentence (2) does not bear the same information as (1). However, both sentences share the same *meaning*.

These remarks will be further developped in "Information and its Relationship to Semantics".

On Semantic Communication

Ignoring semantics, information theory only deals with (literal) informations as just defined. Exclusion of semantics is a strength, by no means a weakness, of information theory. Semantics depends on interpretation rules, which themselves belong to the semantic field. Therefore, a text written in any language, say English, can state that it changes the meaning of the words. Imagine a text which tells that, from now on, the word 'table' will be used in order to mean 'point' and the word 'chair' to mean 'straight line'. Then, the nonsensical sentence 'one and only one chair passes through two tables' becomes an axiom of Euclidean geometry (interestingly, this example is borrowed from the great

[1] We systematically use the acronym 'bit' to designate a binary digit and the word 'shannon', abbreviated as 'Sh', for the binary unit of information, originally named 'bit' by Shannon.

mathematician David Hilbert, not from a linguist). Similarly, during World War II, the BBC broadcast apparently nonsensical messages intended to the Resistance fighters, to whom they were indeed very meaningful. Such remarks suffice to make the concept of meaning extremely difficult to use in a scientific context. The information-theoretic concept of information does not suffer the same basic weakness, although it restricts of course the semantic content of the word ‘information’. Discarding semantics from the very beginning, information theory has cut the Gordian knot.

There is, however, a scientific domain where using the concept of meaning suffers less difficulties and this exception, interestingly, is biology. We may indeed reasonably assume that nature does not intend to fool researchers while such an intent cannot be excluded in human communication.[2] Due to the assumed non-malignancy of nature, a kind of comparatively crude semiotics can be successfully used in biology, free from the endless semantic subtleties that characterize human communication and interaction. For instance, the genetic ‘code’[3] definitely establishes a correspondence between the 3-nucleotide codons of messenger RNA and the amino-acids in proteins so we may say unambiguously, e.g., that the meaning of the codon UAC is the amino-acid thyrosine.

The remark that taking semantics into account increases the difficulty of communication problems, even if malignancy can be excluded, clearly establishes a hierarchy of difficulty between semiotics and information theory. The former is clearly more difficult than the latter. The logical order for applying these sciences to biology would thus be to begin with the latter, which is basically simpler. Then, but only then, could problems involving semantics be attacked. Unfortunately, the historical order has been the reverse. In cruder words, biology has put the cart before the horse.

Information and its Relationship to Semantics

Shannon and the pioneers of information theory had an empirical approach since they did not attempt *defining* information, nor explicating its connection with semantics. They just proposed means for its quantitative measurement. We try here rather naïvely to define information, hopefully shedding some light on the relationship of information and semantics. These remarks are personal and do not express any consensus among information theorists.

Let us consider a digital message, i.e., a sequence of symbols from some finite-size alphabet. Such a message is a mathematical abstraction which needs a physical support for having any interaction with the real world, and especially

[2] Philip Henry Gosse (1816,1888) even attributed this intent to God in his book *Omphalos*, published in 1857, aimed at conciliating the data of geology with the Biblical account of creation.

[3] In the phrase ‘genetic code’ which appeared in the sixties, the word ‘code’ has been given a meaning rather foreign to its earlier use in information theory, that of a correspondence rule between objects of different nature, i.e., nucleotides and amino-acids.

for being communicated. The physical support of a message can take a variety of forms, either material (e.g., ink deposits on a sheet of paper according to conventional patterns, i.e., letters, or holes punched in a card, or local states of magnetization of some ferromagnetic medium, or shallow tiny holes in a compact disk (CD), ...), or consisting of the variation of some physical quantity as a function of time (e.g., air pressure, or electrical current, or electromagnetic field, ...). The material supports are actually used for recording, i.e., communication through time consisting of writing a message to be read later, while the variation in time of a physical quantity can be propagated as a wave hence enables communication through space. Regardless of its support, each of the alphabet symbols needs only be unambiguously distinguishable from the other ones.

A same message can be supported by different physical media which moreover can be converted from one to another. For instance, the message recorded on a computer memory or a CD can be converted into an acoustic wave (i.e., a sound), or emitted as an electromagnetic wave. Similarly, the alphabet size can be changed: a musical record in a computer memory uses the binary alphabet, but a large-size alphabet is needed for converting it into an audible acoustical signal. The message itself can moreover be changed so as to improve its characteristics as regards some desired property. For instance, it may be compressed so that its recording needs less memory space (this is source coding as alluded to above), or on the contrary expanded by a redundant encoding making it resist transmission errors (this is channel coding).

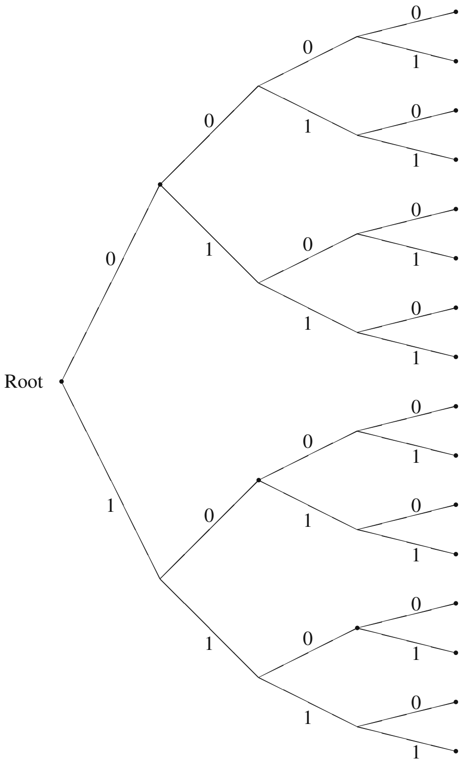
A message can thus exist into a variety of equivalent forms, depending on its possible encoding, alphabet size, and physical support. We refer to the underlying entity which is common to all these forms as *an information*. We may thus define an information as the equivalence class of all the messages which can be converted to each other by changing their encoding, alphabet size, or physical support. Each of these messages will be said to *bear* this information. The equivalence class associated with an information clearly contains infinitely many elements. Any of these messages is a representative of this class. An information is thus an abstract object which manifests itself in the physical world by any of its representatives, or *realizations*.

The following question immediately arises: given some alphabet with an arbitrary physical support, is there a minimal-length realization of a given information? The simplest alphabet, i.e., the binary one, is a natural choice. The problem becomes whether a minimal-length binary message exists within the equivalence class associated with the given information. Shannon's information theory asserts this existence when the given representative is a message generated by a stationary probabilistic source. The algorithmic information theory extends this statement to any message that a universal computer can generate. We'll refer to the minimal binary realization of an information as its *information message*. The fundamental theorem of source coding of Shannon's information theory states that the average length of the information message equals the length of the original message times the source entropy expressed using binary information units, i.e., shannons. In the algorithmic information theory, the length of the information message is used for defining the algorithmic complexity associated with the given information. The first case is much more restrictive than the second one, which may be considered as general. However, the source stationarity enables effectively estimating the probabilities of the sequences it generates, hence its entropy, by making frequency measurements. In the second case, on the contrary, the existence of a minimal-length realization is mathematically proven, but the algorithmic complexity is actually an uncomputable quantity which thus can generally not be evaluated.

Given a stationary source of entropy per symbol H shannons, the fundamental theorem of source coding states that any n-symbol message generated by this source can be transformed by source coding into a binary message of average length at least $\overline{\ell} = nH$ bits. This minimal-length realization of an information generated by a probabilistic source is its information message. It results from optimal source coding, which entails that its bits are probabilistically independent and equiprobable.

The set of all binary messages of length $\overline{\ell}$ can be represented by a tree like that of Fig. 1 with each of its branches labelled with a bit according to some convention. Let us interpret the i-th bit of the information message as the answer to a dichotomic question (i.e., answerable by yes-or-no), '0' meaning

Fig. 1 Binary tree for representing all binary sequences of length 4. An ascending branch represents the bit 0 and a descending one the bit 1



for instance ‘yes’ and ‘1’ meaning ‘no’. Any information message of length ℓ may thus be interpreted as one of the 2^ℓ paths of a binary tree of length ℓ taken from the root to the leaves, the choice of a branch at any fork being made at random with probability 1/2. A path in this tree, i.e., an information message of length ℓ, can be interpreted as *an integer* i, $0 \leq i \leq 2^\ell - 1$. The questions associated with the successive bits of the information message may for instance be those needed for identifying the species to which some given living being belongs. Provided the set of species is ordered according to a binary hierarchical taxonomy, using $\ell = nH$ properly chosen successive dichotomic questions enables distinguishing from each other $2^{nH} = (2^H)^n$ species. The entropy H of the source then measures the ability of the messages it generates to discriminate among objects of some outer world, hence to bear some kind of semantics. The corresponding information quantity is simply the number of binary digits which are needed to represent it.

Establishing a correspondence between each bit of the information message and a dichotomic question makes it eventually identify or represent an object which belongs to some outer world, like living beings as in the above example of taxonomy, making possible to distinguish between them. Then information in the above meaning is given a semantic content according to an external convention which we may refer to as its *meaning*.

This is a rather crude kind of semantics, apparently restricted to representing material objects which can be ordered in a tree-like fashion. However, the possible semantic content can be widely extended if we notice that:

- Describing an outer reality by a binary message is not limited to answering dichotomic questions. Data of various other kind can also be represented. For instance, grouping k bits of the information message into a block may be used to specify that the value of some parameter is one of 2^k predetermined levels. If k is large enough, this amounts to approximately specify the value of a continuously varying parameter. This is for instance currently used in telephony to represent instantaneous values of the speech signal (referred to as samples) by a sequence of bits, a process referred to as ‘pulse code modulation’ (PCM). For frequent enough samples (8 kHz, i.e., a sample every 125 μs) and k as low as 8 (hence $2^8 = 256$ levels), a sufficient quality of speech transmission is achieved.
- Moreover, a relation between material objects can be represented by the same means as the objects themselves, then opening the semantic field to abstract as well as material objects.

Information theory, either Shannon’s or algorithmic, uses the length of the information message as a quantitative measure of information. Since an information message of length ℓ enables distinguishing 2^ℓ different objects, ℓ is a logarithmic measure of the discriminating ability of the information message, regardless of the distinguished objects, a matter of semantics. We may thus understand a quantity of information as the number of semantic instances it can represent, or as the number of dimensions of some space which represents semantic objects. It should be kept in mind that the information quantity is by no means an exhaustive description of an information, just like the mass is only one of the many attributes that a material object possesses.

Notice that in the information-theoretic literature, the word ‘information’ is most often used to mean *measure* or *quantity* of information. For instance, *mutual information* refers to the quantity of information that the output of a channel provides as regards its input. The word *mutual* expresses that it is also the quantity of information that the input of a channel provides about its output. The capacity of a channel is the largest possible mutual information between its input and its output.

Some Remarks About Information Theory

Literal communication being basically simpler, information theory developed as a *mathematical* and *quantitative* science. However, the necessity of implementing the functions of communication revealed unexpected difficulties. This point will be illustrated by comparing quotations from Marcello Barbieri and Claude Shannon.

Barbieri (2008) defines semiosis as the *production* of signs, while Shannon, in the text already quoted in the introduction, writes that ‘the fundamental problem of communication is that of **re**producing at one point [...] a message selected at another point’. Shannon’s formulation implies that the difficulty of engineering communication mainly lays at the *receiving* end. Indeed, since the selected message is unknown at the receiving end, the choice which has been made at the transmitting end must be *inferred*. ‘Producing signs’ is, in a sense, trivial. Inferring what signs have been produced implies that the receiving end deals with the possibly sent signs as chance events, which entails that it needs knowing the repertoire of signs which can occur. Moreover, the intended destination of the produced signs seldom (actually never) escapes outer influences. The signs which are produced are thus not perceived in isolation, but in the presence of *noise*, this word designating the collective result of external influences which can only be dealt with as *random*. The receiving process is thus basically probabilistic, hence has a nonzero probability of failure or error. Reflection about implementation reveals here an overlooked difficulty.

Measuring Information Conforms to Peirce’s Triadic Paradigm

The necessity for the receiver to know the repertoire of signs in use results in the measure of information conforming to Peirce’s triadic paradigm. Let us consider a set $\mathcal{M}$ of M randomly occurring events or messages. Let p_m denote the probability that the particular message m occurs. Since a message of the set $\mathcal{M}$ must occur, we have

$$\sum_{m=1}^{M} p_m = 1. \tag{1}$$

Information theory measures the information quantity brought by the occurrence of the particular message m by

$$h_m = \log(1/p_m) = -\log p_m, \tag{2}$$

which is positive (more precisely, nonnegative) since $p_m \leq 1$. Then the more improbable is m, the more information it bears.

The *entropy* associated with $\mathcal{M}$ is the information quantity brought in the average by the occurrence of one of its elements, namely,

$$H = \sum_{m=1}^{M} p_m h_m = -\sum_{m=1}^{M} p_m \log p_m. \tag{3}$$

This information measure is relevant to the *set* of messages $\mathcal{M}$ as a whole, not to any particular message which belongs to it.

Now consider the information quantity h_m associated with a single message m according to (2). It does not depend on the message m itself, but only on the probability p_m of its occurrence. That the probabilities of all the messages in $\mathcal{M}$ sum up to 1 according to (1) entails that it is not possible to change one of these probabilities without changing others. Hence the information measure that information theory associates with a single event m according to (2) depends on the context in which it occurs, namely, the probabilities of the events in the set $\mathcal{M}$ to which it belongs: in a sense, the information borne by the occurrence of a message depends on the probabilities of the messages which *were not selected* as well as that of the single one which was. Measuring the information of a single event thus needs an *interpretation* in terms of its *context*.

No Definite Information is Borne by a Single Sequence

As an important consequence, it is meaningless to refer to 'the information borne by a sequence' since a single sequence can belong to an infinity of different contexts. For instance, the 8-digit sequence 10101010 bears at most 8 binary information units (shannons) if its digits are assumed to belong to the binary alphabet. However, 0 and 1 are also decimal digits as belonging to the set $\{0, 1, 2, \ldots, 9\}$ (they are actually digits in a numeration system to an arbitrary base b, $b \geq 2$). If the sequence 10101010 is interpreted as decimal, it bears $\log_2 10$ times more information, i.e., approximately 26.6 Sh at most.

An attempt to prove the concept of intelligent design allegedly using information-theoretic arguments, in the book by Pullen (2005), is wrong because single sequences are assumed to bear absolute quantities of information. Incidentally, a major argument in favour of intelligent design is that the probability of a single-nucleotide mutation being of about 10^{-8} per generation, a mutation involving two single-nucleotide mutations, assumed to entail a significant phenotypic difference, has a per generation probability of $(10^{-8})^2 = 10^{-16}$, hence is practically impossible. This would be true only if the two single-nucleotide mutations were independent events. This is not true, however, if both nucleotides belong to a word of a genomic error-correcting code. It is this codeword as a whole which is erroneously chosen if a regeneration error occurs and its symbols are strongly correlated. Ironically, the upholders of mainstream biology rightfully reply that the two single-nucleotide mutations are not independent events but simultaneously ignore the necessity of a genomic error-correcting code (to be recalled later) which is precisely the reason why they are correlated.

Difficulties are Mainly Found at the Receiving End

Realizing the unequal processing difficulty at the transmitting and receiving ends illustrates one of the benefits of an engineering approach pointed out by Benner as quoted in “What Engineering can Bring to Biology”: *implementation* questions how certain functions are performed and reveals unexpected difficulties. In the biological literature, the complexity of processes which take place at the receiving end is typically overlooked. This is especially true for *recognition* processes, the importance of which is capital in the living world. Humans, as most living beings, are very efficient in the tasks of recognition. It is only when engineers tried to design machines able to perform such tasks that it was realized how they were difficult. The laymen think of them as ‘natural’ and completely overlook the needed complexity of the mechanisms which perform recognition. As yet, the machines designed to this end, however sophisticated, remain significantly less efficient than those that nature implemented. Moreover, the best of them use artificial neural networks which need some learning process, i.e., mimic natural devices and processes.

Heredity as Literal Communication

Heredity is a problem of literal communication since it consists of communicating over time some message, the *genome*. The faithful communication of genomes is of capital importance for the living world as a whole.

Let us consider the two functions of the genome, as illustrated in Fig. 2.

As replicating itself, a genome provides (1) another identical genome, where ‘identical’ means that it is conserved at the geological timescale. A genome also (2) instructs the construction of a phenotype. Performing (1) is a problem of purely literal communication, hence fully relevant to information theory. We shall show below the benefits that applying information theory to this problem can provide to genetics. We’ll especially show that, contrary to the current belief, the template-replication paradigm does not suffice to solve it. Performing (2) involves semantics and thus escapes the competence of information theory, so we’ll let it aside. Notice that problem (2) is dealt with in a vast majority of papers. This may be due to the fact that we are phenotypes (with dormant genomes inside them) so we are more or less consciously affected with ‘phenotypocentrism’. Indeed, the two functions are absolutely necessary, but (1) plays the rôle of the egg in the chicken-and-egg dilemma, which may be thought of as the more basic.

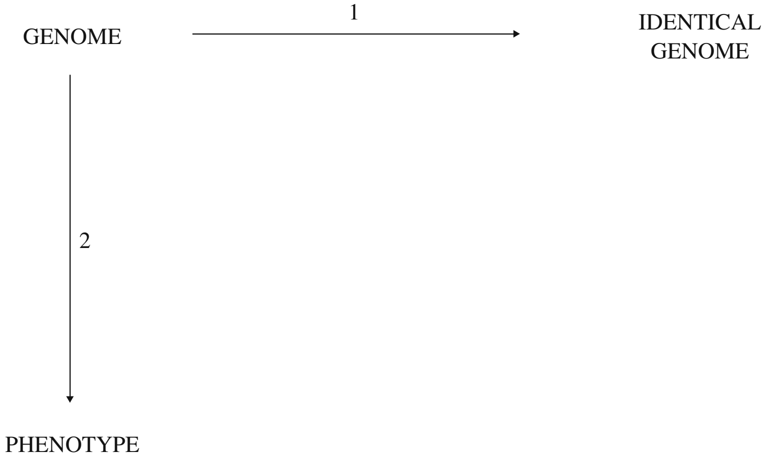


Fig. 2 What a genome generates

Conserving a message using any kind of memory, for instance DNA if it is a genome, is a problem of literal communication (over time) hence fully relevant to information theory. As a communication channel, any permanent memory has a capacity: an upper bound of which can easily be computed (Battail 2006b). It turns out that, except if the error rate is exactly 0, this capacity vanishes exponentially fast. In adamant words, no static memory is permanent. Since genetic mutations occur with nonzero probability, DNA alone cannot conserve the genome. This suffices to refute the template-replication paradigm.

The trouble with template-replication is that copying reproduces the erroneous symbols as well as the correct ones. The genome should not be copied, but made *resilient* to errors by means of an error-correcting code. Then, provided the cumulated number of errors remains less than some threshold, the genome can be *regenerated*. Regeneration is intended to rewrite the genomic message in such a way that:

- the rewritten message strictly satisfies the constraints which define the genomic code;
- and is the closest to the original genomic message.

Contrary to replication, regeneration does not result in a message faithful to the original one. It is however faithful to the genomic code seen as a set of constraints.

Then, the casual errors which may affect the genome are corrected, except if these errors are as numerous as to exceed the error-correcting ability of the code. If such a regeneration error occurs, it results in a widely different genome. For a long enough and properly designed code, and a short enough interval between regenerations, the probability of a regeneration error is extremely small. The crucial importance of conserving genomes necessarily led natural selection to retain a good enough genomic code and a short enough time interval between successive regenerations.

Regeneration in itself does not suffice to ensure the survival of a species: the number of its members should tend to increase, so a genome must be replicated once it has been regenerated. As regards its implementation, regeneration is much more costly than replication in terms of processing complexity.

We thus assume that:

- a genomic error-correcting code exists (main hypothesis);
- this genomic code unequally protects the data, since some of the oldest are the most faithfully conserved. A system of nested component codes is proposed as performing so (subsidiary hypothesis).

Assuming the above main and subsidiary hypotheses to be true, as we did first in Battail (1997), explains very basic features of the living world, e.g.:

- that nature proceeds with successive generations (the number of cumulated errors should not exceed the correcting ability of the code);
- the existence of discrete species directly results from the main hypothesis. Moreover, the subsidiary hypothesis entails that they can be ordered according to a hierarchical taxonomy which coincides with phylogeny;
- the trend of evolution towards increasing complexity, as a result of natural selection operating on the genomic error-correcting codes and favouring efficient codes, which need be longer to be more efficient as shown by information theory.

These topics were dealt with at length in previous works (Battail 2006a, c, 2007, 2008).

The above remarks illustrate the benefit that an external discipline can provide to biology by bringing ideas already foreign to it. This may help detecting and correcting prejudices and even possible misconceptions which remain unquestioned within its own framework. If a discipline tries to solve problems as they are posed within the conventional framework of another discipline, prejudices and misconceptions can be inherited within the problem statements. Then the main advantage of interdisciplinarity, criticism from the outside, is lost. It turns out that the semiotic approach to biology has not had the same critical rôle as information theory: its attention has been restricted to the processes by which *a genome generates a phenotype*. Semantics is central in this problem. Being mainly concerned with semantics, semiotics ‘naturally’ inherited this prejudice from molecular biology.

The goal of synthetic biology, artificial life, is very ambitious, Promethean indeed (Regis 2008). Much more modestly, having an engineering look at functions basic to life can provide biology with the claimed benefits of synthesis, namely, to quote again Benner, ‘forcing discovery and paradigm change’. As fully relevant to information theory, heredity is very interesting in this respect. The verdict of information theory is final: the template-replication paradigm has to be replaced by that of genomic error-correcting code.

Conclusion

Dawkins wrote in *The selfish gene* (Dawkins 1976), 1976:

> We do not know how accurately the original replicator molecules made their copies. Their modern descendants, the DNA molecules, are astonishingly faithful compared with the most high-fidelity human copying processes.

We have shown in earlier works and repeated above that copying is not the function that DNA molecules should implement, which instead must actually involve error-correcting means enabling their regeneration. Doing so, we actually complied with Dawkins's further remark (in *The Blind Watchmaker*, (Dawkins 1986), 1986):

> If you want to understand life, don't think about vibrant, throbbing gels and oozes, think about information technology.

However, the semi-conductor hardware is quite foreign to the enzyme-catalyzed reactions which occur in the cell. It is not at the level of implementation means that information technology resembles life, but as regards the algorithms which are implemented. Thus, we can fully agree with the above quotation only if 'theory' is substituted for 'technology'.

Indeed, biology *must hear* the lessons of information theory. A collaboration between biologists and information theorists will be highly beneficial for both. Maybe what mostly lacks for establishing such a collaboration is that information theory be properly *popularized*.

References

Barbieri, M. (2008). What is biosemiotics? *Biosemiotics, 1*, 1–3.

Battail, G. (1997). Does information theory explain biological evolution? *Europhysics Letters, 40*(3), 343–348.

Battail, G. (2006a) Information theory and error-correcting codes in genetics and biological evolution. In M. Barbieri (Ed.), *Introduction to biosemiotics*. Berlin: Springer.

Battail, G. (2006b). Error-correcting codes and genetics. *tripleC, 4*(2), 217–229. http://triplec.uti.at/.

Battail, G. (2006c). Should genetics get an information-theoretic education? *IEEE Engineering in Medicine and Biology Magazine, 25*(1), 34–45.

Battail, G. (2007) Impact of information theory on the fundamentals of genetics. In G. Witzany (Ed.), *Biosemiotics in transdisciplinary contexts*. Helsinki: Umweb.

Battail, G. (2008). Genomic error-correcting codes in the living world. *Biosemiotics*. doi:10.1007/s12304-008-9019-z.

Benner, S. (2008). Biology from the bottom up. *Nature, 452*(7188), 692–694.

Berrou, C., Glavieux, A., & Thitimajshima, P. (1993). Near Shannon limit error-correcting coding and decoding: Turbo-codes. In *Proc. ICC'93* (pp. 1064–1070). Switzerland: Geneva.

Berrou, C., & Glavieux, A. (1996). Near optimum error correcting coding and decoding: Turbo codes. *IEEE Transactions on Communications, 44*, 1261–1271.

Chaitin, G. (2005). *Metamaths!* New York: Pantheon.

Dawkins, R. (1976). *The selfish gene*. Oxford: Oxford University Press (New edition, 1989).

Dawkins, R. (1986). *The blind watchmaker*. New York: Longman.

Guizzo, E. (2004). Closing in on the perfect code. *IEEE Spectrum, 41*(3) (INT), 28–34.

Pullen, S. W. (2005). *Intelligent design or evolution? Why the origin of life and the evolution of molecular knowledge imply design*. Raleigh: Intelligent Design.

Regis, E. (2008). *What is life? Investigating the nature of life in the age of synthetic biology*. New York: Farrar, Straus and Giroux.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–457, 623–656.

Yockey, H. P. (1992). *Information theory and molecular biology*. Cambridge: Cambridge University Press.

Yockey, H. P. (2005). *Information theory, evolution, and the origin of life*. Cambridge: Cambridge University Press.