CHAPTER 13

INFORMATION THEORY AND ERROR-CORRECTING CODES IN GENETICS AND BIOLOGICAL EVOLUTION

GÉRARD BATTAIL
E.N.S.T., Paris, France (retired), E-mail: gbattail@club-internet.fr

Abstract: As semiotics itself, biosemiotics is concerned with semantics. On the other hand, the scientific study of communication engineering led to the development of information theory, which ignores semantics. For this reason, many biologists thought that it would be useless in their disciplines. It turns out however that problems of communication engineering are met in biology and thus can only properly be dealt with using information theory. As an important example, the faithful transmission of genetic information through the ages is a difficult problem which has been overlooked by biologists. Cumulated errors in the DNA molecule due to radiations and even to its own indeterminism as a quantum object actually perturb its communication through time. A simple information-theoretic computation shows that, contrary to the current belief, the genomic memory is ephemeral at the time scale of geology. The conventional template-replication paradigm is thus not tenable. According to a fundamental theorem of information theory, error-correcting codes can perform almost errorless communication provided certain conditions are met. Faithful conservation of genomes can thus be ensured only if they involve error-correcting codes. Then the genomes can be recovered with an arbitrarily small probability of error, provided the interval between successive generations is as short (at the time scale of geology) as to almost always avoid that the number of cumulated errors exceeds the correcting ability of the code

This paper presents an intuitive outline of information theory and error-correcting codes, and briefly reviews the consequences of their application to the problem of genome conservation. It discusses the possible architecture of genomic error-correcting codes, proposing a layered structure referred to as 'nested codes' which unequally protects information: the older and more fundamental it is, the better it is protected. As regards the component codes of this system, we notice that the error-correcting ability of codes relies on the existence of constraints which tie together the successive symbols of a sequence. It is convenient in engineering to use mathematical constraints implemented by physical means for performing error correction. Nature is assumed to use to this end 'soft codes' with physico-chemical constraints, in addition to linguistic constraints that the genomes need for directing the construction and maintenance of phenotypes. The hypotheses that genomic error-correction means exist and take the form of nested codes then suffice to deduce many features of the living world and of its evolution. Some of these features

M. Barbieri (ed.), Introduction to Biosemiotics, 299–345.

are recognized biological facts, and others answer debated questions. Most of them have no satisfactory explanation in current biology. The theoretical impossibility of genome conservation without error-correcting means makes these consequences as necessary as the hypotheses themselves. The direct identification of natural error-correcting means is still lacking, but one cannot expect it to be performed without the active involvement of practising geneticists. The paper also briefly questions the epistemological status of the engineering concept of information and its possible relation to semantics. Roughly stated, information appears as a necessary container for semantics, providing a bridge between the concrete and the abstract

Keywords: Biological evolution, error-correcting codes, genome conservation, genomic channel capacity, information theory, nested codes, soft codes

1. INTRODUCTION

It has been recognized during the last decades that recording, communication and processing of information play a paramount rôle in the living world, at any spatial and temporal scale, through a wide range of physical and chemical means. Moreover, it has become more and more apparent that the true divide between the non-living and the living things is precisely that the latter make extensive use of information, while the former do not. These statements legitimize the concept of biosemiotics and provide motivation for it. The researchers in the field naturally relied on the already established semiotics, especially following the pioneering works of Ferdinand de Saussure, Roman Jakobson, Charles Peirce, and others. Understandably, the semantic problems are at the heart of biosemiotics, as they are central to 'classical semiotics'.

Quite independently, the scientific study of communication engineering led more than half a century ago to the development of *information theory* (Shannon, 1948). It was intended to solve the technical problems associated with the communication of a message from some sender to some addressee, without any care of its semantic content: a messenger has indeed not to know about the meaning of the message he carries. Only outer characteristics of the physical support of this message (e.g., its spatial dimensions and its weight if it consists of a solid object) are relevant to him. Completely ignoring semantics, information theory introduced a quantitative measure of information perfectly fitted to communication engineering and very successfully used in this field.

One may wonder why information theory has been yet so sparsely used in biology. The early attempts made by biologists to use concepts of the 'classical' information theory (i.e., as introduced and developed by Shannon (Shannon, 1948)) almost invariably came to a sudden end with the remark that the entity referred to in the theory as 'information' is very restrictive with respect to the ordinary meaning of the word, especially insofar as it ignores semantics. They thought that a better

fitted 'organic information' should be used instead. However, they were unable to appropriately define it and they preferred to wait until somebody could do so.

This point of view would be tenable only if no problems of communication engineering were met by living beings. Communication is so familiar to humans that the necessity of physical supports and devices in order to perform it is very often overlooked. Engineers, who have to design systems and devices intended to communicate, are on the contrary fully conscious of this necessity. As an engineer, I observe that the engineering problems of communication are far from being foreign to the living world, so information theory has actually much more to offer to biologists than most of them believe. That information theory impoverishes the concept of information with respect to the common meaning of the word is undeniable. But is it a reason for rejecting it? As regards the definition of fundamental entities in sciences which use a mathematical formalism, it often occurs that *less is more*. It turns out that the admittedly restrictive concept used in information theory probably captures the most important features of information, at least as far as engineering functions like recording and communication are concerned. Moreover, the simplicity of its definition enabled extremely wide and successful developments. By 'successful' I mean not only that it enabled the onset of information theory as a consistent new science but that it has been experimentally confirmed in a striking manner through the countless engineering applications of information theory. At its beginning, however, information theory appeared as weakly connected with engineering practice for lack of an available implementation technology. The solutions that information theory offered to engineering problems looked by far too complicated to be reliably implemented at reasonable costs. But a few decades later the semi-conductor technology had made such progresses that this implementation became possible and fruitful. Although almost all the basic concepts of information theory were already contained in Shannon's work (Shannon, 1948), a very valuable collective experience was gained in its applications, which unfortunately is only shared within the community of communication engineers. I believe that the *a priori* rejection of the classical information theory deprived the biological community of a wealth of potentially useful concepts and results of fundamental importance. Stated in more adamant words, it amounted to throw out the baby with the bathwater. Among these ignored results, those related to the protection against errors have been generally overlooked by biologists. The firmly established, but paradoxical, theoretical possibility of *errorless communication in the presence of errors* is ill-known, as well as the main properties of the technical means which implement it, namely, the *error-correcting codes*. Laymen as well as scientists of other disciplines than communication engineering often ignore their existence, let alone how they work, although they make a daily use of them, e.g., as a necessary ingredient of mobile telephony.

I show indeed, and this is a major topic dealt with in this paper, that the template-replication paradigm is unable to account for the genome conservation. Although phenotypic membranes shield the genome from mechanical and chemical aggressions, errors in the DNA sequence of nucleotides inevitably occur due to

radiations (of cosmic, solar or terrestrial origin), and even because molecules are basically indeterministic quantum objects. Information theory tells the limits of what is possible as regards communication. The limit which is imposed on any communication by the presence of errors is called 'channel capacity'. It defines a horizon beyond which no communication is possible. Merely computing the channel capacity associated with the genomic memory shows that the template-copying paradigm of today's genetics has a far too short horizon in time, hence is unable to account for the faithful transmission of genetic information through the ages. The genomic memory actually appears as ephemeral at the geological time scale. In order to faithfully communicate the genetic information, the genome must act as an error-correcting code and be regenerated from time to time, after an interval as short as to ensure that the number of occurring errors is very unlikely to exceed its error-correcting ability. Besides being a trivially known fact, that nature proceeds by successive generations appears as an absolute necessity in the light of information theory. 'Generation' assumes here the strong meaning of genome *regeneration*. Notice that the parameters which determine the error-correcting ability of the code on the one hand, and those which control the time interval between regenerations on the other hand, are presumably unrelated. Depending on these parameters, a variety of situations as regards the permanency or mutability of species results. Their proper matching results from natural selection, making the time interval as short as to ensure the conservation of the species which we observe in the living world, but long enough to leave room for the variability needed to fit environmental changes.

Conservation of the genome then no longer appears to be the rule and replication errors, the exception. On the contrary, the genome conservation can only result from a dynamic process where error-correcting codes necessarily play a major rôle. This reversal of point of view has deep consequences on our understanding of the living world and the means by which it came to existence, i.e., biological evolution. Many features of the living world and of its evolution can actually be deduced from the main hypothesis that the genomes actually involve error-correcting means, together with the subsidiary one that they involve 'nested codes', i.e., combine several codes into a layered architecture. Some of these features are well-known biological facts, others suggest answers to debated questions. Most of them are not satisfactorily explained by current biology. In a sense, one may rightfully think of such results as speculative as mere consequences of hypotheses. However, the information-theoretical impossibility of genome conservation without error-correcting codes makes these hypotheses necessary, and so are the consequences derived from them.

As regards how nature implements error-correcting codes, I first recall that the ability of a code to correct errors results from the dependence between the symbols of its words which has been induced by the encoding operation. Engineers use to this end constraints of mathematical character because they are well defined and easily implemented by physical devices. Constraints of other kind can be used for the purpose of error correction, however, although mathematical constraints only are actually implemented by engineers due to their convenience. I thus assume

that, at variance with engineers, nature uses physico-chemical constraints which make the successive nucleotides of DNA mutually dependent hence endowed with error-correction ability. I also contemplate constraints of linguistic character: in order to direct the phenotype construction, the genome must use some kind of language which necessarily implies constraints. Notice that the possibility of error correction then appears as a by-product of other biological constraints, according to an approach which is typical of nature and can be referred to as 'tinkering'[1]. No explicit encoding operation is needed. In fact the direct experimental identification of natural error-correcting means is still lacking, but one cannot expect it to be performed without the active involvement of practising geneticists.

2. AN INTUITIVE OUTLINE OF INFORMATION THEORY AND ERROR-CORRECTING CODES

2.1 Shannon's Paradigm, Variants and Interpretations

Before we can deal with the possible use of error-correcting codes for genomic communication, we need to introduce some information-theoretic concepts originating in communication engineering. We shall also propose variants and interpretations of these concepts which hopefully will make them useful in the context of genetics. These concepts and interpretations were to a large extent left implicit in our previous works about the application of error-correcting codes to genetic communication (Battail, 1997–2006), so the present section may be thought of as a key to help their understanding. We shall then provide an overview on error-correcting codes.

2.1.1 Shannon's paradigm

The basic scheme of a communication in the engineering literature (sometimes explicit but more often implicit) is Shannon's paradigm, a variant of which is represented in Fig. 1. A *source* generates an information message (an arbitrary sequence of symbols, a symbol being an element of some given finite set named alphabet[2]) intended to a *destination*. The only link between the source and the destination is a *noisy channel* which propagates the signals representing the symbols of its input message in an imperfect fashion, so a received symbol can differ from the transmitted one (an event referred to as *error*) with nonzero probability. The device labelled *channel encoder* transforms the message that the source generates into the one which enters the noisy channel. The *channel decoder* operates the inverse transformation and delivers to the *destination* a message which is hopefully identical to the one the source generated.

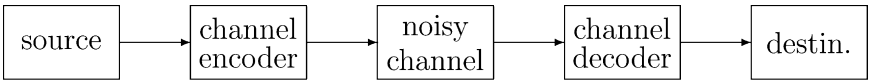

Figure 1. Shannon's paradigm, restricted to the functions of channel encoding and decoding

The encoded message must contain a larger number of symbols than the source output, a property referred to as *redundancy* which is essential for protecting the information message against the channel noise, i.e., to perform *error correction*. According to the *fundamental theorem of channel coding*, if the redundancy is high enough (depending on the channel noise and being the larger, the worse is the channel), the joint operation of the encoder and the decoder can ideally overcome the channel errors. Then the message that the destination receives is identical to the one originally delivered by the source, and error-free communication is achieved. The error probability after decoding cannot exactly equal zero for a message of finite length, but increasing the message length and properly designing the encoder and its associated decoder can make the probability of a decoding error vanish when the message length approaches infinity, regardless of the channel error probability. More on this topic will be found below (Sec. 2.3.2).

Shannon's paradigm is so important as a background in communication engineering that we now give some comments about it. We believe it conveys a spirit which biologists should share with engineers in order to actually realize and exploit the potential of information theory in their own discipline, as well as to be aware of its limits.

2.1.2 Some preliminary comments on Shannon's paradigm

Notice first that the success of Shannon's paradigm in engineering is due to a large extent to its *flexibility*. The borders of the blocks in Fig. 1 should not be considered as rigidly fixed, but on the contrary as possibly redefined at will so as to best fit any problem at hand. We shall try below to use this latitude in order to make it useful in the context of genetics. The only requirement is that everything that perturbs the communication between the source and the destination, to be referred to as *noise*, should be located in the channel. In order to prevent misunderstandings, let us stress that the concepts of information and noise are by no means absolute. On the contrary, the message from the source is considered as bearing information only insofar as it can be used by the *destination*. The error events are interpreted as noise with respect to this particular choice of the information message. The usefulness to the destination is the *only* criterion which differentiates the information message from the noise, meaning that their distinction is *arbitrary*. They may well exchange their rôles. For example, the sun is a source of noise for a receiver intended to the signal of a communication satellite. However, the signal from the satellite perturbs the observation of a radioastronomer who studies solar radiation, hence it is a noise source for this particular destination. When we refer to the usefulness or harm that a sequence of symbols has for the destination, we necessarily consider its *purpose*, hence we cannot avoid some kind of teleology or subjectivity.

The entities at the extremities of the communication link in Shannon's paradigm, namely the source and the destination, are *devices* or *living beings*. In an engineering context, the devices are man-made artifacts so the living beings ultimately involved are human, either directly or indirectly through their engineering products. Insofar as we consider nature as an engineer, we may think of the source and destination

as extended to other living beings or nature-made devices. On the other hand, the channel is a physical device or system.

As it has been described and represented in Fig. 1 Shannon's paradigm is unidirectional, in the sense that it only considers the source sending messages to the destination, but not the destination sending messages to the source. A conversation between humans, for instance, is better represented as a bidirectional scheme. Such a scheme is also used in many cases of interest to communication engineers. Then each of two locations contains a source and a destination merged into a single entity, and two channels are provided between the source of one of the entities and the destination of the other one.

2.1.3 Communication through space or through time

In communication engineering, the source and the destination of Fig. 1 are located at a distance and the message has thus to be communicated through space. It consists of a succession of symbols in time. However, the same scheme can depict the situation where the source and the destination are separated in *time*. Then the message is a sequence of symbols written on some support extending itself in *space* and read later, so the 'channel' of the figure refers in this case to the support on which the message is written. The channel errors result from the degradation that this message possibly suffers during the time interval which separates the instant of writing the message and that of its reading. In genetics, we are concerned with such a communication through time and the support of information is a DNA (or RNA) unidimensional polymer. In such a case, no communication channel can exist from the future to the past (such a channel would violate causality), so only the unidirectional version of Shannon's paradigm, as depicted in Fig. 1, is valid to represent genetic communication. Similarly, if the source is a gene and the destination the molecular machinery which generates a polypeptidic chain, the central dogma of molecular biology asserts that only the unidirectional version of Shannon's paradigm is valid, although no violation of causality would result from the availability of a channel in the reverse direction. For these reasons, we shall in the following exclude the bidirectional scheme and consider only the unidirectional version of Shannon's paradigm as fitted to genetic applications.

2.1.4 Variants of Shannon's paradigm intended to genetic applications

In Fig. 1 which describes a usual engineering situation, the source is clearly identified as the origin of the message. Sending an information message is intentional, which is compatible with the above remark that both the source and the destination are ultimately human beings. We now propose a variant of Shannon's paradigm which better fits the absence of purpose which is typical of the biological context (where, at least, we do not know if there is an intention).

This variant of the basic scheme of Fig. 1 is depicted in Fig. 2, where we just merged the first two blocks of Fig. 1 into a single block (inside a dashed box) labelled 'redundant source', and its last two blocks into a single one (again inside a dashed box) labelled 'redefined destination'. Then the message delivered by the

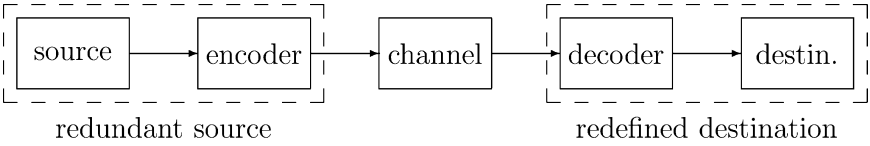

Figure 2. A variant of Shannon's paradigm

redundant source has error-correcting capabilities which can be exploited by the redefined destination.

Let us now consider the case of a genetic message incurring successive replications. Figure 3 depicts the case where two channels are successively used (i.e., where two replications successively occur). The error-correcting properties of the encoded message enable merging the decoder which follows the first channel with the encoder which precedes the second one into a single entity named 'regenerator'. The concept of *regeneration* thus appears as better fitted to the context of genetics than the engineering concept of decoding which refers to an explicit 'information message' to be communicated, and its very possibility relies on the redundancy of the initial encoding.

We may now describe a chain of successive replications incurred by a registered message (e.g., the genetic message of DNA) as in Fig. 4. An original redundant source at left delivers a redundantly encoded message which is written on a first support (labelled 'channel 1'), regenerated in the device labelled 'regen. 1', written again on 'channel 2', etc. The last step considered is the i-th regeneration, where i is a finite but possibly large number.

The redundant source has been depicted in Fig. 2 and regenerators in Fig. 3. If the number of replication steps i is very large, the initial encoded message from

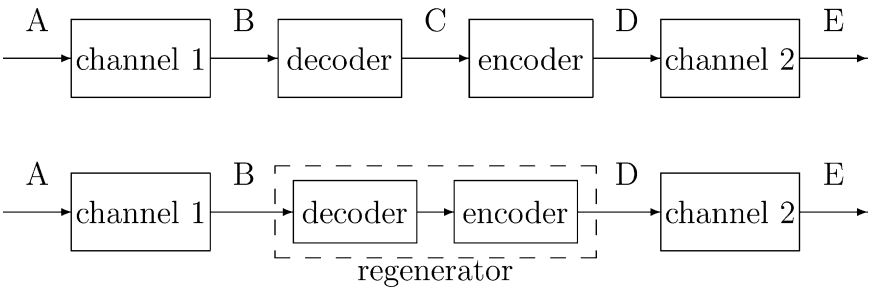

Figure 3. The regeneration function. Channels 1 and 2 have to be used successively. The upper picture is relevant to a conventional error-correcting code. The sequences found at the points designated by letters are: in A, an encoded sequence; in B and E, received sequences; in C, an information message; and in D, the sequence which results from the encoding of the first decoded information message, hence the restored initial sequence if no decoding error occurs. In the lower picture, the decoder and encoder have been merged into a single entity labelled 'regenerator' where the information message no longer appears

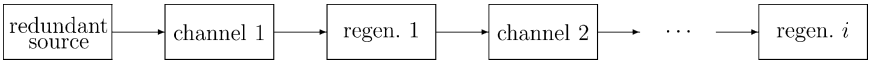

Figure 4. A chain of i successive regenerations

the redundant source is likely to have been modified and even maybe 'forgotten', insofar as regeneration errors likely occurred. As i increases, the current encoded message thus depends less and less on the original redundant source, and more and more on the regeneration errors which occurred, i.e., on a succession of contingent events. In the context of genetics, of course, natural selection operates on the corresponding phenotypes, letting only the fittest ones survive. This illustrates well the ambivalence of regeneration errors. On the one hand, they should occur very unfrequently for the sake of genome conservation, which is necessary for maintening a species and multiplying the number of the individuals which belong to it. On the other hand, they give rise to new phenotypes when they occur, thus exploring at random the field of what is possible. These new phenotypes are the rough material which makes evolution possible, as targets of natural selection.

Remember that, as we noticed above, the rôles of information and noise are relative to each other, entirely depending on their usefulness or harm for the destination in Shannon's paradigm. In the context of genetics, the destination consists of the cell machinery which processes the genetic message borne by a DNA molecule. Therefore, the regeneration errors *create new genetic messages* in a very objective sense. There is no paradox in this statement since there is no difference of nature between information and noise, as we already noticed. Moreover, because we liken the error-correcting codes to the multiple constraints which make DNA strings admissible as genetic messages (see Sec. 5 below), such strings having suffered regeneration errors are just as plausible to the cell machinery as the error-free ones.

2.2 An Outline of Information Theory

We shall now use Shannon's paradigm in order to introduce the main quantities defined by information theory and discuss their relationship. We first introduce the basic measures of information (proper and mutual) and the quantities which are associated with the blocks of Shannon's paradigm: the entropy of the source and the capacity of the channel. We introduce also the basic functions of source coding (intended to replace an initial message by a shorter one bearing the same information) and channel coding (aimed at protecting a message against transmission errors), and show that the entropy and the capacity acquire an operational significance as defining the limits of what is possible in source- and channel-coding, respectively. We also mention an alternative definition of information seen as a measure of complexity which gave rise to the algorithmic information theory (Kolmogorov, Solomonov, Chaitin). It provides another understanding of information but does not really question the validity of the basic classical results. We believe that Shannon's information is better fitted to the actual application

of information theory to sciences like biology but our discussion below of the relationship of information and semantics will also be inspired by the algorithmic information theory.

2.2.1 Basic information-theoretic measures

This section is intended to provide an insight into the information-theoretic quantities used in the paper, especially the channel capacity. Of course, much more on these topics can be found in Shannon's seminal work (Shannon, 1948) and in textbooks like (Gallager, 1968; Cover and Thomas, 1991).

The *information* brought by the occurrence of an event is measured by its *unexpectedness*. If an event x occurs with probability p, then its unexpectedness can be measured by $1/p$ or by some positive increasing function of it, $f(1/p)$. If two events x_1 and x_2 occur independently of each other with probabilities p_1 and p_2, respectively, it is reasonable to measure the information associated with the occurrence of both by the sum of the information measures separately associated with each of them. The probability of the outcome of both events is $p_1 p_2$ since they were assumed to be independent, so the function f is chosen such that

$$f(1/p_1 p_2) = f(1/p_1) + f(1/p_2).$$

The logarithmic function satisfies this equality and has many desirable mathematical properties, especially continuity and derivability, and moreover can be shown to uniquely satisfy a set of axioms which are plausible for an information measure. It has thus been chosen, so the above equality becomes

$$\log(1/p_1 p_2) = \log(1/p_1) + \log(1/p_2) = -\log(p_1 p_2) = -\log(p_1) - \log(p_2).$$

The information associated with the occurrence of the event x of probability p is thus measured by

$$i(x) = \log(1/p) = -\log p.$$

Choosing the base of the logarithms defines the unit of information quantity. Following Shannon, the most usual choice of this base is 2, and the unit is then referred to as the *bit*, an acronym for *bi*nary digi*t*. However, a digit and an information unit are distinct entities. A binary digit does not necessarily bear information; if it does, it bears at most a binary information unit. To avoid any confusion with the more common use of the acronym 'bit' in order to mean 'binary digit', we prefer to rename 'shannon' the binary unit of information, instead of 'bit' as Shannon originally did.

Let us now consider a repetitive random event X having its q possible outcomes denoted by $x_1, x_2, \ldots, x_q$, each occurring with probability $p(x_1) = p_1, p(x_2) = p_2, \ldots, p(x_q) = p_q$, respectively. That one of the q outcomes necessarily occurs results in

$$p_1 + p_2 + \cdots + p_q = \sum_{i=1}^{q} p_i = 1.$$

The quantity of information brought in the average by the occurrence of X is thus

$$(1) \quad H(X) = p_1 \log(1/p_1) + p_2 \log(1/p_2) + \cdots + p_q \log(1/p_q) = -\sum_{i=1}^{q} p_i \log(p_i),$$

a positive quantity if none of the probabilities $p_1, p_2, \ldots, p_q$ is 0 or 1, i.e., if X is actually random. In information theory the quantity $H(X)$ is referred to as the entropy of X, but since the same word is used in physics with a different but related meaning we prefer to name it *prior uncertainty*. It measures the uncertainty which precedes the event and is resolved by its occurrence. The event X which consists of choosing a symbol among an alphabet of size q with probabilities $p_1, p_2, \ldots, p_q$, to be referred to as the *source event*, thus provides an average amount of information equal to $H(X)$. The maximum of $H(X)$ is achieved when $p_1 = p_2 = \ldots = p_q = 1/q$. This maximum equals $\log q$.

When the successive outcomes of X are not independent, the prior uncertainty (entropy) per symbol of a stationary source[3] is defined as the limit:

$$(2) \qquad H = \lim_{n \to \infty} \frac{1}{n} H_n,$$

where

$$H_n = -\sum_{\underline{s}} p(\underline{s}) \log p(\underline{s}),$$

where $\underline{s}$ is any sequence of length n output by the source and $p(\underline{s})$ is its probability. The summation is made over all possible sequences $\underline{s}$. The assumption that the source is stationary suffices to ensure that the limit in (2) exists.

An important property of the entropy of a redundant source is that the set of its outputs can be divided into two disjoint categories: the *typical* and atypical sequences. The number of typical sequences is about q^{nH}, where q is the source alphabet, n is the sequence length and H the entropy of the source expressed using logarithms to the base q. For long enough sequences, the probability that the source output is atypical vanishes, which means that the probability that the source generates a typical sequence approaches 1. Remember that the maximum entropy of a q-ary source is $\log q$ and that $\log_q q = 1$. For a redundant source, i.e., such that its prior uncertainty or entropy differs from the maximum, its value H when using logarithms to the base q is less than 1. The q^{nH} typical sequences are thus a minority among all possible strings of n q-ary symbols, whose number is q^n, but the actual source output sequences almost surely belong to this minority.

Now, a *channel* can be considered as a means for observing the source event X, referred to as its input, through the outcomes $y_1, y_2, \ldots, y_r$ of another event Y, its output, which is probabilistically related to X. By 'probabilistically related', we mean that when the outcome of X is x_i, $1 \leq i \leq q$, then the probability of a particular outcome y_j of Y, $1 \leq j \leq r$, assumes a value $p(y_j|x_i)$ which depends on x_i. Such a probability is referred to as the conditional probability of y_j given x_i.

It will be convenient for us, although it is not necessary, to assume that $r \geq q$. We cannot observe directly the source event which would provide to us the quantity of information $H(X)$, but only the channel output Y which depends on the source event only through the set of conditional probabilities $p(y_j|x_i)$, $1 \leq i \leq q$, $1 \leq j \leq r$. The channel output Y then provides about the source event X a quantity of information equal to the uncertainty which is *resolved* by the outcome of X, $H(X)$, *minus* the uncertainty about X which *remains* when the outcome of Y is known, denoted by $H(X|Y)$. This quantity of information is thus expressed as

$$I(X;Y) = H(X) - H(X|Y), \tag{3}$$

referred to as the *mutual information* of X and Y, where $H(X|Y)$ is given by

$$H(X|Y) = \sum_{i=1}^{q} \sum_{j=1}^{r} p(x_i, y_j) \log p(x_i|y_j),$$

where $p(x_i, y_j) = p(y_j|x_i)p(x_i)$ is the probability of occurrence of both x_i and y_j, referred to as the joint probability of x_i and y_j, and

$$p(x_i|y_j) = \frac{p(x_i, y_j)}{p(y_j)} = \frac{p(x_i, y_j)}{\sum_{i=1}^{q} p(x_i, y_j)},$$

where $p(x_i)$ and $p(y_j)$ are the probabilities of x_i and y_j, respectively. $H(X|Y)$ is referred to as the conditional uncertainty (or conditional entropy in the usual information-theoretic vocabulary) of X given Y. The word 'mutual' expresses the easily proved symmetry of $I(X;Y)$, namely that

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = I(Y;X).$$

The mutual information $I(X;Y)$ is a nonnegative quantity, as expected for a measure of information.

The extension to sequences where the successive outcomes of X are not independent, similar to the one which led to the expression (2) of $H(X)$, is used to define $H(X|Y)$ in this case, hence $I(X;Y)$ according to (3).

Another important extension of the mutual information concerns the case of continuous variables. Let now X denote a real random variable which assumes a value belonging to the infinitesimal interval $(x, x + \mathrm{d}x)$ with probability $p(x)\mathrm{d}x$. The function $p(x)$ is referred to as the probability density function of X and is such that $\int p(x)\mathrm{d}x = 1$, where the integral is taken on the whole real axis. At variance with the case where X assumes discrete values, it is not possible to define its entropy by (1). However, if we formally replace in (1) the summation by an integration, the probability p_i being replaced by the probability density function $p(x)$ of X, we obtain the quantity

$$h(X) = -\int p(x) \log[p(x)]\mathrm{d}x \tag{4}$$

which is referred to as the *differential entropy* of X. This quantity exhibits some of the properties of the entropy of the discrete case, but not all. For instance, it may be negative and depends on the unit which measures X. However, it turns out that the mutual information of two continuous variables can still be defined according to the formula homologous to (3), namely

$$I(X; Y) = h(X) - h(X|Y), \tag{5}$$

where the proper entropies have just been replaced by the differential entropies defined by (4), with $h(X|Y)$ defined as

$$h(X|Y) = \int_x \int_y p(x, y) \log p(x|y) \mathrm{d}x \mathrm{d}y,$$

where $p(x, y) = p(y|x)p(x)$ is the joint probability density function associated with the pair of variables $\{x, y\}$, and

$$p(x|y) = \frac{p(x, y)}{p(y)} = \frac{p(x, y)}{\int p(x, y) \mathrm{d}x},$$

where $p(x)$ and $p(y)$ are the probability density functions of x and y, respectively.

The *channel capacity* $C(X; Y)$ is defined as the maximum, for all probability distributions $\{p_1, p_2, \ldots, p_q\}$ on X, of the mutual information $I(X; Y)$. Its computation is very simple in the discrete case for symmetric channels (like those given as examples) since then the maximum of the mutual information is obtained for equally probable outcomes of X, i.e., $p_i = 1/q$ for any i such that $1 \leq i \leq q$. Then

$$C(X; Y) = H(Y) - H(Y|X) \tag{6}$$

for equally probable outcomes of X. The conditional probabilities involved in the computation of $H(Y|X)$ are the transition probabilities of the channel and $H(Y)$ is easily computed in terms of the transition probabilities since $p(y_j) = \sum_{i=1}^{q} p(y_j|x_i)p(x_i) = (1/q) \sum_{i=1}^{q} p(y_j|x_i)$.

Discrete channels can be represented by diagrams where arrows indicate the transitions from the input alphabet symbols to those of the output alphabet. A transition probability is associated with each arrow. Three such diagrams are drawn in Fig. 5. The first one, **a**, is drawn for $q = 4$, the second one, **b**, for $q = 2$, and the third one pertains to the case where the alphabet symbol considered is a pair of complementary binary symbols (to be interpreted below as nucleotides), their non-complementarity being interpreted as an erasure denoted by ε. The capacity of the channels represented by Figs 5 **b** and **c** will be computed in Sec. 3.3, Eq. (11) and (12), respectively.

Thanks to the extended definition (5) of the mutual information, the capacity of a channel with discrete or continuous input and continuous output can be defined, *mutatis mutandis*. The case of discrete input and continuous output is especially important as describing the typical channel in digital communication, where a finite number of waveforms represent the alphabet symbols and where the channel perturbations are continuous, e.g., thermal noise modelled as Gaussianly distributed.

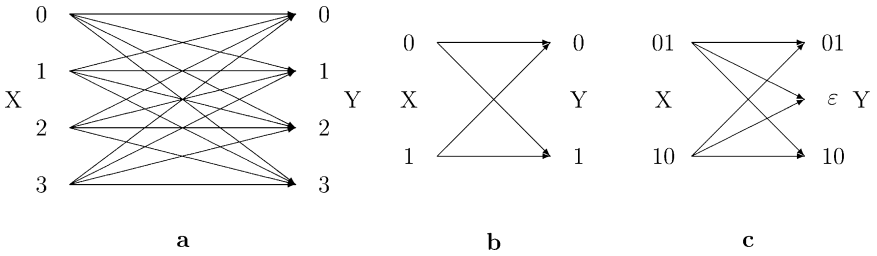

Figure 5. Example of channels. **a**: Transitions for input and output alphabets of size q. The transition probability associated with the horizontal branches equals $1-p$, where p is the errror probability, that associated with any oblique branch is $p/(q-1)$. The figure has been drawn for $q=4$. **b**: Same diagram when $q=2$. The horizontal branches are labelled $1-p$, the oblique ones p. **c**: Complementary binary symbols. The case where the two symbols of the received pair are not complementary (i.e., $Y=00$ or $Y=11$) is interpreted as a third symbol denoted by ε. If the probability of error of a single symbol is p, the probability of a transition from 01 or 10 to ε is $2p(1-p)$, the probability of transition from 01 to 10, or 10 to 01, is p^2, and that associated with the horizontal branches is $(1-p)^2$

2.2.2 Basic coding functions, source coding

The output of a source can be transformed into another fully equivalent string of symbols so as to either reduce its length (compress it), or make it resilient to errors. These encoding operations are referred to as *source coding* and *channel coding*, respectively. The two fundamental theorems of information theory state the limits of what is possible for each of these codings. The shortest possible length after source coding of a string of n q-ary symbols is nH, where H is the source entropy as defined by (1) or (2), expressed using logarithms to the base q, the alphabet size of the source, so we have $H \leq 1$. This means that it is possible to encode the output of a source of entropy H so as to reduce its length by a factor of H and still be able to recover its initial output, while the same output can not be recovered from any string of length reduced by a factor less than H. We shall not develop here the topic of source coding. Let us just notice that the entropy of a source can then be interpreted as the ratio of the length of the shortest possible compressed string which can be attained by source coding to the length of the original source output, the compressed string remaining equivalent to it in the sense that it enables an exact recovery.

As regards channel coding, errorless communication is possible if the source entropy H is less than the channel capacity C but not if it is larger than it. It turns out moreover that source coding increases the vulnerability of the message to errors and that channel coding increases the message length as introducing redundancy, so the main two coding operations appear in a duality relationship. Expressing the limits of what is possible in terms of the information measures defined above, namely, the entropy of the source and the capacity of the channel gives these information measures an operational significance. We shall in Sections 2.3.1 and 2.3.2 discuss at greater length the topic of channel coding, which involves error-correcting

codes. We shall especially insist on the paradoxical but experimentally confirmed possibility of errorless communication in the presence of channel errors.

2.2.3 Algorithmic information theory

Another way of defining information measures is provided by the *algorithmic information theory*. Given a *universal computer* like the Turing machine, historically the first theoretically conceived computer, or any modern general-purpose computer, the basic idea of the algorithmic information theory consists of defining the information measure associated with a string of symbols as the length of the shortest programme which lets the computer generate an output identical to this string (the computer input and output alphabets and that of the string are assumed to be the same, usually binary). This measure is generally referred to as the *algorithmic complexity* of the string. (The length of the programme depends to some extent on the computer, but mainly on the string itself and thus is a valid measure of its actual complexity, at least for sufficiently long strings.) The algorithmic information theory uses the algorithmic complexity as the basic measure of information. A recent book by Chaitin provides an excellent and highly readable account of this topic (Chaitin, 2005).

At first sight, the probabilistic information theory developed by Shannon and the more recent algorithmic information theory have little in common. The former deals with probabilistic ensembles of symbol strings, while the latter seems to work on individual strings. This point of view is however too simplistic. The only objective way to determine the parameters of the ensembles of symbol strings considered by the probabilistic information theory consists of deriving them from frequency measurements, which necessarily involve a number of realizations which is very small as compared with the total number of strings they contain (remember that the total number N of possible binary strings of length n is 2^n; for instance for $n = 1,000$ we have $N = 2^{1,000} \approx 10^{301}$, a number which exceeds by far the estimated number of atoms in the visible universe, namely, 10^{80}). As regards the algorithmic information theory we assumed that, given some string, a programme shorter than it can be found which lets the computer generate an output identical to this very string. Such a programme must involve its description in machine language. Any language is a combinatoric object which implicitly refers to an ensemble of realizations. Although there seems to be little relationship between the information measure associated with a source in the conventional information theory, namely its entropy (1) or (2), and the measure of complexity provided by the algorithmic information theory, it turns out that both are asymptotically related in the sense that, for very long strings, the complexity approaches the entropy. Despite their obvious difference, the conventional information theory and the algorithmic information theory interestingly appear as describing two facets of a same fundamental entity.

The main practical weakness of the complexity as defined by algorithmic information theory is its uncomputability. Given a source output, knowing its complexity would imply the availability of a means to *optimally* compress it (in the sense of source coding, see Sec. 2.2.2). However, when an algorithm is known to compress a

source (i.e., a computer programme which results in the same output as the given source output and is shorter than it), it is impossible to know if this specific algorithm is the shortest possible one. Only an upper limit on the algorithmic complexity is thus known. On the other hand, the availability of accurate probability measures is obviously needed in order to compute the quantities defined by the conventional information theory. Frequency measurements are the ordinary means for obtaining plausible probability distributions but they necessarily rely on comparatively few samples, hence have a limited precision.

2.3 On Error-correcting Codes

2.3.1 An introduction to error-correcting codes

Error-correcting codes have a long and complicated history. The most successful codes yet known, the *turbo codes*, can be considered as practically implementing 'error-free communication', the paradoxical promise of Shannon's information theory made no less than 45 years before the invention of turbo codes. Turbo codes can indeed be described in a very simple way which gives intuitive insight into the reason of their success and ignores the twists and turns which preceded their invention. We shall give below an intuitive description of turbo codes as an introduction to the needed properties of good error-correcting codes but we need first introduce the information-theoretic result that error-free communication is possible as a necessary background.

To begin with, we define as above an alphabet as a given collection of symbols in finite number, say q, referred to as the alphabet size. These symbols can be arbitrary signs or objects provided they can be unambiguously distinguished, like letters, digits, electric voltages, signal forms, or molecules… Due to the necessity that its symbols be distinguishable, the smallest size of an alphabet is $q = 2$. The main properties of codes can be understood if we assume, as we shall do most often in this section, that the alphabet is binary, i.e., that its size equals this minimum, an assumption which entails little loss of generality. The symbols of the binary alphabet will be denoted by 0 and 1.

Let us now define a *word* of length n as a sequence of n symbols from a given alphabet of size q. Since each symbol of a word can assume q distinguishable values, the total number of possible different words is q^n, say 2^n in the binary case. It will be very convenient to represent an n-symbol word as a point in an n-dimensional space, each coordinate of this point being one of the n symbols of the word. For instance, if $q = 2$ and $n = 3$, there are $2^3 = 8$ different possible words, each of which being represented as a vertex of a cube. The useful values of n are much larger, but there is no difficulty in extending this definition to an n-dimensional space with $n > 3$. We may define the *Hamming distance* d between two words as the number of coordinates where their symbols differ. For instance, if $n = 7$, the distance between 1101000 and 0110100 is $d = 4$. We refer to the space endowed with this distance measure as the n-dimensional Hamming space. An error-correcting code is a subset of all possible n-symbol words such that

the minimum distance between any two of its words is larger than 1. The minimum distance between any two different n-symbol words is only 1 since they may differ in a single coordinate, so an error-correcting code is a strict subset of all n-symbol words. The property that not any n-symbol word belongs to the error-correcting code is referred to as *redundancy*. In the case where $n = 3$, we may define a code as having even weight, the weight of a word being defined as the number of symbols '1' it contains. Here is the list of its codewords: 000, 011, 110 and 101. Its minimum distance is $d = 2$. A code with the largest possible minimum distance for $n = 3$, i.e., $d = 3$, only contains two words, for instance 000 and 111. To communicate a message of length k, with $k < n$ to ensure the code redundancy, we must establish a one-to-one correspondence, or *encoding rule*, between the 2^k messages of length k and the 2^k n-symbol words which belong to the code. Little loss of generality results if the message explicitly appears, e.g., in the first k positions in the word (or in any k determined positions). Then the encoding rule is said to be *systematic*, the symbols at the selected k positions are said *information* symbols, and the remaining ones, which are completely determined by the information symbols, are said *check* or redundancy symbols. For instance, with $n = 3$ and $k = 2$, if the 2 first symbols represent the message, we have a single check symbol which results from adding modulo 2 the information symbols (addition modulo 2 is the same as ordinary addition except that 1 plus $1 = 0$ modulo 2).

In a communication system using an error-correcting code, only words belonging to this code may be transmitted. As an example, let us assume that a binary code is used and that an error in the channel consists of changing a 1 into a 0 or vice-versa. Then the channel errors result in a received word which possibly differs from the transmitted one. Moreover, the received word is at a Hamming distance from the transmitted one which equals the number of errors which occurred, say e, to be referred to as the *weight* of the error pattern. For a binary symmetric channel, i.e., if we may characterize it as making an error with a constant probability $p < 1/2$, independently, on each symbol of the word, then the probability of a particular error pattern of weight e is simply $P_e = p^e(1-p)^{n-e}$. For $p < 1/2$, P_e is a decreasing function of e, so a given error pattern is the more probable, the smaller its weight. There is no loss of generality in assuming $p < 1/2$ since, being arbitrary, the labelling of the received symbols by '0' or '1' can always be chosen so that this inequality holds provided $p \neq 1/2$. The case $p = 1/2$ is not relevant since it is equivalent to the absence of any channel. We may thus use as the best possible rule for recovering the transmitted word: *choose the word of the code the closest to the received word*. Its use determines the word of the code which has the most probably been transmitted. This *regeneration* rule is expressed in very simple geometrical terms in the n-dimensional Hamming space thanks to the distance defined between its words.

The mere statement of this rule enables us to understand the most important properties that an error-correcting code must possess in order to be efficient. The words of a code must be far from each other, so they should be very few as compared with all possible n-symbol words, i.e., the redundancy should be large.

But they should also be as evenly distributed in the n-dimensional space as possible, since any concentration of codewords would reduce their mutual distances with respect to the case of a more even distribution. For a given amount of redundancy, endowing a code with this property is by far the most difficult task in the design of an error-correcting code, although its necessity is quite intuitive and its statement is easy. We shall see below that the best known method to provide evenly distributed points in the Hamming space actually consists of choosing them *at random*, as strange as it may look.

2.3.2 *Error-free communication over a noisy channel is theoretically possible*

It was convenient in the above examples to consider small values of the word length n. Let us now go to the other extreme and assume that n is very large. Then, the *law of large numbers* tells that the weight of an error pattern is very probably close to its average, namely np (in other words, the frequency of errors measured in a large sample is with high probability close to the error probability). In geometrical parlance, this means that the received point is with high probability close to the 'surface' (an $(n-1)$-dimensional volume) of the n-dimensional sphere of radius np centred on the transmitted word. If the radius np is smaller than half the minimum distance d between any two words of the code (simply referred to as its *minimum distance*), then clearly the received word is with high probability closer to the truly transmitted word than to any other, so the above regeneration rule succeeds with high probability. Morever, the probability of a regeneration error vanishes as n approaches infinity. On the contrary, if $np > d/2$ a wrong codeword may be closer to the received word, in which case the regeneration rule above fails with very high probability. Notice the paradox: for a given probability p of channel error, increasing the word length n also increases the average number of erroneous symbols in the received word. Nevertheless, increasing n decreases the probability of a regeneration error provided $p < d/2n$. If this inequality holds, *errorless* communication of a message through an *unreliable* channel is possible. This result itself is paradoxical, and nobody imagined it could be reached anyway before its possibility was proved by information theory. It started the researches on error-correcting codes and remained later a very strong incentive to them.

The problem of designing an optimal error-correction code having M words of length n using a q-symbol alphabet for a given channel has no known general solution. However, choosing $M = q^k$ words at random within the n-dimensional space, with $k < n$ to provide redundancy, results in a code close to the optimum. This method, referred to as *random coding*, was used by Shannon in the proof of the fundamental theorem of channel coding (Shannon, 1948). This theorem asserts that 'errorless' communication is possible if, and only if, the information rate $R = k/n$ is less than a limit which depends on the channel error probability p (decreasing as p increases), referred to as the *channel capacity* C (see Sec. 2.2.1). 'Errorless' means that, provided $R < C$, a vanishing probability of error can result from using an adequate (but not explicitly specified) code, as n approaches infinity.

The main virtue of random coding is to statistically ensure that the codewords are as evenly distributed in the Hamming space as possible. Further elaboration of this fundamental theorem led to stronger results which, loosely speaking, tell that an arbitrarily chosen code is good with high probability. In a more adamant style: *all codes are good*. The problem of almost optimum error-correction coding *seems* thus to be solved, and moreover in an unexpectedly simple way.

However, a formidable problem remains. Remember that implementing the regeneration rule above implies to find the codeword the closest to the received word. In the absence of any structure, a code is an arbitrary set of M n-symbol words. There is no other way for implementing this regeneration rule than to compare *each* of the M codewords with any *single* received (possibly erroneous) word to be regenerated. The trouble is that for useful values of the codeword length, i.e., n as large as to make the probability of a regeneration error small enough, M is a huge number. For example, in a binary code with $n = 1,000$ and $R = 1/2$, we have $M = 2^{500} \approx 10^{150}$. (Remember that the number of atoms in the visible universe is estimated to about 10^{80}.) Implementing regeneration when an arbitrary code is used thus bumps against a complexity barrier. This problem cannot actually be solved unless the code is given some structure intended to alleviate the complexity of regenerating its codewords. A large number of codes and code families having a strong mathematical structure were invented, but their results were invariably far from the promise of the fundamental theorem of channel coding, namely error-free communication at a rate close to the channel capacity. Most experts believed that finding good codes having a tractable structure was hopeless due to an intrinsic incompatibility of goodness and structure. This widely shared opinion was summarized in the folk theorem: *all codes are good, except those we can think of.*

This opinion was by far too pessimistic. For instance, I noticed in 1989 that the sole criterion used in order to design a good code was to endow it with a minimum distance *as large as possible*. I criticized this seeming dogma, and suggested that a better criterion could be to look for *random-like* codes, i.e., codes such that the distribution of distances between their words is close in some sense to that of random codes (regardless of their actual minimum distance) but constructed according to a deterministic process (Battail, 1989, 1996). (Analogously, easily generated pseudo-random sequences which mimic truly random sequences are known and widely used in simulation.) Codes designed according to this criterion should have performance close to the optimum.

2.3.3 Error-free communication can be practically approached: turbo codes

In 1993, soon after the random-like criterion was stated, the pessimistic opinion above was definitively ruined with the advent of the *turbo codes* (Berrou et al., 1993; Berrou and Glavieux, 1996). Turbo codes actually meet the random-like criterion, although they were not explicitly designed in order to fulfil it (Battail, Berrou and Glavieux, 1993). Their implementation is comparatively simple and well within the possibilities of current technology. Besides being the best codes presently available, turbo codes have a performance close enough to the theoretical

limit (the channel capacity) to be considered as almost optimal, at least from a practical point of view.

In the brief description of turbo codes to be presented now, we shall restrict ourselves to the binary alphabet, with its symbols denoted by 0 and 1, endowed with the structure of binary field, i.e., where two operations are defined: multiplication (the same as in ordinary arithmetic: $0 \times x = 0$, $1 \times x = x$, for $x = 0$ or 1) and modulo 2 addition to be denoted by $\oplus$ (the same as ordinary addition except that $1 \oplus 1 = 0$). The structure of a turbo encoder is depicted in Fig. 6. (It is astonishing that such a simple device can provide a good approximation of random coding, the implementation of which is of prohibitive complexity.)

We assumed that the necessary redundancy is obtained by generating three symbols every time an information symbol enters the encoder. In other words, the code rate R defined as the number of symbols of the information message divided by the number of actually transmitted symbols is $R = 1/3$. The choice of this particular rate will make turbo codes easy to understand, but several technical means enable generating turbo codes having different rates. For the ease of its description, the encoder is depicted here as having a single input and three outputs, but these outputs are easily transformed into a single one with a symbol rate three times larger, an operation called 'parallel-to-serial conversion'. A sequence U of N binary information symbols enters the encoder. One of the three output sequences is identical to the input sequence U. The other two output sequences, denoted by V_1 and $V_2(\Pi)$, are generated by two, possibly identical, rate-1 encoders. By 'rate-1 encoder' we mean a device which computes a binary output symbol as the sum modulo 2 of the binary symbol which enters it and of certain of the m preceding ones at well defined positions. The 'memory' m is a small number (typically in the range of 3 to 5) in order to avoid an excessive complexity of the decoding devices. Moreover, each of the rate-1 encoders is assumed to be *recursive*, i.e., its output is added modulo 2 to the entering binary symbol. The first rate-1 encoder directly receives the input information sequence U. The second one receives an *interleaved* version $\Pi(U)$ of U where the symbols of U are reordered by a device named interleaver. For example,

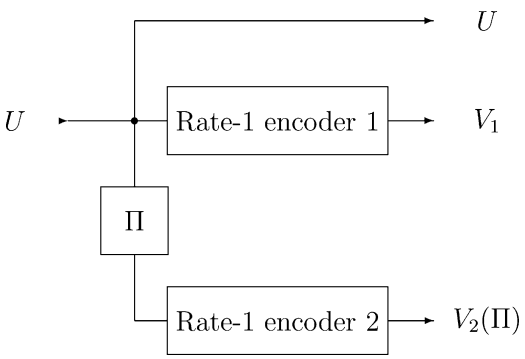

Figure 6. Rate-1/3 turbo encoder

if $N = 7$, the interleaver Π may reorder the symbols initially numbered 1234567 according to the order 3517426. Then $U = 1101000$ results in $\Pi(U) = 0010110$ and $U = 0110111$ in $\Pi(U) = 1101011$. The actually useful values of N are much larger, say 1000 or more, so the number $N! = N \times (N-1) \times (N-2) \cdots \times 1$ of possible different interleavers is huge.

The code generated by this encoder is referred to as 'linear', which means that it only implements the operations of the binary field. A linear code always contains the all-0 word and one easily checks that the set of Hamming distances between all the distinct codewords reduces to the set of its distances to the all-0 word, i.e., to the weights of its words. Due to the assumed recursivity of the rate-1 encoders, an input consisting of a single 1 would result in an output of infinite weight should no truncation occur, but it is actually limited by the size of the blocks considered, namely N symbols. We shall call 'large weight sequences' the sequences which would be of infinite weight should no truncation of the block size occur, and the other ones as 'small weight sequences'. The response of the rate-1 encoders to information messages of weight larger than 1 is not always of large weight, but for a properly designed encoder the proportion of sequences of small weight is no greater than 2^{-m}, m being the encoder memory. For a randomly chosen interleaver Π, the probability that the encoded sequences V_1 and $V_2(\Pi)$ are both of small weight is only 2^{-2m}. The total weight of a codeword so generated is the sum of the weights of U, V_1 and $V_2(\Pi)$. Since the small weight sequences generated by the rate-1 encoders are few, and since moreover the weight of most of them is not very small, we obtain that the 'weight spectrum' of the code, i.e., the set of weights which is obtained for all the 2^N possible input information sequences, closely resembles the set of weights of random sequences of length $3N$, so very few of them have little weight. The turbo encoder has thus generated a 'pseudo-random' code which satisfies the random-like criterion alluded to above.

Interestingly, the turbo encoder of Fig. 6 exhibits the three main properties which can be expected from a good encoding device. Splitting the input into several branches (three in the figure) amounts to repeat each of the entering symbols, which is the most elementary form of *redundancy*. The rate-1 encoders introduce *dependence* between the successive symbols they receive which, jointly with redundancy, is used in the decoding process. And the permutation that the interleaver operates introduces some form of *randomness* which, contrary to Shannon's random coding, is easily undone since the inverse permutation is immediately available. Not surprisingly, a cleverly designed decoding device can use all these features to closely approach the theoretical limit, and we may think of the turbo code scheme as a kind of paradigm.

2.3.4 Decoding turbo codes

Besides its good weight (distance) properties, a very interesting feature of the turbo code scheme is its comparatively easy decoding. More precisely, a reasonably good approximation of its optimal decoding has low complexity. A key concept here is that of *extrinsic information* on a symbol. Let us denote by C_1 the constraints

which tie together the symbols of U and V_1, as created by the first rate-1 encoder. Similarly, we denote by C_2 the constraints due to the second rate-1 encoder, which tie together the symbols of U and $V_2(\Pi)$. Now consider some binary symbol u which belongs to U. Due to C_1, symbols of V_1 and of U (u excepted) contain information about u, besides u itself, which is referred to as its *extrinsic information*. Since it belongs to U, the *same* information symbol u is somewhere in the sequence $\Pi(U)$ which enters the rate-1 encoder generating $V_2(\Pi)$, and we know its location since the interleaver Π is known as a part of the whole encoder. Therefore, due to C_2, symbols of $V_2(\Pi)$ and of U (u excepted) also bear extrinsic information about u.

Let us now consider the corresponding received symbols and sequences. Let us denote by u' the received symbol which corresponds to u and by U', V_1' and $V_2'(\Pi)$ the received sequences which correspond to U, V_1 and $V_2(\Pi)$, respectively, 'received' meaning that the channel noise makes the symbols erroneous with a certain probability. Due to the channel noise, the receiver does not know u with certainty but only its *a priori* probability. It is intended to evaluate the probabilities $\Pr(u = 0)$ and $\Pr(u = 1) = 1 - \Pr(u = 0)$, and to take as decoding decision about u the binary value which corresponds to the largest of these two probabilities, in terms of the sequences U', V_1' and $V_2'(\Pi)$. Besides the known *a priori* probability that $u = 0$, the receiver can reassess the probability $\Pr(u = 0)$ in terms of the extrinsic information due to C_1, using algorithms which exploit the code constraints. These algorithms are easy to implement if the memory m of the rate-1 encoders is not too large. The reassessed probability of error is less than the initial one. The receiver can also use the extrinsic information associated with C_2. The interleaving performed by Π makes the extrinsic information associated with C_1 independent from that associated with C_2, so the probability $\Pr(u = 0)$ as reassessed in terms of C_1 can be used as the *a priori* probability for reassessing the same probability in terms of C_2. Moreover, and this is the most powerful tool for decoding, this process can be *iterated*, with the newly reassessed probability in terms of C_2 being used as the *a priori* probability for a reassessment in terms of C_1, and so on. This process is repeated as many times as needed (there are criteria for stopping this iteration).

If the channel is not too bad (in more formal words, if the code rate is smaller than some threshold which is itself smaller than the channel capacity), this iterated decoding process converges to a 'hard decision' (i.e., all the reassessed symbol probabilities approach 0 or 1) which is very likely the best possible decoding decision. The precise analysis of this process is not easy, and the design of the interleaver so as to optimize the overall performance remains an open problem. However, the decoding mechanism as a whole is well understood and the performance of turbo codes so decoded is much closer to the theoretical limit, i.e., the channel capacity defined in Sec. 2.2.1, than previously obtained by the use of other codes. The capacity thus can be considered as actually reached for most practical purposes. The reader is referred to (Guizzo, 2004) for an excellent description of the turbo codes in non-technical terms and the history of their invention. The word 'turbo code' which was coined by Berrou and Glavieux to designate these codes

was actually inspired by the iterated decoding process where the result of a first decoding is used again in the next step, in a way reminiscent of a car turbo charger which uses its own exhaust to force air into the engine and boost combustion. This iteration process can be interpreted as a kind of feedback.

3. CONSERVING THE GENOME NEEDS ERROR CORRECTION

After these lengthy but necessary preliminaries, we are now able to apply concepts from information theory and error-correcting codes to genetics. To begin with, we compute the probability of error of a nucleotide as a function of time, and then the corresponding capacity as the main parameter which measures the genome ability to communicate information through time. This computation shows that the genomic capacity decreases exponentially fast down to zero due to the accumulated errors, hence that the genome is ephemeral at the time scale of geology. The faithful communication of genetic information can thus only rely on error-correcting codes.

3.1 The Genome as a Sequence of Symbols

Applying information-theoretic results to genomes implies as a first step the identification of their alphabet. The quaternary alphabet {A, T, G, C} having as symbols the DNA nucleotides may seem obviously relevant, but experimental data show that genomes, or regions of genomes, are more or less '(G+C) rich'. The (G+C) density is even assumed to have an important genetic rôle (Forsdyke, 1996), and how such an offset with respect to an equal frequency of nucleotides is conserved through time needs moreover to be explained. Using instead the binary alphabet {R, Y} which only keeps the chemical structure of the nucleotides (purine R, double-cycle molecule, i.e., A or G, or pyrimidine Y, single-cycle molecule, i.e., T or C) presumably better fits reality since the genomes are experimentally known to have the same average number of purines and pyrimidines. We shall in the sequel make all calculations with an alphabet size equal to some number denoted by q, but for the purpose of illustration we shall assume that the binary alphabet {R, Y} is considered, i.e., $q = 2$.

The integrity of a genome is mainly threatened by chemical reactants and radiations. Cellular and nucleic membranes can provide an adequate shielding against chemical agents, but not against radiations of solar and cosmic origin, or due to natural radioactivity. Moreover, the DNA molecule belongs to the quantum world according to two of its dimensions but, as a long string of nucleotides, it extends itself in the third dimension at the macroscopic scale. It can support a definite information only provided its intrinsic indeterminism as a quantum object is corrected by genomic codes.

To take into account the possibility of genomic symbol errors, let us now consider a situation where a symbol from an alphabet of size q has been chosen to bear some information but may, or not, be replaced by (or changed into, or received as) another symbol of the same alphabet[4], an event to be referred to in general as

a *transition*, or to an *error* when it results in a symbol different from the initial one. Assuming that a given symbol is randomly subjected to error with a constant probability per unit of time, ν, we shall compute in Sec. 3.2 its probability of error as a function of time. We shall then use this result to show that the corresponding channel capacity decreases exponentially fast down to zero as time tends to infinity.

The computation presented in Sec. 3.2 only concerns a single type of errors, where a symbol different from the correct one is substituted for it. Errors consisting of an erasure (a symbol is not recognized as belonging to the alphabet), a deletion (a symbol has been removed from the message) or an insertion (an extraneous symbol has been appended) may occur. Our restriction to errors by substitution is mainly motivated by the fact that this case has been much more studied by information and coding theoretists than the other ones, and that the design and implementation of error-correcting codes for deletions and insertions is significantly more difficult. Taking account of other types of error than substitutions would complicate the discussion although presumably not entailing very different conclusions. Moreover, taking account of other types of errors can but worsen the situation, except as regards the erasures which are milder than substitutions. Even if all errors consisted of erasures, an utterly optimistic assumption, the capacity would still exponentially decrease down to zero as a function of time, as shown in Sec. 3.3.3.

3.2 Symbol Error Probability as a Function of Time

Remember that, as discussed in Sec. 2.3.1, the typical information-bearing event in any communication system is the choice of a particular symbol among a given alphabet. To assess the communication performance of a genome, let us consider its nucleotides as symbols of an alphabet of size q. Let us assume that such a symbol incurs an error during the infinitesimal time interval $\mathrm{d}t$ with probability $\nu\mathrm{d}t$, where ν is a frequency provisionally assumed to be constant. We assume that an error affecting a symbol consists of replacing it by another one, chosen with probability $1/(q-1)$ among the other $q-1$ symbols of the alphabet (this is the worst probability distribution in case of an error by substitution, according to information theory).

Let $P(t)$ denote the probability that a given symbol differs from the initial (correct) one at some time $t \geq 0$. The given symbol is identical to the initial one with probability $1-P(t)$, and in this case the probability of error increases during the interval $(t, t+\mathrm{d}t)$ by an amount of $\nu\mathrm{d}t$. But if the given symbol is already in error, an event of probability $P(t)$, the probability of error *decreases* by an amount of $\nu\mathrm{d}t/(q-1)$ since the newly occurring error can recover by chance the initial symbol. We can thus express the probability $P(t+\mathrm{d}t)$ as

$$P(t+\mathrm{d}t) = P(t) + \nu\mathrm{d}t[1-P(t)] - \nu\mathrm{d}t\frac{P(t)}{q-1} = P(t) + \nu\mathrm{d}t\left[1 - \frac{q}{q-1}P(t)\right].$$

This equality is equivalent to the differential equation

$$P'(t) = \nu\left[1 - \frac{q}{q-1}P(t)\right], \tag{7}$$

where $P'(t)$ denotes the derivative of $P(t)$ with respect to time. Its solution satisfying the initial condition $P(0) = 0$ is

$$P(t) = \frac{q-1}{q}\left[1 - \exp\left(-\frac{q}{q-1}\nu t\right)\right]. \tag{8}$$

Figure 7 represents this error probability when $q = 2$.

The slope of the graph of $P(t)$ at the origin, $P'(0)$, equals ν, and $P(t)$ tends to the horizontal asymptote $P(\infty) = (q-1)/q$. This asymptotic behaviour for t approaching infinity means that after a long enough time the given symbol no longer depends on the initial one and becomes random with uniform probability over the alphabet.

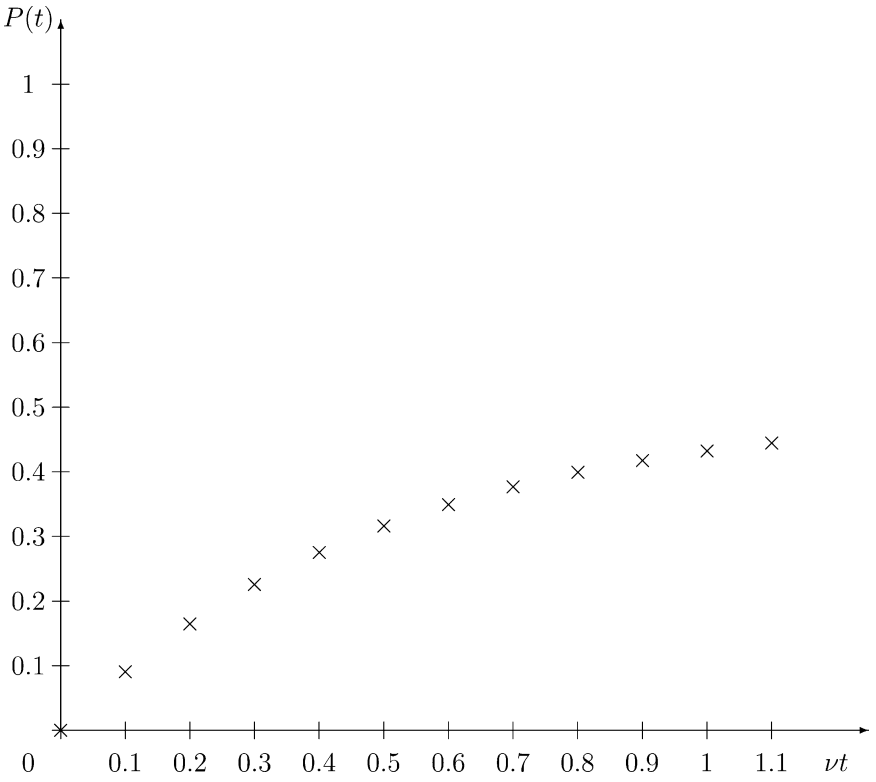

Figure 7. Probability of error as a function of time t in the presence of a constant frequency of errors ν. The alphabet was assumed to be binary

If we now consider a genome consisting of a sequence of n symbols, each of them being independently affected by errors with probability $P(t)$, the average number of erroneous symbols at time t in this genome is then $N_e(t) = nP(t)$. Replacing $P(t)$ by its expression (8) results in

$$N_e(t) = nP(t) = \frac{(q-1)n}{q}\left[1 - \exp\left(-\frac{q}{q-1}\nu t\right)\right].$$

If the sequence considered is a word of an error-correcting code of length n and minimum distance d, remember that it will be corrected with a probability the larger, the larger is n, which moreover approaches 1 as n tends to infinity, provided $N_e(t) < d/2$.

If the symbol error frequency varies as a function of time, say $\nu(t)$, one should just substitute $\nu(t)$ for ν in (7). No simple solution like (8) can then in general be obtained, but $P(t)$ remains an increasing function of time since (7) shows that its derivative $P'(t)$ is positive, and it tends to the same asymptotic value $P(\infty) = (q-1)/q$. If $\nu(t)$ remains larger than some known value ν_0, then (8) written for $\nu = \nu_0$ provides a lower bound $P_0(t)$ to the actual probability of error, and the capacity computed in terms of $P_0(t)$ is larger than the actual capacity.

3.3 Capacity Associated with Nucleotides

3.3.1 Capacity of a nucleotide in a single-strand genome

As stated in Sec. 2.2.1, information theory defines the *channel capacity* as the largest amount of information which can be communicated in the average by a symbol. It equals $\log q$ in the absence of errors (with the base of the logarithms defining the information unit), but it also accounts for the information loss due to a nonzero probability of error. Its importance results from the fundamental theorem which states that errorless communication using an (n, k) error-correcting code (each word of which is n-symbol long and bears k information symbols) can be achieved asymptotically for n approaching infinity if, and only if, the information rate k/n of the code is less than the channel capacity expressed using q-ary units.

In the case of q-ary symbols affected with probability p by errors of the type specified above, this capacity, computed using (6), reads:

$$\begin{aligned} (9) \qquad C_q &= \log_2 q + p\log_2 p + (1-p)\log_2(1-p) - p\log_2(q-1) \\ &= \log_2 q + p\log_2\left(\frac{p}{q-1}\right) + (1-p)\log_2(1-p) \end{aligned}$$

shannons per symbol (remember that we name 'shannon' the binary unit of information; see Sec. 2.2.1). The subscript q in C_q is intended to remind the alphabet

size. Letting in the above expression $p = P(t)$ as given by (8), thus assuming again that ν is constant, results in

$$(10) \qquad C_q(t) = \frac{q-1}{q}\left[1-\exp\left(-\frac{q}{q-1}\nu t\right)\right]\log_2\left[1-\exp\left(-\frac{q}{q-1}\nu t\right)\right] + \frac{1}{q}\left[1+(q-1)\exp\left(-\frac{q}{q-1}\nu t\right)\right] \log_2\left[1+(q-1)\exp\left(-\frac{q}{q-1}\nu t\right)\right]$$

which expresses the capacity of the genomic channel as a function of time in the presence of a constant error frequency ν. Notice that the error probability $P(t)$ and hence the capacity $C(t)$ depend on time in (8) and (10) through the product $\nu t = \tau$, a dimensionless quantity which can be interpreted as a measure of time using $1/\nu$ as unit. The formula (10) accounts for the degradation of the channel capacity due to the accumulation of errors. It decreases from $\log_2 q$ for $\tau = 0$, with a slope equal to $-\infty$, down to 0, *exponentially* for τ approaching infinity.

Let us assume again that the relevant alphabet is binary (say, {R, Y}, the purine/pyrimidine chemical structure of a nucleotide). The capacity given by (10) for $q = 2$, namely

$$(11) \qquad C_2(t) = \frac{1}{2}\{[1-\exp(-2\nu t)]\log_2[1-\exp(-2\nu t)] + [1+\exp(-2\nu t)]\log_2[1+\exp(-2\nu t)]\},$$

has been plotted in terms of $\tau = \nu t$ in Fig. 8 where it is represented by crosses. Similar shapes would be obtained with alphabets of other size. For a single-strand genome, a binary error-correcting code can be used in order to ensure errorless communication (asymptotically for large n) provided $C_2(t)$ remains larger than its rate k/n, hence if t is small enough.

3.3.2 Capacity of a pair of nucleotides in a double-strand genome

A pair of complementary nucleotides in double-strand DNA has however a larger capacity. If it happens that the two available nucleotides are not complementary, it is clear that one of them is in error although which is wrong is not known, so no decision about the value of one of them can be taken. This case is referred to in information theory as an 'erasure'. It is less harmful than a wrong decision since it warns that the considered symbol is unreliable. Taking account of such erasures results in an increased capacity as exploiting the informational equivalence of the complementary strands. Let us assume that the errors independently affect the nucleotides of a complementary pair and let p denote the error probability of a single nucleotide. An error occurs only when both nucleotides are simultaneously in error, an event of probability p^2, and an erasure when one of them is in error but

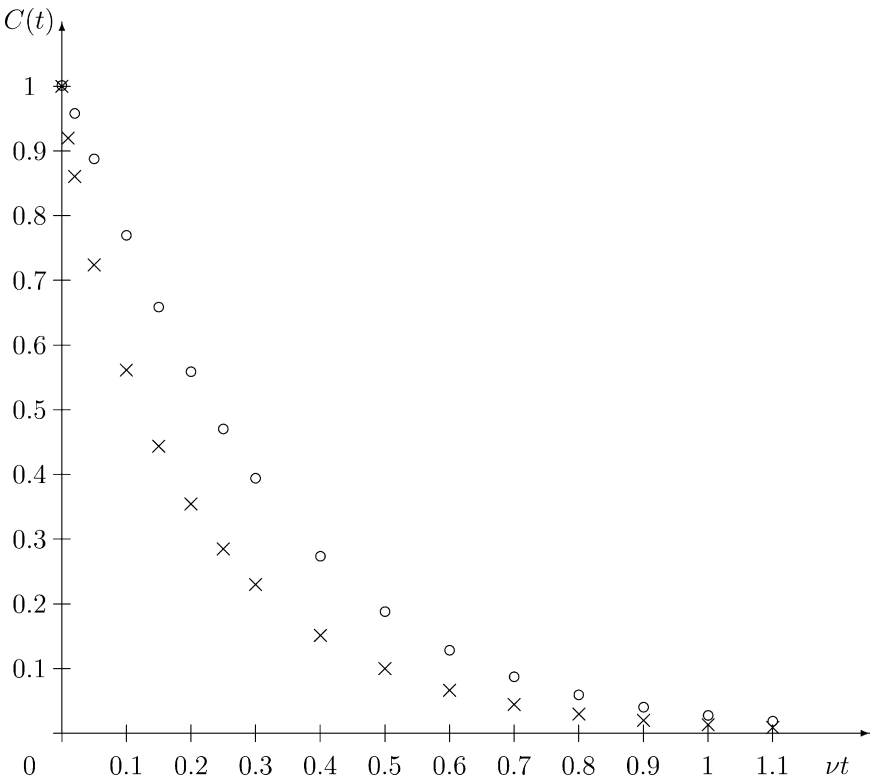

Figure 8. Genomic capacity as a function of time t in the presence of a constant frequency of errors ν, in shannons (binary units) per symbol. The alphabet is assumed to be binary. Points represented by crosses refer to a single DNA strand, while those represented by circles are plots of the capacity taking into account the availability of the two complementary DNA strands

the other one is not, which occurs with probability $2p(1-p)$. Let $C_{q,\mathrm{ds}}$ denote the capacity of this channel, where the subscript 'ds' stands for 'double strand'.

Assuming again for the purpose of illustration that the alphabet is binary, i.e., {R, Y}, the capacity $C_{2,\mathrm{ds}}$, computed using again (6), is expressed in shannons as:

$$C_{2,\mathrm{ds}} = (1-2p+2p^2)[1-\log_2(1-2p+2p^2)]+2[p^2\log_2 p+(1-p)^2\log_2(1-p)],$$

or, after replacing p with the probability of error at time t, $P(t)$, expressed in terms of νt according to (8):

$$\begin{aligned}(12)\qquad C_{2,\mathrm{ds}}(t) = \frac{1}{2}\{&[1-\exp(-2\nu t)]^2\log_2[1-\exp(-2\nu t)]\\ &+\ [1+\exp(-2\nu t)]^2\log_2[1+\exp(-2\nu t)]\\ &-\ [1+\exp(-4\nu t)]\log_2[1+\exp(-4\nu t)]\}.\end{aligned}$$

This capacity is also plotted in terms of $\tau = \nu t$ in Fig. 8 (circles). As expected, it is larger than the single-strand capacity $C_2(t)$ given by (11). The slope near the origin is -2 instead of $-\infty$. When τ approaches infinity, $C_{2,\mathrm{ds}}$ approaches $2C_2$. Taking into account the availability of two complementary DNA strands thus results in a moderate improvement of the capacity (by a factor less than, and asymptotically equal to, 2). Remember that the capacity $C_{2,\mathrm{ds}}(t)$ given by (12) measures the largest possible information rate per symbol of the genetic channel for the binary alphabet {R, Y}, and that an (n, k) error-correcting code (each word of which is n-symbol long and bears k information symbols) can provide errorless communication (asymptotically for n approaching infinity) only if $k/n < C_{2,\mathrm{ds}}(t)$.

3.3.3 Capacity in the presence of erasures only

The above capacities were computed assuming that an error affecting the genome consists of substituting a wrong symbol for the correct one. A milder kind of error would consist of simply *erasing* it, i.e., identifying it as not belonging to the alphabet. A very optimistic assumption would be that only erasures may occur. Let v denote the probability of an erasure. One easily shows that the corresponding capacity is $(1-v)\log_2 q$ shannons. The same reasoning as in Sec. 3.2, but assuming that once a symbol has been erased it remains so, shows that if the probability that an erasure occurs within the infinitesimal time interval $\mathrm{d}t$ is $\nu \mathrm{d}t$ where the frequency ν is assumed to be constant, then the probability of erasure as a function of time is simply $v(t) = 1 - \exp(-\nu t)$. The capacity as a function of time is thus, in the single-strand case:

$$C_{q,\mathrm{er}} = \exp(-\nu t)\log_2 q,$$

which is again an exponentially decreasing function. Using the double strand structure would reduce the probability of erasure to $v^2(t) = [1 - \exp(-\nu t)]^2$, i.e., that of simultaneous erasures on both strands, finally resulting in the capacity:

$$C_{q,\mathrm{er,ds}} = \exp(-\nu t)[2 - \exp(-\nu t)]\log_2 q$$

which asymptotically equals twice the single-strand capacity so it still exponentially decreases down to 0 when t approaches infinity.

3.4 How can Genomes be Conserved?

3.4.1 General consequences of the capacity calculations

The curves of Fig. 8 (or those which can be drawn when other assumptions about the alphabet size are made) clearly show that the capacity becomes negligible after a time interval close to $1/\nu$, meaning that the genomic channel is completely inefficient at the time scale of geology for plausible values[5] of ν. Means for

regenerating the genome must be available and necessarily take the form of error-correcting codes endowing the genome with the necessary property of *resilience to errors*. The genome regeneration must moreover be performed after a time interval as small as to avoid the genomic channel to degrade beyond the code correction ability (see Sec. 2.3.2).

The results plotted in Fig. 8, based on the capacity computations of Sec. 3.3, thus show that the genomes quickly (at the geological time scale, of course) bear less and less information about the original message in the absence of a regeneration process. Conservation of the genome is not the rule and error is not the exception. This implies a reversal of the onus of proof: it is the conservation of distinct genomic features which needs to be explained. We shall develop this remark below (Sec. 4.1) but we may already stress that it plainly contradicts a basic assumption of today's genetics, underlying almost all its arguments but left implicit as probably believed obvious.

3.4.2 Main and subsidiary hypotheses

That genomic error-correcting codes exist will be referred to in the sequel as our *main hypothesis*, although information theory shows it is necessary, not merely speculative. A subsidiary hypothesis must furthermore be introduced in order to fit properties of the living world as well as nature's approach. It consists of assuming that a genomic code combines several codes according to a layered architecture referred to as *nested codes*.

The assumption that genomes are words of error-correcting codes is tenable only if they are redundant, as such codes should be. In information theory, 'redundancy' does not merely mean that several copies of something are available but the far more general property that the number of symbols which are used in order to represent the information exceeds that which would be strictly necessary. Genomes are in fact extremely redundant since the number of distinct genomes of length $n = 133$, 4^{133}, approximately equals the estimated number of atoms in the visible universe, namely, 10^{80}. In the absence of redundancy, the number of possible genomes of length n would be 10^n, an inconceivably large number for n of a few millions as in bacteria, let alone for n of a few billions as in many plants and animals. Even the shortest genomes, that of viruses, are a few thousands of nucleotides long and thus still appear as very redundant.

3.4.3 Subsidiary hypothesis: nature uses 'nested codes'

Our subsidiary hypothesis is that nature uses nested codes. By 'nested codes', we mean a system which combines several codes as follows (it is convenient here to assume that the encoding rule is systematic, as defined in Sec. 2.3.1). A first information message is encoded according to some code. Then, a second information message is appended to the codeword which resulted from the first encoding, and the resulting message is encoded again using another code.

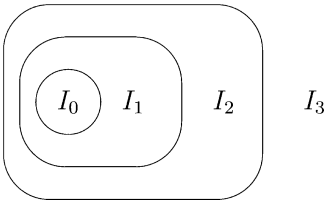

Figure 9. The fortress metaphor. A code is represented as a closed wall which protects what is inside it. I_0, I_1, I_2 and I_3 are successive information messages. I_0 is protected by 3 codes, I_1 by 2 codes, I_2 by a single code and I_3 is left uncoded

This process is repeated several times, the last information message being left uncoded. The resulting nested system is depicted in Fig. 9 using the fortress metaphor where each code is represented as a wall which encloses the information message it contains, assuming a three-layer scheme. Clearly, an information message is the better protected, the closer to the centre it is in this picture. Of course, the walls represent here purely abstract obstacles to errors seen as attackers. Notice that a very efficient protection of the most central information does not demand very efficient individual codes: the multiplicity of enclosing walls provides a much higher degree of safety than each of them separately.

Notice that we assume that the different coding layers appeared successively in time, meaning that the nested structure of ancestral forms of life has a number of layers less than that of more recent ones. The appearance of the successive coding layers may well, at least for certain of them, coincide with the 'major transitions in evolution' (Maynard Smith and Szathmáry, 1995; Barbieri, 2003). The onset of the nested structure can be understood since DNA can be replicated. If its copy is appended to the initial genome instead of being separated from it, then an increase in the genome length results and the longer genome can evolve so as to increase its error-correcting capability. Using old means to create new ones, as assumed here, is a typical feature of nature's approach often referred to as *tinkering*.

The hypothesized nested structure is plausible if we notice that certain parts of the genome like the *HOX* genes are conserved with astonishing faithfulness in many animal species, with the consequence that the organization plans of the corresponding phenotypes are very faithfully conserved. At variance with the extreme permanency of the *HOX* genes, however, it turns out that some genomic variability has advantages as witnessed by the evolutive success of sex as a means for creating new combinations of alleles. It is thus likely that genomic information is unequally protected against errors, and the nested structure is the simplest way to do so. Moreover, since we assumed that the codes appeared successively in time, the genomic information is the better protected, the older it is, so that the variability mainly concerns the most peripheral layers of the nested structure. We now undertake to draw consequences from the main and subsidiary hypotheses we made.

4. HOW GENOMIC ERROR-CORRECTING MEANS FIT IN WITH THE LIVING WORLD AND ITS EVOLUTION

We claim that our main and subsidiary hypotheses explain many features of the living world and of its evolution. We do not feel exaggerated to say that they explain very basic properties of life, including some left unexplained by today's biology for lack of relevant answers, or even for not realizing that they need an explanation. Other consequences of our hypotheses provide arguments about controversial points, especially those discussed in Sec. 4.3. We first examine the consequences of the computations presented in Sections 3.2 and 3.3, showing that the genome conservation needs frequent enough regenerations. Then we consider the consequences of our hypotheses as regards the living world and, since its present state cannot be understood without considering the way it came into existence, as regards biological evolution.

4.1 Genome Conservation and Time Between Successive Regenerations

We have seen in Sections 3.2 and 3.3 that the probability that a nucleotide is in error, hence the corresponding capacity, are functions of time: the probability of error increases, and the capacity exponentially decreases down to zero as time increases. An obvious consequence of this fact is that the genomes must act as words of an error-correcting code (our main hypothesis) and, moreover, that they must be regenerated (in the sense of Sec. 2.3.1) before the accumulated errors result in uncorrectable error patterns. As a consequence, the time interval between two regenerations must be short enough.

The minimum distance between the codewords in the system of nested codes, assuming the subsidiary hypothesis to hold, is the minimum distance d of the code which corresponds to the outmost wall of Fig. 9, hence to the species level. The time interval t_r between successive regenerations should be such that the average number of errors $nP(t_r)$ in a codeword (where n is the codeword length and $P(t)$ denotes the symbol error probability as a function of time as in Sec. 3.2) remains significantly less than half the minimum distance of the code, $d/2$, so that regeneration is almost surely successful. Similarly, but now for the best possible code of a given rate k/n, the capacity $C(t_r)$ should remain larger than this rate.

The genome conservation depends on an external parameter: the nucleotide error frequency ν, and two internal parameters. A parameter of the genome itself: its correcting ability as a code, measured by its minimum distance d, on the one hand; and a parameter of the living being which contains it: the time interval between regenerations, t_r, on the other hand. Both internal parameters may be assumed to have evolved independently, their mutual matching in the extant living world resulting from natural selection. We may assume that evolution eventually resulted in a proper adjustment of these parameters for most species, but let us consider what happens in case of a mismatch. Let us first assume that t_r is too large. Then regeneration errors are frequent so living beings in this situation belong to unstable

species, with many new forms appearing and disappearing in comparatively short time intervals. The Cambrian explosion may have resulted from a mismatch of this type. If on the contrary the time interval t_r is too small, then we have very conservative species which can remain identical to themselves during long periods but lack flexibility to evolve in the presence of a changing environment, hence risk to get extinct for lack of fitness. Of course, the actual picture is much more complicated if we take account of our subsidiary hypothesis that the error-correcting codes assume the form of nested codes. Roughly speaking, we may think that the more numerous are the code layers, i.e., the more recent is the considered genome, the higher is the global error-correcting ability of the code, meaning that more variation may be accepted and even favoured in the most peripheral layers. Therefore we may expect that more recent, better protected genomes can accept a time interval between regeneration much longer than that of earlier and less protected genomes. If we assume that the lifetime of individuals is equal to the time interval t_r (in the case of bacteria) or proportional to it (e.g., for animals), we may thus explain why it is much shorter for ancestral forms of life than for highly complex more recent beings.

We do not actually know when the regeneration process takes place. In animals with sexual reproduction one may plausibly assume that it occurs during meiosis. Then, regeneration coincides with generation and, besides being a trivial biological fact, that nature proceeds by successive generations appears as an absolute necessity in the light of information theory. But other mechanisms may be contemplated. For instance, the recent finding in *Arabidopsis thaliana* of 'non-Mendelian inheritance' (Lolle et al., 2005) could be explained by assuming that, in this species and probably in other plants, the regeneration process is not systematically performed each time the genome of a gamete is replicated, but is sporadically triggered from the outside by some kind of stress.

4.2 Discreteness and Hierarchical Taxonomy of the Living World

We now assume that evolution resulted in a proper adjustment of the two parameters d and t_r which control the genome conservation and the mutation rate. The hypothesis that genomic error-correcting codes exist immediately implies that the genomes are far from each other in terms of the Hamming distance, separated by at least the minimum distance d of the code. If we except the small individual differences due to the information left uncoded, genomes are at least at a distance d from their closest neighbours, which implies the *existence of distinct species*. In the absence of genomic error-correcting properties, the living world would appear as a collection of chimeras.

The picture becomes more complicated but more realistic when we take into account the subsidiary hypothesis of a system of nested codes. Let d_0 denote the minimum distance of the most central (ancestral) code. Geometrically, this means that the messages which pertain to the initial information I_0 can be represented by points which are initially at least d_0 apart in the Hamming space. The further

encoding by a code with minimum distances d_1 replaces each of these points by a constellation of points centred on it but the distance of the points representing the messages of I_0 becomes at least $d_0 + d_1$. After a third encoding, the points corresponding to I_0 become $d_0 + d_1 + d_2$ apart, etc. The points which correspond to I_1 are only $d_1 + d_2$ apart, those corresponding to I_2 only d_2 apart. Every time a new encoding is performed, the minimum distance between the points representing the previously encoded words is enhanced by the minimum distance of the new code.

We just described the succession of events which resulted in the construction of the nested code system. Simultaneously to the construction of this system, regeneration errors occur at random and are the more frequent, the distance between the points in the Hamming space is the lesser. But the points are the more distant in this space, they represent words which belong to the more central layers of Fig. 9. A large distance implies that the corresponding regeneration error pattern has larger weight, thus presumably gives rise to a phenotype more different from the original than an error pattern of smaller weight[6]. Clearly, besides the discreteness of species which results from the main hypothesis, the layers of Fig. 9 delineate a *hierarchical taxonomy* among them which results from the subsidiary hypothesis.

But why should the multiple layers of the nested codes appear successively in time? Appending a new error-correcting code to those already in use results in a diminished probability of error, hence in an increased permanency, so it provides an *immediate* evolutive benefit. Indeed, doing so increases both the length and the redundancy of the overall code and the increase of these parameters reduces the probability of regeneration (decoding) error. At the same time, increasing the genome length gives more room for specifying the phenotypes, which may thus be more complex, hence potentially better fitted as regards natural selection in its conventional form. Appending a new code thus both immediately improves the genome permanency and indirectly enables enhancing the phenotype fitness in subsequent evolution. The next section develops these remarks in more general terms.

4.3 Consequences of the Hypotheses as Regards Evolution

Besides the necessity of using error-correcting codes so as to ensure the faithful conservation of genomes, we see that consequences of our hypotheses, which assume that error-correcting codes are actually implemented in the process of transmitting the genomic information and moreover take the form of a system of nested codes, closely match known features of the living world. They also hint at features of the biological evolution.

4.3.1 Trend of evolution towards complexity

The subsidiary hypothesis of a nested structure is not even necessary to explain the trend of evolution towards complexity, a puzzling feature for present biological theories. Our main hypothesis alone implies the trend of evolution towards complexity as a mere consequence of the rather paradoxical information-theoretic

fact that the longer the code, the smaller can be made the error probability after decoding, even if the code rate remains constant. Hence increasing the genome size can result in increasing its permanency.

We saw in Sec. 2.3.2 above that the error-correcting codes are means for performing *reliable* communication over *unreliable* channels. Here, 'reliable' is intended to mean that the error probability can be made as small as desired, regardless of the initial error rate, by increasing the length of the codewords, subject to the necessary condition that the codes are *redundant* enough. This key property is not merely a paradoxical theoretical result, but it is fully supported by experiment as countless engineering applications of error-correcting codes were made possible by the tremendous progress of semi-conductor technology. As a striking example, mobile telephony would simply not exist without the availability of sophisticated long codes. If nature uses efficient enough codes (and we may safely assume that the Darwinian mechanisms resulted in almost optimal codes, as products of evolution having a prominent rôle in the genome conservation), then we may think that increasing the genome length results in diminishing the error rate of the genome replication, hence increasing its permanency. However, increasing the genome length while keeping the redundancy rate constant increases the quantity of information which is borne by the genome, thus giving room for specifying more complex (and, thanks to natural selection, better fitted) phenotypes. Indeed, although information theory ignores semantics, information can be thought of as a *container for semantics* (see Sec. 7 below). The availability of more information thus enables to specify more phenotypic features, so basic results of information theory explain the yet poorly understood trend of evolution towards an increased complexity.

4.3.2 Evolution proceeds by jumps

The hypothesis that the genomes behave as words of error-correcting codes, hence are distinctly far apart in the Hamming space, entails that, as resulting from regeneration errors, mutations change genomes to distinctly different ones, which implies that evolution proceeds by jumps. The view of evolution which we propose is thus clearly saltationist, giving an unequivocal answer to this rather controversial point.

4.3.3 Genetic information has a random origin

The accumulation of errors tends to make the genomic message less and less dependent on the original one. The information-theoretic quantity which measures this dependence, the *channel capacity*, has been computed as a function of time in Sec. 3.3 and plotted in Fig. 8. As time increases, it exponentially decreases down to zero. If an error-correcting code is present, the genomic message is exactly regenerated provided the correcting ability of the code is not exceeded, which occurs with high probability if the genome regeneration (decoding) is performed at short enough time intervals. The genomic message only varies when a regeneration error occurs. Such an event is very unfrequent, but it results in a burst of at least d erroneous symbols when it occurs (d denotes as above the minimum distance of the genomic code), the new genome thus becoming significantly different from

the initial one. The genomic code then ensures the conservation of this 'wrong' genome exactly as it does for the initial 'correct' one. Instead of a genome gradually becoming less and less dependent on the original genome due to the occurring errors, we obtain that it remains a long time faithfully conserved but suddenly becomes markedly different from the original when a regeneration error occurs. Next regeneration errors increase the difference in discrete steps. Continuing this process during a long enough time has thus the same ultimate consequence on the genome as if no error-correcting code is used: the original genomic message is progressively forgotten, but according to a much slower pace depending on the time interval between regenerations. Another difference is that, when an error-correcting code is employed, the genomes resulting from replication errors are conserved as efficiently as the original one was. Then each genome, whether original or affected by errors, remains identical to itself during an average time interval the average of which depends only on the probability of a decoding error. Each decoding error may be thought of as originating a separate species (excluding errors occurring in the most peripheral, uncoded layer of the nested codes scheme, which only account for differences between individuals of a same species). Another important consequence of our hypotheses is that the extant genomic information originated from replication errors since the original DNA molecule is presumably forgotten for long but, of course, these products of *chance* were strongly filtered by the *necessity* of natural selection acting on the corresponding phenotypes. Only information at the most central position in the nested codes system, hence very old and fundamental, is a possible remnant of the common origin of the extant living beings.

5. GENOMIC ERROR-CORRECTING CODES AS 'SOFT CODES'

5.1 Defining Soft Codes

It would be naïve to expect that the error-correcting codes that nature uses closely resemble those that engineers design. The latter are defined as a set of words which obey constraints expressed by deterministic mathematical equalities. Looking for error-correcting codes of natural origin, we were led to the concept of 'soft code', where the constraints may be expressed as inequalities or forbidding rules as well as mathematical equalities, and may be probabilistic as well as deterministic. Having thus extended the concept of error-correcting codes, we may think of the many mechanical, steric and chemical constraints obeyed by the DNA molecule, or the protein for which it 'codes', as inducing soft codes. Even linguistic constraints may be considered since in a sense the genome describes the construction of a phenotype, which needs some kind of language.

5.2 Potential Genomic Soft Codes

We gave elsewhere a rather comprehensive list of the potential genomic soft codes which result from the several constraints which the genome obeys (Battail, 2005).

For the sake of completeness we list here more briefly these soft codes and then give more emphasis on comments.

A first kind of potential soft codes are associated with structural constraints of DNA. As a sequence of nucleotides, a DNA molecule is clearly subjected to mechanical and chemical constraints due to its spatial structure, its bonding with proteins like histones and, in eukaryotes, its packing in nucleosomes and higher-order structures. Researches in this direction even suggested more precisely that the DNA molecule as packed in chromatin can be interpreted as a kind of 'soft turbo code', in both prokaryotes and eukaryotes (Carlach, 2005). Genomes (especially the human one) often contain very short sequences (typically 3-base long) which are repeated thousands or even millions of times. Such sequences bear almost no information. Such 'junk' DNA may however play a rôle in an error-correction system as separating along the DNA strand more informative sequences which, due to the 3-dimentional structure of the DNA molecule, may be spatially close to each other and share mechanical or chemical constraints (a function which loosely resembles that of interleaving used in engineering). Interestingly, interleaving has an important function in turbo codes, as described in Sec. 2.3.3. That this 'separator' conserves its structure of a short motif repeated many times hints at a function which needs to be maintained, in contradiction with the word 'junk' used to qualify it. Similarly, the conservation of the (G+C) density at a value different from the average 1/2 which would be expected from pure randomness, must be explained as resulting from some kind of error-correcting means.

In the portions of the genome which specify proteins, i.e., in genes in a restricted sense, the sequence of codons (triplets of nucleotides) is furthermore constrained as are the proteins themselves: the structural constraints of proteins induce soft codes on the sequence of codons which correspond to the amino-acids according to the 'genetic code'[7]. Physiologically active proteins are not fully characterized by the sequence of amino-acids (the polypeptidic chain) that the sequence of gene codons specifies. They are made of a number of 3-dimensional substructures (α helices, β sheets, which are themselves included into higher order structures named 'domains') which impose strong constraints of steric and chemical character on proteins. Moreover, proteins owe their functional properties to their folding according to a unique pattern, which implies many chemical bonds (especially disulphur bridges) between amino-acids which are separated along the polypeptidic chain but close together in the 3-dimensional space when the protein is properly folded. The sequence of amino-acids is thus subjected to many constraints, which in turn affect the codons through the inverse 'genetic code'. Due to the universal rôle of DNA for specifying proteins, such constraints must be present in any living being.

At a high level of generality, we mentioned above that soft codes may be induced by linguistic constraints, too. We already noticed that the message which is needed for unambiguously identifying a biological species and even an individual inside it is very much shorter than the actual genomes, even the shortest ones like those of viruses (see Sec. 3.4.2). Genomes are thus highly redundant, a necessary condition for them to possess significant error-correcting properties. From another

point of view, this redundancy has rather obvious reasons: the genome does not merely identify a living being. Modern biology interprets it as a *blueprint* for its construction. The genome of any living being needs actually contain the *recipe* for its development and its maintenance. Besides those parts of the genome which direct the synthesis of proteins, i.e., the genes in a restricted sense, and the associated regulatory sequences which switch on or off their expression (i.e., make the gene direct or not the synthesis of the protein it specifies), the genome must somehow *describe* the succession of operations which results in the development and the maintenance of a living thing. This demands some kind of *language*, and a language involves many morphological and syntactic constraints which may be interpreted as generating redundant soft codes having error-correcting capabilities. Moreover, the linguistic constraints appear at several different levels, so a written text assumes the structure of nested soft codes which we were led to hypothesize for the genetic message. In order to illustrate the error-correcting ability of languages, notice that we can correctly perceive the spoken language in extremely noisy acoustic surroundings like vehicules or crowded places. By 'correctly perceive', we do not mean to grasp the meaning, which concerns the semantic level, but simply recover without error the uttered speach as a sequence of phonemes. It turns out that the individual phonemes are identified with a large error probability, but the linguistic constraints together with the high processing power of the human brain eventually result in errorless communication in the presence of noise. We can say that the daily experience of a conversation experimentally proves the ability of the human language, as a highly redundant soft code, to behave like good error-correcting codes designed by engineers.

The number and variety of constraints indeed suggest that many potential genomic error-correcting mechanisms actually exist, which moreover are organised as nested soft codes. The resulting system of nested soft codes closely resembles Barbieri's organic codes (Barbieri, 2003), although it is merely intended to cope with the necessity of protecting the DNA molecule against radiations and quantum indeterminism which no phenotypic shielding can ensure. Barbieri's concept of organic codes, on the other hand, does not refer to the necessity of error correction but results from a deep reflection on biological facts. He considers as an organic code the correspondence which exists between unidimensional[8] strings of completely different molecules (as a famous example, think of the relationship between triplets of nucleotides and the 20 amino-acids which make up proteins, referred to as the 'genetic code'). This correspondence does not result from any physical or chemical law, but can be considered as a pure convention or artifact, just like conventional rules are found in linguistic or engineering. Such rules are maintained thanks to 'semantic feedback loops' (Battail, 2005). According to our point of view, the specific constraints of each of the strings which are tied together by a correspondence rule act as soft codes with error-correcting ability. Barbieri's organic codes actually assume the structure of nested codes. Especially significant in this respect is Fig. 8.2 in (Barbieri, 2003), p. 233, to be compared with Fig. 9 above which uses the fortress metaphor to illustrate the concept of nested codes. This rather unexpected

convergence provides a mutual confirmation of both points of view, which appear as complementary. This may also be thought of as an illustration of 'tinkering' as a typical method of nature, where some biological objects are used to perform functions completely different from those they initially performed. However, since many biological problems take the chicken-and-egg form, a word like 'multivalence' could be more appropriate than 'tinkering' (although less picturesque) in the absence of a clear knowledge of the chronology.

5.3 Some Further Comments about Genomic Soft Codes

Soft codes do not exactly fit the properties of error-correcting codes which were described in Sec. 2.3.1. Since their definition involves probabilistic constraints and constraints expressed as inequalities, the mutual distances between their words become random, and especially the minimum distance d which accounts to a large extent for the performance of a code. On the other hand, when discussing in Sec. 3.4.2 the consequences of our hypotheses we assumed that the properties of genomic error-correcting codes were those of conventional codes. This may be thought of as a simplifying assumption. One may moreover argue that, if the soft codes combined into the nested scheme are numerous enough, and if moreover their words are long enough, then the law of large number results in a small-variance minimum distance which can rightfully be approximated by a deterministic quantity.

Let us also notice that the soft code concept implies that the biological constraints are also those which enable error correction, at variance with the uncoded case but also with that of hypothetic codes obeying purely mathematical constraints. This may mean that the genomes which are admissible as words of a genomic error-correcting code also specify viable phenotypes. If this is true, decoding (regeneration) errors produce viable, possibly hopeful, monsters. This remark makes rather plausible the explanation of the Cambrian explosion which we suggested in Sec. 4.1.

6. IDENTIFICATION OF GENOMIC ERROR CORRECTION MEANS

6.1 Indirect Evidence of Genomic Error Correction Means

6.1.1 Spectral and correlation properties of genomes

The experimental analysis of DNA sequences has shown they exhibit long-range dependence. First of all, their power spectral density has been found to behave as $1/f^{\beta}$, asymptotically for small f, where f denotes the frequency and β is a constant which depends on the species: roughly speaking, β is the smaller, the higher the species is in the scale of evolution; it is very close to 1 for bacteria and significantly less for animals and plants (Voss, 1992).

Another study of DNA sequences first restricted the quaternary alphabet of nucleic bases {A, T, G, C} to the binary one {R,Y} by retaining only their chemical structure, purine or pyrimidine (as we did above, too). An appropriate wavelet

transform was used to cancel the trend and its first derivative. Then the autocorrelation function of the binary string thus obtained has been shown to decrease according to a power law (Audit et al., 2002). This implies long-range dependence at variance with, e.g., Markovian processes which exhibit an exponential decrease. Moreover, in eukaryotic DNA the long-range dependence thus demonstrated has been shown to depend on structural constraints (Audit et al., 2002). The double-strand DNA is actually wrapped around histone molecules acting as a spool (making up together a 'nucleosome'), which implies bending constraints along the two turns or so of the DNA sequence in each nucleosome.

The $1/f^{\beta}$ behaviour of the spectrum and the long-range dependence of the DNA sequence restricted to the purine/pyrimidine alphabet are of course compatible with each other. Moreover, they both denote (at least if further conditions are fulfilled) the existence of a fractal structure, meaning that the DNA sequence is in some sense self-similar. In other words, a basic motif is more or less faithfully repeated at any observation scale. We may therefore think of the message borne by the DNA strand as resulting from 'multiple unfaithful repetition' which could in principle enable the use of many low-reliability replicas of the basic motif symbols in order to get reliable decisions for the purpose of regeneration. This implies a very large redundancy, indeed an obvious property of the DNA message which we already noticed. The existence of such a regeneration process, possibly approximated by majority voting, is as yet a conjecture. It remains to determine whether, and how, nature implements regeneration based on long-range dependence at some stage of the DNA replication process (Battail, 2003). Moreover, the long-range dependence is compatible with the turbo code structure which has been conjectured to exist in genomes (Carlach, 2005).

6.1.2 Distance properties of eukaryotic genes under evolutive pressure

Forsdyke formulated in 1981 (Forsdyke, 1981) the interesting idea that in eukaryotic genes the introns are made of check symbols associated with the information message borne by the exons so as to make up words of a code in systematic form (as defined in Sec. 2.3.1). The literature generally states that introns are more variable than exons. A counterexample was however provided in 1995, again by Forsdyke, who experimentally found that the exons are more variable than the introns in genes which 'code' for snake venoms (Forsdyke, 1995).

It turns out that both the generally observed greater variability of introns and Forsdyke's counterexample can be explained by the assumption that the system of exons and introns actually acts as an error-correcting code in systematic form where the exons constitute the information message (which directs the synthesis of a protein) and the introns are made of the associated check symbols. Interpreted as a regeneration error, a mutation occurs in favour of a codeword at a distance from the original word equal to the minimum distance of the code or slightly larger. If the exons 'code' for a protein of physiological importance, which is by far the most usual case, the evolutive pressure tends to the conservation of this protein so the regeneration errors are mostly located in introns. If however the evolutive pressure

tends to make the protein highly variable, as in the arms race of snakes and rodents, then the regeneration errors will be mostly located in exons and the introns will be conserved (Battail, 2004). Strictly speaking, this does not prove that exons and introns together constitute a codeword in systematic form. At least, we can say that the experimental evidence does not disprove this statement.

6.2 Lack of Direct Identification of Genomic Codes

Error-correction means are necessary for counteracting the lack of permanency of the genome pointed out in Sec. 3.3. We showed moreover in Sec. 3.4.2 that assuming their existence enables to derive a number of properties which the living world actually possesses, some of them being so familiar and general that biologists did not even try to explain them. We just mentioned above indirect experimental evidence of this existence. The direct identification of genomic error-correcting codes would be highly desirable as an experimental proof of their existence, but it is still lacking. A necessary condition for identifying these codes is of course that geneticists look for them, which implies their active involvement. Moreover, succeeding in this task needs more than superficial knowledge and understanding of error-correcting codes and information theory (Battail, 2006).

6.3 Identifying the Regeneration Means: an Open Problem

The problem of genomic regeneration (decoding) is left for future researches. The principle of the regeneration can be stated: the genome replication process aims at creating a new genome, hence subjected to all the constraints that a genome should obey. On the other hand, it should replicate the old genome which presumably suffered errors. These conflicting requirements must be solved in favour of the *constraints*. Since we used constraints of biological origin to define the genomic codes, obeying these constraints amounts to correct errors. We may thus think of the replication process as necessarily performing regeneration by *providing the approximate copy of the old genome which best fits the genomic constraints.* Replacing 'old genome' by 'received codeword' in the above statement just describes the engineering function of regeneration, as defined in Sec. 2.1.4. An intriguing feature of regeneration as implementing this rule is that its operation demands that the regenerator (decoder) possesses a full description of the constraints at any level, whether they are of linguistic character or originate in physico-chemical properties of molecular strings. This is again a chicken-and-egg problem, and it is impossible (and maybe meaningless) to know what came first: the description of constraints or the onset of molecular strings with their physico-chemical properties.

As regards the implementation of regeneration, it must be stressed that the full knowledge of a code does not *ipso facto* entail that adequate means for its decoding are known. Moreover, there exist several decoding processes for a given code which more or less approximately implement the optimum rule stated in Sec. 2.1.4

which are generally the more complex, the closer to optimality. Still more than the identification of genomic error-correcting codes, that of the means actually implemented by nature for their decoding (i.e., genome regeneration) is thus difficult and challenging. Remember that we used above the human language as an example to illustrate the error-correcting properties of certain soft codes: the means implemented in the brain for this task are presumably very complex and still unknown. Also, it is possible that existing mechanisms believed to perform 'proof-reading' actually implement some kind of genome regeneration. (Incidentally, proof-reading can only check that the copy is faithful to the original, hence correct errors intrinsic to the replication process. It is of no use if the original itself suffered errors.)

A rather fruitful and general interpretation of decoding consists of noticing that each symbol in an encoded message is represented in two ways. Obviously, on the one hand, by itself; but on the other hand, less obviously, by other symbols due to the constraints which tie together the encoded symbols and provide 'extrinsic information' about the considered symbol as introduced in Sec. 2.3.4. In good codes, the constraints enable to compute each symbol in terms of many other ones, according to many different combinations which provide *replicas* of it (Battail et al., 1979; Battail, 2003). In the presence of errors, an erroneous symbol can be corrected by majority voting or possibly by an improved decision rule where the replicas are weighted in terms of their *a priori* probability of error, and moreover taking into account the possible presence of a same symbol in several replicas. An erroneous symbol is then corrected with high probability if the error rate is as small as to let few replicas be wrong. However, a single erroneous symbol combined with others to compute a replica suffices to make wrong this replica, so the correction process becomes inefficient if the error rate increases too much.

7. ON THE EPISTEMOLOGICAL STATUS OF INFORMATION AND ITS RELATION TO SEMANTICS

7.1 The Word and the Concept of Information

The purpose of Shannon was to develop *a theory of communication*, as stated in the title of his seminal papers (Shannon, 1948). Soon after their publication, this theory has however been referred to as *information theory*, and some regret that the word 'information' has been used to designate it. I think that it was necessary to name what is communicated or, more precisely, what can be quantitatively measured in what is communicated. Despite its vagueness and polysemy, the word 'information' is quite acceptable to this end provided one reminds that, once a word belongs to the vocabulary of science, its meaning becomes precise at the expense of more or less severe semantic restrictions (Servien, 1931). It turns out that, in its restricted scientific meaning, information is an entity of its own.

Information theory uses easily understood and well defined information measures, but does not define the very concept of information. I try now to outline such a definition, consistent of course with information theory. I even attempt to clarify

the relationship of information and semantics. This subject is still debated. While the review of information theory given in Sec. 2 expresses a consensus among information theoretists, I give here my rather naïve personal point of view as a mere contribution to the debate about the epistemological status of information and its relation to semantics. These personal views have been moreover strengthened and made more precise with the help of the algorithmic information theory, especially as discussed in (Chaitin, 2005). This book presents in nontechnical terms powerful arguments sheding light on the very meaning and limits of scientific explanations, which necessarily rely on a finite quantity of information. (A very short outline of the algorithmic information theory has been given in Sec. 2.2.3 above.)

7.2 An Information as an Equivalence Class Needing a Physical Inscription

Let us consider first an arbitrary string of symbols from some alphabet. We may transform it by channel coding into a longer string which will resist errors within certain limits, or by source coding so as to shorten it. In both cases, the initial string can be exactly recovered since the encoding and decoding operations are assumed to be stricly reversible. Since the size of the output alphabet of an encoder can differ from the input one, the encoding operations can also perform alphabet conversions. Moreover, the physical support of the encoded messages is itself arbitrary. What is needed is only that a number of states of some physical system equal to the alphabet size can be unambiguously distinguished, regardless of the phenomena which are involved. We may thus think of *an* information as the equivalence class of all the messages which can be transformed into one another by any source and channel coding operation, using any alphabet and borne by any physical support.

Defining an information as an equivalence class is rather abstract and I disagree with the statement made by Landauer in (Landauer, 1991) and many subsequent papers that 'information is physical'. I consider however, without contradiction, that recording or communicating an information necessarily involves some physical support bearing one of the members of the equivalence class to which it belongs, as a string of symbols of some alphabet. I thus believe that information has no objective existence without a physical support, at variance with a more or less implicit idealistic point of view. I believe that even the information recorded and dealt with in the human mind has a physical support in neuronal devices and processes inside the brain, although they remain still largely unknown. The embodiment of information into some support, which is necessary for its recording, communication and processing, will be referred to below as its *physical inscription*.

7.3 Information as a Container for Semantics

Defining an information as an equivalence class, we have been able to avoid any reference to semantics, in accordance with the total separation between information

and semantics that information theory assumed since its beginning. That information theory has been fully developed without having recourse to any notion of semantics *a posteriori* legitimates the initial separation, and any reflection about the relationship of information and semantics must account for it.

Having identified *an* information with an equivalence class, semantics results from associating some meaning with some information according to some convention. As an example, assume that I want to identify a plant. To know the species to which it belongs, I shall use a plant guide which contains a series of questions about the shape of the leaves, the number of petals of the flowers, and other distinctive features. These questions are dichotomic, i.e., can be answered by yes or no. To each binary information associated with an answer corresponds some semantic content which results in differentiating plant species. Clearly, the more numerous the questions, the larger the set of different species which can be distinguished. If the questions are independent, answering n questions enables distinguishing 2^n different species. A set of n dichotomic questions bears an information measured by at most n binary units or shannons (see Sec. 2.2.1) which means in a sense that more information can contain 'more' semantics. Of course a quantitative measure of semantics is meaningless in general, but this example shows that semantics can be associated with information by conventions and moreover that more information enables to make finer semantic distinctions. In a computer context as in (Chaitin, 2005), we may similarly consider that each symbol of a binary programme asks the computer to make one operation among two possible ones, which is equivalent to answer a question by yes or no. The shortest possible programme for a given computer output (the length of which defines its algorithmic information measure, see Sec. 2.2.3) then answers independent dichotomic questions. Information appears here as a *container for semantics*, providing a bridge between the physical world through its necessary physical inscription, and the world of abstraction through the conventions which associate meaning and information.

That information acts as a container for semantics is maybe best illustrated by a counterexample. It dates back to 1670 and its author is Molière. In his play *Le Bourgeois gentilhomme*, Monsieur Jourdain is a bourgeois infatuated with nobility. His young neighbour Cléonte is in love with his daughter, but cannot expect to marry her unless Monsieur Jourdain believes he is noble. He thus disguises himself as the son of the Turkish sultan. Not only he wears Turkish attire, but he uses a forged, allegedly Turkish, language that his servant Covielle, similarly disguised, is assumed to translate into French. Here is a very short excerpt of the play:

Cléonte (in Turkish attire, playing the sultan's son and speaking to Monsieur Jourdain)
— *Bel-men.*
Covielle (Cléonte's servant, playing a Turkish interpreter)
— *Il dit que vous alliez vite avec lui vous préparer pour la cérémonie, afin de voir ensuite votre fille, et de conclure le mariage.* (He says that you should go fast with him to prepare yourself for the ceremony, in order to later see your daughter, and to celebrate the marriage.)

The audience intuitively perceives that something is wrong and laughs: indeed, so many semantic instances cannot be contained in a message bearing so little information.

8. CONCLUSION

Probably the most important contribution of this paper is the statement that the paradigm of genome replication by copying a template is wrong and should be replaced by that of genome regeneration based on its intrinsic error-correcting properties. For cultural reasons, biologists and communication engineers have little interaction, although both communities abundantly use the word 'information' (but rarely agree on the *concept* of information). Wrongly believing that conservation of the genome is the rule and error, the exception, biologists consider natural selection as the sole factor which shapes the living world. For instance, the book *Mendel's demon* by Mark Ridley is entirely devoted to the problem of the genome conservation and evolution but does not contemplate other mechanisms than template-replication and natural selection (Ridley Mark, 2000). The possibility of any intrinsic error-correcting ability of genomes is not alluded to, and the book contains no reference at all to the literature on error-correcting codes. Similarly, Lolle et al. simply look for RNA templates in order to explain their observations (Lolle et al., 2005), although RNA is notoriously less stable than DNA.

Biologists actually have much to gain in learning information theory (Yockey, 2005; Battail, 2006). The deep change of point of view it would provide can generate new means for understanding the living world and prompt a vast amount of researches in unexpected directions. That its basic paradigm is refuted by information theory shows that a complete renewal of genetics is needed, a wide and demanding task. Although we established above that the existence of genomic error-correcting codes is an unavoidable consequence of the faithful conservation of genomes, almost nothing is known as yet as regards the error-correcting means that nature actually implements. Even though some plausible assumptions can be made about the constraints which provide the genomic error-correction ability, how they are actually used to the benefit of the genome conservation remains entirely to be discovered. The multiplicity of codes combined into the hypothesized nested structure hints at a variety of encoding constraints and regeneration processes. It is thus a very wide field of research that information theory opens to genetics and evolution theory. Besides biology itself, communication engineering should in turn learn much from the eventual understanding of these mechanisms, just like aeronautic engineering learned much from knowing how birds and insects fly. A collaboration between communication engineers and biologists should thus be highly beneficial for both communities. Let us wish that the present paper can help them in setting up such a fruitful interaction.

NOTES

[1] One may even think of the various biological constraints as by-products of the necessary error-correcting means.
[2] We shall in the following denote by q the number of symbols of the alphabet, and they will be referred to as q-ary. Most examples will use the binary alphabet, i.e., $q = 2$.
[3] 'Stationary' means that the probabilities involved in the source operation do not vary with time.
[4] Or an alphabet of larger size which contains all the symbols of the initial alphabet plus some others. An example of this case is provided in Sec. 2.2.1 where an erasure is represented as a third symbol appended to the binary alphabet.
[5] The lack of a reliable estimate of the error frequency ν unfortunately forbids to quantitatively exploit these results.
[6] We assume here that the more different are genomes, the more different are the corresponding phenotypes. A kind of isomorphism between the genomes and the phenotypes is thus assumed although it can only be approximative. The same assumption legitimates the use of the Hamming distance for reconstructing phyletic trees, a current biological practice.
[7] We use quotes, here and in the sequel, in order to express that it is not truly a code in the information-theoretic sense, but rather a mapping in the mathematical vocabulary.
[8] Unidimensionality is a common feature of messages in engineering and in biology. It appears as necessary for unambiguous semantic communication.

REFERENCES

Audit B, Vaillant C, Arneodo A, d'Aubenton-Carafa Y, Thermes C (2002) Long-range correlation between DNA bending sites: relation to the structure and dynamics of nucleosomes. J. Mol. Biol. 316: 903–918

Barbieri M (2003) The organic codes. Cambridge University Press Cambridge, UK

Battail G, Decouvelaere M, Godlewski P (1979) Replication decoding. IEEE Trans. Inf. Th. IT-25(3): 332–345, May 1979

Battail G (1989) Construction explicite de bons codes longs. Annales Télécommunic. 44(7–8): 392–404, July-August 1989

Battail G, Berrou C, Glavieux A (1993) Pseudo-random recursive convolutional coding for near-capacity performance. Proc. GLOBECOM'93, Communication Theory Mini-Conference, Houston, USA, Nov. 29-Dec. 2, 4: 23–27

Battail G (1996) On random-like codes. Information Theory and Applications II J.-Y. In: Chouinard P. Fortier Gulliver TA (eds) Lecture Notes in Computer Science No. 1133, pp 76–94, Springer

Battail G (1997) Does information theory explain biological evolution? Europhysics Letters 40(3) Nov. 1st: 343–348

Battail G (2001) Is biological evolution relevant to information theory and coding? Proc. ISCTA'01, Ambleside, UK, 343–351 July 2001

Battail G (2003) Replication decoding revisited. Proc. Information Theory Worshop 03, Paris, France, Mar.–Apr. 2003, 1–5

Battail G (2004) An engineer's view on genetic information and biological evolution. Biosystems 76: 279–290

Battail G (2004) Can we explain the faithful communication of genetic information? DIMACS working group on theoretical advances in information recording: 22–24 March 2004

Battail G (2005) Genetics as a communication process involving error-correcting codes. Journal of Biosemiotics 1(1): 103–144

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

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

Berrou C, Glavieux A (1996) Near optimum error correcting coding and decoding: turbo codes. IEEE Trans. on Communications 44: 1261–1271, Oct. 1996

Carlach J-C (2005) Comparison of structures and error-detecting/correcting properties between DNA molecules and turbo codes, private communication

Chaitin G (2005) Metamath! Pantheon Books, New York

Cover TM, Thomas JA (1991) Elements of information theory. Wiley, New York

Forsdyke DR (1996) Different biological species 'brodcast' their DNA at different (G+C)% "wavelengths". J. Theor. Biol. 178: 405–417.

Forsdyke DR (1981) Are introns in-series error-detecting sequences? J. Theor. Biol. 93: 861–866

Forsdyke DR (1995) Conservation of stem-loop potential in introns of snake venom phospholipase A_2 genes. An application of FORS-D analysis. Mol. Biol. and Evol. 12: 1157–1165

Gallager RG (1968) Information theory and reliable communication. Wiley, New York

Guizzo Erico (2004) Closing in on the perfect code. IEEE Spectrum INT-41(3): 28–34, March 2004

Landauer R (1991) Information is physical. Physics Today. May 1991, 23–29

Lolle SJ, Victor JL, Young JM, Pruitt RE (2005) Genome-wide non-mendelian inheritance of extra-genomic information in *Arabidopsis*. Nature 434(7032): 505–509, March 24, 2005

Maynard Smith J, Szathmáry E (1995) The major transitions in evolution. Oxford University Press, Oxford, UK

Ridley Mark (2000) Mendel's demon: [gene justice and the complexity of life]. Weidenfeld & Nicholson, London

Servien Pius (1931) Le langage des sciences. Payot, Paris

Shannon CE (1948) A mathematical theory of communication. BSTJ 27: 379–457, 623–656, July and October 1948. These papers have been reprinted, with comments by Weaver W as a book entitled The mathematical theory of communication. University of Illinois Press, Chicago 1949

Voss RF (1992) Evolution of long-range fractal correlation and $1/f$ noise in DNA base sequences. Phys. Rev. Lett. 68: 3805–3808, June 1992

Yockey HP (2005) Information theory, evolution, and the origin of life. Cambridge University Press, New York