Can we explain the faithful communication of genetic information?

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

Abstract

Genomes are conserved with astonishing faithfulness through geological ages. To explain this fact, we made the hypothesis that the genetic message borne by DNA involves powerful error-correction means, consisting of 'nested codes' which provide a stronger protection to the most basic (and oldest) parts of the genomic message. It turns out that, assuming very long codes and performance close to that promised by the channel coding theorem of information theory, many basic features of the living world and of biological evolution can be explained: the existence of discrete species, the possibility of their taxonomy, and the trend of evolution towards complexity. Moreover, answers to controversial points are proposed, e.g., that evolution proceeds by jumps, which suggests a nonDarwinian mechanism for the origin of species. We assume that the hypothesized error-correction means use 'soft codes', a soft code being defined as a set of constraints which link symbols together. Not only deterministic constraints expressed by operations in mathematical structures like finite fields are considered, but also probabilistic constraints, or rules stating incompatibilities. We only require that these constraints modify the conditional probability of any symbol given subsets of other ones. Then, decoding (or, more precisely, regeneration) consists of reassessing the symbol probabilities so as to take account of all their mutual constraints. We examine how soft codes can arise from mechanical, geometrical and chemical constraints obeyed by the DNA molecule as well as, inside a gene, by the protein that DNA specifies by means of the 'genetic code'. Moreover, the genome being a 'blueprint' for the development and maintenance of a living thing, it must use some kind of language, implying morphologic and syntactic constraints which can be interpreted as soft codes. Although the paper is mainly speculative, we show that experimental works agree with the hypotheses made. Some decoding (regeneration) principles are outlined but the detailed mechanisms for their implementation, which involve chemical means, remain to be discovered.

1 Introduction

The support of genetic information is the deoxiribonucleic acid (DNA), a long unidimensional polymer bearing nucleic bases, which are small molecules of only four different types, denoted A (adenine), T (thymine), G (guanine) and C (cytosine). Each nucleic

base acts as a symbol of the quaternary alphabet {A,T,G,C}, and the genetic message consists of a sequence of such molecules. DNA is thus a *molecular* memory, at variance with *all* man-made memory devices where the bearers of the alphabet symbols have a much larger physical size (although the progress of technologies makes it decrease as the years pass, as exemplified by [1, 2, 3]). The genetic message could thus be expected to be especially fragile as subject to quantum effects, thermal noise and radiations of various origins. In sharp constrast, however, its longevity widely outperforms that of man-made memories since the genome of extant living things faithfully bears information dating back to hundreds of million years and even more. As a striking example, *HOX* genes which determine features of a living being as fundamental as its organization plan are common to a large number of animal species, including humans and flies.

Only the hypothesis that the mechanisms of DNA replication involve powerful error-correcting codes which make it resilient to errors can conciliate the seeming weakness of DNA and the astonishing longevity of the message it bears. Being products of evolution, these codes are assumed to be almost optimized and they must be very long so as to make the error probability low enough. This hypothesis was formulated in [4, 5] and [6], together with the auxiliary one that the hypothesized error-correcting means consist of nested codes. Then it was shown that it could explain biological facts as basic as, for instance, the existence of distinct species, the possibility of their taxonomy and the trend of evolution towards complexity. The existence of some kind of error correcting codes in the genome, at the molecular level or involving short codes, was also suggested in [7, 8, 9, 10]. The interesting idea that introns are made of check symbols associated with the message borne by the exons was formulated in [11]. However, the search for a simple linear code described in [12] was unsuccessful[1]. On the other hand, biological error-correcting mechanisms foreign to the genome replication were discovered (see, e.g., [13]). The role of codes in biology, although not explicitly seen as error-correcting ones, has been dealt with in [14]. The references given above are not exhaustive, but it is clear nevertheless that comparatively few works have been done on biological coding systems. This paucity is surprising. Since the discovery by Avery, McCarty and MacLeod [15] that DNA is the bearer of hereditary information, and the subsequent discovery of its double-helix structure by Franklin, Watson and Crick [17, 16], it has been recognized that the genetic message consists of the sequence of nucleic bases in DNA. Since the science of digital sequences is an important part of information theory, it should have become a major tool in molecular biology, but it has not been so. The parochialism in sciences has to be blamed for that, and the difficulties of communicating results from a scientific discipline to another one should not be overlooked. It seems especially difficult to let results from an engineering discipline be known by the upholders of a fundamental one.

The need for genomic error-correcting means will be reviewed in Section 3, as well as the consequences of their hypothesized existence. In Section 4, we shall introduce the concept of *soft code*, which weakens but extends the engineering concept of error-correcting code so as to better fit the biological context. Then, we shall look for possible genomic soft codes in Section 5. Before dealing with these topics, we shall consider in Section 2 why a mutual benefit should result from a collaboration of biology and engineering.

[1] This negative result is very questionable; see Sec. 5.1.

2 Biology and engineering, a needed collaboration

Having produced all living beings, with their incredible variety of forms and functions although the basic components are shared by all, nature appears as an engineer of extremely broad competence. It thus seems that human engineers should be deeply interested in nature's achievements, and understanding the engineering aspects of life should be a major concern for biologists. However, even though nature's achievements very often outperform those of human engineers, the methods used by nature on the one hand, and human engineers on the other hand, are in sharp contrast.

At variance with human engineers, nature does not use purposeful design, but 'tinkering', exhaustive search and natural selection. It ignores time limitation. Continuity of life is its sole (but tough) major constraint. There is also a broad difference between nature and engineers as regards spatial and temporal scales: engineers design and build objects of large physical size within a short time, and these objects have short lifetimes. The most basic properties of living things depend on objects at the molecular scale, and the time scale of nature extends to that of geology, i.e., up to billion years.

As regards the difference in performance, nature's achievements are outstanding. In almost any case, they are understood only insofar as human engineers invented similar solutions to a problem that nature solved aeons ago. Nature achievements very often outperform what human engineers can do, and even appeared untill recently as far beyond their reach (think of the autonomous motion and spontaneous behaviour of animals). Having genuine self-repair capabilities, living beings are moreover much more flexible and resisting to degradation than the products of human engineering. No wonder if during millenia humans could account for these differences only by the existence of supernatural entities.

The progresses of science and technology however reduced to some extent the gap between nature's and engineers' achievements. It became increasingly clear that the mechanisms of life do not rely on mystic properties like a 'vital force' but just obey the laws of physics and chemistry. As regards the time scale, the deep past is better and better understood, first thanks to Darwin's theory of evolution, and nowadays due to the avalaibility of physical means to estimate the ages of rocks and fossils, and to the possibility of establishing the phyletic trees of living species based on genome comparisons. As regards the physical size, objects at the molecular level can now be 'seen' and 'handled', and even 'manufactured' as nanotechnologies are becoming a reality. Again, [1, 2, 3]) are impressive examples of this trend. The main distinctive features of the living beings are their extreme complexity, which is unmatched in the nonliving world, and (not independently) the rather obvious but still overlooked fact that, besides matter and energy, they receive and transmit information, and they heavily rely on its transfer and conservation. This last point too has no equivalent outside the living world and appears as the specific mark which radically differenciates the living world from the nonliving one. It makes biology especially relevant to information theory, thus challenging information engineers and prompting biologists to use information theory as a main tool.

3 Hypothesizing genomic error correction

3.1 The need for genomic error correction

Perhaps the most convincing argument for the need of genomic error-correcting means is the fact that mutations, i.e., errors in the genome replication due to chemical agents or radiations, are responsible for ageing and certain diseases like cancers. If the error rate in communicating genomic information has noticeable effects at the scale of the lifetime of an individual, then the accumulation of errors during periods million times longer would simply make genetic communication impossible.

Moreover, if we look at the litterature on chromosomes and cellular division, on the one hand, and that on the performance of DNA replication on the other hand, the former appears as describing messy, involved and unreliable mechanisms; however, outstanding faithfulness of DNA replication is reported in the latter. This sharp contrast strongly suggests that error-correction mechanisms needed in order to correct replication errors actually exist. Many 'proof-reading' mechanisms[2] are known, but true error-correcting codes are necessary to reach the needed performance level since proof-reading can just ensure that the copy is faithful to the original, i.e., can only correct the errors due to the replication process itself, not the errors of external origin that affect the original. As a molecule, DNA is subject to mechanical, chemical and radiative aggressions. It is shielded against mechanical and chemical aggressions by membranes (the cellular membrane for all living beings, and furthermore the nucleic membrane in the case of eukaryotes whose genome thus benefits from a two-level protection). However, only error-correcting codes can efficiently protect against errors arising from radiations. These are of solar and cosmic origin, or due to natural radioactivity.

Although it is difficult to understand what such a figure really means, the error rate of DNA replication is reported to be of about 10^{-9} per nucleic base and per replication for higher animals. It is far greater (10^{-3} and even more) for some genes of viruses and bacteria. This difference between more or less complex living beings is itself difficult to understand without hypothesizing that more efficient error-correcting means exist in higher living beings than in bacteria and viruses. In turn, this assumption is consistent with the difference of the corresponding genome sizes and the proven result of information theory that the longer the encoded message, the more efficient can be the correction of errors.

To summarize, the outstanding performance of the genetic communication process cannot be accounted for unless assuming the existence of highly efficient error-correction means making genomes resilient to errors. They should thus go far beyond elementary encoding means and short codes, and we believe that error correction is not only possible but *necessary*. It is a *major* factor in genetics, not a dispensable one. This opinion is corroborated by the fact that, as will be reviewed in Sec. 3.2, very fundamental properties of evolution and of the living world can be derived from the mere hypothesis that the genome includes an error correction system with performance close to the information-theoretic limit, i.e., the channel capacity.

[2]Based on the duplication, in complementary form, of the sequence of nucleotides in the double-helix structure and the assumption that damages on one string can be corrected in terms of the other one.

Not only errors resulting from substitution of a wrong nucleic base to another one have to be considered, but possibly those due to erasures, deletions and insertions. We shall nevertheless limit ourselves to the errors due to substitutions because this case has been extensively studied by engineers, although the deletions and insertions are at least equally important.

It would be naïve to expect that the hypothesized natural error-correction means closely resemble those the engineers design. This paper is intended to outline what such natural error-correcting means may look like. A key concept will be that of *soft code*, which weakens but extends the engineering concept of code, and we will moreover try to outline the concept of *nested* soft codes. It will become clear that, in search for genomic error correction means, looking for soft codes is interesting in that any constraint that affects the sequence of nucleic bases in DNA (either mechanical, geometrical, chemical, physiological, ...) can be thought of as contributing somehow to the genomic message regeneration. Another interesting point will be that such codes are met also in linguistics, although to the best of our knowledge the potential error correcting capabilities of languages has not been scientifically studied. That natural languages have such capabilities is nevertheless an obvious lesson of daily experience. Some kind of language is needed for communicating hereditary information, which as a soft code implies error-correction capabilities.

3.2 Consequences of the hypothesized existence of genomic error correction

We briefly review consequences of the hypothesized existence of genomic error-correcting means, which were already stated in [4, 5] and [6]. These consequences are in agreement with biological facts, or suggest answers to still controversial questions.

1 - Existence of discrete species. According to our viewpoint, it is a direct consequence of the distance structure of an error-correcting code. The existence of discrete species is generally considered as a fact, although precisely defining the concept of species is difficult and still controversial.

1' - Taxonomy is possible. It is a consequence of the auxiliary hypothesis that the genome error-correcting means consist of nested codes. This is generally considered as a fact, or at least as an essential working hypothesis in biology.

2 - Evolution proceeds by jumps (is 'saltationist'). This is a still controversial issue. According to our viewpoint, it is a direct consequence of the distance structure of an error-correcting code. It implies that natural selection does not act on close variants of existing beings, but on mutants with a genome at a distance from the original one equal at least to the minimum distance of the code, which depends itself on the code level in the assumed system of nested codes. It results in a non-Darwinian mechanism for the origin of species, reminiscent of the 'hopeful monster' hypothesized by Goldschmidt.

3 - There exists a trend of evolution towards complexity. If understood as implying that species having a larger genome than the previously existing ones appeared in many instances during the process of evolution, it may be considered as an experimental fact, although a larger genome does not imply an increase in complexity but only makes it possible (see Sec. 5.5). It can be interpreted as a consequence of the hypothesis that error-correcting means exist in the genome. Indeed, a longer genome is an evolutive burden

as regards the speed of replication but is advantageous as enabling a more efficient error correction, according to the channel coding theorem of information theory, so its net effect may be to increase the genome *permanency* (as defined in [5]) and therefore to provide an evolutive benefit.

4 Soft codes: definition and properties

4.1 Soft codes vs conventional codes

Let us first consider a linear binary block code $\mathcal{C}(n,k)$ of length n and dimension k as a typical example of a conventional error-correcting code. It has two possible equivalent definitions:

i) as a subset of the set of vectors $\mathbb{F}_2^n$ over the binary field $\mathbb{F}_2$ such that any codeword $c \in \mathcal{C}(n,k)$ belongs to a list of 2^k words, $k < n$, which can be obtained by giving the information vector $\underline{u} \in \mathbb{F}_2^k$ all possible values and computing $\underline{c}$ according to

$$\underline{c} = \underline{u}\mathrm{G}, \tag{1}$$

where G, an $n \times k$ binary matrix, is the generator matrix of the code. Vectors $\underline{u}$ and $\underline{c}$ are represented as row matrices. Any binary vector $\underline{v} \in \mathbb{F}_2^n$ thus belongs or not to $\mathcal{C}$.

ii) as a subset of $\mathbb{F}_2^n$ such that any binary vector $\underline{c} \in \mathbb{F}_2^n$ belongs to $\mathcal{C}(n,k)$ if and only if it obeys the check equation

$$\underline{c}\mathrm{H}^{\mathrm{t}} = \underline{0}, \tag{2}$$

where H, an $n \times (n-k)$ binary matrix, is the parity-check matrix of the code. Superscript t denotes matrix transposition and $\underline{0}$ is an $(n-k)$ row matrix all the entries of which are 0. Given a binary vector $\underline{v} \in \mathbb{F}_2^n$, it belongs or not to $\mathcal{C}$ depending on $\underline{v}\mathrm{H}^{\mathrm{t}}$ being zero or nonzero: a codeword $\underline{c} \in \mathcal{C}(n,k)$ is thus characterized by the fact it obeys the constraint expressed by (2).

Then, the matrix equality (1) describes the encoding process, while (2) is useful for detecting errors, which is often the first step of an error-correction process. The two matrices G and H are closely related to each other since, when G is put in the systematic form

$$\mathrm{G} = [\mathrm{I}_k |\ \mathrm{P}]\,,$$

H reads

$$\mathrm{H} = \left[\mathrm{P}^{\mathrm{t}} \middle|\ \mathrm{I}_{n-k}\right]$$

where I_k denotes the unity matrix of order k and P is some $(n-k) \times k$ matrix. This close relationship makes definitions i) and ii) of $\mathcal{C}$ not so different.

If one wants to extend the concept of error-correcting code, however, definition ii) is most convenient since other constraints than those expressed by deterministic mathematical equalities can be used. For instance, probabilistic constraints can be contemplated, as well as constraints expressed by incompatibilities or forbidding rules. Such extensions of error-correcting codes will be referred to as 'soft codes'.

Of course, the dichotomic property that any binary vector $\underline{v} \in \mathbb{F}_2^n$ belongs or not to $\mathcal{C}$ is lost if $\mathcal{C}$ is a soft code. One can only assign a probability distribution to any $v \in \mathbb{F}_2^n$

as induced by the constraints which define the soft code. The concept of 'information message' vanishes (hence the encoding operation becomes meaningless[3]) and there is no longer a simple encoding rule like (1). As compared with an ordinary error-correcting code, the usual parameters like the dimension k, the distribution of distances and the minimum distance are no longer defined or become random variables. The function of decoding can no longer be thought of as intended to recover an information message and we shall instead define as more appropriate the function of 'regeneration', to be introduced later. Before we deal with the function of regeneration, let us illustrate the concept of soft code by a few examples.

4.2 Some examples

We shall only consider two examples which are mainly intended to show some connection between the concept of soft code and the conventional one.

Our first example is the set of words of an error-correcting code seen through a noisy channel. As a very simple error-correcting code, let us consider the (6,3) binary systematic linear code (it is the (7,4) Hamming code shortened so as to become 'threshold decodable') having as generator matrix

$$\mathrm{G} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}, \tag{3}$$

and thus as parity-check matrix

$$\mathrm{H} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}. \tag{4}$$

Assume that this code is seen through a noisy channel, say a binary symmetric channel with error probability $p < 1/2$. Let $e_1, \ldots, e_6$ denote the binary errors which affect the code symbols $c_1, \ldots, c_6$. We may think of the set of words $\{c'_1 = c_1 \oplus e_1, \ldots, c'_6 = c_6 \oplus e_6\}$ as a soft code which, instead of the deterministic parity-check equations that the codewords of $\mathcal{C}$ verify, namely

$$\left.\begin{array}{rcl} c_1 \oplus c_2 \oplus c_4 & = & 0, \\ c_2 \oplus c_3 \oplus c_5 & = & 0, \\ c_1 \oplus c_3 \oplus c_6 & = & 0, \end{array}\right\} \tag{5}$$

can be characterized by the probabilistic constraint

$$\Pr(h_1 = 0, h_2 = 0, h_3 = 0) = q, \tag{6}$$

where

$$\begin{array}{rcl} h_1 & \stackrel{\Delta}{=} & c'_1 \oplus c'_2 \oplus c'_4, \\ h_2 & \stackrel{\Delta}{=} & c'_2 \oplus c'_3 \oplus c'_5, \\ h_3 & \stackrel{\Delta}{=} & c'_1 \oplus c'_3 \oplus c'_6, \end{array}$$

[3]That purposeful encoding is not needed answers some criticisms against the use of Shannon's paradigm made in [9] so as to model the process of synthesis of a protein in terms of the DNA message borne by a gene.

and where $q = (1-p)^2(1-4p+6p^2)$ has been calculated in terms of the channel error probability p using the trellis associated with code $\mathcal{C}$. Note that we must consider the joint probability since the parity checks h_1, h_2 and h_3 are not independent, as having symbols in common.

As another example linking soft and conventional codes, we shall consider soft output decoding of, say, a binary code of length n. Soft output decoding means that the decoding role has been extended to reassess the probability of all the codewords [18]. In this case, the result of decoding is (ideally) the list of all codewords, each affected with its reassessed probability. This may be thought of as a soft code where the probabiity of an n-tuple is 0 if this n-tuple is not a codeword, but is nonzero for each codeword and specific to it.

4.3 Decoding and regeneration

4.3.1 A convenient formalism

We now look at the problems of decoding. We shall restrict ourselves in the following to the binary alphabet, which will enable us to use the simple formalism which associates with any binary random variable b its *algebraic value*, i.e. the real number a defined as:

$$a \triangleq \ln \frac{\Pr(b=0)}{\Pr(b=1)}. \tag{7}$$

When the probabilities in (7) are conditional probabilities, for instance when b is the output of a noisy channel, a is itself a random variable. According to (7) its mean is

$$\mathrm{E}[\mathrm{sign}(a)] = \mathrm{E}[(-1)^b] = 1 - 2\Pr(b=1) = t(a),$$

where:

$$\mathrm{t}(\cdot) \triangleq \tanh(\cdot/2) = \frac{\exp(\cdot)-1}{\exp(\cdot)+1}. \tag{8}$$

The algebraic value a of a random binary variable b is a real number. Its sign represents the most probable decision $\hat{b}$ according to

$$\mathrm{sign}(a) = (-1)^{\hat{b}}$$

and its magnitude measures the reliability of this decision, since

$$|a| = \ln \frac{b=\hat{b}}{b\neq\hat{b}}.$$

If the probabilities in (7) are just those of a channel output, the algebraic value a is referred to as an *a priori* algebraic value. If these probabilities are computed in terms of *a priori* algebraic values so as to take account of the constraints of a code, the resulting algebraic value is referred to as an *a posteriori* algebraic value and denoted by a capital letter (at variance with the usual notation where capital letters denote random variables and lower case ones realizations; here, both a and A are random variables, but the former is a channel output while the latter is that of a decoder). Computing an *a posteriori* algebraic value is

referred to as a *soft decision*. Keeping only its sign is referred to as a *hard decision*. The soft decision thus involves, besides a mere binary decision, an assessment of its reliability which is lost in a hard decision.

Extension of this formalism to larger-size alphabets endowed with the finite field structure is rather straightforward and has been used for decoding nonbinary linear codes [19, 20], but will not be dealt with here.

4.3.2 The case of mere repetition

The usefulness of this formalism is clear if we consider the simple case of a mere n-fold *repetition*. Let us consider an equiprobable binary symbol u which is repeated, for instance, $n = 3$ times. A codeword (c_1, c_2, c_3) thus obeys the coding constraint $c_1 = c_2 = c_3 = u$. If c_1, c_2 and c_3 are transmitted using independent binary symmetric channels[4] having p_i $(0 < p_i \leq 1/2,\ i \in \{1, 2, 3\})$ as error probabilities, then the optimum decision rule consists of computing:

$$A = \sum_{i=1}^{n} a_i = a_1 + a_2 + a_3, \tag{9}$$

where $a_i = (-1)^u \ln \frac{1-p_i}{p_i}$ is the *a priori* algebraic value associated with the i-th channel output. Then the quantity A computed by (9) is the *a posteriori* algebraic value of the optimum (maximum likelihood) binary decision U on u so we have:

$$A = (-1)^U \ln \frac{1-P}{P}, \tag{10}$$

where P is the probability that this decision is wrong [21, 19]. The quantities a_1, a_2 and a_3 are available as the outputs of ideal (matched filter) binary demodulators. The decoder which implements (9) is of the soft-input soft-output (SISO) type i.e., its inputs as well as its output are real numbers which represent both a binary decision (borne by the sign) and the estimate of its reliability (the magnitude).

4.3.3 Threshold decodable code

This decidion rule can be extended to more complicated cases, for instance, the still very simple (6,3) code considered above. We can solve the parity-check relations (5) in terms of c_1, which results in $c_1 = c_2 \oplus c_4$ and $c_1 = c_3 \oplus c_6$. This means that $c_2 \oplus c_4$ and $c_3 \oplus c_6$ repeat c_1 so it is possible to apply the above decision rules to the 3 received *replicas* of c_1: c'_1, $c'_2 \oplus c'_4$ and $c'_3 \oplus c'_6$, where the prime denotes the received symbols. The situation is not very different from the three-fold repetition of a single bit considered above, except that the code rate is now $1/2$ instead of $1/3$. We shall refer to the first replica as the 'trivial' one (it is the symbol to be decoded itself) and the other two as 'compound replicas', made of a combination of other symbols.

The 'soft decision' rule consists of applying (9) to the corresponding *a priori* algebraic values, namely a_1 which is associated with c_1, and the *a priori* algebraic values associated

[4] Or when c_1, c_2 and c_3 are successive outputs of a memoryless channel the individual reliabilities of which are separately measured.

with the sums modulo 2 $c'_2 \oplus c'_4$ and $c'_3 \oplus c'_6$. Let a_2, a_4, a_3 and a_6 denote the *a priori* algebraic values of c'_2, c'_4, c'_3 and c'_6, respectively. One easily shows that the algebraic value a associated with the sum modulo 2 of two received binary symbols of algebraic values a_i and a_j is such that

$$t(a) = t(a_i)t(a_j) \tag{11}$$

where $t(a)$ has been defined in (8). Then (11) results in

$$a = \ln \frac{1 + t(a_i)t(a_j)}{1 - t(a_i)t(a_j)} = \ln \frac{\exp(a_i + a_j) + 1}{\exp(a_i) + \exp(a_j)}. \tag{12}$$

Using this expression enables writing the *a posteriori* algebraic value A_1 associated with c_1 in terms of *a priori* algebraic values as

$$A_1 = a_1 + \ln \frac{1 + t(a_2)t(a_4)}{1 - t(a_2)t(a_4)} + \ln \frac{1 + t(a_3)t(a_6)}{1 - t(a_3)t(a_6)}, \tag{13}$$

where A_1 is the *a posteriori* algebraic value associated with c_1. Similar expressions for decoding c_2 and c_3 are easily derived.

This case is especially simple since compound replicas of the information bits have no symbol in common. It is the soft-input soft-output version of Massey's threshold decoding [22].

4.3.4 General case

For an arbitrary (n, k) conventional code $\mathcal{C}$, we may define the *a posteriori* algebraic value of the i-th binary symbol as the ratio of the probability that it belongs to the set of words of the code $\mathcal{C}$ having 0 as i-th symbol, to be denoted $\mathcal{C}_{i0}$, to that of belonging to $\mathcal{C}_{i1}$, the set of words of the code $\mathcal{C}$ having 1 as i-th symbol. Writing it in terms of the *a priori* algebraic values results in

$$\begin{aligned} A_i &= \ln \frac{\sum_{\underline{c} \in \mathcal{C}_{i0}} \exp(-\sum_j c_j a_j)}{\sum_{\underline{c}' \in \mathcal{C}_{i1}} \exp(-\sum_j c'_j a_j)}, \\ &= a_i + F_i(a_1,\ a_2,\ \ldots,\ a_n). \end{aligned} \tag{14}$$

$F_i(\cdot,\ \cdots,\ \cdot)$ does not depend on a_i and is complicated except for particular cases like (13); it expresses the *a posteriori* algebraic value of the replica of the i-th symbol which combines all the other symbols according to the code constraints. It thus represents the 'extrinsic information' in the vocabulary of turbo codes [23]. Decoding algorithms which use a trellis to represent the code constraints may be thought of as means to compute this function [24, 21, 25, 26]. Notice that the *a posteriori* algebraic values involved in the decision rule (14) are those of the n encoded symbols, not only of the information symbols if the code is systematic, so this decision rule remains valid even if information symbols are no longer well defined, which is the case for soft codes, provided it is properly extended.

The decoding rule (14) deals separately with each symbol, and it turns out that the word consisting of the hard decisions on each *a posteriori* algebraic value is not necessarily

a codeword. This decoding rule can be simplified by only keeping in the numerator and denominator of the fraction in (14) the largest of the exponential terms which results in

$$\begin{aligned} A'_i &= \min_{\underline{c}' \in \mathcal{C}_{i1}} (\sum_j c'_j a_j) - \min_{\underline{c} \in \mathcal{C}_{i0}} (\sum_j c_j a_j), \\ &= a_i + F'_i(a_1,\ a_2,\ \ldots,\ a_n) \end{aligned} \tag{15}$$

where $F'_i(a_1,\ a_2,\ \ldots,\ a_n)$ does not depend on a_i and, for a linear code, is a sum of *a priori* algebraic values. This highly simplified rule provides only an approximate *a posteriori* algebraic value of each symbol, but it turns out that taking hard decisions on each of the approximate *a posteriori* algebraic values A'_i in (15) results in the optimally decoded word [21, 19].

Let us now look at the case of a soft code. We may in principle extend the decoding rule (14) to a soft code characterized by the set of constraints K as

$$A_i = \ln \frac{\sum_{\underline{c} \in \mathcal{F}_2^n} \Pr(c_i = 0 | \underline{c}, K)}{\sum_{\underline{c} \in \mathcal{F}_2^n} \Pr(c_i = 1 | \underline{c}, K)}.$$

Let us redefine $\mathcal{C}_{i0}$ and $\mathcal{C}_{i1}$ as $\mathcal{C}_{i0} \stackrel{\Delta}{=} \{\underline{c} \in \mathbb{F}_2^n | c_i = 0\}$ and $\mathcal{C}_{i1} \stackrel{\Delta}{=} \{\underline{c} \in \mathbb{F}_2^n | c_i = 1\}$. Using Bayes rule and the algebraic value formalism then results in

$$\begin{aligned} A_i &= \ln \frac{\sum_{\underline{c} \in \mathcal{C}_{i0}} \Pr(\underline{c}|K) \Pr(c_i = 0)}{\sum_{\underline{c}' \in \mathcal{C}_{i1}} \Pr(\underline{c}'|K) \Pr(c'_i = 1)}, \\ &= a_i + F_i^K(a_1,\ a_2,\ \ldots,\ a_n) \end{aligned} \tag{16}$$

where, similarly to (14), $F_i^K(\cdots)$ does not depend on a_i and expresses the extrinsic information about the i-th symbol. Again, this function is in general very complicated but can be widely simplified by keeping only the largest of the terms in the numerator and denominator of the fraction in (16).

4.3.5 The regeneration function

We already noticed that decision rules (14), (15) and (16) can be used to reassess the probabilities of all the symbols of the word, not only the information symbols of a systematic code. They may thus be interpreted as performing *regeneration* rather than decoding. In the case of a soft code, this concept should be substituted for that of decoding since the notion of information symbols vanishes. However, this function is not specific to soft codes.

Let us assume that we have to communicate some redundantly encoded message through the sucessive use of two channels, as depicted in Fig. 1. At the output of the first channel, we may use a decoder so as to restore the information message, and if the encoding-decoding system is properly designed it will incur an arbitrarily small error probability. The decoder output, i.e., the restored information message, must then be encoded again before transmission over the second channel. The combined operation of encoding and decoding may be thought of as a single regeneration operation, which succeeds with large probability in restoring the initial encoded sequence by means of exploiting the

coding constraints. Although the operation of encoding loses almost any significance for soft codes, restoration of the transmitted sequence thanks to the coding constraints remains possible in a probabilistic sense, so the regeneration of the input sequence remains meaningful although decoding in its usual acceptation does not. The use of several successive channels typically depicts the successive genome replications, so regeneration is fully relevant to this case.

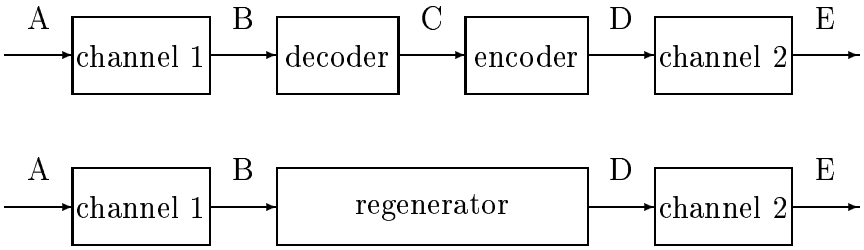

Figure 1: The regeneration function. Channels 1 and 2 have to be used successively. The upper picture is relevant to a conventional error-correcting code. The sequences found at the points designated by letters are: in A, an encoded sequence; in B and E, received sequences; in C, an information message;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

Moreover, we may think of decoding a conventional error-correcting code as a two-step process. Consider the space of all n-tuples. For long enough words, the point which represents the received word is close to the surface of the sphere centered on the transmitted word and having as radius the expected number of errors as a mere consequence of the law of large numbers. Maximum likelihood decoding consists first of determining the point which represents the closest codeword to the point representing the received word, and then to output the information message which labels this point. The first step may be interpreted as regenerating the original codeword, and the second one is almost trivial due to the one-to-one correspondence between codewords and information messages.

Decoding error-correcting codes is based on the assumption that the constraints obeyed are known with certainty, while the received word not only can differ from the transmitted one due to the channel operation, but actually differs from it with high probability in conditions which make the encoding-decoding process efficient. The decoding process thus looks for a word which obeys the constraints, and the received word just *approximately* indicates where to look for. Thus, faithful reconstruction of the transmitted word is achieved through unconditionally relying on the constraints, the received word giving only an indication of proximity. The same can be done in the case of soft codes, except that now the constraints are no longer necessarily deterministic. (Exactly knowing a probabilistic constraint implies no contradiction.)

In the absence of a thorough study of soft codes, we may assume for convenience that, although the main parameters which determine the performance of a conventional

code, for instance its distance distribution and especially its minimum distance, become random when transposed to a soft code, the main properties of error-correcting codes are not fundamentally altered so they remain approximately relevant to the biological soft codes. The consequences of error-correcting means consisting of soft codes will thus not be significantly different from those of conventional codes, which leaves the conclusions of Sec. 3.2 essentially valid. Besides a convenient assumption, it may be fairly close to reality if both the code lengths considered are large and the soft code is specified by many independent constraints, as a consequence of the law of large numbers.

4.4 Nested soft codes

Nested codes can be more easily described in the case of conventional binary systematic codes. We assume that a first information message I_0 of length k_0 is encoded according to a code $\mathcal{C}(n_0, k_0)$. Then, a second message I_1 of length k_1 is appended to the codeword which resulted from the first encoding, and it is encoded again by a code $\mathcal{C}(n_1, n_0 + k_1)$. This process is repeated t times. The last information message I_t is left uncoded. This process is depicted in Fig. 2 using the fortress metaphor where each code is depicted as a wall which encloses its encoded information message, for $t = 3$.

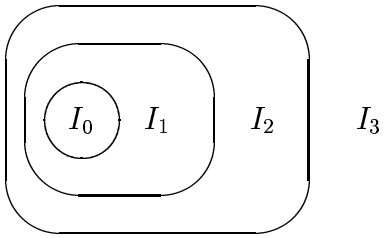

Figure 2: The fortress metaphor. A code is represented as a closed wall which protects what is inside it. I_0, I_1, I_2 and I_3 are successive information messages. I_0 is protected by 3 codes, I_1 by 2 codes, I_2 by a single code and I_3 is left uncoded.

The case of nonsystematic codes is less simple but any linear code can be put in systematic form without loss of generality. Another mean for obtaining nested codes consists of using different alphabets of size increasing in terms of the encoding level. Many variants are possible.

Defining *nested soft codes* is more difficult since then the concept of information message vanishes in this case. We may think of the nested code concept in general terms: the i-th encoding creates dependency between the results of $i-1$ previous encodings, regardless of the codes and the alphabets which are used. Similarly to the case of conventional codes, the alphabet may differ from an encoding to another one.

5 Looking for genomic soft codes

Thanks to the concept of soft code, any constraint that affects the sequence of nucleic bases in DNA (either mechanical, geometrical, chemical, physiological, ...) can be thought of as

contributing somehow to its regeneration. The biological literature is rich in descriptions of such constraints. Each of them implies a peculiar soft code in the assumed nested coding system. We shall review some of such constraints and their hypothesized or experimentally observed consequences as regards regeneration of the DNA message. Prior to doing so, we must look at the problem of identifying the alphabet.

5.1 Identification of the alphabet

In engineering problems, the alphabet is almost always unambiguously defined, being given as a parameter. This is not the case for the hypothesized error-correcting codes of DNA where the alphabets themselves have to be determined. We put here 'alphabets' to the plural since we consider nested soft codes, and we already noticed that the successive encodings involved can use different alphabets.

An apparently obvious choice is that of the quaternary alphabet {A, T, G, C}, but it should be endowed with some mathematical structure. This approach was used by Liebovitch et al. [12], who assumed that this structure was $\mathbb{Z}/4\mathbb{Z}$, the ring of integers modulo 4. This choice is arbitrary and the negative result of [12] just means that no error-correcting code using this alphabet was found. A proper linear code would have been searched for if the Galois field $\mathbb{F}_4$, defined as an extension of the binary field based on the primitive polynomial $1 + x$, were used instead.

It is not sure that such an approach is good, even if $\mathbb{F}_4$ is used rather than $\mathbb{Z}/4\mathbb{Z}$. The connection which the concept of soft code establishes between the physical and chemical constraints and the error-correcting properties suggests to look at alphabets having a physico-chemical significance. In this respect, it is much more relevant to consider that any quaternary symbol simultaneous belongs to two binary alphabets. First, the alphabet {R,Y} where the symbols are the chemical structure of the corresponding nucleic base, namely purine (2-cycle molecule, A or G), denoted R, or pyrimidine (single-cycle molecule, T and C), denoted Y. Second, the alphabet {2H,3H} where 2H represents the couple of complementary nucleotides A–T which are tied together by two hydrogen bonds (H-bonds), and 3H the other couple, namely G–C, where the nucleotides are tied together by three H-bonds. The alphabet {R,Y} corresponds to nucleic bases of different physical size, while the second one, {2H,3H}, indicates how strongly a nucleic base is tied with the complementary one. Then, Forsdyke interpreted a sequence of quaternary symbols as simultaneously bearing two independent binary codes, one over the alphabet {R,Y} and the other one over {2H,3H} [27]. According to the 'second Chargaff parity rule', the first code is balanced, i.e., the two symbols R and Y have the same frequency, like almost all codes designed by human engineers. On the contrary, the code over the alphabet {2H,3H} is not balanced since the frequency of its symbols varies from a species to another one and, for long and inhomogeneous genomes like the human one, from a region to another one inside the genome. It could be interpreted as a kind of 'density modulation' which maybe is read at several scales. The different number of hydrogen bonds of the two symbols implies that this density modulation results in a variation of the bonding energy between the two DNA strands.

Other constraints are naturally expressed in terms of other alphabets. For instance, constraints induced on DNA by the structural properties of the proteins for which it

'codes' are likely to involve triplets of nucleic bases, i.e., the codons of the genetic 'code'. Genes themselves can even be considered as the symbols of an alphabet [28, 29]. The successive use of alphabets of different size is a mean of implementing nested codes, as mentioned in Sec. 4.4.

Closely related to the alphabet identification is the problem of source coding, or compression, of a DNA sequence. The attempts to apply standard source coding algorithms to DNA sequences have been yet quite unsuccessful [29], maybe because they can only deal with short-term dependence. A full understanding of the hypothesized nested codes system would be necessary for performing efficient source coding of DNA sequences, then consisting of 'undoing' the channel-coding process. Such source coding is thus likely to imply several levels and alphabets. Could a source coding system perform efficiently on a DNA sequence, then it would be possible to restore the concept of an information message associated with it.

5.2 Constraints from the spatial structure of DNA

The alphabet which is relevant here is more likely to be {R,Y} as introduced in the previous section, but the alphabet {2H,3H} may also be relevant since the ease of separating the two DNA strands is an important factor during the replication process.

Among the constraints which affect the message borne by a DNA molecule, we may first think of mechanical and chemical constraints due to the spatial structure of the DNA molecule, its bonding with proteins like histones and especially its packing in nucleosomes and higher-order structures (when such structures exist, i.e., in eukaryotes).

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 [31].

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. 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 [32]. 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 related to structural constraints [32]: 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 (more precisely, the 165 base-pair-long) 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 for decoding of many low-reliability replicas of the basic motif symbols, hence to reliable decisions on these symbols based on many unreliable replicas. This implies a very large redundancy, indeed an obvious property of the DNA message. The existence of such a decoding process, possibly approximated by majority voting, is as yet a conjecture. It remains to determine if, and how, nature implements a decoding process based on long range dependence at some stage of the DNA replication process [30].

One may wonder why the decoding process does not turn this unfaithful repetition into a faithful one by correcting the 'wrong' symbols. We may explain why it is not so by assuming that other soft codes also exist with independent constraints. Then, an actual symbol of the DNA message results from a compromise between the constraints of the several soft codes in which it is involved.

5.3 Constraints induced on DNA inside the genes by the structure of proteins

Physiologically active proteins do not reduce to the polypeptidic chain that the sequence of codons of the gene 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 on proteins. Moreover, proteins owe their functional properties to their folding according to a unique pattern, which implies many chemical bonds between amino-acids which are separated along the polypeptidic chain but close together in the 3-dimensional space when the protein is properly folded. For instance, many proteins having enzymatic function fold into a globular shape. Despite their weakening through the inverse genetic 'code', i.e., the one-to-many correspondence between the amino-acids and the codons (triples of nucleic bases), the constraints of steric and chemical character obeyed by proteins induce constraints on the corresponding DNA. Due to the central role of genes in directing the synthesis of proteins, these constraints must be present in the genome of any living being, whether it is a prokaryote or a eukaryote.

5.4 Is a gene involving exons and introns a kind of systematic codeword?

The interesting idea that introns are made of check symbols associated with the message borne by the exons was formulated by Forsdyke in 1981 [11]. The literature generally states that introns are more variable than exons. A counter-example was however provided in 1995 by Forsdyke, who experimentally found that the exons are more variable than introns in genes which 'code' for snake venoms [33].

It turns out that both the generally observed greater variability of introns and Forsdyke's counter-example can be explained by the assumption that the system of exons and introns actually acts as a systematic error-correcting code where exons constitute the information message (which directs the synthesis of a protein) and introns are made of the associated check symbols. Interpreted as a decoding error, a mutation occurs with

large probability 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, it may be expected that only mutations with a few errors within the exons, hence having no or little incidence on the protein, will survive natural selection. Few errors being located in the exons, most of them will affect the introns since the total number of errors is at least equal to the minimum distance of the code.

The situation is completely different in the case of genes which 'code' for snake venoms. The typical preys of snakes are rodents. Snakes and rodents are involved in an 'arms race': some rodents incur mutations which provide an immunity to snake venom, the population of rodents with such mutations increases as they escape their main predators, and the snakes are threatened with starvation unless mutations in their own genes make their venom able to kill mutated rodents [33]. The genes which 'code' for snake venoms are thus under 'high evolutive pressure': natural selection favours mutated genes producing proteins as different as possible from the original one. In terms of the Hamming distance, much of the difference should thus be located in the exons. The total number of errors in exons and introns being roughly constant for a given code (equal to the minimum distance or slightly larger), introns are much less variable.

5.5 Linguistic constraints

In the absence of redundancy, the number of different genomes of size 3×10^9 (the approximate length of the human genome) would be $4^{(3\times 10^9)}$, a number which defies imagination. A genome of a few hundred base pairs would be enough to uniquely characterize not only all past and extant species, but all individuals having once belonged or presently belonging to them. Even the shortest genomes are thus highly redundant.

The contrast between the comparative brevity of the message which is needed for unambiguously identifying a biological species and an individual inside it on the one hand, and the length of actual genomes on the other hand, has rather obvious reasons. The genome role is by no means restricted to identify a living being since modern biology interprets it as a *blueprint* for its construction. The genome of any living being actually contains 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*. Biologists do not yet know this language although some of them claim in newspapers that they 'decipher' or 'decrypt' genomes. In a sense, many of them deny its existence when they dub 'junk DNA' every part of the DNA outside the genes and their regulatory sequences: they declare useless what they do not understand. But, on the other hand, they consistently use the metaphor of a written text to explain the role of the genetic message, at least in popular science books like [34] and many others. This metaphor is quite convincing, but the consequences it can have on the genome conservation are overlooked: indeed, it turns out that a language involves many morphological and syntactic constraints which may be interpreted as 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. Of course, it remains to understand how these error-correcting capabilities are exploited. Indeed, recent researches use tools of formal linguistics in order to describe the genomes and proteins [35, 36] but ignore the error correction problem. Moreover, the concept of dependence is actually shared by information and coding theory on the one hand, and formal linguistics applied to genomes and proteins on the other hand. As another interesting analogy, natural languages have undergone evolution, and the same kind of methods as used for studying biological evolution have been applied to that of languages.

The connection just outlined between semantics and error-correcting capabilities implies that a longer genome is not only useful to decrease the error probability, but also provides room for 'more semantics' and thus enables specifying more complex beings. An important and useful tenet of information theory is the separation between information and semantics. However, the hypothesized error-correction mechanisms based on linguistic constraints heavily rely on the genome being a blueprint for the construction and maintenance of a living being, so one could thus consider the error-correction property of the genetic message as, at least partially, a by-product of its semantics. But this is only a facet of the question. One can equally well argue that the error-correction property is the main feature of this message, since without it no transmission of hereditary characters would be possible and life could not have developed. Then, the construction and maintenance of living things would be a mere projection in the physico-chemical world of the abstract properties of the genetic message which enable error correction. This is a hen-and-egg problem, as often met in biology.

5.6 Further comments on regeneration and copying

The genome replication process results in a new genome. As such, it is submitted to all the constraints that a genome should obey. On the other hand, it should replicate the old genome which presumably suffered errors, so these conflicting requirements must be solved in favour of the *constraints*. But we likened the constraints of biological origin with soft codes, so obeying constraints is equivalent to correcting errors. Replacing the word 'genome' by 'codeword' in the above statement just describes the technical function of regeneration, defined above as the first and more important step of any decoding process. 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.*

The assumption that the error-control system in DNA is a by-product of prior biological constraints matches a characteristic of nature's method, namely, the use of some preexisting hardware to perform a function other than the initial one, which can be referred to as tinkering.

6 Conclusion

We think that information theory can answer the question asked in the title of this paper. The concept of soft code may be helpful to understand how the many constraints of biological character which affect the genome also involve error-correcting capabilities. Of course, the precise underlying mechanisms at work remain to be identified. Progresses along this line cannot be expected unless information and coding theoretists collaborate not only with biologists and especially with molecular genetists, but also with chemists and linguists. Clearly, a broad interdisciplinary effort is needed if the ideas outlined above are to be developped.

Erwin Chargaff wrote in 1979 a paper entitled "How genetics got a chemical education" [37], where he complained that genetists were so reluctant to accept the consequences of the discovery that DNA was the actual bearer of genetic information (by Avery et al. [15], 35 years earlier) that a more appropriate title of his paper could have been "How genetics refused to get a chemical education". We hope that the title "How genetics got an information-theoretic education", paraphrasing Chargaff, will eventually match reality.

References

[1] M. Mansuripur, "Macro-molecular data storage with petabyte/cm^3 density, highly parallel read/write operations, and genuine 3D storage capability", DIMACS working group on theoretical advances in information recording, 22–24 March, 2004.

[2] W. Coene, "Coding and signal processing for two-dimensional optical storage", DIMACS working group on theoretical advances in information recording, 22–24 March, 2004.

[3] G. Cherubini, "A nanotechnology-based approach for highly-parallel, ultra-dense data storage", DIMACS working group on theoretical advances in information recording, 22–24 March, 2004.

[4] G. Battail, "Does information theory explain biological evolution?," *Europhysics Letters*, Vol. 40, No. 3, pp. 343–348, Nov. 1st, 1997.

[5] G. Battail, "Is biological evolution relevant to information theory and coding?", *Proc. ISCTA '01*, pp. 343–351, Ambleside, UK, 15–20 July 2001.

[6] G. Battail, "An engineer's view on genetic information and biological evolution", IPCAT2003 (5-th International Workshop on Information Processing in Cells and Tissues), Lausanne, Sep. 8-11, 2003, Pre-proceedings, pp. 431–448.

[7] G. Cullmann & J.-M. Labouygues, The logic of the genetic code, 1983, *Biosystems*, Vol. 16, pp. 9–29.

[8] J. Rzeszowska-Wolny, "Is genetic code error-correcting?", *J. Theor. Biol.*, Vol. 104, pp. 701–702, 1983.

[9] H. P. Yockey, *Information theory and molecular biology* Cambridge University Press, 1992.

[10] D.A. Mac Dónaill, A parity code interpretation of nucleotide alphabet composition, 2002, *Chem. Communic.*, Vol. 18, pp. 2062–2063.

[11] D.R. Forsdyke, "Are introns in-series error-detecting sequences?", *J. Theor. Biol.*, Vol. 93, pp. 861–866, 1981.

[12] L.S. Liebovitch, Y. Tao, A.T. Todorov, and L. Levine, "Is there an Error Correcting Code in the Base Sequence in DNA?", *Biophysical Journal*, Vol. 71, pp. 1539–1544, 1996.

[13] E.E. May, M.A. Vouk, D.L. Bitzer, D.I. Rosnick, "Coding theory based models for protein translation initiation in prokaryotic organisms", IPCAT2003 (5-th International Workshop on Information Processing in Cells and Tissues), Lausanne, Sep. 8-11, 2003, Pre-proceedings, pp. 371–389.

[14] M. Barbieri, *The Organic Codes*, Cambridge University Press, Cambridge, UK, 2003.

[15] O. Avery, M. McCarty, and C. MacLeod, "Studies of the chemical nature of the substance inducing the transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus Type III", *J. Exp. Med.*, Vol. 79, pp. 137–158, 1944.

[16] R.E. Franklin, R.G. Gosling, "Molecular configuration in sodium thymonucleate", *Nature*, Vol. 171, No. 4356, pp. 740–741, 25 Apr. 1953. Reprinted in *Nature*, Vol. 421, No. 6921, pp. 400–401, Jan. 23 2003.

[17] J.D. Watson, F.H.C. Crick, "Molecular structure of nucleic acids", *Nature*, Vol. 171, No. 4356, pp. 737–738, Apr. 25, 1953. Reprinted in *Nature*, Vol. 421, No. 6921, pp. 397–398, Jan. 23, 2003.

[18] G. Battail, "Le décodage pondéré en tant que procédé de réévaluation d'une distribution de probabilité," *Annales Télécommunic.*, Vol. 42, No. 9–10, pp. 499–509, Sep.-Oct. 1987.

[19] G. Battail, M. Decouvelaere and P. Godlewski, "Replication decoding", *IEEE Trans. Inf. Th.*, Vol. IT-25, No. 3, pp. 332–345, May 1979.

[20] G. Battail, "Décodage pondéré par résolution d'un système d'équations implicites analogiques", *Annales Télécommunic.*, Vol. 45, No. 7–8, pp. 393–409, Jul.-Aug. 1990.

[21] G. Battail and M. Decouvelaere, "Décodage par répliques", *Annales Télécommunic.*, Vol. 31, No. 11–12, pp. 387-404, Nov.-Dec. 1976.

[22] J.L. Massey, *Threshold decoding*, MIT Press: Cambridge, MA, 1963.

[23] C. Berrou, A. Glavieux and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: turbo-codes", in: *Proc. ICC'93*, pp. 1064–1070, Geneva, Switzerland, 1993.

[24] L.R. Bahl, J. Cocke, F. Jelinek and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate", *IEEE Trans. Inf. Th.*, Vol. IT-20, pp. 284–287, Mar. 1974.

[25] G. Battail, "Pondération des symboles décodés par l'algorithme de Viterbi," *Annales Télécommunic.*, Vol. 42, No. 1–2, pp. 31–38, Jan.-Feb. 1987.

[26] J. Hagenauer and P. Hoeher, "A Viterbi algorithm with soft-decision outputs and its applications", GLOBECOM'89, Dallas, Texas, pp. 1680–1686, Nov. 1989.

[27] D.R. Forsdyke, http://post.queensu.ca/ forsdyke/

[28] S.A. Kauffman, *The origins of order*, New York: Oxford University Press, 1993.

[29] O. Milenkovic, "The information processing mechanism of DNA and efficient DNA storage", DIMACS working group on theoretical advances in information recording, 22–24 March, 2004.

[30] G. Battail, "Replication decoding revisited", *Proc. Information Theory Worshop 2003*, Paris, 31 Mar.–4 Apr. 2003, pp. 1–5.

[31] R.F. Voss, "Evolution of long-range fractal correlation and $1/f$ noise in DNA base sequences", *Phys. Rev. Lett.*, Vol. 68, pp. 3805–3808, June 1992.

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

[33] D.R. Forsdyke, "Conservation of stem-loop potential in introns of snake venom phospholipase A_2 genes. An application of FORS-D analysis", *Mol. Biol. and Evol.*, Vol. 12, pp. 1157–1165, 1995.

[34] R. Dawkins, *The Blind Watchmaker*, Longman, 1986.

[35] R.N. Mantegna, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, S. Simons, and H.E. Stanley, "Linguistic features of noncoding DNA sequences", *Phys. Rev. Lettr.*, Vol. 73, pp. 3169–3172, 1994.

[36] D.B. Searls, "The language of genes", *Nature*, Vol. 420, No. 6912, pp. 211–217, Nov. 14, 2002.

[37] E. Chargaff, "How genetics got a chemical education", *Ann. New York Acad. of Sc.*, Vol. 325, pp. 345–360, 1979.