tripleC 4(2): 217-229, 2006
ISSN 1726-670X
http://tripleC.uti.at

Error-correcting codes and genetics

Gérard Battail

Ecole Nationale Supérieure des Télécommunications, Paris (retired). E-mail: gbattail@club-internet.fr

Abstract: The conservation of genetic information through the ages can not be explained unless one assumes the existence of genomic error-correcting codes, our main hypothesis. Shielding by phenotypic membranes does not protect the genomes against radiations and their own quantum indeterminism. The cumulated errors then make the genomic memory ephemeral at the time scale of geology. Only means intrinsic to the genome itself, taking the form of error-correcting codes, can ensure the genome permanency. According to information theory, they can achieve reliable communication over unreliable channels, so paradoxical it may look, provided some conditions are met. The experience of communication engineers witnesses their high efficiency. As a subsidiary hypothesis, we assume moreover they take the form of `nested codes', i.e., that several codes are combined into a layered structure which results in an unequal protection: the older and more fundamental parts of the genomic information are better protected than more recent ones. Looking for how nature implements error-correcting codes, we are led to assume that they rely on the many physical, steric, chemical and linguistic constraints to which the DNA molecule and the proteins for which they code are subjected. Taking account of these constraints enables to regenerate the genome provided the number of accumulated errors remains less than the correcting ability of the code, i.e., after a short enough time.

Based on these hypotheses, fundamental results of information theory explain basic features of the living world, especially that life proceeds by successive generations, the discreteness of species and their hierarchical taxonomy, as well as the trend of evolution towards complexity. Other consequences are that evolution proceeds by jumps and that the genomic message originates in random regeneration errors. That basic results of information theory and error--correcting codes explain biological facts left unexplained by today's biology confirms the necessity of our hypotheses. The direct experimental identification of genomic error-correcting codes and regeneration means still lacks, however, but it would obviously require the active collaboration of practicing geneticists.

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

1 Introduction

The transmission of genetic information through the ages is considered in this paper from the point of view of *communication engineering* and *information theory*. A striking discrepancy is observed between the statement that mutations, seen as errors in the genome replication due to chemical reactants and physical agents like radiations, are mainly responsible for the ageing of living beings, on the one hand, and the main tenet of genetics that genomes are faithfully conserved through the ages at the geological scale, on the other hand. The times involved are immensely different since the lifetime of individuals is extremely short at the geological scale. The next section presents results of simple information-theoretic computations which show that if the error frequency is as high as to make mutations have any noticeable effect at the scale of a lifetime, then the accumulated errors make the genomic communication simply impossible at the geological time scale. This blatant contradiction cannot be solved unless assuming, as we did in (Battail, 1997) and in subsequent works, e.g., (Battail, 2004a), that *error-correcting* codes ensure the conservation of genetic information through the ages as making them intrinsically *resilient to errors*. Then the genome conservation implies a regeneration process and not the mere replication of a template, in plain contradiction with the paradigm currently in force in today's genetics. We shall refer to the assumption that error-correcting codes make the genetic information resilient to errors as our *main hypothesis*. We are moreover led to make the *subsidiary hypothesis* that several genomic error-correcting codes are combined into a layered architecture, to be referred to as `nested codes', which provides an unequal protection in favour of the older and deeper information. Based on these hypotheses, fundamental results of information theory about error-correcting codes suffice to explain several basic features of the living world: that nature proceeds by successive generations; the discrete character of the living world and especially the existence of distinct species; the possibility of their hierarchical taxonomy; the trend of evolution towards complexity. We believe that these features match those of the living world, so our hypotheses can explain very basic properties left unexplained by today's biology. Other consequences of the hypotheses are that evolution proceeds by jumps and that genomic information has a random origin. Probably many others can be found.

2 Genomic error-correcting codes are needed

2.1 Origin of genomic errors

First of all, why do genomes suffer errors? Their integrity is mainly threatened by chemical reactants and radiations. Phenotypic membranes can provide an adequate shielding against chemical agents, but not against radiations of solar and cosmic origin, or due to natural radioactivity. Moreover, DNA is a strange molecule which belongs to the macroscopic world in one of its dimensions but to the submicroscopic world in the other two. It can support a definite information only provided its intrinsic indeterminism as a quantum object is corrected by genomic codes. As strings of nucleotides, their codewords belong to the macroscopic dimension.

2.2 Symbol error probability and genomic channel capacity as functions of time

Any representation or communication of information relies on the use of an alphabet. The basic information-bearing event is the choice of one of its symbols. We define an *alphabet* as a given collection of a finite number *q* of symbols referred to as the alphabet size. These symbols are arbitrary signs or objects which can be unambiguously distinguished from each other, like letters, digits, electric voltages, signal forms, or, in genetics, molecules like nucleotides or amino-acids. The smallest possible size of an alphabet is $q=2$.

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, an event to be referred to in general as a *transition*, or to an *error* when it results in a symbol which differs from the initial one. Let us assume that a given symbol is randomly subjected to error with a probability ν per unit of time. When an error occurs, we assume that the $q-1$ symbols other than the correct one are equally probable. Then it is easily shown that the probability of symbol error, as a function of time, obeys the differential equation $P'(t)=\nu[1-P(t)q/(q-1)]$ where $P'(t)$ denotes the derivative of $P(t)$ with respect to time. If ν is constant, the solution to this equation which satisfies the initial condition $P(0)=0$ is

$$P(t)=[1-\exp(-\nu t q/(q-1))](q-1)/q \,.$$

This probability of error is an increasing function of time which tends to the asymptotic value $(q-1)/q$ as *t* approaches infinity. Notice that the error probability $P(t)$ only depends on time through the product $\nu t=\tau$, a dimensionless quantity which can be interpreted as a measure of time using $1/\nu$ as unit. The average number of errors in a string of *n* such symbols at time *t* is thus $N_e(t)=nP(t)$. Even if the probability of error per unit of time ν is not constant, the average number of errors still increases with time and tends to the same asymptotic value $n(q-1)/q$.

The set of all transitions possibly incurred by a symbol is referred to as the *channel* and its *capacity* is the largest quantity of information it can transmit in the average. Besides the alphabet size, it depends on the transition probabilities. It is the most important information-theoretic measure related to communication in the presence of errors. Computing the channel capacity $C_q(t)$ which corresponds to the cumulated probability of error $P(t)$ above results in a function of τ which decreases from $\log_2 q$ for $\tau=0$, where its slope equals minus infinity, down to 0, *exponentially* for τ approaching infinity. The expression of the channel capacity accounts for its degradation due to the accumulation of errors. Other assumptions than the substitution of a wrong symbol to the correct one can be made, especially if we consider the possible lack of complementarity of a pair of nucleotides in double-strand DNA as equivalent to the erasure of a symbol. Then, the capacity is asymptotically twice larger so it still decreases exponentially down to zero for τ approaching infinity. The derivation and exact expression of the cumulated error probability, as well as that of the corresponding capacity, will appear in (Battail, 2006b).

For applying these results to genomes, the quaternary alphabet {A, T, G, C} having as symbols the DNA nucleotides would seem obviously relevant, but considering 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) maybe better fits reality. Whatever the chosen alphabet, the main fact is the exponential decrease of the channel capacity down to zero which makes it negligible after a time interval of a few times $1/\nu$. The lack of a reliable estimate of the error frequency ν unfortunately forbids a quantitative exploitation of this result. We can just notice that it is currently stated that errors in the genome replication are responsible, at least partially, for ageing and diseases like certain cancers. That these disorders occur within the lifetime of individuals clearly shows that ν is rather large, so $1/\nu$ is small at the geological time scale. The genomic channel thus becomes completely inefficient at the time scale of geology.

The fundamental theorem of channel coding asserts that, within precise limits to be stated below, *errorless communication* can be performed despite the occurring errors thanks to *error-correcting codes*. The engineers's experience in implementing such codes in countless applications practically checks the validity of this paradoxical theoretical statement. Since the genome loses more and more information about the original message as time goes, due to the cumulated errors, it can be conserved only provided specific processes perform its regeneration, using means which must rely on genomic error-correcting codes. Then the genomes are endowed with the necessary property of *resilience to errors*. Due to the fast degradation of the genomic message, it must be regenerated after a time interval as small as to avoid that the genomic channel capacity becomes degraded beyond the code correcting ability. It is why nature proceeds with successive generations. 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 nonrandom genomic features which needs to be explained. We shall develop this remark below (Sec. 6) 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 believed obvious.

2.3 A short introduction to error-correcting codes

For introducing error-correcting codes, let us 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 *n*-symbol words is q^n. It is very convenient to represent an *n*-symbol word as a point in an *n*-dimensional space, having the *n* symbols of the word as coordinates. For instance, if $q=2$ (with symbols denoted by 0 and 1) 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 representation to an n-dimensional space with $n>3$. We may define the *Hamming distance* between two words as the number of coordinates where they differ. For instance, for $n=7$, the distance between 1101000 and 0110100 is 4. We refer to the space endowed with this distance measure as the n-dimensional Hamming space. The minimum distance between any two different n-symbol words is only 1 since they may differ in a single coordinate. An error-correcting code is a subset of all possible n-symbol words such that the minimum Hamming distance d between any two words of the code, referred to as the *minimum distance* of the code, is larger than 1, so it 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*. For instance, if $n=3$ we may define a code made of all words having even weight, the weight of a word being defined as the number of symbols `1' it contains. Its codewords are: 000, 011, 110 and 101 and 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. For larger values of the codelength n, the minimum distance d can be made the larger, the smaller the code rate $R=k/n$ (i.e., the larger the code redundancy) and, of course, the larger is n.

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 over a channel where an error 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 and which moreover is at a Hamming distance from it equal to the number of errors which occurred, say e, 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 given 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 an error pattern is the more probable, the smaller its weight. There is no loss of generality in assuming $p<1/2$ since the labelling of the received symbols by `0' or `1' is arbitrary, thus it can always be chosen such that this inequality holds provided p is different from 1/2. The case $p=1/2$ is not relevant since it is equivalent to the absence of any channel.

The rule for most probably recovering the transmitted word is thus: *choose the word of the code the closest to the received word*. Its mere statement enables 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 Hamming 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.

The above *regeneration* rule succeeds in recovering the actually transmitted word if the weight of the error pattern e is less than half the minimum distance of the code: a sufficient condition of the regeneration success is $e<d/2$. Indeed, if this inequality is satisfied, the point which represents the received word (i.e., affected by errors) is closer to the actually transmitted one than to any others. If e is equal to or larger than $d/2$, the regeneration can fail with a probability which increases and tends to 1 as e grows beyond $d/2$.

2.4 Error-free communication is possible in the presence of errors

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, 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' of the *n*-dimensional sphere of radius *np* centred on the transmitted word (this surface is an $(n-1)$ -dimensional volume). If the radius np is smaller than half the minimum distance *d* of the code, then clearly the received word is with high probability closer to the 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 is often closer to the received word so the regeneration rule above generally fails and, as the word length *n* approaches infinity, the probability of a regeneration error approaches 1. We thus have a regeneration rule which fails with very low probability if $p < d/2n$, but with very high probability if $p > d/2n$. The transition between the two behaviours is the sharper, the larger *n*. 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 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 up to now a very strong incentive to their continuation.

The problem of designing an optimal error-correcting code using a *q*-symbol alphabet and having *M* words of length *n* has no known general solution for a given channel. 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 for a given value of the rate $R = k/n$. 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 decreases as the channel error probability *p* increases: this limit is the *channel capacity C* already introduced in Sec. 2.2. `Errorless' means that, provided $R < C$, a vanishing probability of error can result from using adequate (but not explicitly specified) codes as their length *n* approaches infinity. The main virtue of random coding is to ensure that, statistically, 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.

It is far less simple, however, if one looks at the decoding side. 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 any *single* received (erroneous) word to be regenerated with *each* of the *M* codewords. 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 length $n = 1{,}000$ and information rate $R = 1/2$, we have $M = 2^{500}$ or approximately 10^{150} . (For comparison, 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 an information rate close to the channel capacity. Most experts believed that finding good codes with a tractable structure was hopeless due to an intrinsic incompatibility of goodness and structure, an opinion summarized in the folk theorem: *All codes are good, except those we can think of.*

It turns out that this opinion was by far too pessimistic. For instance, we 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*. We 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 are known, and widely used in simulation, which mimic truly random sequences.) Codes designed according to this criterion should have a performance close to the optimum. In 1993, soon after the random-like criterion was proposed, the pessimistic opinion quoted above was definitively ruined with the advent of the *turbo codes* (Berrou et al., 1993,1996) (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). Turbo codes actually meet the random-like criterion, although they were not explicitly designed in order to fulfil it (Battail et al., 1993). Their implementation is comparatively simple and well within the possibilities of current technology. Besides being the best codes presently available, turbo codes perform closely enough to the theoretical limit (the channel capacity) to be considered as almost optimal, at least from a practical point of view. The channel capacity then appears as defining the limit of what is possible as regards errorless communication, practically as well as theoretically. It is why we centred our discussion of Sec. 2.2 on this most important parameter of the genomic channel.

3 Natural implementation of error-correcting codes

3.1 Genomes are redundant

The assumption that genomes are words of error-correcting codes implies that they are redundant. In information theory, `redundancy' does not merely mean that several copies of a message 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 4^{15} is approximately equal to 10^{9}, so a genome of about 15 nucleotides would suffice to uniquely specify all past and extant species (of course, according to a rough and disputable estimation), meaning that a genome length of less than 100 nucleotides would suffice to uniquely label each individual within any of the past and extant species. If we compare this figure with the actual length of genomes, even the shortest ones (that of viruses) appear as very redundant. In the absence of redundancy, the number of possible *n*-nucleotide genomes 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.

3.2 Nature uses `nested codes'

The assumption that nature uses genomic error-correcting codes is our main hypothesis. The subsidiary hypothesis that nature uses nested codes should furthermore be made. By `nested codes', we mean a system which combines several codes into a layered architecture. 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 by another code. This process is repeated several times, the last information message being left uncoded. Notice that a very efficient protection of the oldest and most central information does not demand very efficient individual codes: the multiplicity of the codes provides a much higher degree of safety than each of them separately.

The nested codes concept arose from noticing that certain parts of the genome like the *HOX* genes (hence the organization plans of the corresponding phenotypes) are conserved with astonishing faithfulness in many animal species. At variance with this extreme permanency, 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 codes scheme is the simplest way to do so. Moreover, we assumed that the codes appeared successively in time, the genomic information being the better protected, the older it is, so that the variability mainly concerns the most peripheral layers of the nested codes scheme.

3.3 Genomic error-correcting codes as `soft codes'

It would be naïve to expect that the error-correcting codes that nature uses closely resemble those designed by engineers. The latter are defined as a set of words which obey constraints expressed by deterministic mathematical equalities easily implemented using physical devices. 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 defining soft codes. Even linguistic constraints may be considered since the genome describes the construction of a phenotype, which needs some kind of language. We gave in (Battail, 2005) a rather comprehensive list of the potential genomic soft codes which result from the several constraints which the genome obeys, briefly recalled here.

A first kind of soft codes are those which 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 especially its packing in nucleosomes and higher-order structures (when they exist, i.e., in eukaryotes).

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' (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.) Physiologically active proteins 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. Moreover, proteins owe their functional properties to the folding of the polypeptidic chain according to a unique pattern, which implies many chemical bonds (especially disulphur bridges but also weaker ones) 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 role of DNA for specifying proteins, such constraints must be present in any living being.

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 much shorter than the actual genomes, even those of viruses (see Sec. 3.1). This high redundancy has rather obvious reasons: the genome rôle is by no means restricted to identify a living being but it acts as a *blueprint* for its construction 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, the genome must somehow *describe* the succession of operations which results in the development and the maintenance of phenotype. This demands some kind of *language*, involving many morphological and syntactic constraints which may be interpreted as generating soft codes having error-correcting capabilities. Moreover, the linguistic constraints appear at several different levels according to the same structure of `nested soft codes' which we were led to hypothesize for the genetic message. The error-correcting ability of languages is manifest in the ability of the spoken human language to be literally perceived in extremely noisy acoustic surroundings. It turns out that the individual phonemes are identified with a large probability of error, but the linguistic constraints together with the high processing power of the human brain eventually result in errorless communication despite 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. It consists of the correspondence between unidimensional strings of completely different molecules like, for instance, the relationship between triplets of nucleotides and the 20 amino-acids which make up proteins, referred to as the `genetic code'. (Unidimensionality is a common feature of messages in engineering and in genetics. It appears as necessary for unambiguous semantic communication.) Such a correspondence does not result from any physical or chemical law, but can be considered as a pure convention or artifact, just like rules in linguistics or engineering. These rules are maintained thanks to `semantic feedback loops' (Battail, 2005). Barbieri's organic codes moreover assume the nested codes structure (see Fig. 8.2 in (Barbieri, 2003), p. 233).

4 Identification of genomic error-correction means

4.1 Indirect evidence of genomic error-correction codes

4.1.1. Spectral and correlation properties of genomes

The experimental analysis of DNA sequences has shown they exhibit long-range dependence. Their power spectral density has been found to behave as $1/f^{\beta}$ asymptotically for small values of the frequency f, where 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 with the alphabet of nucleotides restricted to the binary one {R,Y} as we did above has shown that their autocorrelation function decreases 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 related to structural constraints of nucleosomes.

4.1.2. Distance properties of eukaryotic genes under evolutive pressure

As early as 1981, Forsdyke suggested that the introns in eukaryotic genes are made of check symbols associated with the information message borne by the exons (Forsdyke, 1981). The literature generally states that introns are more variable than exons, but a counterexample was provided in 1995, again by Forsdyke, 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 assuming that an error-correcting system exists. Interpreted as a regeneration 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 the most usual case, the evolutive pressure tends to its conservation so the erroneous symbols after a mutation 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 erroneous symbols after regeneration will be mostly located in exons and the introns will be almost conserved (Battail, 2004b).

4.2 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. 2.2 and we show in Sec. 5 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 would be highly desirable, but it is still lacking.

4.3 Identifying the regeneration means: an open problem

The problem of genomic regeneration (decoding) is left for future researches. Its principle 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 biological constraints to define genomic soft codes, obeying constraints amounts to correcting errors. We may thus think of the regeneration process as necessarily *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 results in the definition of regeneration from an engineering viewpoint, given in Sec. 2.3. An intriguing feature of regeneration as implementing this rule is that its operation demands that the decoder possesses a full description of the constraints at any level, including those which originate in physico-chemical properties of molecular strings.

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. For a given code, there exist several decoding processes which more or less approximately implement the optimum rule stated in Sec.2.3; the more complex, the closer they are to optimality. Still more than the identification of genomic error-correcting codes, that of the means actually implemented by nature for their regeneration is thus difficult and challenging. Remember that we used above the human language as an example to illustrate the error-correcting properties of soft codes, but the means implemented in the brain for performing this task are presumably very complex and still unknown. Also, it is likely 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.

5 Consequences of the hypotheses

5.1 Nature proceeds by successive generations

That nature proceeds by successive generations is a direct consequence of our main hypothesis. The fundamental theorem of channel coding tells that error-free communication is possible only provided an error-correcting code is used, having an information rate k/n less than the channel capacity (remember that $k=\log_q M$ for an M-word code, as defined in Sec. 2.4). The information rate is constant for a given code. On the other hand, we have seen in Sec. 2.2 that the capacity of the genomic channel is a fast decreasing function of time. Regeneration of the genome must thus be performed before the genomic channel capacity has become too low for enabling the code to correct the occurring errors, and must further be repeated at short enough time intervals.

The average time interval between regenerations should be matched to the correction ability of the code. If this interval is as long as to often exceed the limit set by the minimum distance of the code, then regeneration errors will frequently occur and result in low-permanency phenotypes markedly different from each other. Maybe a phenomenon like the Cambrian explosion could be explained by a mismatch between the average regeneration interval and the actual efficiency of the available error-correcting codes, this interval being too long for ensuring an almost certain regeneration. The genetic factors which control the regeneration interval on the one hand, and the efficiency of the error-correcting codes on the other hand, are probably quite independent so their matching, as an evolutive advantage, could only result from the Darwinian selection. Also, the recent finding in *Arabidopsis thaliana* of `non-Mendelian inheritance' (Lolle et al., 2005) could be explained by assuming that, in this species and maybe in other plants, the regeneration process does not systematically coincide with the genome replication, but is sporadically triggered by some kind of `stress'.

5.2 Discreteness and hierarchical taxonomy of the living world

The hypothesis that genomic error-correcting codes exist immediately implies that the genomes are far from each other in terms of the Hamming distance. This would be obvious in the case of a single code having a large minimum distance *d*. Then genomes are in the Hamming space at this distance from their closest neighbours, which implies the existence of distinct species (as opposed to chimeras).

The picture becomes more complicated but more realistic when we take into account our subsidiary hypothesis. Every time a new encoding has been performed in the process of constructing the nested codes system, the minimum distance between the points representing the previously encoded words has been enhanced by the minimum distance of the new code. Once this system is built (and in fact simultaneously to its construction), 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, the more central the layer to which they belong. A large distance implies that the corresponding regeneration error pattern has a large weight, thus gives rise to a phenotype more different from the original than an error pattern of smaller weight. (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.) Clearly, besides the discreteness of species which results from the main hypothesis, the layers of the nested codes system delineate a hierarchical taxonomy among them which results from the subsidiary hypothesis.

5.3 Trend of evolution towards complexity

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, which provides an *immediate* evolutive benefit. The hypothesis of a nested codes structure is not even necessary to explain the trend of evolution towards complexity. It actually appears as a mere consequence of the rather paradoxical information-theoretic fact that the longer the code, the smaller can be made the regeneration error probability. Hence increasing the genome size can result in increasing its permanency. 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 role in the genome conservation), then we may think that increasing the genome length results in diminishing the probability of a regeneration error, hence increases its permanency. Moreover, 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, after being filtered by natural selection, better fitted) phenotypes. Indeed, although information theory ignores semantics, information can be thought of as a *container for semantics*: the availability of more information enables to specify more phenotypic features. As regards the epistemological status of information, we believe that it has no existence without a *physical* support and that it acts as a container for *semantics*. We may thus think of information as a *bridge* between the concrete and the abstract: genomes are *concrete* objects which bear the *abstract* recipe for developping and maintaining concrete phenotypes.

5.4 Answering debated questions

5.4.1. Evolution is saltationist

The hypothesis that the genomes behave as words of error-correcting codes, hence are distinctly far apart in the Hamming space, entails that mutations resulting from regeneration errors change genomes into distinctly different ones, so evolution proceeds by jumps.

5.4.2. Genetic information has a random origin

The accumulation of errors tends to make the genomic message less and less dependent on the original one (see Sec. 2.2). If an error-correcting code is present, the genomic message is exactly regenerated provided the correcting ability of the code is not exceeded and only varies when a regeneration error occurs. Such an event is very unfrequent but results in a burst of at least *d* erroneous symbols when it occurs (*d* denotes the minimum distance of the genomic code). The genomic code then ensures the conservation of this `wrong' genome exactly as it does for the initial `correct' one. Without error-correcting properties, the genome would gradually become less and less dependent on the original genome due to the occurring errors. If endowed with error-correcting properties, 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 ultimate consequence that the original genomic message is progressively forgotten, but according to a much slower pace, depending on the time interval between regenerations, when error-correcting means are used. 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 a time interval which depends only on the probability of a regeneration error. Each regeneration error may be thought of as generating 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). In contrast, an uncoded system would give rise to a world of *chimeras*, not of discrete species. Another important consequence of our hypotheses is that all the extant genomic information originated in 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.

5.5 Further comments

Although one may deem that the results of this section are speculative as relying on hypotheses, the theoretical impossibility of genome conservation without error-correcting means makes these hypotheses necessary, and so are their consequences. 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.

6 Conclusion: the genome conservation is a dynamic process

To conclude, let us stress the epistemological importance of the results presented at the beginning of this paper (Sec. 2.2). In the presence of a steady error frequency, they show that the average number of errors which would affect the genome if it were uncoded is an increasing function of the time elapsed which makes the genomic capacity, in the information-theoretic sense, decrease exponentially fast down to zero. Conservation of the genomic information thus demands that the genome be endowed with an error-correcting code and moreover that it be regenerated after a time interval such that the number of accumulated errors is very unlikely to exceed the error-correction capability of the code. Then, rather paradoxically, *conservation* of the genome appears as a *dynamic* process. (Lewis Carroll's Red Queen is here an illuminating metaphor.)

That the conservation of genetic information, far from being systematically secured, results from a dynamic process appears as a total reversal with respect to the traditional point of view. It leads to ask new questions. For instance, a usual argument is that that part of the genome which specifies how the phenotype is built and maintained is conserved because natural selection precisely has the phenotypes as targets. But, at least in many eukaryotes, only a small fraction of the genome is known to have this role. The remainder has no known function and is often dubbed `junk DNA'. How is ensured its conservation? Even short motifs repeated a large number of times, which bear negligibly few information, are conserved. How and why it is so should be understood. What needs indeed to be explained is why any portion of the genome departs from being purely random. While the conventional genetics tacitly assumes that the genome conservation is the rule, the genome would fast become random, hence devoid of any structure, unless specific means ensure its conservation. All the symbols of a codeword both participate in and benefit from error correction. Any departure from randomness in DNA can be accounted for only by error-correction mechanisms. Hence the conservation of `junk DNA', as superfluous as it may look, means that it benefits from error correction, which also implies it has an active role in it and thus should not be qualified as `junk'.

Things we can observe are those which are conserved. One may wonder why the consequences of such an obvious statement have been (and still are) so often overlooked. Within the human time scale permanency seems to be an intrinsic property of macroscopic objects but within the time scale of geology this generally does not remain true. Permanency is erroneously believed to be a trivial property of things, maybe as an unthought extension of our daily experience. At the geological time scale, conservation is however the exception and not the rule. According to Darwin, conservation of a living thing depends on its ability to get food, escape predators and pathogenic agents, and, as a species, to reproduce itself. Stressing the importance of the genome (remember that its very existence was unknown to Darwin), modern neo-Darwinians made a step more in interpreting the phenotype as subordinate to ensuring the genome conservation by shielding it against physical and chemical aggressions. We still make a step in the same direction in pointing out that the message borne by the genome should moreover contain means for its own conservation, in the form of intrinsic error-correcting codes which extend the genome protection to other kind of error-inducing agents, especially radiations. Indeed, the faithful conservation of a DNA molecule, a submicroscopic object in two of its dimensions, is not conceivable at the time scale of geology without intrinsic means which ensure it. It is definitely impossible to deal with it as if it were a permanent solid body and the conventional paradigm of template replication is wrong.

The above discussion made the genome memory appear as ephemeral at the geological time scale in the absence of corrective means, which implies that the genome conservation is not the rule and needs to be ensured by a dynamic process. Among the many questions which arise from this statement, the problems of identifying genomic error-correcting codes and the means for genome regeneration have to a large extent been left open. We would like to emphasize that no progress can be expected in these directions unless geneticists get interested, and even educated, in information theory (Yockey, 2005; Battail, 2006a). Needless to say, we think that rejecting information theory as did a majority of biologists many years ago (arguing that its concept of information is too restrictive, especially as ignoring semantics) was just throwing out the baby with the bathwater. We believe that information theory can be an extremely useful conceptual tool not only in genetics (as we tried to show it above) but in biology as a whole, provided a much closer collaboration of information engineers and biologists can be set up. The wish that it be so concludes the paper.

References

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

Barbieri, Marcello (2003) *The Organic Codes*, Cambridge University Press, Cambridge, UK.

Battail, Gérard (1989) *Construction explicite de bons codes longs.* In: Annales des Télécommunic., Vol.44, No. 7-8, pp. 392-404.

Battail, Gérard (1996) *On random-like codes*. In: Lecture Notes in Computer Science, Springer, No. 1133, pp. 76—94.

Battail, Gérard (1997) *Does information theory explain biological evolution?* In: Europhysics Letters}, Vol. 40, No. 3, pp. 343-348

Battail, Gérard (2004a) *An engineer's view on genetic information and biological evolution*". In: Biosystems, Vol. 76, No. 1-3, pp. 279-290.

Battail, Gérard (2004b) *Can we explain the faithful communication of genetic information?*, DIMACS working group on theoretical advances in information recording, 22-24 March.

Battail, Gérard (2005) *Genetics as a communication process involving error-correcting codes*. In: Journal of Biosemiotics, Vol. 1, No. 1, pp. 103-144.

Battail, Gérard (2006a) *Should genetics get an information-theoretic education?* In: IEEE Engineering in Medicine and Biology Magazine, Vol. 25, No. 1, pp. 34-45.

Battail, Gérard (2006b) *Information theory and error-correcting codes in genetics and biological evolution.* In: Introduction to Biosemiotics, Ed: Barbieri,Marcello, to be published.

Battail, Gérard, Berrou, Claude, and Glavieux, Alain (1993) *Pseudo-random recursive convolutional coding for near-capacity performance*. In: Proc. GLOBECOM'93, Communication Theory Mini-Conference, Vol. 4, pp. 23—27.

Berrou, Claude, Glavieux, Alain, and Thitimajshima, Punya (1993) *Near Shannon limit error-correcting coding and decoding: turbo-codes*. In: Proc. of ICC'93, Geneva, Switzerland, pp. 1064—1070.

Berrou, Claude, and Glavieux, Alain (1996) *Near optimum error correcting coding and decoding: turbo codes*, IEEE Trans. on Communications, Vol. 44, pp. 1261—1271.

Forsdyke, Donald R. (1981) *Are introns in-series error-detecting sequences?* In: J. Theor. Biol., Vol. 93, pp. 861—866.

Forsdyke, Donald R. (1995) *Conservation of stem-loop potential in introns of snake venom phospholipase* A_2 *genes. An application of FORS-D analysis*. In: Mol. Biol. and Evol., Vol. 12, pp. 1157-1165.

Guizzo, Erico (2004) *Closing in on the perfect code*. In: IEEE Spectrum, Vol. 41, No. 3 (INT), pp. 28-34.

Lolle, S.J., Victor, J.L., Young, J.M., and Pruitt, R.E. (2005) *Genome-wide non-mendelian inheritance of extra-genomic information in Arabidopsis*. In: Nature, Vol. 434, No. 7032, pp. 505-509.

Shannon,Claude E. (1948) *A mathematical theory of communication.* In: BSTJ, Vol. 27, pp. 379-457 and 623-656.

Voss, R.F. (1992) *Evolution of long-range fractal correlation and* $1/f$ *noise in DNA base sequences*. In: Phys. Rev. Lett., Vol. 68, pp. 3805-3808.

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