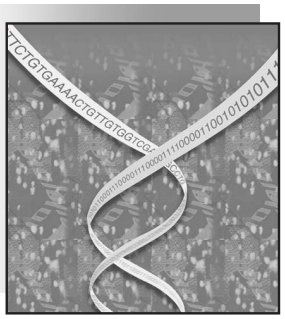

Should Genetics Get an Information-Theoretic Education?

Genomes as Error-Correcting Codes

BY GÉRARD BATTAIL

The main contribution of engineering to biology and medicine has mainly been as yet ancillary, e.g., providing instrumentation in fields like imagery and assisted diagnosis, which enables investigating reality far beyond the range accessible to the senses alone, thus widely improving the vision and control that biologists, physicians, and surgeons can have on living things. We shall outline here another potential contribution of engineering that is completely different and has few precedents, that of providing a theoretical framework and conceptual tools to biologists. This article is intended to show that engineering concepts can help account for the prominent role of information in life phenomena.

Living beings are open systems which exchange matter, energy, and information with the outer world, with each others, and in the process of their own operation. They receive information from the outer world, and information circulates inside themselves and between individuals at any scale, from the molecular level to that of ecosystems and beyond. Moreover, they are constructed and maintained using the genetic information they receive from their ancestors since the very beginning of life, some 3.5 billion years ago or maybe even earlier. In its historical development, biology slowly integrated the sciences of matter and energy, namely chemistry and physics, which enabled it to go far beyond the mere description of the living world, thereby acquiring a truly scientific status. Last but not least, the less obvious importance of information in the life phenomena has been recognized much later than that of matter and energy. Modern biology assigns an ever increasingly important role to information, but the science relevant to it, information theory, has not been integrated into biology yet as were chemistry and physics much earlier. It is indeed a much younger science since its birth can be dated to 1948 when Shannon's papers were published [1]. Its strong connection with communication techniques, its mathematical formalism, and some conceptual difficulties made its impact on other sciences rather limited. For several decades, it looked as a rather abstract matter with little impact even on communication techniques for its lack of proper implementation means. However, with the tremendous development of the semiconductor technology (which also started in 1948!), the applications of information theory are by now countless. As a striking example, mobile telephony would simply not exist without information theory and the coding techniques it generated. The many succcessful applications of information theory now provide strong experimental proof of its validity in the field of communications. Although it became the most powerful conceptual tool avaible in this field, it remains almost unknown outside the communication engineers' community. It shaped much of the present way of living, but few people realize this even though everybody makes daily use of its engineering products.

That information theory is largely unknown in the biologists' community does not mean that the importance of information in the living world is overlooked. An increasing number of biological papers are devoted to many forms of information recording and transfer, and the word *information* has become ubiquitous in the biological literature. However, it is used most often with a loose meaning. Many biologists seem to ignore that the scientific concept of information generated a science having reached maturity. We believe that no real progress will result from the recognition of the prominent role of information in life phenomena unless information theory is integrated into biology as physics and chemistry had been.

Among the many domains of biology where information plays a prominent role, we shall restrict ourselves to the communication of genetic information through the ages. Starting from the very fundamental question How is genetic information faithfully communicated?, we hypothesize that nature developed error-correcting codes since the origin of life. As powerful tools available to communication engineers, error-correcting codes are, paradoxically, reliable communication over unreliable channels. Some of the most important results of information theory are statements about these codes, especially concerning the attainable limits of their performance. It turns out that aside from answering the above question, our hypothesis also sheds light on the process of biological evolution and on the structure of the living world. We first formulated it and discussed its biological impact in [2]. Subsequent papers were devoted to refine this hypothesis and to better understand its consequences. We also tried to identify the error-correcting means involved and to understand how they are implemented [3]–[5]. Besides the interest of this topic of its own, we also think of it as exemplifying

 0739-5175/06/$20.00©2006IEEE

the mutual benefits that a collaboration between information theorists and biologists could provide.

Genetics was at its beginnings a rather abstract science. Later, the discovery by Avery et al. [6] that DNA is the bearer of genetic information, and the subsequent discovery of its double-helix structure by Franklin, Watson, and Crick [7], [8], gave a chemical content to the concept of gene. In 1979, Chargaff wrote a paper entitled "How Genetics Got a Chemical Education" [9], where he complained that genetists were so reluctant to accept the consequences of the discovery that DNA was the actual bearer of genetic information (35 years earlier) that a more appropriate title of his paper could have been "How Genetics Refused to Get a Chemical Education." On the interrogative mode, the title of this article is an allusion to Chargaff's. Besides the chemical structure of DNA, we believe it is time to consider the abstract framework of information theory as appropriate to genetics.

There is another reason why this title is relevant. It stresses the need for an education; that is, genetists should make an effort to assimilate a topic that is rather foreign to their traditional culture. A superficial knowledge of the *results* of information theory does not suffice. Only a deep enough understanding of the *topic* itself, including its paradoxes, can be fruitful since adaptation is required before information theory can be of any use in genetics (and, more generally, in biology). The problems of terminology then become of paramount importance, since the same words are often used with different meanings in genetics and in information theory. As an example, the "genetic code" is not truly a code in the information-theoretic sense. Extreme attention must therefore be be paid to the property of vocabulary.

The Faithful Communication of Genetic Information: A Crucial Question

A Model of Genetic Communication

When we first attempted to study evolution at the light of information theory, we found that high-quality popularizing books dealing with genetics and biological evolution, especially those authored by Dawkins [10], [11], contain a very simple model of genetics and biological evolution: the genome is made of deoxiribonucleic acid (DNA), a long unidimensional polymer bearing nucleic bases (or nucleotides), 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. The genome can be replicated. Each genome is housed inside a phenotype which shields it against outer perturbations which would destroy it if left unprotected. The development of the phenotype is controlled by the DNA message itself, which directs a succession of protein syntheses through the "genetic code" (we use quotes here and in the sequel as a reminder that this is a mapping rather than a code in the engineering meaning). This succession of protein syntheses results in the construction of a phenotype through an extremely complicated and still poorly understood process. The phenotypes are subject to natural selection, so the only remaining genomes host surviving phenotypes.

We may thus think of the communication of genetic information through the ages as a recording and copying process. An initial written message has been copied several times, its copies themselves have been copied, and this process has been repeated. It reminds how texts written in antiquity were made available to us thanks to generations of monks. In a sense, however, this metaphor is misleading. Ink strokes on parchment are macroscopic objects involving a huge number of molecules so they may be expected to strongly resist degradation. Contrasting with any man-made memory element, the bearer of the genetic message is a single DNA molecule. Belonging to the submicroscopic world, it may be thought of as highly vulnerable to degradation by mechanical, chemical, and radiative agents, and since it is relevant to quantum physics, it can be described only in terms of probabilities. Contrary to any expectation, the genetic message has however unmatched longevity since for instance the *HOX* genes which determine the organization plan of living beings are shared by, e.g., humans and flies, which diverged from a common ancestor hundreds of millions of years ago. Explaining this very paradoxical longevity, which is the cornerstone of molecular genetics, has been our major goal.

DNA should be shielded by membranes, rather obviously, because a protection against mechanical constraints and chemical reactants is necessary. But radiations of cosmic or solar origin, or due to natural radioactivity, are pervasive threats against DNA integrity. At a still more fundamental level, the DNA molecule is a quantum object which cannot bear a precisely defined message unless it is protected against its own indeterminism. Protection against radiations and indeterminism cannot be provided by outer devices like a membrane but must be intrinsic to the genome itself. As a consequence, the idea that the phenotype is the sole target of natural selection is not tenable. The genome must itself be subject to it with regard to its capacity to resist errors (here we encounter a statement formulated by some biologists who introduced the concept of "genome phenotype," a seeming oxymoron [12]). The problem of natural selection should be restated to include the hypothesized existence of an error-correcting system intrinsic to the genome.

Reformulating the Model

When a phenotype is destroyed in the process of natural selection, the genome it hosts disappears. But a genome also ceases existing if an error transforms it into another one. Therefore, survival of any genome implies that its replication be as reliable as possible, aside from it hosting a well-fitted phenotype.

Although they properly recognized the importance of maintaining the genome integrity in the evolutive success of a species, biologists did not realize how difficult it is to perform at the scale of geological times. Dawkins wrote about the needed accuracy of replication [10], pp. 16–17:

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

He expresses his astonishment about such a high reliability but does not question the ways to obtain it nor the consequences which may result from their use. Communication engineers know that the answer lies in the use of error-correcting codes and that the price to pay for correction ability is redundancy, which should be high enough to make the transmission rate less than the channel capacity, a fundamental limit set by

The word information has become ubiquitous in the biological literature.

information theory. It is why we suggested that information theory could be relevant to biological evolution [2].

Resuming in [3] the argument of [10], p. 24, we considered the two antagonistic properties of fecundity and permanency as being both beneficial to the genome conservation. *Fecundity* refers to the rate at which the genome replicates itself. We named *permanency* the ability of a genome to remain strictly identical to itself whether it has been replicated or not, thus combining the two properties of longevity and copying-fidelity that Dawkins distinguished. The material for making copies of the genomes is available in limited quantity so only the most successful ones in maintaining and replicating themselves will survive. This is the most basic form of natural selection. The most abundant, hence eventually surviving, genomes will then be those which optimally combine permanency and fecundity. Clearly, the balance between these two properties can be very different from one species to another, and the selective success can mainly rely on a high fecundity (as for viruses and bacteria) or on a high permanency (as for very complex beings like animals and plants).

For a given replication mechanism, the shorter the genome, the higher the fecundity. However, the genome size has a lower limit because it must specify the machinery and process for its own replication (we exclude here the smallest genomes, those of viruses, which are devoid of a replication machinery of their own and must use that of host cells to replicate themselves). The attainable survival ability, which results from fecundity alone, therefore, has an upper limit. Simultaneously starting the replication process at several places would probably not successfully increase fecundity for short genomes, because complicating the replication process would need a longer genome for specifying its machinery, with a presumably detrimental effect on the overall replication speed. Only already long genomes could benefit from such a strategy, insofar as the genome part devoted to specify the replication machinery is small with respect to the remainder.

If error-correcting codes are used in the genome replication process, they are themselves products of evolution. Due to their key role in the genome conservation and the efficiency of the Darwinian process, we may safely assume that natural selection made them evolve to come very close to the limit of what is possible, so the error correction means that we hypothesized are endowed with the properties of theoretically optimal codes, hence exhibiting the seemingly paradoxical behavior of a decreasing error rate as the codeword length increases (see below). Then, increasing the length of a genome increases its permanency. Moreover, no upper limit is set to the genome length, so increasing it appears as a way for a genome to endlessly enhance its ability to survive natural selection.

To summarize, *permanency* as just defined measures the ability of a genomic message to survive in its physical, chemical, and biological environment, including that of resisting radiations and its own indeterminism. Therefore, the survival of a genome does not only depend on the ability of the phenotype it hosts to exploit its physical and ecological environment, escaping predators and resisting pathogenic agents, but also, and more fundamentally, on the error-correcting means it developed. This just generalizes the concept of natural selection to encompass the most pervasive and ubiquitous threats to the genome integrity. At variance with the traditional view that the phenotype is the sole target of natural selection (as in Dawkins's model recalled above), this point of view extends the concept of natural selection to the genome itself.

Aging, Mutations, and Variability of Genome Conservation

A very strong argument for the need of genomic error-correcting means (perhaps the most convincing one) is the fact that mutations, i.e., errors in the genome replication due to chemical agents or radiations, are responsible for aging and certain diseases like cancers. Had the error rate in communicating genomic information noticeable effects at the scale of the lifetime of an individual, the accumulation of errors during periods million times longer would simply make genetic communication—hence life—impossible.

Moreover, if we look at the litterature on chromosomes and cellular division and the literature on the performance of DNA replication, 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 mechanisms needed for correcting replication errors actually exist. 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, many "proofreading" mechanisms are known. However, they can at best ensure that the copy is faithful to the original. In other words, they can correct the errors which occur within the replication process but not those that may affect the original itself. Faithful copying thus does not adequately describe the function of replication. The needed property, which only error-correcting codes can provide, is resilience to casual errors.

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 greater $10^{(-3)}$ per year and even more, which amounts to a rate per replication some hundred times larger than that of higher beings) for some genes of viruses and bacteria. This large 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. And this assumption itself is consistent with the difference of the corresponding genome sizes and the result

of information theory alluded to above that the longer the encoded message, the more efficiently errors can be corrected.

Although a high longevity is an advantage for a particular genome, replication errors are necessary for the evolution process as a whole. They should be as rare as possible in order to keep the identity of a particular genome, but they play a major role in evolution since each error which changes the information borne by the genome generally results in a change in the corresponding phenotype, referred to as a *mutation* (except for "synonymous" ones which transform a codon into another one, which specifies the same amino acid). One may think of mutations as randomly exploring the set of possible phenotypes, the chosen phenotypes being further filtered by natural selection.

We developed the basic ideas originally expressed in [2] into two main directions: first, analyze the consequences of the hypothesis that error-correcting codes are involved in the genome replication process and compare them with known features of the living world; second, try to identify the error-correcting mechanisms that nature implements. Another goal was to convince genetists that information theory and error-correcting codes could be useful to their discipline. Indeed, little progress in the two directions mentioned above could be expected without the active collaboration of biologists, especially as regards experimental works which are crucially needed in order to validate or refute hypotheses. This goal is far from being reached, and it is why the research presented here remained as yet speculative.

As regards the first direction, no contradictions were found between the hypothesis that natural genomic error-correcting means exist and the properties of the living world. On the contrary, it seems to account for a number of facts, especially of evolution, that conventional theories fail to explain. The subsidiary hypothesis that nature uses nested codes (to be defined below) moreover explains that a hierarchical taxonomy is possible. In the second direction, the concept of *soft code,* which both weakens and widely extends the engineering concept of error-correcting code, also to be defined below, enables associating error-correcting properties with constraints on DNA and proteins, thus suggesting that many potential genetic error-correcting systems actually exist. The problem becomes having a precise understanding of whether and how they are implemented, i.e., how the dependence induced by the constraints between the nucleic bases is actually used to regenerate the genome. (Incidentally, we found that in the absence of an explicit encoding process, *regeneration* better describes what in engineering terms is referred to as *decoding*.) Before we develop these topics, we must give an insight on error-correcting codes: what they are, how they work, and what their main properties are.

An Outlook on Error-Correcting Codes

Introduction and Geometrical Representation

Error-correcting codes appear in the engineering literature as a highly mathematical topic, which gives little hold on intuition. However, we believe that their understanding does not need a big mathematical apparatus, as examplified by [13], which describes the most successful codes yet known, the turbo codes, in nonmathematical terms. We now try to state the principles behind error-correcting codes in simple words, using a geometrical representation that is very helpful in getting an intuitive insight and that is also mathematically rigorous. We shall also provide an introduction to the concepts of regeneration and soft code that are used below.

Let us first define an alphabet of size q as a collection of q different symbols which may be any signs or objects that can be unambiguously distinguished, like letters, digits, electric voltages, signal forms, molecules, etc. The smallest size of an alphabet is $q = 2$, and the main properties of codes can be understood if we assume, as we shall do most often in this section, that the alphabet is binary with its symbols denoted by 0 and 1.

Let us now define a word of length n as a sequence of n symbols from an alphabet of size q. Each of its symbols can assume q distinguishable values, so the total number of possible different words is q^n (2^n in the binary case). It will be very convenient to interpret an n-symbol word as defining a point in an n-dimensional space, each of its coordinates being one of the n symbols. For instance, if $q = 2$ and $n = 3$, there are $2^3 = 8$ different possible words, each representing a vertex of a cube. The useful values of n are much larger, but there is no difficulty in extending this definition to $n > 3$. Inside this n-dimensional space, we may define the Hamming distance d between two words as being the number of coordinates where their symbols differ. For instance, if $n = 7$, the distance between 1101000 and 0110100 is $d = 4$. An error-correcting code is a subset of all possible n-symbol words such that the minimum distance between any two of its words is larger than 1. Two n-symbol words may differ in a single coordinate, so an error-correcting code is a strict subset of the set of all n-symbol words. The property that no n-symbol word belongs to the error-correcting code is referred to as *redundancy*. In the case where $n = 3$, we may define a code as containing only words with an even number of symbols "1" (of even weight), namely, 000, 011, 110, and 101. The minimum distance between two of its words 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.

In a communication system using an error-correcting code, only words belonging to this code may be transmitted. As an example, consider a binary code 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 is at a Hamming distance from it equal to the number of errors which occurred, say e, to be referred to as the *weight* of the error pattern. For a binary symmetric channel, i.e., where an error occurs with a constant probability $p < 1/2$, independently, on each symbol of the word, the probability of an error pattern of weight e is simply $P_e = p^e(1-p)^{(n-e)}$ which, for $p < 1/2$, is a decreasing function of its weight e. (Assuming $p < 1/2$ does not restrict generality, since the labeling of the received symbols by 0 or 1 is arbitrary, so it can be chosen such that this inequality holds.) In order to determine the codeword which has most probably been transmitted, we may use as a rule: Choose the codeword the closest to the received word. This rule is expressed in very simple geometrical terms thanks to the definition of a distance in the n-dimensional space; its implementation will be referred to as *regeneration*.

The mere statement of this rule enables us to understand the most important properties that an error-correcting code must possess in order to be efficient. Its words must be far

from each other, so they should be very few as compared with all possible n-symbol words (its *redundancy* should be high). But the words should also be as evenly distributed in the n-dimensional space as possible, since any concentration of codewords would reduce their mutual distances with respect to the case of a more even distribution. For a given amount of redundancy, endowing a code with this property is by far the most difficult task in the design of an error-correcting code, although its necessity is quite intuitive and its statement is easy.

Errorless Communication Is Possible Over a Noisy Channel

It was convenient in the above examples to consider small values of the word length n. Let us now go to the other extreme and assume that n is very large. Then, the law of large numbers tells that the weight of an error pattern is very probably close to its average, namely np (in other words, the frequency of errors measured in a large sample is with high probability close to the error probability). In geometrical parlance, the received point is with high probability close to the "surface" [an $(n-1)$-dimensional volume] of the n-dimensional sphere of radius np centered on the transmitted word. If the radius np is smaller than half the minimum distance d between any two words (simply referred to as the *minimum distance* of the code), then clearly the received word is with high probability closer to the truly transmitted word than to any other, so the above regeneration rule succeeds with high probability. Moreover, 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 and the regeneration rule above generally fails. As the word length n approaches infinity, the probability of a regeneration error approaches 1. The regeneration rule thus fails with low probability if $p < d/2n$ but with high probability if $p > d/2n$. The transition between the two behaviors 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 in itself 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. We hypothesize that the faithful communication of genomic information precisely uses this possibility, with the genome replication actually consisting of its regeneration as just described.

Designing Optimal or Nearly Optimal Error-Correcting Codes

No general solution is known to the problem of designing an optimal error-correcting code for arbitrary values of n, p, and the alphabet size q, so the search for such a code may look hopeless. It is, however, possible to approximately (exactly as n approaches infinity) solve a closely related problem. In geometrical terms, choosing M points at random within the n-dimensional space, M an arbitrary integer, results in a code close to the optimum, regardless of the channel error probability p. Shannon used such random coding in the proof of the fundamental theorem of channel coding [1], which asserts that "errorless" communication is possible if, and only if, the information rate R is less than a limit referred to as the *channel capacity* C. The information rate is defined as $R = (\log M)/n$, where M is the number of codewords, and the logarithms are to the base q. The capacity C depends on the channel error probability. (The definition of the information rate follows from the fact that, without redundancy, q^k different k-symbol messages can be written with an alphabet of size q, so the availability of M codewords is equivalent to that of all k-symbol messages, with $k = \log M$. Little generality is lost if we assume that k is an integer. The redundancy rate is defined as $1 - R$.) For instance, the capacity of the binary symmetric channel considered above is $C = 1 + p\log p + (1-p)\log(1-p)$, where the logarithms are to the base 2. "Errorless" means that, provided $R < C$, a vanishing probability of error can result from using adequate (but not explicitly specified) codes as n approaches infinity. 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*, therefore, to be solved and, moreover, in an unexpectedly simple way.

It *seems*, but it is far from being so because a formidable problem remains. Remember that implementing the regeneration rule above implies to find the codeword the closest to the received word. In the absence of any structure, a code is an arbitrary set of M n-symbol words. There is no other way for implementing the rule than to compare *each* of the M codewords with the single received word to be regenerated. The problem is that for useful values of the codeword length (i.e., n) that are large enough to make the probability of a regeneration error small enough, M is huge. For example, a binary code with $n = 1,000$ and $R = 1/2$ contains $M = 2^{500} \approx 10^{150}$ words. 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 regeneration complexity.

A large number of codes and code families having a strong mathematical structure were invented, and the literature on such error-correcting codes is plentiful (see, for instance, [14] and the impressive bibliography it contains). However, the results obtained were invariably far from the promise of the fundamental theorem of channel coding. Most experts believed that finding good codes having a tractable structure was hopeless due to an intrinsic incompatibility of goodness and structure. This widely shared opinion was summarized in the folk theorem: All codes are good, except those we can think of.

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 dogma and suggested that a better criterion could be to look for randomlike codes with the distribution of distances between their words close, in some sense, to that of random codes (regardless of their actual minimum distance) but constructed according to a deterministic process [15], [16]. (Analogously, easily generated pseudorandom sequences, which mimic random sequences, are known and widely used in simulation.) Codes designed according to this criterion should have performance close to the optimum.

Soon after it was proposed, in 1993, the pessimistic opinion above was definitively ruined with the advent of turbo

codes [13], [17], [18]. Turbo codes actually meet the randomlike criterion, although they were not explicitly designed to this end [19]. Their implementation is comparatively simple and well within the possibilities of current technology. Besides being the best codes presently available, turbo codes perform so close to the theoretical limit (the channel capacity) as to render them almost optimal, at least from a practical point of view.

Introducing Soft Codes

It would be naïve to believe that error-correcting codes of natural origin would closely resemble those produced by human engineering. We think that they should be more flexible and versatile than man-made codes. We propose to both weaken and extend the concept of error-correcting code to better fit the specific needs of genomic error correction.

Broadly speaking, there are two alternative ways for specifying an error-correcting code. First, give a construction rule that associates with any k-symbol message an n-symbol word, with $n > k$ to provide the necessary redundancy. Second, define constraints that are exclusively satisfied by the words of the code. Again, imposing constraints restricts the code to a subset of all n-symbol words, hence providing redundancy. In whatever way a code of length n is defined, it possesses the dichotomic property that any n-symbol word belongs or does not belong to it. The codes used in engineering are generally defined by their construction rule (which is implemented in the encoding operation) from which specific constraints are easily derived and used in the decoding (regeneration) process. Both the construction rule and the constraints are expressed in deterministic mathematical terms. For extending the concept of code to genetics, we propose starting from the specification of a code by its constraints. We assume they can be expressed as incompatibilities or forbidding rules or in probabilistic terms, aside from being possibly expressed as deterministic mathematical equalities. For example, constraints can be imposed on the DNA strand by folding properties or induced by constraints on the proteins (the synthesis of which the DNA directs). In this extended meaning, the codes will be referred to as *soft codes*. We introduced this concept in [3] and assumed that the hypothesized genomic error-correcting codes are of this kind. We tried to somewhat refine its definition in further papers [4], [5]. With constraints expressed in probabilistic terms, the dichotomic property that a word belongs or not to a given code is lost. The main parameters of a code, like its minimum distance, then become random variables.

What little we lose in precision when considering soft codes, we gain very much in flexibility and generality since any constraint which directly or indirectly affects the DNA molecule implies some error-correcting ability. These constraints result in dependency between the symbols of the words, so the knowledge of certain symbols enables reassessing the probabilities of others. Since many such genomic constraints exist, the problem is to identify the means which implement regeneration rather than to find naturally implemented encoding processes. Encoding has become implicit but the actually crucial problem is regeneration (decoding, in engineering words). The optimum regeneration rule stated above then becomes: Choose the string of nucleotides obeying the genomic constraints the closest to the one to be replicated.

To illustrate the soft code concept, we used in [5] examples from the error-correcting technique, showing that it is relevant to the analysis of decoding processes. We shall try below to identify genomic soft codes. In order to illustrate the soft code concept, we now consider an example foreign to both engineering and molecular biology: natural languages. These languages involve strings of symbols (phonemes for the spoken language, letters of some alphabet for the written one) that are subjected to many constraints. The properties of the vocal tract severely restrict the combinations of phonemes that can be uttered, thus creating phonetic constraints (and inducing morphological constraints in the corresponding written texts). Among all the combinations of phonemes (or letters) that obey such constraints, only a small fraction are words of a given natural language. Let us refer to this constraint as *lexical*. The words of a language can themselves be combined according to syntactic rules specific to it, although they are possibly rooted in the human brain structure. At a still different level, meanings are associated with the words of any language and combining words according to the syntactic rules results in propositions. Correct propositions as regards these rules can be devoid of any meaning if they fail to obey semantic constraints (e.g., "the cat swept the red theorem" is both syntactically correct and meaningless). The constraints of fundamental nature due to properties of the vocal tract or the human brain, plus the conventional ones which are shared among a linguistic community, restrict the allowed strings of phonemes or letters to a very small subset of all unconstrained strings made of the same phonemes or letters. In other words, any natural language is a highly redundant soft code.

But what about error-correction capabilities? A conversation is such a trivial experience that we do not wonder at its success. Indeed, it almost always results in literal understanding even in the presence of a high noise level as in a street, a car, or a plane. Moreover, even in quiet acoustical surroundings, individual phonemes are identified with a large error rate although meaningful sentences made of the very same phonemes are unambiguously understood. We may thus think of the literal understanding of a language as a decoding process. Furthermore, a language is defined by distinct constraints acting at several hierarchical levels. For instance, phonetic constraints, which are due to the structure of the vocal tract, are more fundamental than constraints specific to a given language, which are social conventions inherited from history. We shall refer below to such a structure as a system of *nested soft codes*. Our daily experience thus witnesses the error-correcting ability of a natural language, although the precise decoding or regeneration mechanisms involved are essentially unknown. They are implemented in the human brain but escape consciousness.

Going back to DNA coding, errors resulting from substitution of a wrong nucleic base to another one should not only be considered but also those due to erasures, deletions, and insertions. We shall nevertheless limit ourselves to the substitution errors because this case has been extensively studied by engineers, although deletions and insertions are at least as important in genetics. Error-correcting codes against this type of errors can be designed with properties similar to those of codes against substitution errors but they were much less studied. Similarly, in the absence of a thorough study of soft codes, we may assume for convenience that the main

properties of error-correcting codes are not fundamentally altered and thus remain approximately relevant to the biological soft codes, 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 consequences of using soft codes as error-correcting means will thus not be significantly different from those of conventional codes as discussed above. Besides being convenient, this assumption may be fairly close to reality, as a consequence of the law of large numbers, if both the code lengths considered are large and the overall code is specified by many independent constraints.

Having stated the necessary basic concepts, we are now able to more precisely formulate the hypotheses regarding genomic error-correcting codes, then compare the consequences which can be derived from them with known biological facts, and even use them as predictive tools to help deciding on debated issues.

Hypotheses and Their Consequences

Main Hypothesis

Our main hypothesis has been already stated: it consists of assuming the existence of error-correcting means that behave like the theoretically optimal ones, i.e., provide a regeneration error probability that decreases as the code length increases and vanishes as it tends to infinity. A necessary condition for their existence is the presence of redundancy. The number of different genomes of some given length n would be 4^n in the absence of redundancy. Even for the shortest genomes, those of viruses, n is at least of several thousands, so 4^n is a number so large that it defies imagination. In contrast, we may evaluate the total number of past and present species to about $10^9 \approx 4^{15}$, so a genome made of 15 nucleic base pairs would suffice to specify all past and extant species. A comparison with the actual genome lengths (ranging from a few million base pairs for bacteria and up to 1 billion base pairs and more for plants and animals) shows that the actual redundancy rate is very high, so the genomes can be far apart from each other in terms of the Hamming distance. That it is actually so explains a striking feature of the living world, namely, that in their own space, genomes are very sparse, so discrete species exist. Uniquely specifying each individual within each of these species would require several tens of nucleotides more, resulting in a genome length less than 100 base pairs.

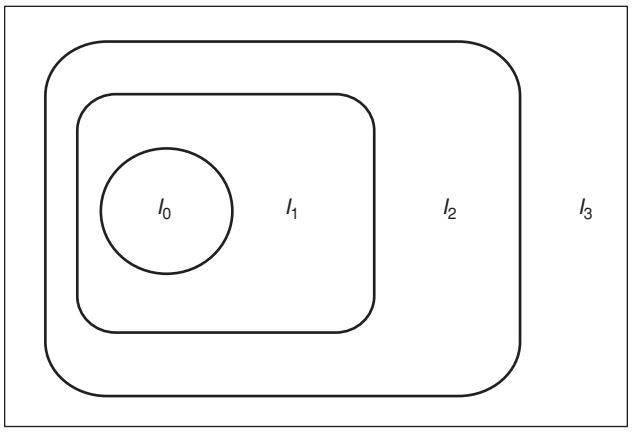

Fig. 1. The fortress metaphor: A code is represented as a closed wall that protects what is inside it. I_0, I_1, I_2, and I_3 are successive information messages; I_0 is protected by three codes, I_1 by two codes, I_2 by a single code, and I_3 is left uncoded.

The existence of some kind of error-correcting codes in the genome, at the molecular level or mostly involving short codes, was also suggested in [20]–[23]. The interesting idea that introns are made of check symbols associated with the message borne by the exons was formulated in [24]. The search for a simple linear code described in [25] was unsuccessful, but this negative result is questionable (see below). On the other hand, biological error-correcting mechanisms foreign to the genome replication were discovered (see, for example, [26]). The role of codes in biology has been stressed in [27]. However, Barbieri's concept of organic codes results from a deep reflection on biological facts but does not refer to the necessity of error correction.

Subsidiary Hypothesis and Nested Codes

We must introduce a subsidiary hypothesis before proceeding further. We were led to formulate it because the assumed genomic error-correction means need to provide an unequal error-protection. If we look at the features of living beings, we see that some are conserved with an extreme faithfulness, as witnessed by the permanency of certain genes like the *HOX* genes, but that other features are much more variable. As a means for introducing genomic variation, sexuality has moreover been favored by evolution in most of the living species. To account for these facts, we were led to assume that the hypothesized error-correcting system consists of nested codes. Notice that a similar scheme has independently been used by Barbieri to describe the organic codes [27].

Nested codes can be more easily described in the case of conventional systematic codes $C(n, k)$, i.e., where a k-symbol information message is encoded into a longer n-symbol word where the k-symbol message explicitly appears in a set of k defined positions. We assume that a first information message I_0 of length k_0 is encoded according to a code $C(n_0, k_0)$. Then, a second message I_1 of length k_1 is appended to the codeword that resulted from the first encoding, and encoding again by a code $C(n_1, n_0 + k_1)$ is performed. This process is repeated t times. The last information message I_t is left uncoded. This process is depicted in Figure 1 with the fortress metaphor, where each code is depicted as a wall that encloses its encoded information message for $t = 3$.

The component codes of a nested codes system may use different alphabets. Defining *nested soft codes* is more difficult since the concept of information message vanishes in this case. We may think of the nested code concept in more general terms: the ith encoding creates dependency between the results of $i - 1$ previous encodings, regardless of the alphabets and the codes which are used. Our above example of a natural language actually illustrates a system of nested soft codes. As an example of genomic nested codes, we may think of constraints induced on the genome by that of proteins as defining a basic soft code; in eukaryotes, constraints due to the wrapping of the DNA double strand in nucleosomes are superimposed and define a more peripheral soft code (see below). Besides assuming that a genomic error-correcting code is made of several nested codes, we furthermore assume that it was built in successive steps where the codes appeared in the order of the

Our hypothesis also sheds light on the process of biological evolution and on the structure of the living world.

layers, beginning with the innermost one. In other words, we assume that the encodings according to codes $C(n_0, k_0)$, $C(n_1, n_0 + k_1)$, etc., appeared successively in the geological times or, referring again to Figure 1, that the walls were constructed successively, beginning with the innermost one. According to this viewpoint, older genetic information is better protected than more recent information. When genomic variability is needed, it should correspond to the periphery of the nested codes scheme depicted in Figure 1. As a generator of variability, sexuality should operate at this level. If (as many believe) it provides a defence against infectious agents, this variability is necessary to match the high genomic variability of viruses and bacteria.

Consequences of the Hypotheses as Regards Evolution and the Living World

The main arguments to be developed now rely on the assumption of a kind of similarity between living beings and the corresponding genomes considered as codewords, especially regarding their distance properties. In other words, we assume that the genomic space to some extent provides an image of the living world as we perceive it, i.e., that of phenotypes. Although this similarity is rather fuzzy and difficult to make more precise, we notice that it is implicit in many current biological approaches where phenotypes are compared in terms of distances of the corresponding genomes, e.g., for building phyletic trees. We shall give below further arguments in favor of this similarity.

In our earliest work on genomic error-correcting means [2], we believed that a relationship could exist between the genome length and the permanency of a species. It seemed that some species known for their very long genomes were also among the less variable ones (e.g., lungfishes or newts). The permanency being for optimal codes the greater the larger the codeword**,** we thought that the same relationship was likely to exist between the permanency of a species and the length of its genome. Implicit in this belief was the assumption of a constant redundancy rate. The recent improvements in the knowledge of genomes of many species make them appear as highly nonhomogeneous, so this relationship is questionable. The beings with the smaller genomes, especially the pathogenic agents, are actually more variable than beings with longer genomes.

Going back to the assumed similarity of the "phenotypic space" and the genomic space, and moreover assuming that genomes use error-correcting means in the form of nested codes, we may adequately account for the discreteness of species and the existence of a taxonomic hierarchy. In contrast, a world of living beings with uncoded genomes would not exhibit such a hierarchy and no taxonomy would be possible. The fact that we live in a world of discrete and taxonomically ordered species, not in a world of chimeras, is a strong argument in favor of our hypotheses in the (hopefully provisional) absence of direct experimental proofs. Indeed, a world of chimeras has been described in [28]: that of bacteria. This does not contradict the above statements; however, since our subsidiary hypothesis of time-successive nested codes leads to identifying the degree of evolution with the number of nested code levels of the genomes, we may expect that the amount of coding is less—hence the structure of distinct species is less strong—in the (ancestral) bacteria than in complex (more recent) beings like plants and animals.

A rather puzzling feature of the living world also finds a simple explanation in our hypotheses. It is the trend of evolution towards increased complexity. We may consider as an experimental fact that species having a larger genome than the previously existing ones appeared in many instances during the process of evolution. 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 can be to increase the genome permanency (as defined in [3] and above) and, therefore, to provide an immediate evolutive benefit. An increased genome length does not necessarily imply an increase in complexity, but it provides room for it. More complexity in turn enables improving the evolutive fitness of phenotypes and, hence, should be favored by natural selection.

Another simple consequence of our hypotheses is that evolution proceeds by jumps (i.e., is *saltationist*), a still debated issue. It is a straightforward 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 produced by regeneration errors, hence having a genome at a distance from the original one at least equal to the minimum distance of the code. It hints at a non-Darwinian mechanism for the origin of species, reminiscent of the "hopeful monster" hypothesized by Goldschmidt. With our subsidiary hypothesis, this distance itself depends on the code level in the assumed system of nested codes. Moreover, it accounts for the fact that evolution proceeds along phyletic trees with more frequent branchings, the lowest the level inside the nested codes system, since the probability of a regeneration error is higher the smaller the distance between genomes.

Looking for Genomic Error-Correcting Codes

Searching for genomic error-correcting codes in the form of soft codes amounts to listing the several constraints that the genome obeys, each of them being a component soft code in the assumed nested codes system. We shall below consider first those which are directly associated with structural

constraints of DNA, then those induced in DNA by structural constraints of proteins, and finally constraints which result from the role of the genome to direct the construction of a phenotype. Before dealing with these topics, some remarks concerning the alphabets will be useful.

Identifying the Alphabets

In engineering problems, the alphabet is often given as a parameter and is endowed with some a priori mathematical structure. This is not the case for the hypothesized genomic error-correcting codes where the alphabets themselves and their possible mathematical structure have to be determined. We make *alphabets* plural here since we consider nested soft codes, and we already noticed that their component codes can use different alphabets.

An apparently obvious choice is that of the quaternary alphabet {A, T, G, C}, but with what mathematical structure is it endowed? Liebovitch et al. [25], for instance, answered this question assuming it to be the ring of integers modulo 4. This choice is arbitrary and the usual structure considered in the literature for defining a linear code is that of a Galois field. It is only when the alphabet size q is a prime that the addition rule modulo q and that of the Galois field are identical.

Even with a mathematical structure more appropriate in engineering terms, such an approach is questionable as involving an arbitrary choice. The connection that the concept of soft code establishes between the physical and chemical constraints and the error-correcting properties suggests looking at alphabets having a physicochemical significance. In this respect, it is much more relevant to consider that any quaternary symbol simultaneously belongs to two independent codes over the following two binary alphabets: 1) the alphabet {R, Y}, whose symbols are the chemical structures of nucleic bases, namely, purine (two-cycle molecule, A or G) denoted R, or pyrimidine (single-cycle molecule, T or C) denoted Y and 2) 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} [29]. 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 one region to another inside the genome. It could be interpreted as a kind of density modulation, which perhaps is read at several scales. The different number of hydrogen bonds of the two base pairs implies that this density modulation results in a variation of the bonding energy between the two DNA strands in the double helix.

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." An alphabet size of $4^3 = 64$ could be considered, but dealing with the synonymous codons that "code" for the same amino acid as a single symbol (resulting in a 21-symbol alphabet) directly translates the constraints on the amino acids into constraints on DNA. Genes themselves can even be considered as the symbols of an alphabet [30], [31]. The successive use of alphabets of different sizes is a means for implementing nested codes, as already noted.

Soft Codes Associated with Structural Constraints of DNA

The alphabet which is relevant here is more likely to be {R, Y} as introduced in the previous section, namely based on the distinction purine/pyrimidine. The alphabet {2H, 3H} may also be relevant because the ease of separating the two DNA strands is an important factor during the replication process.

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 spatial frequency and β is a constant which depends on the species. Roughly speaking, β is smaller the higher the species is on the scale of evolution; it is very close to 1 for bacteria and significantly less for animals and plants [32].

Another study of the mutual dependence in DNA sequences only considered the binary alphabet {R, Y}. An appropriate wavelet transform was used to cancel the trend and its first derivative. The autocorrelation function of the binary string thus obtained has been shown to decrease according to a power law [33]. This implies long-range dependence at variance with, for example, Markovian processes, which exhibit an exponential decrease. Moreover, in eukaryotic DNA, the long-range dependence demonstrated has been related to structural constraints due to the packing of the double-strand DNA into nucleosomes where it is wrapped around histone molecules acting as a spool, which implies bending constraints along the two turns or so of the DNA sequence in each nucleosome [33].

The $1/f^{\beta}$ behavior of the spectrum and the long-range dependence of the DNA sequence restricted to the {R, Y} 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. Therefore, we may think of the message borne by the DNA strand as resulting from multiple unfaithful repetition, which could, in principle, enable the use of many low-reliability replicas of the basic motif symbols for the purpose of regeneration, in terms of which reliable decisions can be taken. This implies a very large redundancy, an obvious property of the DNA message. The existence of such a regeneration process, possibly approximated by majority voting, is as yet a conjecture. It is as yet to be determined whether, and how, nature implements regeneration based on long-range dependence at some stage of the DNA replication process [34]. One may wonder why the regeneration process does not turn this unfaithful repetition into a faithful one by correcting the '"wrong" symbols. We may explain why it is not necessarily so by the existence of other soft codes having independent probabilistic constraints within the assumed nested codes system. Then, the most probable symbol of the actual DNA message results from a compromise between the constraints of the several soft codes in which it is involved.

Soft Codes Induced by Structural Constraints of Proteins

Proteins are not fully described by the polypeptidic chain that the sequence of codons of a gene specifies. They owe their functional properties to their folding according to a unique pattern, which implies many chemical bonds (especially disulphur bridges) between amino acids that are separated along the polypeptidic chain but close to each other in the three-dimensional (3-D) space when the protein is properly folded. For instance, many proteins with an enzymatic function fold into a globular shape. Moreover, proteins are most often made of a number of 3-D substructures (α helices and β sheets, which are themselves included in higher-order structures named *domains*). These substructures impose strong geometrical, steric, and chemical constraints on the sequence of amino acids, which in turn induce constraints on the corresponding DNA. Due to the central role of genes in directing the synthesis of proteins, such constraints are present in the genome of any living being, whether it is a prokaryote or a eukaryote.

Interpreting a Gene with Exons and Introns as a Kind of Systematic Codeword

Forsdyke suggested in 1981 that introns are made of check symbols associated with the message borne by the exons [24]. The literature generally states that introns are more variable than exons, but a counterexample was provided in 1995 by Forsdyke, who experimentally found that the exons are more variable than introns in genes which "code" for snake venom [35].

It turns out that both the generally observed greater variability of introns and Forsdyke's counterexample can be explained by the assumption that the system of exons and introns actually acts as 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 regeneration error, a mutation occurs with large probability in favor 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, 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. The total number of errors is at least equal to the minimum distance of the code. If few errors are located in the exons, most of them must affect the introns..

The situation is completely different in the case of genes that "code" for snake venom. Rodents are the typical prey of snakes. Snakes and rodents are involved in an "arms race." Some rodents incur mutations that 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 [35]. The genes which "code" for snake venom are thus under high evolutive pressure, since natural selection favors mutated genes producing proteins as different as possible from the original ones. In terms of the Hamming distance, much of the difference should therefore be located in the exons. With the total number of errors in exons and introns being roughly constant for a given code, introns are much less variable. These properties of eukaryotic genes are precisely those which can be expected from genes acting as systematic error-correcting codes, but the encoding and regeneration processes remain unknown. It is not even known whether the distance properties of these genes are actually used for error correction. Clearly, discovering the encoding and regeneration mechanisms at work here needs the active collaboration of biologists.

A Possible Role of "Junk" DNA

Genomes (especially the human genome) often contain very short sequences (e.g., three bases long), which are repeated thousands or even millions of times. Such sequences bear almost no information. Such "junk" DNA may, however, play a role in an error-correction system as separating along the DNA strand more informative sequences which, due to the 3-D structure of the DNA molecule, may be spatially close to each other and share mechanical or chemical constraints (a function which loosely resembles that of interleaving used in the coding technique).

On the other hand, the most successful encoding scheme available to engineers is that of turbo codes [17], [18], which can be interpreted as combining three main functions [36]: replication (repeating a symbol), interleaving (permuting a sequence of symbols), and rate-1 encoding (computing output symbols in terms of a sequence of input symbols), as depicted in Figure 2. Each of the blocks of this figure performs one of the three functions which may be expected from a good encoder, namely, providing redundancy, randomness, and mutual dependence, respectively. Replication is the sole function that produces redundancy. The other functions convert mere repetition into distributed redundancy, which is much more efficient regarding error correction. We may thus interpret the scheme of Figure 2 as a kind of paradigmatic encoder. The junk DNA made of a short sequence repeated many times may play the same role as an interleaver. We may think of it as separating along the DNA strand sequences which, due to the 3-D structure of the DNA molecule, are spatially close to each other and can share mechanical or chemical constraints (see Forsdyke [24]). Although the efficiency of such a separator is poor in terms of redundancy (compared with a true interleaver), we already noticed that the genomes are characterized by a very high redundancy, so genomic redundancy may be thought of as "cheap." In engineering, on the contrary, redundancy often has a cost which limits it to moderate amounts. We notice, moreover, that it is not too difficult to imagine how such an encoder has been generated through the ages, since the separator, if we let it replace the interleaver in Figure 2, results from a sequence being repeated, which is the most basic function of DNA.

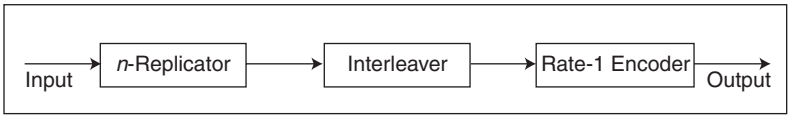

Fig. 2. A schematic representation of a rate $1/n$ turbo encoder. The box labeled n-replicator represents a device which successively delivers n times its input symbol. The interleaver changes the order of the symbols in its input sequence, and the rate-1 encoder outputs symbols which combine a number of its successive input symbols. The n-replicator is the sole of the devices in the scheme to generate the necessary redundancy.

Soft Codes from Linguistic Constraints

We stressed above the contrast between the comparative brevity of the message which is needed for unambiguously identifying a biological species (and even an individual inside it) on the one hand, and the length of actual genomes on the other hand. This contrast has rather obvious reasons since the genome role is by no means restricted to identify a living being: 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 direct 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 its phenotype. This demands some kind of language. Biologists do not yet know it 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 [11] and many others. This metaphor is quite convincing, but its consequences in terms of genome conservation are overlooked. Indeed, any language involves many lexical, syntactic, and semantic constraints that may be interpreted as soft codes having error-correcting abilities (as we argued above for human languages). Moreover, they appear at several different levels and thus assume 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 abilities are exploited. Current researchers already use tools of formal linguistics (which shares the concept of dependence with information and coding theory) in order to describe the genomes and proteins [37], [38] but ignore the error-correction problem.

The connection just outlined between linguistics and error-correcting ability implies that a longer genome is not only useful to decrease the error probability but also provides room for more semantics and, therefore, 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 phenotype, so one could consider the error-correction ability 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 this correction ability is its main feature, since without it no transmission of hereditary characters would be possible and life could not have developed. Then, the construction and maintenance of phenotypes would be a mere projection in the physicochemical world of the abstract properties of the genetic message that enable error correction. This is a hen-and-egg problem, as often met in biology. Interestingly, the similarity of the phenotypic and genomic spaces we were led to assume above may have its roots in this relationship.

Biology and Engineering: A Needed Collaboration

Nature obviously appears as an engineer of very broad competence, and its achievements are outstanding. Therefore, human engineers should be deeply interested in the products of nature's engineering, i.e., living things. Similarly, understanding the engineering aspects of life should be a major concern for biologists. However, the methods used by nature on the one hand, and human engineers on the other hand, exhibit a sharp contrast which may explain why biologists and engineers do not more closely collaborate. 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 very difficult) 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, especially the genome and the cell replication machinery, and the time scale of nature extends to that of geology, i.e., up to billions of years. Having genuine self-repair capabilities, living beings are moreover much more flexible and resistant to degradation than the products of human engineering, and nature's achievements often outperform what human engineers can do. It turns out, moreover, that they are understood in almost any case only insofar as human engineers invented similar solutions to problems that nature solved eons ago. That the methods of nature and engineers are so markedly different is perhaps why we can learn so much from nature. Clearly, exaustive search is not a good method for purposely designing an object within some prescribed short time, but it guarantees the absence of any bias. In constrast, no human engineer can claim to be completely free from prejudice.

The main distinctive features of living beings are their extreme complexity, which is unmatched in the nonliving world, and (not independently) the fact that, besides matter and energy, they receive and transmit information and heavily rely on its transfer and conservation for their construction and maintenance. This last point also has no equivalent outside the living world and appears as the specific mark which radically differenciates it from the nonliving world. It makes biology especially relevant to information theory, thus prompting biologists to use information theory as a main tool and challenging information engineers to get interested in biology.

Conclusions

The question of how genetic information is faithfully communicated clearly needs to be answered. Dealing with the genome as if it were a permanent object, like those of our daily lives at our time scale, is not tenable. Information theory and the experience gained by engineers for designing and implementing error-correcting codes will help to answer this question properly. The above speculations were intended to this aim but could only rely on published biological works. Many works on the genome were aimed at understanding how it directs the construction of a phenotype but, unfortunately, fewer were devoted to the way it replicates itself. It may be rather futile, however, to question how the genome produces a phenotype if we do not first understand how the genome produces a genome.

If genetics eventually gets an information-theoretic education, there is little doubt that unexpected error-correcting means will be discovered and that our understanding of evolution, and, therefore, of the living world, will be deeply improved. The speculations presented here are but provisional steps in this direction. We may safely predict that, in this field as in others, nature will reveal itself more inventive and efficient than human engineers. To quote Jerome Wiesner, "No one is visionary enough to match reality." The extreme importance of information in the living world even suggests that getting an information-theoretic education should be widely beneficial to biology as a whole.

Gérard Battail graduated from the Faculté des Sciences (1954) and Ecole Nationale Supérieure des Télécommunications (ENST) in 1956, both in Paris, France. He joined the Centre National d'Etudes des Télécommunications (CNET) in 1959. He worked there on modulation systems and especially on frequency modulation, using fundamental concepts of information theory to understand its behavior in the presence of noise, namely, the threshold effect. In 1966, he joined the Compagnie Française Thomson-Houston (later called Thomson-CSF), where he acted as a scientific advisor to technical teams designing radioelectric devices. There he interpreted channel coding as a diversity system for designing decoders, especially soft-input ones. He also worked on source coding, frequency synthesizers, mobile communication, and other problems related to the design of industrial radiocommunication devices. In 1973, he joined ENST as a professor. He taught modulation, information theory, and coding there. He also had research activities in the same fields with a special emphasis on adaptive algorithms regarding source coding and, for channel coding, on soft-in, soft-output decoding of product and concatenated codes. He was led to criticize the conventional criterion of maximizing the minimum distance of a code and instead proposed a criterion of closeness of the distance distribution with respect to that of random coding. Some of these ideas are at the root of the invention of turbo codes. After his retirement in 1997, he started working on applications of information theory to the sciences of nature. He has especially investigated the role of information theory and error-correcting codes in genetics and biological evolution.

Battail has applied for many patents, written many papers, and participated in many symposia and workshops. He also authored a textbook on information theory published by Masson in 1997. He is a member of the Société de l'Electricité, de l'Electronique et des Technologies de l'Information et de la Communication (SEE) and of the IEEE. Before his retirement, he was a member of the editorial board of the *Annales des Télécommunications*. From 1990–1997, he was the French official member of Commission C of URSI (International Radio-Scientific Union). From June 2001–May 2004, he served as associate editor at large of *IEEE Transactions on Information Theory*.

Address for Correspondence: Gérard Battail, la Chanatte, le Guimand, F-26120 Chabeuil, France. E-mail: gbattail@club-internet.fr.

References

[1] C.E. Shannon, "A mathematical theory of communication," *BSTJ*, vol. 27, pp. 379–457, pp. 623–656, 1948.
[2] G. Battail, "Does information theory explain biological evolution?," *Europhys. Lett.*, vol. 40, no. 3, pp. 343–348, Nov.1997.
[3] G. Battail, "Is biological evolution relevant to information theory and coding?," in *Proc. ISCTA '01*, Ambleside, UK, 2001, pp. 343–351.
[4] G. Battail, "An engineer's view on genetic information and biological evolution," *Biosystems*, vol. 76, no. 1–3, pp. 279–290, 2004.
[5] G. Battail, "Can we explain the faithful communication of genetic information?," presented at the DIMACS working group on theoretical advances in information recording, Mar. 22–24, 2004.
[6] 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.
[7] R.E. Franklin and R.G. Gosling, "Molecular configuration in sodium thymonucleate," *Nature*, vol. 171, no. 4356, pp. 740–741, 25 Apr. 1953.
[8] J.D. Watson and F.H.C. Crick, "Molecular structure of nucleic acids," *Nature*, vol. 171, no. 4356, pp. 737–738, Apr. 25, 1953.
[9] E. Chargaff, "How genetics got a chemical education," *Ann. New York Acad. of Sci.*, vol. 325, pp. 345–360, 1979.
[10] R. Dawkins, *The Selfish Gene*. Oxford, UK: Oxford Univ. Press, 1976.
[11] R. Dawkins, *The Blind Watchmaker*. Harlow: Longman, 1986.
[12] D.R. Forsdyke, "Selective pressures that decrease synonymous mutations in Plasmodium falciparum," *Trends in Parasitology*, vol. 18, pp. 411–418, 2002.
[13] E. Guizzo, "Closing in on the perfect code," *IEEE Spectr.*, vol. 41, no. 3 (INT), pp. 28–34, Mar. 2004.
[14] F.J. MacWilliams and N.J.A. Sloane, *The Theory of Error-Correcting Codes*. Amsterdam: North Holland, 1977.
[15] G. Battail, "Construction explicite de bons codes longs," *Annales Télécommunic.*, vol. 44, no. 7–8, pp. 392–404, July-Aug. 1989.
[16] G. Battail, *On Random-like Codes*, Lecture Notes in Computer Science No. 1133. New York: Springer, 1996, pp. 76–94.
[17] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo-codes," in *Proc. ICC'93*, Geneva, Switzerland, May 1993, pp. 1064–1070.
[18] C. Berrou and A. Glavieux, "Near optimum error correcting coding and decoding: Turbo codes," *IEEE Trans. Commun.*, vol. 44, pp. 1261–1271, Oct. 1996.
[19] G. Battail, C. Berrou and A. Glavieux, "Pseudo-random recursive convolutional coding for near-capacity performance," in *Proc. GLOBECOM'93*, *Commun. Theory Mini-Conf.*, Houston, TX, 1993, vol. 4, pp. 23–27.
[20] G. Cullmann and J.-M. Labouygues, "The logic of the genetic code," *Biosystems*, vol. 16, pp. 9–29, 1983.
[21] J. Rzeszowska-Wolny, "Is genetic code error-correcting?," *J. Theor. Biol.*, vol. 104, pp. 701–702, 1983.
[22] H.P. Yockey, *Information Theory and Molecular Biology.* Cambridge, UK: Cambridge University Press, 1992.
[23] D.A. Mac Dónaill, "A parity code interpretation of nucleotide alphabet composition," *Chem. Commun.*, vol. 18, pp. 2062–2063, 2002.
[24] D.R. Forsdyke, "Are introns in-series error-detecting sequences?," *J. Theor. Biol.*, vol. 93, pp. 861–866, 1981.
[25] L.S. Liebovitch, Y. Tao, A.T. Todorov, and L. Levine, "Is there an error correcting code in the base sequence in DNA?," *Biophys. J.*, vol. 71, pp. 1539–1544, 1996.
[26] E.E. May, M.A. Vouk, D.L. Bitzer, and D.I. Rosnick, "Coding theory based models for protein translation initiation in prokaryotic organisms," *Biosystems*, vol. 76, no. 1–3, pp. 249–260, 2004.
[27] M. Barbieri, *The Organic Codes*. Cambridge, UK: Cambridge Univ. Press, 2003.
[28] L. Margulis and D. Sagan, *Microcosmos, Four Billion Years of Evolution from Our Microbial Ancestors*. New York: Summit Books, 1986.
[29] D.R. Forsdyke home page [Online]. Available: http://post.queensu.ca/forsdyke/
[30] S.A. Kauffman, *The Origins of Order*. New York: Oxford Univ. Press, 1993.
[31] O. Milenkovic, "The information processing mechanism of DNA and efficient DNA storage," presented at DIMACS working group on theoretical advances in information recording, Mar. 22–24, 2004.
[32] 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.
[33] 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.
[34] G. Battail, "Replication decoding revisited," in *Proc. Information Theory Workshop 2003*, Paris, France, pp. 1–5.
[35] D.R. Forsdyke, "Conservation of stem-loop potential in introns of snake venom phospholipase A2 genes. An application of FORS-D analysis," *Mol. Biol. Evol.*, vol. 12, pp. 1157–1165, 1995.
[36] J.J. Boutros, "Asymptotic behavior study of irregular turbo codes," in *Proc. DSP'2001*, Sesimbra, Portugal, Oct. 2001.
[37] 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. Lett.*., vol. 73, pp. 3169–3172, 1994.
[38] D.B. Searls, "The language of genes," *Nature*, vol. 420, no. 6912, pp. 211–217, Nov. 2002.