Biosemiotics (2008) 1:57–73
DOI 10.1007/s12304-008-9005-5

The Hypothesis of a Genetic Protolanguage: an Epistemological Investigation

Gregory Katz

Received: 03 November 2007 / Accepted: 07 December 2007 /
Published online: 8 February 2008

Abstract Progress in molecular biology has revealed profound relations between linguistic and genomic sciences, mainly through advances in bioinformatics. The structural symmetries between biochemical and verbal syntaxes raise the question of their origins: did they emerge independently, or did one arise from the other? Does the genetic code contain the traces of a protolanguage, a universal grammar whose gradual evolution and successive mutations progressively led to the polymorphism of natural languages? To explore this question, we review the isomorphism of the genetic code and verbal codes from lexical, syntactic, semantic and pragmatic standpoints. We discuss the limits of these symmetries and their anthropomorphic connotations. We observe the gradual evolution of species and languages according to parallel mechanisms, and the genetic roots of the physiology of language. In conclusion, we hypothesize that human observers may not be projecting linguistic frameworks onto genomic structures. Rather, it could be their linguistic faculties that reflect the grammatical structure of genetic code.

Keywords Protolanguage · Language evolution · Nucleic acid · Polymorphism · Phylogenesis · Natural languages · Biolinguistics · Bioinformatics · Biosemiotics · Genetics · Anthropomorphism

> "When I first came across linguistic terms in the biological literature, I said to myself: we need to check whether this is just a manner of speech, a metaphoric usage, or whether there is something deeper here. I must say that what biologists have done is quite legitimate from a linguistic standpoint, and in fact we can take things even further."
>
> Roman Jakobson, 1968

G. Katz (✉)
ESSEC Business School, Paris-Singapore, Avenue Bernard Hirsch, 95021 Cergy, France
e-mail: katz@essec.fr

Introduction

Founded in 1866, the influential Société de Linguistique de Paris was devoted to "the study of languages," but it explicitly mentioned in its statutes that it "will not allow any communications concerning either the origin of language or the creation of a universal language" (article II). Today, however, because the sequencing of the human genome has unlocked a universal code for the living, these questions need to be addressed. At the frontier of the language of biology and the biology of language, linguistic metaphors are now deeply entrenched in the vocabulary of molecular biology. Since the discovery of the DNA double helix, linguistic concepts have both impacted on the theoretical models used in genomics and proved their heuristic value at an experimental level (Fox-Keller 1995).

While most geneticists are still reluctant to lend credit to the linguistic model of the genome—though they are willing to use it—linguists such as Jakobson (1973) noticed a striking similarity between the structure of the genetic code and the fundamental principles of natural languages. Inspired by his work, the anthropologist Claude Lévi-Strauss (1971) saw in the genetic code the most primitive model of all forms of language, a universal prototype emerging from nature before culture appeared: a protolanguage from which natural languages may have derived through the history of evolution. The structural isomorphism of the genetic and verbal grammars indeed raises the question of their origins: did they emerge independently, or did one originate from the other? Do their respective grammars contain the trace of a common core, or are they the product of radically distinct evolutions?

This article discusses three issues: (1) Do the analogies between the structures of genetic code and the structures of language reveal a true isomorphism? (2) Can the genetic code be considered a protolanguage? (3) What consequences do these analogies have in terms of epistemological modeling? To address these questions, we probe into the symmetries between genetic and verbal codes. We review the isomorphism of the two codes from lexical, syntactic, semantic and pragmatic standpoints. We discuss the limits of these analogies and their anthropomorphic connotations. We observe the gradual evolution of species and languages according to parallel mechanisms, and the genetic roots of the physiology of language. In conclusion, we hypothesize that human observers may not be projecting linguistic frameworks onto genomic structures; rather, it could be their linguistic faculties that reflect the grammatical structure of the genetic code.

Chomsky and the Universal Grammar

The 1950s witnessed the discovery of the structure of DNA and more widely the advent of the modern era of molecular biology (Aitchison 1999). Those years were also a time of sweeping revolution in linguistics, mainly because of the work of Noam Chomsky. Going beyond the classical approach, which consisted of establishing an inventory of linguistic utterances, Chomsky explored the mechanisms by which they are produced. By seeking to identify the universal foundation of all languages, he established the principle of a new form of generative grammar—a set of syntactic structures—that could explain the tremendous creativity of linguistic production (Chomsky 1957). His quest for a "universal grammar" and his

mathematical models deeply influenced language theory, particularly in relation to the analysis of the invariant structure contained in the core of all languages, beyond their intrinsic variations (Keenan and Stabler 2003). These works had an impact on the information sciences in terms of the recognition of syntactic structures, the interpretation of computer languages, and the processes for understanding natural languages (Lenneberg 1967; Jurafsky and Martin 2000). Chomsky's hierarchy of linguistic classes was particularly effective in stratifying formal languages according to their power of expression and their mathematical and computational complexity (Searls 2002). Chomsky's influence also extended to the cognitive sciences (Chomsky 2001), analytical philosophy and even critical literature (Chomsky 2004, 2005).

The cross-disciplinary dialogue between life sciences and language sciences took a singular turn with the famous debate between the linguist Roman Jakobson, the biologist François Jacob and the anthropologist Claude Lévi-Strauss (Jakobson 1968). By comparing the combinatorial structure of verbal and genetic codes, the human sciences and the biological sciences have since succeeded in identifying the profound convergences between human and biochemical languages, thereby opening up a whole new field of analysis at the interface of their respective disciplines.

Convergence and Isomorphism

In the 1980's biologists attempted to apply the works of Chomsky and Jakobson to molecular biology. The initial results showed that these linguistic models were compatible with those used for biological sequences (Head 1987). Since then, analogies between the evolution of human language and some aspects of the emergence of cellular life have often been discussed in the literature of both linguistics and evolutionary biology. An obvious similarity relates the modular structure of words and grammars to the hierarchical levels of organization of biological molecules, especially of biopolymers such as DNA and proteins. Another example is the analogy between the dualism knowledge/utterances in linguistics and the dualism genotype/phenotype in biology. These parallelisms are often presented as a surprising isomorphism of systems on completely different scales, perhaps reflecting our need to search for recurring patterns in natural phenomena. Yet a deeper scrutiny of this analogy may help us to revisit assumptions and results in both fields and suggest new modes of investigation. This may even reveal common evolutionary mechanisms that are responsible for shaping emergent properties of complex systems, independent of their relative or absolute scales.

The open-ended richness of combinatorial structures is one of the most powerful features characterizing human language, as well as the genetic apparatus in a cell. Somehow, in both cases, elementary units (small molecules in one field of research, simple sounds or elementary concepts in the other) ended up forming long sequences encompassing very complex functions. Common to both disciplines also is the puzzle of what may have been the role of the intermediate-length sequences that must have been present at some point along the evolutionary path. From lexical and biochemical standpoints, half a sentence is in general as useless as half a protein. Hypotheses about possible forms of protocells (Segrè 2000) and protolanguages (Bickerton 1990) need to be formulated in order to address this conundrum.

Interestingly, these early stage transitions may have involved a process of specialization, by which units possessing an initial broad function gradually diversified and gave rise to variants that relate to narrower functional classes.

A particularly interesting commonality of the two evolutionary processes described here is the presence of multiple levels of representation. In a living cell, the reactions transforming different metabolites into each other have a representation in the structures of the various enzymes that catalyze these reactions. In turn, these enzymes have a clear mapping in the genes that code for their amino acid sequence. Similarly, in human language, one can identify mappings between neuronal firing patterns, sound waves and graphical signs. Both in the cellular networks and the structure of human language, multiple representations of the same functional units seem to be present. This multiplicity, or redundancy, of "memory devices" exchanging information with each other may have played a major role in the evolution of the combinatorial nature of human and biochemical languages (Segrè 2002).

Structural Analogies

Human and biochemical languages share several essential characteristics: both are structured, hierarchical, flexible and recursive. By this we mean that they are: (1) Structured in the sense that an utterance is not just a random juxtaposition of units, but that in some way it indicates the relations between these units. (2) Hierarchical in the sense that there are structural levels within the structures themselves. (3) Flexible in the transformational sense that there are many different ways to express the same meaning by moving units around and restructuring sentences according to certain rules. (4) Recursive in the sense that the same rules and structures may recur at different levels in the hierarchy, so that a structure may contain a substructure that is another instantiation of the same structure, in theory repeated ad infinitum (Johansson 2005).

These structural levels appear to be common to both the genetic and verbal codes. Somehow, in both cases, elementary units form long sequences encompassing highly complex functions. In both cases, the systems are organized through an arrangement of distinct and distinctive signs that can either break down into lower-level units or combine to create complex units (Segrè 2000). In linguistics and genetics, a common structural model has emerged based on a hierarchy of the integration levels of meaningful units. In addition, there is a striking correspondence between the two systems at each level of the structural hierarchy. Jakobson established correspondences between nucleotide and letter, codon and word, and gene and sentence (Jakobson 1973). He analyzed the dual articulation of the genetic and verbal languages, which consist of units endowed with meaning, based on discrete subunits which in themselves contain no inherent meaning. "This dual articulation is a property that, among all communication systems, is found only in the verbal code and the genetic code. The isomorphism between these two codes is deeply rooted in the very principles of their mechanism" (Jakobson 1974, p. 67).

Classically, linguistics distinguish between several structural levels: (a) a lexical level where words that are chained linearly are recognized and characterized; (b) a syntactic level where words organize themselves into a hierarchic system according to grammatical rules; (c) a semantic level where meaningful representations are

attributed to syntactic and lexical structures; (d) a pragmatic level where language fits into a global context that establishes interrelations between sentences through a statement or a dialogue. These four linguistic levels may correspond to the four levels commonly used in molecular biology: sequence, structure, function and role (Searls 2002).

This symmetry extends to many points, including the way messages are delimited. Specific signals indicate the start and end of coordinated genetic systems and the limits between genetic segments within these systems. François Jacob called these signals "punctuation signs" or "commas" (Jacob 1966). In the linguistic model, Jakobson stressed that they correspond to the delineation processes used in the phonological division of a statement into sentences, and of sentences into clauses and parts. They are border signals (Grenzsignale), i.e. the limits of the informative message (Jakobson 1973; Trubetskoy 1936).

The strict co-linearity of the time sequence in encoding and decoding operations characterizes both verbal language and genetic code, since it translates a nucleic chain into a proteic chain. Here again we are dealing with a linguistic concept and term that has been quite naturally borrowed by biologists who, when matching the original messages against their peptide translation, detect synonymous codons. One of the functions of verbal synonyms in communication is to avoid a partial homonym. Jacob wondered whether a similar reason could perhaps explain the choice between synonymous codons. Such redundancy would allow a degree of flexibility in the writing of heredity (Jacob 1965).

The Protolanguage Hypothesis

How should we interpret these isomorphic characteristics? Given such a large number of symmetries, Jakobson dared to formulate a bold hypothesis: verbal code could be the distant heir of genetic code, whose syntactic foundations serve as its model. The deeper structure of natural languages could derive from a biochemical ancestor embedded in the living cell. "Individual patterns of speech have a facet that allows us to presume the possibility of a genetic endowment. In fact, our speech contains inalienable and inalterable characteristics whose main origin lies in the lower part of the vocal apparatus, that which is located between the abdomen-diaphragm area and the pharynx" (Jakobson 1973). Yet this linguistic physiology—from Broca's area to the glottis—is genetically programmed (Fitch 2005).

It is surprising that children learn to speak their mother tongue so easily. The generative grammar theory introduced by Chomsky identifies universal mechanisms at the deepest level that locate this innate ability in the human brain (Chomsky 1972). From a biological standpoint, the hypothesis of a hereditary capacity to learn any language implies that such capacity must be encoded in our chromosomes (Lieberman 1984; Jerne 1984). From this perspective, couldn't this code, which is inscribed in DNA, contain a universal grammar that is common to all natural languages?

The main reason we are loath spontaneously to accept this hypothesis is that we are used to conceiving of language as a product of culture as opposed to nature. What connection could there be between a combination of chemical units such as that of our genome and the language we use to express ourselves? The intermediate

levels needed to pass from one code to the other require an infinite number of tiny mutations that are difficult for our mind to envision. In addition, the different forms of biochemical and verbal signifiers spawn confusion. Yet, whether we are dealing with a Braille character, a pictogram, an acoustic image or a genetic unit, the linguistic sign is always arbitrary (Saussure 1908; Monod 1970; Stegmann 2004). Likewise, if you replace the wooden pieces of a chess set with ivory ones, the system is unaffected by this external change. Substituting the form of the signifiers has no effect on the grammar of the game.

Jakobson's hypothesis is based on three observations: (1) Our linguistic faculties are rooted in the vocal apparatus, i.e. in a specific physiology; (2) Like the other organs in our organism, this physiology of language is genetically programmed; (3) There is a deep isomorphism between the grammar of the genetic code and the grammar of verbal codes. According to Jakobson, these three observations allow him to hypothesize the existence of a protolanguage embedded in our genes whose model is at the origin not only of our linguistic faculties but of the linguistic polymorphism that has evolved over time as well (Jakobson 1968). This original code could have passed through various evolutionary phases that progressively modified its means of expression but not its intimate grammar: initially nucleic, then proteic, later physiological, it finally reached the verbal stage. Inspired by the works of Jakobson, Claude Lévi-Strauss also defended the idea of a universal language embedded in the genome. According to him, the genetic code serves as the "absolute prototype from which, at another level, articulated language retrieves the model" (Lévi-Strauss 1971).

Biological Evolution and Linguistic History

Despite the deep symmetries between genetics and linguistics, many biologists and linguists today still consider that there is neither a direct nor indirect relationship between phenomena as remote as the evolution of species and the evolution of language. Likewise, they believe there is no connection between the structure of genetic code and the universal grammar of verbal codes. According to these biologists, these communication phenomena are inherent in each form of life and do not derive from each other. In other words, molecules, cells, organs, organisms and populations develop independent communication systems which are not the result of any common prototype or protolanguage. The isomorphism of genetic code and verbal code could, in fact, be simply an isolated coincidence that does not follow any evolutionary logic.

Yet in The Descent of Man, Darwin himself pointed out that the evolution of the various languages and that of the diverse species, which both developed through a series of gradual processes, are strangely parallel (Darwin 1871). Already in The Origin of Species, he recognized a profound symmetry between the evolution of species and that of languages:

> "If we possessed a perfect pedigree of mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world; and if all extinct languages, and all intermediate and slowly changing dialects, were to be included, such an arrangement would be the only possible one. Yet it might be that some ancient

> language had altered very little and had given rise to few new languages, whilst others had altered much owing to the spreading, isolation, and state of civilization of the several co-descended races, and had thus given rise to many new dialects and languages. The various degrees of difference between the languages of the same stock would have to be expressed by groups subordinate to groups; but the proper or even the only possible arrangement would still be genealogical; and this would be strictly natural, as it would connect together all languages, extinct and recent, by the closest affinities, and would give the filiation and origin of each tongue" (Darwin 1872, chap. XIV).

This genealogical tie between all languages refers directly to the idea of an original code whose linguistic model was reproduced by variation and adaptation according to the laws of evolution. Darwin considered that there was "one primordial form, into which life was first breathed," "some one prototype" from which all organic beings descend (Darwin, ibid). In the furtherance of this idea, the hypothesis of a genetic protolanguage also assumes the existence of a universal prototype common to languages and species, both governed by the same laws of evolution. The genetic code could be this prototype that predisposes forms of life to develop polymorphic and variable systems of expression that are nonetheless based on a minimalist grammar, at once primary and universal.

By studying the evolution of languages, linguistic history gradually came to consider language phenomena as an "inheritance, a will, an instruction from the past and projecting into the future" (Jakobson 1968, p. 12). Here again there is a profound symmetry between verbal code and genetic code: while stable over time, they both vary through history according to processes that are analogous to the evolution of species (Mayr 1966; Lewontin 1974). For instance, the mechanisms that govern the evolution of words themselves are similar to the mutation and recombination processes analyzed in biology (Searls 2001). How are we to understand this tension between "stability" and "variability"? It would be a mistake to accuse linguists and biologists of being illogical or paradoxical because they attribute these two contradictory qualities to the evolution of languages and species. By opposing these two terms, they underscore that languages and species are, each in their own way, intangible but not inalterable. Is this sufficient, however, to trace the phylogenesis of languages to a primordial genetic root?

Emergence and Phylogenesis

Well before the works of Chomsky, language historians attempted to explain the transmission of languages through the hypothesis of an Indo-European protolanguage (Aitchison 1999). Analogies between the evolution of species and languages have inspired many authors since, lending support to arguments against creationist theories (Pennock 1999). Cavalli-Sforza conducted an in-depth analysis on how population genetics could help us to understand the evolution of languages from a demographic and phylogenetic standpoint (Cavalli-Sforza 1988). This meticulous work made it possible to establish a genealogical tree of human languages from their origin (Ruhlen 1994).

The comparison of vocabularies, called lexicostatistics, was the principal method used by linguists to retrace the genealogical evolution of languages (Campbell 1999), mainly by searching for a "minimalist program" (Chomsky 1995). The compilation of the lexical cores across different languages is analogous to the search in biology for a "minimal genome", i.e. the combination representing the smallest number of genes liable to engender life (Mushegian 1999). As for the emergence of the first grammatical forms, the mystery remains whole. Linguists agree that some "undefined" event occurred in the evolutionary process. This event occurred just once, it would appear, apparently quite recently, since no trace of language has been found until approximately 50,000 years ago. The neuroanatomic features required for linguistic faculties, however, appear to have emerged roughly 150,000 years earlier (Chomsky 2003). Before then, there is little archaeological evidence of linguistic faculties (Davidson 2003).

Bickerton believes that the transition from pidgin to Creole could reveal the evolutionary process through which a form of communication devoid of grammar could eventually produce a complete language form (Bickerton 1990, 1995). According to him, Homo erectus developed a "protolanguage," i.e. a form of expression that connected utterances to pre-existing concepts. He emphasizes, however, that any animal equipped with a sufficiently developed representation system could also have acquired these linguistic faculties. A single mutation that coincided with the transition of Homo erectus to Homo sapiens allowed the creation of language based on the protolanguage. Was this mutation genetic? What triggered it?

Protolanguage Definitions

Definitions of the protolanguage have been formulated in several different ways. Generally speaking, a protolanguage was defined as a form of expression in which words are merely grouped in short utterances, with no grammatical support. Its characteristics are: no grammatical words, no long-range dependency within the sentence, no inflection, and no consistent order. Protolanguage is what we settle for when we are in linguistic trouble (Dessalles 2006). It is a precursor of language, an intermediate skill between spontaneous primate communication and language proper, which is universally used in our species.

Two opposite definitions of "protolanguage" have been proposed: one synthetic, the other holistic. According to the synthetic approach (Bickerton 1998; Jackendoff 2002), the protolanguage had symbols that could be used to convey atomic meanings, and these proto-words could be strung together in ad hoc sequences. Language developed from such a protolanguage through the synthesis of these words into increasingly complex, formally structured utterances. On the other hand, the holistic approach suggests that words emerge from longer, entirely arbitrary strings of sounds—non-compositional utterances—via a process of fractionation. Such holistic utterances initially have no internal structure. They represent whole messages. The idea is that over time chance phonetic similarities are observed between sections of utterances, and if similar meanings can be ascribed to these strings, then "words" emerge (Wray 1998; Arbib 2005).

The holistic and synthetic approaches to protolanguage—as well as their critics (Bickerton 2003; Smith 2006; Tallerman 2007)—both refer to a human protolan-

guage. But why should the origin of natural languages be found only in humankind? The hypothesis of a genetic protolanguage breaks with this anthropocentric approach, suggesting that the emergence of a linguistic prototype occurred long before the pre-lingual era, and may date as far back as the end of the pre-biological era. Since genetic grammar is universal in the living world, in some respects it is consistent with Bickerton's hypothesis that linguistic faculties could have emerged in other animal species. It also concurs with Chomsky's hypothesis of a universal grammar inherent in all natural languages.

Beyond their formulations and intrinsic differences, protolanguage hypotheses all study the combinatorial processes of discrete elements enabling the emergence of more elaborate semantic structures. How this combinatoriality emerged is probably the central issue of language evolution (Fitch 2005; Bowie 2006). Combinatorics of discrete elements is not limited, however, to language and heredity. The concept seems to be at work in nature every time a wide diversity of structures is achieved with a limited number of materials. Nevertheless, the point is not to create complexity from simplicity, as occurs in Mendeleïev's table of periodic elements. The deeper analogy between what one finds in genetics and linguistics lies in the fact that the combination of elements that are devoid of meaning, and simple, not only results in something more complex but, more importantly, in something that contains a certain meaning. "The analogy between genetics and linguistics occurs at the level of meaning, and we cannot avoid using this concept of meaning to properly define the analogy" (Lévi-Strauss 1968, p. 18).

Transmitter & Receiver

Many biologists consider that syntactic analogies between genetics and linguistics reach their limit on the semantic level because the nature of genetic code does not contain any components capable of "understanding" the overall message. The communication function of the two systems features obvious differences. Linguistics studies the message transmitted by a transmitter to a receiver. According to Jacob, however, there is no such thing in biology: no transmitter and no receiver. "No one ever wrote the well-known heredity message that is transmitted from one generation to the next. It came into being on its own, slowly, painstakingly, through the vicissitudes of the reproductions that underlie evolution. No one really receives the message either" (Jacob 1974, p. 200).

"The semantics of the genetic message leads to the controversial debate over the "teleological" or "teleonomic" dimension of living systems" (Pittendrigh 1958). When Monod and Jacob realized the teleological connotations of their metaphors, they contested the linguistic model after supporting it. They claimed—falsely—that they had never really supported it, that it was just a confusion of terms. Monod was eventually embarrassed and forced to clarify his statement: "I simply committed a confusion of language, borrowing terms from linguists to describe what we consider to be as mechanical as a machine. (…) For biologists, the mechanics of the code is comparable to a photocopier, not a language" (Monod 1974, p. 76).

Like many other biologists, Henri Atlan also denies the idea of semantics in genetic systems. A disciple of Shannon's theory of communication which removes all form of meaning from the concept of information (Shannon and Weaver 1949),

Atlan focuses exclusively on the physico-chemical properties of the genetic code and rejects all semiotic dimensions of the genomic message (Atlan 1999). He simply distinguishes artificial programs and natural programmes: artificial programs (computer) have a meaning that comes from a programmer located outside the system; inversely, natural programmes (genetic) have no programmer. The meaning we attribute to them comes from within the system. In other words, the genetic programme has no programmer other than itself, which justifies the theory of the self-organization of life (Atlan 1972). According to this autopoietic approach, only "creative chance" could be at the origin of the genome and its remarkable syntactic organization remains, despite appearances, devoid of all semantics (Kjosavik 2006; Schurz 2007). But would it not be wonderful if the genetic code emerged from nothingness, and if, as Bernard Shaw put it, "a swamp of amoebae, with time, became the French Academy?"

Beyond the controversies over the origin of life (Carrier 2005), the hypothesis of a genetic protolanguage reverses the debate over the use of linguistic metaphors in biology (Abel and Trevors 2006). Language was always believed to be a product of culture. Could it be a product of nature? According to the genetic protolanguage hypothesis, human observers may not be projecting linguistic frameworks onto genomic structures. Rather, it could be their linguistic faculties that reflect the grammatical structure of genetic code. In other words, the hypothesis of a genetic protolanguage could be more than just an anthropomorphic metaphor. The genetic code may represent "the Code of Codes" (Kevles and Hood 1992), i.e. the original matrix of all natural languages. This hypothesis led to Jakobson's censorship, because his approach emphasized the teleological properties of genetic code (Jakobson 1974). The Harvard professor was denied publication of an article submitted in 1973 to The New York Review of Books on the grounds that he was advocating a teleological approach that challenged the prevailing neo-Darwinian interpretation (Kay 2000).

Form and Information

At the centre of the debate, the life sciences and human sciences are divided over the semantic dimension of DNA. This division is not over the existence of the code itself: it concerns the existence of a decoder. Indeed, while it is easy to identify a human brain as an interlocutor capable of understanding human language, it is harder to conceive of an anthropomorphic interlocutor when it comes to genetic language. In this case, nucleic acids contain an embedded non-verbal message, and the interlocutors are not linguistic subjects, but rather semiotic entities. Yet does this mean that the code has no semantic content?

Linguistics deals mainly with the path of a discourse, i.e. the alternating roles of the sender/receiver who answers his/her interlocutor. There are, however, analogous systems in biology. According to the immunologist Niels Jerne, the deeper connection between linguistics and immunology lies in the immune system's vast repertoire. This repertoire is not a vocabulary made of words, but a lexicon of sentences that can answer any one of the sentences expressed by the multitude of antigens the immune system may encounter (Jerne 1984). Indeed, life is teeming with messengers: antibodies, hormones, neurotransmitters, etc. Through lock-and-

key binding, at every level, transmitters and receivers answer each other with mutual signs that trigger feedback reactions. Incomparably more subtle than a simple traffic light system, life offers evidence of an intense communicative relationship between the whole and its parts, dedicated to a constant quest for a balance between variability and stability. While metabolic reactions do not exactly solve the mystery of the transmitter and the receiver, they do confirm when messages have been received properly. The performative dimension of the message attests to the existence of meaningful information (Benichou 2002).

There is an ongoing controversy as to whether the genome is a representing system (Shea 2006). Does the interpretation of molecular structures show that genetic information has intrinsic semantics? The role of the encoding process consists of transforming the information of a medium with a single spatial dimension (nucleic acid) into a three-dimensional superstructure (protein). Genomic expression organizes the deployment of a primary structure into a secondary and then a tertiary structure. The stereochemical configuration of proteins attests to a meaningful transfer of information, since a higher level of complexity is reached, with a richer informational content (Monod 1970). The information–formation–function scheme expresses a coded relationship between a nested combinatorial order (genotype) and an integrated spatial arrangement (phenotype). In other words, the morphogenesis of life reveals the deployment of a semantic order. This is a relatively modest result, however, since there are two separate issues at hand: knowing if there is a meaning, and understanding what that meaning is. To answer the second question the life sciences must analyze the genetic code at a deeper level using hybrid techniques at the crossroads of the information sciences and molecular biology.

Nucleic Acid Linguistics

In the 1980s several workers began to follow various threads of Chomsky's legacy in applying linguistic methods to molecular biology. Early results included the demonstration of the utility of grammars in capturing not only informational but also structural aspects of macromolecules (Searls 1988). From this work there followed a series of mathematical results concerning the linguistics of nucleic acid structure (Searls 1989, 1992, 1993). These results derive from the fact that a folded RNA secondary structure entails pairing between nucleotide bases that are at a distance from each other in the primary sequence, establishing relationships that in linguistics are called "dependencies". The most basic secondary-structure element is the stem-loop, in which the stem creates a succession of nested dependencies.

In the light of these practical consequences of linguistic complexity, a significant finding is that there exist phenomena in RNA that in fact raise the language even beyond the context-free. The most obvious of these are so-called non-orthodox secondary structures such as pseudoknots, which are pairs of stem-loop elements in which part of one stem resides within the loop of the other. This configuration induces cross-serial dependencies in the resulting base pairings, requiring context-sensitive expression. Predictably, given this further promotion in the Chomsky hierarchy, the need to encompass pseudoknots within secondary-structure recognition and predication programmes has significantly complicated algorithm design (Lyngso 2000; Searls 2002).

The usefulness of grammars for understanding the informational and structural dimensions of macromolecules has also been demonstrated (Searls 1988). A series of mathematical results concerning linguistics and the structure of nucleic acids was derived from these works (Searls 1989, 1992, 1993). Using formalisms called tree-adjoining grammars and their variants—which are considered to be mildly context-sensitive and relatively tractable—it is possible to encompass a wide range of RNA secondary structures (Uemura et al. 1999). Additionally, new types of grammars have been invented to deal with such biological examples (Searls 1995; Rivas and Eddy 2000). Natural languages seem to be beyond context-free as well, based on linguistic phenomena entailing cross-serial dependencies, although in both domains such phenomena seem to be less common than nested dependencies. Thus, by one measure at least, nucleic acids may be said to be at about the same level of linguistic complexity as natural human languages.

The modulation of genetic information flows was also clarified through an analysis of the mechanisms governing RNA interference (RNAi). From the pairing of two RNA segments, the resulting RNAi neutralizes the expression of the corresponding nucleic sequence and prevents synthesis of the encoded protein. By combining two complementary sequences, the gene becomes silent and the corresponding protein is not expressed (Fire et al. 1998). This catalytic process is behind the semantic modulation of the genome: the expression of meaningful sequences is only possible if other sequences are silenced (Fire et al. 2006). Faced with this alternation of expressions and silences, the fundamental questions of biolinguistics return to the fore: what are the minimal properties required to build a hierarchically structured system of representation and expression? What fundamental factors enable the activation of such a system (Hauser et al. 2002)? “To connect the dots is no trivial problem” (Chomsky 2005, p. 12).

Genomics and Literary Linguistics

To gain a deeper understanding of the genetic grammar and make advances in genome sequencing, researchers have also drawn analogies between genomics and literary linguistics. Literary linguistics refers to stylistic study, textual analysis and literary criticism. While foreign to hard sciences such as molecular biology, at certain levels this discipline shares a common ground with the methods used in bioinformatics to compare texts, identify subtle relations or understand textual variations, including through the use of quantitative methods (Barnbrook 1996). The work of the linguist George K. Zipf allowed a mathematical law to be established which made it possible to analyze the occurrences and frequency of the words in a text, from which the fractal nature of language was assumed (Zipf 1949; Mandelbrot 1983). These fractal natures are also observed at the deepest levels of molecular biology, in particular for analyzing the frequency of oligonucleotides, the size of gene families, protein distribution, RNA folding, and even the levels of expression of genes (Mantegna 1994; Huynen and van Nimwegen 1998; Harrison and Gerstein 2002; Qian et al. 2001; Schuster et al. 1994; Hoyle et al. 2002).

Textual criticism has objectives and instruments that are similar to those of bioinformatics. Word frequency and figures of speech are used in literature through clustering methods, but they are also based on methods used in biology, such as

neural networks or genetic algorithms applied to experiments on DNA micro-arrays (biochips; Yandell and Majoros 2002; Searls 2001). Likewise, the branch of textual criticism called "stemmatics" traces the origin and accuracy of ancient texts through the many corrections and copies made over the ages by scribes, the fragmentary sources used, the translators, etc. For manuscripts copied numerous times by scribes, mathematical models were developed to trace the phylogenesis of the text back to its primitive version, with methods similar to those used in genomics (Barbrook et al. 1998; Platnick and Cameron 1977). These illustrations underscore the degree to which linguistics and genomics share more than just methods: they also use symmetrical techniques. At a deeper level, does this epistemological convergence imply that these two disciplines, while using distinct and complementary approaches, are exploring one and the same object?

Conclusion

In this post-Darwinian era, progress in molecular biology seems bound to linguistic metaphors, though biologists attempt to reject their symbolism. The genetic code contains a dual mystery: the cipher, and the order that deciphers it. The status of genetic code becomes a sort of epistemological chimera: like a centaur, with its animal body and human torso, the genome crystallizes a hybrid structure whose semantic expression is at once biochemical and physiological (Bastide 1985). After tackling cybernetic models, geneticists are now attempting to understand the relationship between the inept embedded in the genotype and the unfit expressed in the phenotype. Since the sequencing of the human genome, the biological sciences are discovering that the key to understanding the living no longer lies in breaking down the material structure of macromolecules into their deepest elements, but rather in eliciting the immaterial relationship between those elements (Danchin 2002). Based on immaterial combinatorics, the genetic code must be interpreted as a code capable of converting chemistry into syntax, messaging into a message, and signals into signs. The twenty-first century has opened up a new epistemological era in the life sciences where the classic structure–function approach is shifting toward a new code-meaning scheme.

After exploring the structural symmetries between the genetic and verbal codes, we conclude that the linguistic concepts used in biology are more than just heuristic metaphors. Though tainted by anthropomorphism, they may refer to a sophisticated form of protolanguage whose genetic grammar could have gradually mutated into several stages of expression: nucleic, proteic, physiological, verbal. The verbal stage, which seems specific to humankind, may not be the final evolutionary level of this universal grammar. Breaking with an anthropocentric approach, the genetic protolanguage hypothesis suggests a Copernican reversal: human observers may not be projecting linguistic frameworks onto genomic structures; rather, it could be their linguistic faculties that reflect the grammatical structure of genetic code. This universal genetic grammar would clarify why the evolutionary mechanisms specific to languages and to species are similar. It would also help to explain their polymorphisms and the physiological basis of natural languages.

Despite the scepticism and criticisms it may incur, the genetic protolanguage hypothesis is gradually gaining support from the growing amount of findings in

genomics and proteomics, mainly thanks to advances in bioinformatics and biolinguistics. From an epistemological standpoint, this hypothesis may reconcile several theoretical models on the origin of language. Its confirmation from an experimental standpoint could enable significant advances in research into the existence of a universal grammar. Such research would also shed light on the process through which linguistic faculties emerge, and help us to understand the symmetries between the phylogenesis of languages and species. If this hypothesis were one day to be verified, linguistics would become a branch of biology.

References

Abel, L., & Trevors, T. (2006). More than metaphor: Genomes are objective sign systems. *Journal of Biosemiotics*, *1*(2), 253–267.

Aitchison, J. (1999). *Linguistics*. Chicago: NTC/Contemporary Publishing.

Arbib, M. (2005). From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics, *Behavioral and Brain Sciences*, *28*, 105–167.

Atlan, H. (1999). *Vers de nouveaux paradigmes en biologie*. Paris: Inra.

Atlan, H. (1972). *L'Organisation biologique et la Théorie de l'information*. Paris: Hermann.

Baldi, P., Brunak, S. (2001). *Bioinformatics: The machine learning approach*. Cambridge, MA: MIT Press.

Barnbrook, G. (1996). *Language and computers*. Edinburgh: Edinburgh University Press.

Barbrook, A., Howe, C. J., Blake, N., & Robinson, P. (1998). The phylogeny of the Canterbury Tales. *Nature*, *394*, 839.

Barbieri, M. (2007). Introduction to biosemiotics. Berlin: Springer.

Bastide, F. (1985). Linguistique et génétique, Actes Sémiotiques, 33.

Berkeley, G. (1710). A treatise concerning the principles of human knowledge. In J. Dancy (Ed.), Oxford: Philosophical Texts.

Benichou, G. (2002). *Le Chiffre de la vie: réconcilier la génétique et l'humanisme*. Paris: Seuil.

Bickerton, D. (1990). *Language and species*. Chicago, IL: University of Chicago Press.

Bickerton, D. (1995). *Language and human behaviour*. London: University College London Press.

Bickerton, D. (1998). Catastrophic evolution: The case for a single step from protolanguage to full human language. In J. R. Hurford, M. Studdert-Kennedy & C. Knight (Eds.), *Approaches to the evolution of language: social and cognitive bases* (pp. 341–358). Cambridge: Cambridge University Press.

Bickerton, D. (2003). Symbol and structure: A comprehensive framework for language evolution. In M. H. Christiansen & S. Kirby (Eds.), *Language evolution* (pp. 77–93). Oxford: Oxford University Press.

Bowie, J. (2006). The evolution of meaningful combinatoriality. Evolution of Language, 6th International Conference, Rome 12–15 April, 2006.

Broca, P. (1875). *Instructions craniologiques et craniométriques*. Paris: Masson.

Campbell, L. (1999). *Historical linguistics: An introduction*. Cambridge, MA: MIT Press.

Carrier, R. C. (2005). The argument from biogenesis: Probabilities against a natural origin of life. *Biology & Philosophy*, *19*(5), 739–764.

Cavalli-Sforza, L. L. (2000). *Genes, peoples and languages*. New York: North Point Press.

Cavalli-Sforza, L. L. (1988). Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences*, *85*, 6002–6.

Chomsky, N. (2001). *New horizons in the study of language and mind*. Cambridge: Cambridge University Press.

Chomsky, N. (2005). Three factors in language design. *Linguistic Inquiry*, *36*(1), 1–22.

Chomsky, N. (1995). *The minimalist program*. Cambridge (MA): MIT Press.

Chomsky, N. (2004). Beyond explanatory adequacy. In A. Belletti (Ed.), *The cartography of syntactic structures, vol. 3, Structures and beyond*. Oxford: Oxford University Press.

Chomsky, N. (2003). *On nature and language*. Cambridge: Cambridge University Press.

Chomsky, N. (1957). *Syntactic structures*.The Hague: Mouton.

Chomsky, N., Miller, G. (1963). *Introduction to the formal analysis of natural languages, in Handbook of Mathematical Psychology*. New York: Wiley.

Chomsky, N. (1972). *Language and mind*. New York: Harcourt Brace Jovanovich.

Danchin, A. (2002). *The Delphic boat: What genomes tell us*. MA: Harvard University Press.

Darwin, C. (1871). *The descent of man*. New Jersey: Princeton University Press.
Darwin, C. (1872). *The origin of species*. PA: University of Pennsylvania Press (1959).
Davidson, I. (2003). The archeological evidence of language origins: States of art. In M. Christiansen, & S. Kirby (Eds.), *Language evolution*. Oxford: Oxford University Press.
Dessalles, J.-L. (2006). From protolanguage to language: model of transition*Marges linguistiques*, *11*, 142–152.
Durbin, R., Krogh, A., Mitchison, G., & Eddy, S. (1988). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge: Cambridge University Press.
Fire, A., Xu, S. Q., Montgomery, M. K., Kostas, S. A., Driver, S. E., & Mello, C. C. (1998). Potent and specific genetic interference by double stranded RNA in *Caenorhabditis elegans*. *Nature*, *391*, 806–811.
Fitch, W. T. (2005). The evolution of language: A comparative review. *Biology & Philosophy*,*20*(2–3), 193–203.
Fire, A., Mello, C. C., & Nobel Lecture (2006). October 2, Karolinska Institutet, Stockholm.
Fox-Keller, E. (1995). *Refiguring life: Changing metaphors in twentieth-century biology*. New York: Columbia University Press.
Harrison, P. M., & Gerstein, M. (2002). Studying genomes through the aeons, protein families, pseudogenes and proteome evolution.*Journal of Molecular biology*, *318*, 1155–1174.
Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*, 1569–1579.
Hauser, M. D. (1996). *The evolution of communication*. Cambridge, MA: MIT Press.
Head, T. (1987). Formal language theory and DNA: An analysis of the generative capacity of specific recombinant behaviours. *Bulletin of mathematical biology*, *49*, 737–759.
Hoyle, D. C., Rattray, M., Jupp, R., & Brass, A. (2002). Making sense of microarray data distributions. *Bioinformatics*, *18*, 576–584.
Huynen, M. A., & van Nimwegen, E. (1998). The frequency distribution of gene family sizes in complete genomes. *Molecular biology and evolution*, *15*, 583–589.
Jacob, F. (1965). *Leçon inaugurale au Collège de France*. Paris: Collège de France archives.
Jacob, F. (1966). Genetics of the bacterial cell. *Science*, *150*, 1464–1470.
Jacob, F. (1970). *La Logique du vivant*. Paris: Gallimard.
Jacob, F. (1971a). The logic of life, Princeton Science Library.
Jacob, F. (1971b). Le modèle linguistique en biologie, Nouvelle Critique, Paris, Octobre.
Jacob, F. (1974). Le modèle linguistique en biologie, Critique, Paris, Éd. de Minuit, 320.
Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford: Oxford University Press.
Jakobson, R. (1968). Vivre et parler, Les Lettres françaises, n° 1221.
Jakobson, R. (1973). La linguistique et les sciences naturelles, Essais de linguistique générale, Paris, Éd. de Minuit.
Jakobson, R. (1974). Vie et langage, Dialectiques, Paris, Presses Universitaires de France, no 7.
Jerne, N. (1984). The generative grammar of the immune system. Nobel Lecture, Dec 8, 1984.
Johansson, S. (2005). *Origins of Language—Constraints on hypotheses*. Amsterdam: Benjamins.
Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing*. Upper Saddle River, NJ: Prentice Hall.
Kay L. (2000) *Who wrote the book of life: A history of the genetic code* CA Stanford University Press
Keenan, E., & Stabler, E. (2003). Linguistic invariants and language variation. 12th International Congress. Logic Methodology and Philosophy of Science, LMPS'03, Oviedo, Spain, August 7–13, 2003.
Kevles, D., & Hood, L. (1992). *The code of codes: Scientific and social issues in the human genome project*. Cambridge: Harvard University Press.
Kjosavik, F. (2006). From symbolism to information? Decoding the gene code.*Biology & Philosophy*, *22* (3), 333–349.
Lenneberg, E. (1967). *Biological foundations of language*. New York: Wiley.
Lévi-Strauss, C. (1968). Vivre et parler, Les Lettres françaises, no 1221.
Lévi-Strauss, C. (1971). *L'Homme nu*. Paris: Plon.
Lewontin, R. (1974). *The genetic basis of evolutionary change*. NY: Columbia University Press.
Lieberman, P. (1984). *The biology and evolution of language*. Cambridge, MA: Harvard University Press.
Lin, J., & Gerstein, M. (2000). Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Research*, *10*, 808–818.
Lyngso, R. (2000). RNA pseudoknot prediction in energy-based models. *Journal of Ccomputational Biology*, *7*, 409–427.
Mandelbrot, B. (1983). *The fractal geometry of nature*. San Francisco: Freeman.

Mantegna, R. N., et al. (1994). Linguistic features of noncoding DNA sequences. *Physical Review Letters*, *73*, 3169–3172.
Mayr, E. (1966). *Animal species and evolution*. Cambridge: Massachusetts.
Mayr, E. (1961). Cause and effect in biology. *Science*, *134*, 1504–1506.
Monod, J. (1970). Le Hasard et la Nécessité, Paris, Éd. du Seuil.
Monod, J. (1974). *L'Unité de l'homme, Colloque de Royaumont*. Paris: Seuil.
Mushegian, A. (1999). The minimal genome concept. *Current Opinion in Genetics & Development*, *9*, 709–714.
Pennock, R. T. (1999). *Tower of Babel: The evidence against the New Creationism*. Cambridge, MA: Bradford MIT Press.
Pittendrigh, C. (1958). *Behavior and evolution*. New Haven, Connecticut: Yale University Press.
Platnick, N. I., & Cameron, H. D. (1977). Cladistic methods in textual, linguistic and phylogenetic analysis. *Systematic Zoology*, *26*, 380–385.
Qian, J., Luscombe, N. M., & Gerstein, M. (2001) Protein family and fold occurrence in genomes: Power-law behaviour and evolutionary model. *Journal of Molecular Biology*, *313*, 673–681.
Rivas, E., & Eddy, S. (2000) The language of RNA: A formal grammar that includes pseudoknots. *Bioinformatics*, *16*, 334–340.
Ruhlen, M. (1994). *On the origin of languages: Studies in linguistic taxonomy*. CA: Stanford University Press.
Saussure Ferdinand, (de), (1908). Cours de linguistique générale, Paris, Payot, 1972.
Schurz J. (2007) Probability and evolution. Why the probability argument of Creationists is wrong. *Journal for General Philosophy of Science*, *38*(1), 163–165.
Schuster, P., Fontana, W., Stadler, P. F., & Hofacker, I. L. (1994). From sequences to shapes and back: A case study in RNA secondary structures. *Proceedings of the Royal Society of London*, *B255*, 279–284.
Searls, D. (2002). The language of genes.*Nature*, *420*, 211–217.
Searls, D. (2001). Reading the book of life. *Bioinformatics*, *17*, 579–580.
Searls, D. (2001). Mining the bibliome. *Pharmacogenomics journal*, *1*, 88–89.
Searls, D. (2001). From Jabberwocky to genome: Lewis Carroll and computational biology. *Journal of Comparative Biology*, *8*, 339–348.
Searls, D. (1999). *Mathematical support of molecular biology*. (117–140). Providence, RI: American Mathematical Society (edited by F.-C. Roberts, & V. Waterman).
Searls, D. (1995). String variable grammar: A logic grammar formalism for DNA sequence. *J. Logic Program*, *24*, 73–102.
Searls, D. (1993). *Artificial intelligence and molecular biology, Ch. 2.* (pp. 47–120). Menlo Park, CA: AAAI Press (edited by L. Hunter).
Searls D. (1992). The linguistics of DNA. *American Scientist*, *80*, 579–591.
Searls, D. (1989). Logic programming. Proceedings of the North American Conference. (pp. 189–208). Cambridge, MA: MIT Press (edited by E. Lusk, R. Overbeek).
Searls, D. (1988). Proceedings of the 7th National Conference on Artificial Intelligence. (pp. 386–391). Menlo Park, CA: AAAI Press.
Segre' (2000). Compositional genomes: Prebiotic information transfer in mutually catalytic noncovalent assemblies. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 4112–4117.
Segre' (2002). Language, genes and the evolution of combinatorics. Evolution of Language: 4th International Conference, Harvard University.
Shannon, C., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.
Shea N. (2006). Representation in the genome and in other inheritance systems.*Biology & Philosophy*, *22* (3), 313–331.
Smith, K. (2006). The protolanguage debate: Bbridging the gap? In A. Cangelosi, A. D. M. Smith & K. Smith (Eds.), *The evolution of language*. Proceedings of the 6th International Conference, (pp. 315–322).
Snel, B.B., Bork, P., & Huynen, M. A. (2000). Genome phylogeny based on gene content. *Nature Genetics*, *21*, 108–110.
Stegmann, U. E. (2004). The arbitrariness of the genetic code. *Biology & Philosophy*, *19*(2), 205–222.
Tallerman, M. (2007). Did our ancestors speak a holistic protolanguage. *Lingua*, *117*(3), 579–604.
Tatusov, R. L., Galperin, M. Y., Natale, D. A., & Koonin, E. V. (2000). The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, *28*, 33–36.
Tekaia, F., Lazcano, A., & Dujon, B. (1999). The genomic tree as revealed from the whole proteome comparisons. *Genome Research*, *9*, 550–557.

Trubetskoy, N. (1936). Die phonologischen Grenzsignale. Proceedings of the 2nd International Congress of Phonetic Sciences, Cambridge, 1936.

Uemura, Y., Hasegawa, A., Kobayashi, S., & Yokomori, T. (1999). Tree-adjoining grammars for RNA structure prediction. *Theoretical computer science*, *10*, 277–303.

Wray, A. (1998). Protolanguage as a holistic system for social interaction. *Language and Communication*, *18*, 47–67.

Yandell, M. D., & Majoros, W. H. (2002). Genomics and natural language processing. *Nature Reviews. Genetics*, *3*, 601–610.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Boston, MA: Addison-Wesley.