The Nucleation of Semantic Information in Prebiotic Matter

Bernd-Olaf Küppers

Abstract The analysis of the inherent context-dependence of genetic information suggests that there are evolutionary mechanisms which are independent of the processes of environmental adaptation and yet are able to push prebiotic matter towards functional complexity. In this regard, the extension of information space, by random prolongation of the primary structure of biological macromolecules, must have played a decisive role in the origin of life. On the one hand, the extension of information space is tantamount to an increase in the syntactic complexity of potential information carriers, which in turn is a prerequisite for the nucleation and evolution of semantic information. On the other hand, the increase in the dimensionality of information space expands the number of possible pathways of evolutionary optimisation and thereby improves the possible choices that can be made by progressive evolution. Alongside the optimisation of evolutionary optimisation itself, there are principles of evolutionary dynamics that direct the formation of functional order in prebiotic matter. Since these principles are constitutive for the proto-semantics of genetic information, they may be regarded as the elements of the semantic code of evolution.

Contents

B.-O. Küppers (✉)
Friedrich Schiller Universität Jena, Jena, Germany
e-mail: bernd.kueppers@uni-jena.de

Current Topics in Microbiology and Immunology (2016) 392: 23–42
DOI 10.1007/82_2015_454

Published Online: 29 July 2015

1 Life = Matter + Information

The present-day understanding of living matter is based essentially on two fundamental assumptions, which are the epistemic guidelines of modern biology:

1. Living matter differs from non-living matter by its high degree of functional order. The transition from non-living to living matter is assumed to be a continuous one. This implies that there is no intrinsic difference between these two forms of matter.
2. The overarching concept for the understanding of living matter is the Darwinian theory of natural selection and evolution.

The claim that there is a continuous transition from non-living to living matter requires closer specification. First of all, we must think of it as a quasi-continuous transition, since matter itself is not a continuous substance. However, more important: Even if the transition is a quasi-continuous one, we still cannot draw a sharp borderline between non-living and living matter. For purely logical reasons, it is impossible to find a definition that expresses an intrinsic difference between the living and the non-living and which at the same time is free of tautology, i.e. of life-specific notions. Instead, the problem of defining life becomes a normative one; that is, the definition will always depend upon the particular paradigm that we regard as appropriate for an understanding of the phenomena of life (Küppers 2000).

The working hypothesis that the transition from non-living to living matter is a continuous one has also an important consequence for the methodology of modern biology. This is because it implies that in living matter the physical and chemical laws are valid, without any exceptions. Moreover, it follows that no additional laws are necessary for a deeper understanding of the phenomena of life. This, however, does not exclude the possibility that the laws of physics and chemistry operate in living matter as a special case—like for example Ohm's famous law, which is adhered to in an electrical circuit as a special case of the general laws of electrodynamics. "Special case" laws operate owing to the special organisation of matter, but they are not an inherent characteristic of matter itself. An important example from biology is the principle of natural selection (see below).

The consistent application of the idea that all life phenomena can in principle be reduced to the basic processes of physics and chemistry is known as the "reductionistic" research program of biology. Although this research program has been exceptionally successful in the past, it has been criticised again and again. Yet behind all the criticism hides a fundamental misunderstanding: the allegation that physics and chemistry still retain the naïve mechanistic view of Nature that was held at the end of the eighteenth century. However, this allegation is wrong. During the last two centuries, physics and chemistry have undergone perpetual change and have extended their theoretical concepts beyond a simple mechanistic understanding of matter.

One of the most important changes took place during the past decades during the development of the so-called structural sciences (Küppers 2000). This new branch of science has arisen within the framework of the analysis of complex systems in Nature and society. The structural sciences pursue the goal of studying the abstract and overarching structures of reality, independently of the question of whether they are found in natural or artificial systems, in non-living or living matter. The best-known examples of this type of science are cybernetics, information theory, systems theory and game theory. Other disciplines—such as network theory, synergetics, complexity theory and the theory of fractals, to mention but a few—enrich the classical reservoir of structural sciences and are increasingly permeating the basic concepts of physics and chemistry as well.

Among all the structural sciences, the theory of information is of central importance for the theoretical understanding of biology, since all basic processes of life are instructed by information. Even the classical concept of Darwinian evolution received a firmer foundation under the influence of modern information theory, which makes the origin of life appear in a new light (Küppers 1990).

With regard to the all-encompassing role of information in living matter, it seems to be justified to rephrase the famous evolutionary dictum

> "Nothing in biology makes sense except in the light of evolution" (Dobzhansky 1973)

in the apodictic assertion

> "Nothing in biology makes sense except in the light of information" (Küppers 2000).

Although the concept of information has become the most important and successful concept for the theoretical understanding of living matter, it is very often called into question. Information, so the criticism, is a notion taken from our cultural world. It has its origin in human communication and can by no means be applied to natural objects. This is to say: matter and information are incommensurable notions and are essentially alien to each other.

However, the criticism does not hold. Information can perfectly well be reduced to physical terms and be applied to natural objects such as genes (Küppers 1992). To shed some light on this, we have to focus on the organisation of living matter. This organisation consists of a hierarchy of material boundaries at all levels of biological complexity (Küppers 1990). The notion of "boundaries" is borrowed from physics. In physics, the term "boundaries" normally denotes the constraints upon the system, like the walls of a gas container or the movement of a bead on a wire. In traditional physics, those boundaries are considered to be "contingent", i.e. they are neither random nor determined by laws. They can be as they are, but they could also have another form. If we change, for example, the walls of a gas container within moderate limits, this will not have any serious influence on the physical processes going on in the system. In contrast to systems of that kind, the boundaries of "functional" systems are exceptional in the sense that they are "non-contingent" properties of the system (Küppers 1992). This means that such systems are very critically dependent upon their boundaries, so that even a marginal change in the boundaries may lead to the collapse of the system's functional properties.

In other words, non-contingent boundaries are highly selective constraints upon the action of natural laws. They restrict all conceivable natural processes to those that are actually operating in the system. This is exactly the physical meaning of the notion of information in biology. It expresses the fact that all essential processes of a living system are instructed by specific physical boundaries, which are encoded in the detailed molecular structure of the genome.

From this point of view, a physical theory of the origin of life has to explain how under prebiotic conditions non-contingent physical boundaries could originate from contingent ones. Since the expectation value of a specific boundary condition—i.e. of the appearance of a macromolecule that carries biological information—is extremely small, specific boundary conditions could not originate by pure chance. However, the statistical analysis of this problem shows that such boundary conditions may appear through the selective self-organisation of matter (Küppers 1987). This suggests that the key for a physical understanding of the origin of genetic information may be sought in the physical foundation of the principle of natural selection.

2 Natural Selection of Information

For a long time, the Darwinian principle of natural selection seemed to be a physical riddle. Natural selection was considered either to be a tautology ("survival of the survivor") or to be a specific property of living matter that could not be reduced to the known principles of physics and chemistry. In fact, until the middle of the last century, the reproductive self-maintenance of living matter—obviously a necessary prerequisite for natural selection—was an unknown property in physics. The breakthrough came only with the epoch-making discovery of the molecular structure of DNA, which demonstrated that the capability of living beings to reproduce themselves is not an irreducible property of living matter, but rather a direct consequence of the physical and chemical properties of the genetic material. Two decades later, it was also demonstrated that Darwinian selection and evolution among molecules is in fact possible, and that such processes can even be simulated in the test tube under cell-free conditions (Mills et al. 1967).

These discoveries opened the door to a physical foundation of the principle of natural selection (Eigen 1971). Here, the first important step was to set up a model system, one that provides reproducible physical conditions for the investigation of the elementary processes of molecular evolution. Such a system has been termed an "evolution reactor". This is an experimental device that best can be compared to an idealised "prebiotic soup". The concept of the evolution reactor was initially a drawing-board idea that served the theoretical study of molecular evolution (Eigen and Schuster 1979; Küppers 1979, 1983). Later, it was also realised as a biotechnological instrument for the design of biochemical substances by means of artificial evolution.

The evolution reactor is essentially a chemical flow reactor, in which a population of self-reproducing biopolymers (e.g. nucleic acids) is competing for nutrients, i.e. energy-rich building-blocks (Fig. 1). Defined reaction conditions can be set up in the system by regulating the overall concentration of biopolymers as well as the supply of energy-rich building-blocks (monomers). Such conditions correspond largely to the experimental conditions under which Darwinian evolution among molecules has been demonstrated in the test tube (Kramer et al. 1974; Mills et al. 1967; Spiegelman 1971).

For a mathematical treatment of the selection process, one has to specify the model systems in more detail. Let us assume that the population inside the reactor consists of i different macromolecular sequences of equal chain length v, whose population numbers per unit volume we denote by x_i. We further assume that in the reactor vessel the total population $Z = \sum_i x_i$ is extremely small in comparison with the number n of all conceivable sequences. The assumption $Z \ll n$ complies with the conditions that presumably prevailed on the primitive Earth. Under this condition, the expectation value of a particular sequence is vanishingly small; it is therefore impossible that the initial distribution, existing in the system at the beginning, could have included all possible sequences.

Moreover, the reaction vessel is assumed to allow the inflow of energy-rich monomers and the outflow of energy-deficient monomers, as outlined in Fig. 1. In principle, there are two possibilities to exercise an experimental control over the

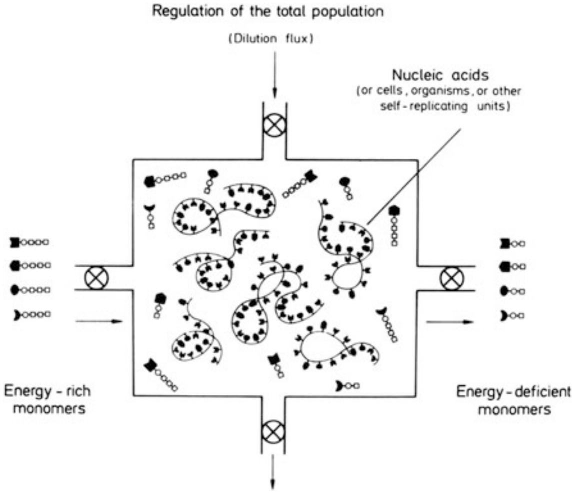
Regulation of the total population
(Dilution flux)
Nucleic acids
(or cells, organisms, or other self-replicating units)
Energy - rich monomers
Energy - deficient monomers

Fig. 1 A model system for the study of molecular self-organisation and precellular Darwinian evolution. In the reactor, there are biological macromolecules (nucleic acids) that are subject to permanent growth and decay. Growth takes place by the consumption of energy-rich monomers that are perpetually supplied to the system from outside (*left*). On the *right*, the energy-deficient decay products are perpetually removed from the system. A variable dilution flux (*top* to the *bottom*) allows the population to be adjusted and—for example—kept at a constant level. From Küppers (1990)

system by retarding or limiting its growth. One can either keep constant the overall population of macromolecules (CP conditions) or the flow rates of the various energy-rich materials (CF condition). In order to take into account Darwin's central idea of the selection mechanism, we introduce the CP assumption in our model system. Thus, the total population is held constant by a dilution flux that is unspecific, i.e. affects equally all substances present. We further assume separate rates of formation and decay of the various competing macromolecular species. In other words, we make the (in this case reasonable) approximation that the formation and decay of biological macromolecules are independent of one another.

We denote the amplification rate of the species x_i as A_i and its decay rate as D_i. The parameters A_i and D_i may depend on the concentrations x_j of other species. Finally we take account of mutability, in that we assume that only a part of the new copies of a particular sequence is error-free. The proportion of correct copies is expressed by a quality factor Q_i. This factor is dimensionless and lies by definition within the range $0 \leq Q_i \leq 1$.

The following equations (a full explanation can be found in Chaps. 1 and 3–5 of this book, as well as in Eigen 1971 and Küppers 1983) describe the change of the variables x_i:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = (A_i Q_i - D_i)x_i - \sum_{j \neq i} \Phi_{ij} x_j - \bar{E}(t) x_i \quad (i,j = 1, \ldots, k), \tag{1}$$

where $\bar{E}(t)$, defined by

$$\bar{E}(t) = \sum_i (A_i Q_i - D_i) x_i \Big/ \sum_i x_i, \tag{2}$$

is the average rate of production of all molecular species.[1]In Eq. (1), $\bar{E}(t)x_i$ denotes a decay term that expresses the contribution made by the ith species to the turnover of individuals in the stationary state (Z = constant). The summed term $\sum_{j \neq i} \Phi_{ij} x_j$ is the contribution to the population number of the master sequence made by all mutant species $x_{j \neq i}$ as a consequence of "back mutation".

The set of Eq. (1) generally describes the kinetics of a reaction system characterised by the properties metabolism, self-reproduction and mutability.

1. *Metabolism* is expressed by the terms $\sum_i A_i x_i$ and $\sum_i D_i x_i$, which describe the turnover from energy-rich to energy-deficient monomers. In other words: The system is open with respect to a flow of matter and energy in the form of activated monomers.
2. *Self-reproduction* is expressed by the form of the reaction equations, in which the rate of formation of a molecular species x_i is proportional to its

[1] Strictly speaking, the $x_i(t)$ should be treated as discrete variables, and the differential equations should be replaced by difference equations. However, this would not alter the conclusions in any significant way, so for the sake of simplicity, we retain the continuous variables.

concentration, independently of how the kinetic parameters A_i and D_i depend on the concentrations x_j of the other species.

3. *Mutability* is expressed by the quality Q_i, which for real systems always fulfils the condition $0 < Q_i < 1$.

Metabolism, self-reproduction and mutability are all necessary conditions for a system being able to undergo evolution.

Let us take a closer look at the mechanism of selection and evolution. The parameters A_i, Q_i and D_i can be condensed into the quantity

$$W_i = A_i Q_i - D_i. \tag{3}$$

It is justified to denote W_i as the *selection value* of the species x_i. This is demonstrated by the following consideration, which is based on a simplification of Eq. (1). Let us assume that the reverse mutations $\sum_{j \neq i} \Phi_{ij} x_j$, which contribute to the species x_i, are negligible (as is the case for long chains). Making use of definition (3), we simplify the selection Eq. (1) and obtain

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = (W_i - \bar{E}(t))x_i \quad (i = 1, \ldots, k). \tag{4}$$

From this, the principle of natural selection follows directly as an extremum principle: All molecular species x_i whose selection value lies below $\bar{E}(t)$ are formed at a negative rate; that is, they die out. All species with a selection value above $\bar{E}(t)$ have positive rates of formation; that is, they increase in number. In consequence of this segregation process, the threshold $\bar{E}(t)$ is displaced to higher and higher levels. As a result, more and more species fall below the "threshold" value $\bar{E}(t)$ and die out. Selection equilibrium is reached when $\bar{E}(t)$ is equal to the greatest selection value in the population, that is,

$$\bar{E}(t) \rightarrow W_{\max}. \tag{5}$$

In selection equilibrium, $\bar{E}(t)$ is constant with respect to time:

$$\bar{E}(t) = W_{\max}. \tag{6}$$

Thus, the principle of selection is revealed within the framework of our model system as a physically justifiable extremum principle.

In the consideration above, the backflow terms $\sum_{j \neq i} \Phi_{ij} x_j$ have been neglected. However, we get the same results if we consider individual species not in isolation, but rather together with their accompanying mutant spectrum. In this case, the target of selection is not only the species with the greatest selection value (often called the "master sequence"), but rather the master sequence including its whole "tail" of mutants. This distribution is termed "quasi-species".

Mathematically, the selection equations of quasi-species are obtained by diagonalisation of the linear system of Eq. (1) after integrating factor transformation (for details see Chap. 3 of this book). If we denote the quasi-species by y_i and the corresponding eigenvalues by λ_i, then the selection kinetics are described by the following set of equations:

$$\frac{\mathrm{d}y_i}{\mathrm{d}t} = \left(\lambda_i - \bar{\lambda}(t)\right) y_i \quad (i = 1, \ldots, k). \tag{7}$$

This is structurally equivalent to the set of Eq. (4). However, in contrast to Eq. (4), which describes the selection kinetics of a single species x_i, Eq. (7) now describes the selection kinetics of the quasi-species y_i, i.e. the master sequence and its mutant distribution.

Ultimately, we come to the conclusion that the principle of natural selection is by no means an irreducible property of living matter. Rather, it is a consequence of physical and chemical laws, and it becomes manifest if matter has self-reproductive properties and is subject to limitation of growth. This result has also been verified by the extensive experimental investigations of the evolution of biological macromolecules in vitro (Kramer et al. 1974; Küppers 1979; Mills et al. 1967; Spiegelman 1971). These have demonstrated that a substantial part of the reductionist research program is sound.

3 Evolutionary Optimisation of Information

Let us consider the selection process described by Eq. (1) in more detail. For this purpose, it will be useful to introduce the concept of sequence space. This is a mathematical space whose coordinates cover all sequence alternatives of a given sequence of signs. In the abstract case, the signs are binary digits. In the present case, the signs are the monomers from which a biopolymer is composed. If we consider a biological macromolecule of length ν, which is built up from μ classes of monomers, the total number of sequence alternatives n—and thus also the number of dimensions of the sequence space—is given by

$$n = \mu^{\nu}. \tag{8}$$

If population (or concentration) numbers are assigned to the "coordinates" of sequence space, one obtains the population (or concentration) profile. Alternatively, one can construct a "value profile" by assigning to each coordinate in sequence space the corresponding selection value W_i. The resulting "fitness landscape" is depicted in Fig. 2 in a greatly simplified manner. A precise mathematical description of the construction and the topological properties of sequence space can be found elsewhere (see for example Eigen 2013). From an information-theoretical point of view, sequence space may also be regarded as an information space, in which the selection of genetic information takes place.

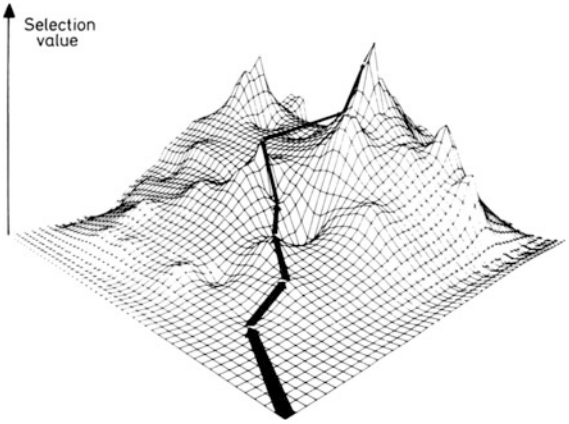

Fig. 2 Schematic representation of the adaptive surface in the sequence space. If all possible sequences n of a biological information carrier are represented as "coordinates" in sequence space and the selection value of the corresponding species is plotted over the appropriate coordinate, then an n-dimensional "mountain-range" profile is obtained, as shown here in a simplified, three-dimensional representation. The evolutionary origin of information then corresponds to a process of optimisation that leads from a low (local) peak to a higher (local) peak. From Küppers (1990)

Using this concept, we can describe the selection of a quasi-species as a condensation in information space (Eigen 2013). However, the resulting selection equilibrium is metastable. Whenever a species x_{i+1} appears that is selectively more favoured than the (hitherto) dominant species x_i, the original steady state collapses and a new selection equilibrium, characterised by the higher selection value of the now dominant species x_{i+1}, is attained. Thus, in the course of time, the system passes through a sequence of metastable selection equilibria, which can be described by a sequence of inequalities

$$W_{\mathrm{max}_1} < W_{\mathrm{max}_2} \cdots < W_{\mathrm{opt}}. \tag{9}$$

Here, $W_{\mathrm{max}_\mathrm{i}}$ is the selection value of the species x_i that dominates the selection equilibrium.

The physical significance of relation (9) can easily made clear by reference to a fitness landscape built upon sequence space (Fig. 2). Here the parameter $W_{\mathrm{max}_\mathrm{i}}$ represents a local maximum of adaptation, that is, a peak in the value profile. Equation (9) restricts the evolutionary optimisation of the system, insofar as it defines a gradient in sequence space to which the route of optimisation is tied. In its evolutionary development, the system can only proceed along a route that takes it, starting from a local maximum, to a higher local maximum.[2]

[2] Strictly, this applies for the case of deterministic selection equations, in which fluctuations in the population are not taken into account.

From the foregoing discussion, we can draw the conclusion that the selection value of genetic information is clearly defined by the ability of a biological information-carrier to reproduce itself as fast as possible while maintaining high accuracy and stability. These are the conditions under which the selection value of a molecular species is maximised.

In the simplest case, W_i depends only upon the physical parameters P_k of the environment, such as temperature and energy flow. In a more general case, ecological coupling—such as the dependence upon the population size $x_{j \neq i}$ of other species—may appear. Thus, in general, W_i is a function that depends not only upon the parameters A_i, Q_i and D_i but also upon $x_{j \neq i}$ and P_k, that is:

$$W_i = W_i(x_{j \neq i}, P_k). \tag{10}$$

As expected, selection values reflect the complexity of living systems, including the complexity of their environment. It is therefore not surprising that the quantities W_i can only be specified physically for comparatively simple macromolecular systems. However, living systems are so complex—even at the lowest organisational levels—that their selection values can at best be given as phenomenological quantities.

The fact that selection values for living beings cannot be calculated explicitly has often given rise to the conjecture that Darwin's principle of the "survival of the fittest" is a mere tautology, because—it is said—the term "fittest" is defined alone by the fact of having survived ("survival of the survivor"). This is indeed the case for the non-Darwinian models of "neutral selection" (Kimura 1983). In these systems, in which all species are assumed to have the same selection values, a fluctuation in the population number of a certain species can amplify itself and finally reach the size of the whole population.

However, in the Darwinian model of natural selection, the situation is quite different. This becomes clear on examination of evolution in sequence space. The selection principle would be a tautology if the population profile, which represents the fact of "survival", and the value profile, which represents the "fitness" landscape, were to turn out to be identical. However, the physical analysis of Darwinian selection systems has shown that (as a rule) population profiles and value profiles possess different structures, which disproves the supposition of a tautology in the selection principle. This is seen, for example, in the case where one species in a population has the greatest selection value, but is present in a concentration lower than that of the mutant distribution arising from it. This is always the case when a dominant species reproduces itself with such a high error rate that just one copy is reproducibly preserved. At any moment, the stationary-state proportion of the selectively poorer mutants in the total population is greater than the proportion of the master copy, even though the mutants, seen as individuals, naturally die out again. The tautology asserted to lie behind the selection principle is thus falsified; its seeming existence is due to the extreme complexity of living systems and the resulting limits of their predictability.

The concept of sequence space allows further conclusions with regard to the origin and evolution of genetic information. In cases where selection values depend only on environmental conditions, the structure of value space remains the same as long as the environment does not change. However, the assumption of a constant environment is an idealised condition: it cannot even be realised at the level of molecular evolution, since all individuals of a population of molecules contribute, with their physicochemical properties, to the environmental conditions of the population. Every change in the composition of a population must therefore lead to a change in the environment. In addition, the selection value of every single species will as a rule depend on the population variables of the other species taking part in selection, so that every evolutionary step changes the structure of the value profile (Eq. 10). This in turn means that goal and goal-directedness are interdependent. Since the elementary events (mutations) that lead to evolutionary changes are completely indeterminate, every evolutionary process is historically unique. This makes it clear that the molecular theory of evolution predicts only the *generation of genetic information as such*, but it does not predict the detailed outcome of evolution, as manifested in the *content of the genetic information*.

Equation (3), which defines the selection value, contains no details about the functional properties, which contribute to the parameters A_i, Q_i and D_i. Although it describes the "value" of an information carrier in a selection competition, its semantics are completely restricted to the aspect of selectivity. Yet the selection value, specified by Eq. (3), tells us nothing about the forces that determine the highly developed and differentiated semantics expressed in the functional complexity of living matter. For this reason, the semantics of genetic information are usually explained within the Darwinian theory of evolution a posteriori by appeal to plausibility. However, for a deeper understanding of the origin and evolution of genetic information, we need an approach to the semantic aspect of information that goes beyond the mere aspect of selectivity.

4 The Context-Dependence of Semantic Information

Semantic information is defined as "valued" information. The use of this expression already indicates that access to the semantic aspect of information must be sought by the receiver, which evaluates the information. In the widest sense, the receiver represents the "context" of the information. The context-dependence of information is a universal aspect of any kind of information. This is due to the fact that information in an absolute sense does not exist (Küppers 1995). Information obtains its meaning only in relation to its context. This is no less true of genetic information, which becomes operational—i.e. unfolds its meaningful content—only under certain physical and chemical conditions.

Because information in an absolute sense does not exist, each recipient needs some background information as a reference frame within which to evaluate the content of the information received. Even the task of identifying a piece of information "as" information requires some prior information, or prior knowledge, on the part of the recipient. This immediately raises a further question: How much additional information is necessary in order to understand a given piece of (meaningful) information (Küppers 2013)?

At first glance, this question seems unanswerable, as it involves the problematic concept of "understanding". It is all the more surprising that an exact solution to this problem is nonetheless possible. However, to reach this solution one has to restrict the consideration to the minimum condition required for any kind of understanding (Küppers 2010). This minimum condition is the fact that the receiver must first register the information before the actual process of understanding can begin. This requirement is self-evident. It applies for every process of communication, independently of whether the communication takes place in a natural or an artificial system.

Let us analyse the consequences with regard to semantic information, written down in the letters of a human language. Even a superficial view of language reveals that any meaningful sequence of letters has an aperiodic structure. The reason for this is clear: only the use of aperiodic sequences opens up enough space for human language to code information and thus makes the unlimited richness of human language possible at all. If language were to use more or less periodic sequences of letters for coding information, the potential information space would be more or less empty.

This thought can be deepened by using the concept of algorithmic information theory, which has been developed within the framework of computer science. The core of algorithmic information theory is a measure of information that is linked to the "complexity" of a sequence of digits or symbols. A sequence of binary digits is to be considered as complex when the sequence cannot be compressed significantly, i.e. when there is no algorithm, shorter than the sequence itself, from which the sequence can be derived (see for example Chaitin 1987). According to this idea, the complexity K of a binary sequence s is given by the length L of the shortest program P of a computer C from which s can be generated:

$$K_{\mathrm{C}}(s) = \min_{C(P)=s} L(P). \tag{11}$$

In this definition, the complexity depends upon the *degree of incompressibility* of a sequence of binary digits. Two aspects of this definition must be emphasised: (1) The notion of "complexity", as introduced here, is completely equivalent to the notions of "aperiodicity" and "randomness". (2) The transition between non-complex sequences and complex sequences is obviously a continuous one.

Within algorithmic information theory, the measure of the information content of a message is its complexity, which in this case means the aperiodicity of its syntax. It is, as it were, the irreducible "bulk" of information that is contained in the message. If the complexity of a sequence of digits or symbols is at a maximum, then there is no algorithm that would be shorter than this sequence and by means of which the sequence or a missing part of it could be reconstructed. In this sense, the sequence is aperiodic or irregular. If, in contrast, the sequence is periodic (or largely periodic), then its inherent regularity would allow it to be compressed—or, if a part of the sequence were already known, this would allow the other part to be generated.

From this point of view, meaningful information in human language is always associated with aperiodic sequences. However, this statement should not be inverted! Not every aperiodic arrangement of letters in human language represents a meaningful sequence, nor does the aperiodicity of the syntax imply a random origin of the associated information. In short, the degree of aperiodicity is only a measure of the complexity of semantic information.

The assertion that semantic information is always encoded in aperiodic sequences will have important consequences for the recipient of this information (Küppers 2013). Since in this case there is no algorithm that allows the reconstruction of the whole sequence on the basis of a fragment of this sequence, the recipient must be in possession of the entire sequence, before the actual process of understanding its content can commence. In short, this means that the mere act of registration of an item of semantic information by a receiver demands that a certain quantity of information be already present with the recipient, and that this information has at least the same degree of complexity as the information that is to be understood.

This conclusion is generally valid. It remains unaffected by the fact that every language possesses syntactic rules according to which the words of the language are assembled into correctly formed sentences (Küppers 2013). Such rules only restrict the set of aperiodic sequences that can carry meaning at all. But they do not allow any inference to be made about the content itself. One can express this result in another way: semantic information cannot be compressed without loss of a part of its meaning. Of course, a piece of information may sometimes be reduced to its bare essentials, as done in telegram style or in boulevard newspapers, but some information is always lost in this process. In general, however, the loss of information is compensated for by a certain pre-knowledge of meaningful communication, which the receiver of this information possesses thanks to his cultural background, experience, prior agreements, etc.

The above conclusions rest totally on the assumption that semantic information is associated with random sequences. However, this is indeed a mere assumption, which we formulated on the basis of a plausibility consideration. We cannot prove it in any strict sense. Thus, it may be possible that there exist hidden algorithms that are able to generate a piece of semantic information, but which we have not discovered or identified so far. As soon as such a compact algorithm was found, however, our whole chain of arguments would break down.

Nevertheless, we can ascribe a high probability to our hypothesis by virtue of the fact that almost all binary sequences are random. It can easily be demonstrated in the following way. Let us consider all binary sequences of length v. Since the transition from random to non-random sequences is a continuous one, we must define a limit for randomness. Thus, we define all sequences with a complexity of—let us say—$K \geq v - 10$ as random. To this class of sequences belong all sequences which cannot be compressed by more than 10 bits.

We now ask how many sequences have a complexity below the threshold $K = v - 10$ and which could in principle be able to generate a sequence of complexity

$K \geq v - 10$. It is obvious that there are 2^1 sequences of complexity $K = 1$ with this property, 2^2 sequences of complexity $K = 2$, … and 2^{v-11} sequences of complexity $K = v - 11$. The number of all algorithms of complexity $K < v - 10$ thus adds up to

$$\sum_{i=1}^{v-11} 2^i = 2^{v-10} - 2. \tag{12}$$

As no algorithm with $K < v - 10$ can generate more than one binary sequence, there are fewer than 2^{v-10} ordered binary sequences. These make up one 2^{10}th of all v-membered binary sequences. This means that among all binary sequences of length v only about every thousandth sequence is non-random with a complexity $K < v - 10$. Thus, the overwhelming large fraction of all binary sequences indeed comprises random sequences.

To summarise the result: In order to understand a piece of information, one invariably needs a quantity of background information that has at least the same degree of complexity as the information that is to be understood. This finding gives the context-dependence of semantic information a highly precise form. It is generally valid, independently of the way in which the information is stored.

The foregoing conclusions can be applied immediately to the semantics of genetic information. This is not least because the sequence of nucleotides in the genome represents semantic information in the same way as the letters of a written text do.

As in the case of human language, a crucial feature of the genetic program is the aperiodicity of the sequence of nucleotides in the genome.[3] This in turn means that there are no hidden algorithms, i.e. no life-specific laws, that are able to order the monomers in a biological macromolecule in such a way that a piece of semantic information will originate (Küppers 1990). Since a random synthesis of meaningful information is excluded as well, only the Darwinian concept of evolution remains to explain the origin of semantic information in prebiotic matter (Küppers 1990).

In this regard, however, the Darwinian concept seems to leave an explanatory gap. This becomes clear if one takes into consideration the inherent context-dependence of information. Thus, from an information-theoretical point of view, evolution by adaptation must be regarded as a kind of communication between the sender and the receiver of information. The "sender" is the biological macromolecule with its information content, while the "receiver" of this information is the environment. The environment in turn represents an external source of information, which "evaluates" the content of genetic information according to the capability of the information carrier to survive under conditions of selection competition. However, within the Darwinian understanding of evolution, the

[3] That the aperiodicity must be the source for the complexity of living matter was already recognised at the dawn of molecular biology and led to the conjecture that chromosomes must have the structure of an "aperiodic crystal" (Schrödinger 1944).

environment, which directs the adaptation, is usually thought of as a *biotic* environment. Yet this environment is itself already enriched with semantic information. So where does this bulk of information come from? At the beginning of evolution no functional complexity, no meaningful information was present that could have provided a reference frame for progressive adaptation. The only environment that prevailed on the primordial earth was a world of physical laws and contingent boundaries. Thus, we must seek other principles of evolutionary dynamics—ones that act independently of the processes of adaptation and nevertheless push prebiotic matter towards the nucleation of functional complexity.

5 Overcoming Information Barriers

In the very early phase of the evolution of biological macromolecules, when a translation apparatus was not yet present, the target of selection could at best have been the phenotypic properties of the macromolecules themselves. This kind of selection, however, will most probably lead to a decrease of the chain length of potential information carriers, rather than to an increase. For, if there is no additional information encoded in the molecules, an increase in the chain length has no selective advantage under the constraint of fast reproduction. On the contrary, larger chain lengths would impede rapid reproduction and therefore lead to a selective disadvantage.

This conjecture was confirmed by the serial transfer experiment with the genome of the phage Q_β, where selection was aiming exclusively at the phenotypic properties of the genome (Mills et al. 1967). Thus, the chain length of the RNA was only preserved insofar, as it has to fulfil certain structural prerequisites for the reproduction by the enzyme. These are, in particular, the recognition signals of the template for the Q_β-replicase (Küppers and Sumper 1975). Consequently, the major part of the genetic information stored in the genome became eliminated in favour of fast reproduction.

Moreover, the serial transfer experiment underlines our previous conclusion that pure selection—i.e. selection decoupled from environmental adaptation—is not sufficient for the nucleation of semantic information. Instead, pre-existing semantic information may even become eliminated for the benefit of fast reproduction. Thus, in the early phase of molecular evolution, the nucleation of semantic information could only come about if there existed a principle that acted against the evolutionary tendency to reduce the complexity of macromolecular sequences.

In fact, alongside the mechanism of adaptation, there is a driving force of evolution that does not depend on the environment, but which nevertheless may lead to an increase in the chain lengths of biological macromolecules. This driving force turns out to be a special property of sequence space, in which information originates. It is related to the fact that, independently of the environmental conditions, the process of evolutionary optimisation becomes more effective in a high-dimensional space than in a low-dimensional one. The reason for this is that a high-dimensional space opens up more possible pathways for optimisation than does a low-dimensional space, as is evident from Fig. 3. Consequently, in a high-dimensional space, there will be a greater probability of the evolutionary optimisation to avoid a dead end, where the system is captured in a local equilibrium of selection (Eigen 2013). This in turn means that any random extension of the chain length ν of an information carrier will have a positive effect on the optimisation process itself. At the same time, an increase in the chain length of the sequences will lead to an increase in their capacity to encode meaningful information.

This finding is only apparently in contradiction to the result of the serial transfer experiment, in which under the constraints of fast reproduction, short RNA sequences have a selective advantage over long ones. In evolution it is quite common for antagonistic principles to act together in the evolutionary optimisation process without cancelling each other out. In fact, this kind of interaction seems to be indispensable for the nucleation of genetic information, as suggested by the concept of quasi-species (Eigen and Schuster 1979). During the very early stage of molecular self-organisation, when no proteins were present to catalyse the reproduction of potential information carriers, the quality of self-reproduction must be assumed to have been very low. This in turn places a fundamental limitation upon the amount of information that can be transferred reproducibly from one generation to the next.

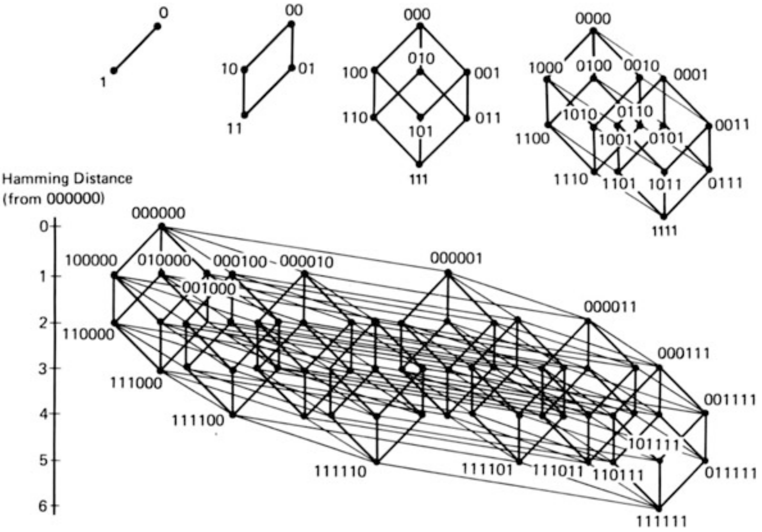

Fig. 3 Sequence space for binary sequences of chain length = 1, 2, 3, 4 and 6. The sequences are arranged according to the Hamming distance $d(ij)$, defined as the number of different positions between sequence i and sequence j. From Volkenstein (1994)

Detailed analysis of the error threshold has shown (Eigen and Schuster 1979) that, in a self-reproducing unit, the maximum number $v_{\max}$ of molecular symbols that can be transferred reproducibly across generations is given by equation

$$v_{\max} = \frac{\ln \sigma_m}{1 - \bar{q}_m}, \tag{13}$$

where $1 - \bar{q}_m$ is the average error rate per symbol and σ_m, defined by

$$\sigma_m = \frac{A_m}{\bar{E}_{k \neq m} + D_m}, \tag{14}$$

is the superiority factor of the species x_m, i.e. the advantage in growth of the master sequence x_m over its mutants $x_{k \neq m}$.

In the very early phase of prebiotic evolution, enzyme-free replication of RNA most probably involved per-digit error rates of 5 × 10^{-2}, which allows—depending on σ_m—the reproducible transfer of nucleic acid sequences between 14 and 60 nucleotides long (see Table 1). This amount of information is just enough to code for proteins with a rudimentary catalytic function, but is far from being sufficient for the formation of sophisticated functional order in prebiotic matter.

Thus, an important step in the solution of the problem of the origin of genetic information was the finding that the information barrier, which is a consequence of the error threshold, can be surmounted by the hypercyclic organisation of biological

Table 1 The amount of information v_{max} that can be transferred reproducibly from one generation to the next, depending on the quality of the reproduction rate. From Eigen and Schuster (1979)

Digit error rate $1 - \bar{q}_m$	Superiority σ_m	Maximum digit content $v_{\max}$	Molecular mechanism and example in biology
5 × 10^{-2}	2	14	Enzyme-free RNA replication[a] t-RNA precursor, v = 80
	20	60	
	200	106	
5 × 10^{-4}	2	1386	Single-stranded RNA replication via specific replicases phage Q_β, v = 4500
	20	5991	
	200	10597	
1 × 10^{-6}	2	0.7 × 10^6	DNA replication via polymerases including proofreading by exonuclease *E. coli*, v = 4 × 10^6
	20	3.0 × 10^6	
	200	5.3 × 10^6	
1 × 10^{-9}	2	0.7 × 10^9	DNA replication and recombination in eukaryotic cells vertebrates (man), v = 3 × 10^9
	20	3.0 × 10^9	
	200	5.3 × 10^9	

[a] Uncatalyzed replication of RNA never has been observed to any satisfactory extent; however, catalysis at surfaces or via not specifically adapted proteinoids (as proposed by S.W. Fox) may involve error rates corresponding to the values quoted

macromolecules (Eigen 1971; Eigen and Schuster 1979). The cyclic coupling of self-reproducing information carriers into a hypercycle forces the competing information units to cooperate, which in turn leads to mutual stabilization of their information content and thereby to an increase in the total amount of information. The hypercycle, which combines competition with cooperation, is an important example of the balanced action of antagonistic principles in the early evolution of genetic information.

6 Deciphering the Semantic Code of Evolution

The hypercyclic organisation of nucleic acids and proteins must be considered as the archetype of genetic organisation, which is based on the principle of cooperation. With progressing evolution, other principles of evolutionary dynamics come into play, which refine the structure of genetic organisation. Besides cooperation, these principles include self-regulation, efficiency, recombination, flexibility, stability and others. They make possible the coexistence of information carriers, the overcoming of information barriers, resistance against perturbations and the integration of advantageous information. Although these principles are partly in conflict with each other, they act together in a well-balanced relationship, which allows the formation and evolutionary optimisation of functional order in prebiotic matter. They determine the proto-semantics of genetic information by fixing the functional frame for the development of genetic information. The detailed content of this information then emerges from the processes of adaptation to the environment.

In a certain sense, which is to be explained in more detail, the above principles can be conceived as elements ε_k of a semantic code of evolution C_{sem}, defined by

$$C_{\text{sem}} = \{\varepsilon_k\} \quad (k = 1, \ldots, n). \tag{15}$$

The general idea of a semantic code has been developed within the framework of structural sciences (Küppers 2013). It serves the purpose of getting a strict access to the semantic aspect of information. However, in contradistinction to the usual understanding of the notion of "code", the semantic code does not provide any rules for the assignment of symbols or sequences of symbols to another source of symbols. Instead, the semantic code represents the value scale that a recipient applies to a piece of information that he is going to decode in respect of its meaning.

Strictly speaking, the semantic code represents an evaluation scale that, by superimposition and specific weighting of its elements, restricts the value that the information has for the recipient and in this way becomes a measure of the meaning of the information. If the elements ε_k have the weights P_{jk} for the evaluation of a piece of information I_j by the recipient, then an adequate measure for the semantic value $\varepsilon(I_j)$ would be a linear combination of the weighted elements ε_k:

$$\varepsilon\left(I_j\right) = \sum_k p_{jk}\varepsilon_k \quad \text{with} \quad \sum_k p_{jk} = 1. \tag{16}$$

This measure, in turn, has the same mathematical structure as the entropy H of a message source $\{I_j\}$.

However, in place of the weighted messages, Eq. (16) now contains the weighted values ε_k of a chosen message I_j. In the limiting case, where the only value a recipient attaches to a message is its novelty, given by the expectation value of this message, Eq. (16) reduces to the information measure of classical information theory. At the same time, the number k is a measure of the fine structure of the evaluation scale: the greater k is, the sharper—i.e. the more highly differentiated—is the evaluation by the recipient.

The information value $\varepsilon\left(I_j\right)$ is a relative and subjective measure insofar as it depends upon the evaluation criteria applied by the recipient. However, for all recipients who use the same elements of the semantic code, and who for a given message I_j assign the same weights to these elements, $\varepsilon\left(I_j\right)$ is an objective quantity. Equation (16) describes the different value aspects, contributing to the overall value, which has some information for a receiver. In this abstract form, Eq. (16) applies to all systems in which semantic information is evaluated.

However, the evaluation of semantic information by a receiver requires some "pre-information". In human communication, the value scale is a highly specific one; it depends upon the recipient's prior knowledge, prejudices, desires and expectations. Therefore, the successful exchange of meaningful information requires standardisation of the framework of mutual understanding. This standardisation is achieved by precise co-ordination between the individuals of a community. However, in natural systems that exchange information, there is no explicit agreement about the value scale. Instead, the value scale is ultimately given—as in the case of Darwinian evolution—by the internal principles of evolutionary optimisation and the prevailing environmental conditions.

The semantic code, as introduced by Eq. (16), contains the elements that contribute (according to their weights) to the formation of functional organisation in prebiotic matter. The weights depend upon the type of organisation and the degree of evolutionary optimisation. Since the elements of the semantic code constitute the nucleation and improvement of genetic organisation itself, they epitomise in the true sense the creativity of natural evolution.

Finally, there arises the question of whether one can ascribe numbers to the weights of the different elements of the semantic code in the evolution of functional order. In view of the tremendous complexity of living matter and its physical boundaries, this seems to be an impossible task. Nevertheless, the elementary processes of natural evolution can be studied in the test tube under the idealised and reproducible conditions of controlled experiments. To this end, evolution reactors have been built, and these may also prove suitable for unravelling experimentally the semantic code of evolution and may thus lead to a deeper understanding of the general principles of the evolution of life.

References

Chaitin GJ (1987) Algorithmic information theory. Cambridge University Press, Cambridge

Dobzhansky T (1973) Nothing in biology makes sense except in the light of evolution. Am Biol Teach 25:125–129

Eigen M (1971) Selforganisation of matter and the evolution of biological macromolecules. Naturwissenschaften 58:465–523

Eigen M (2013) From strange simplicity to complex familiarity. Oxford University Press, Oxford

Eigen M, Schuster P (1979) The hypercycle. Springer, Berlin

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Kramer FR, Mills DR, Cole PE, Nishihara T, Spiegelman S (1974) Evolution in vitro: sequence and phenotype of a mutant RNA resistant to ethidium bromide. J Mol Biol 89:719 (1974)

Küppers B-O (1979) Towards an experimental analysis of molecular self-organization and precellular Darwinian evolution. Naturwissenschaften 66:228–243

Küppers B-O (1983) Molecular theory of evolution. Springer, Berlin (reprinted 1985)

Küppers B-O (1987) On the prior probability of the existence of life. In: Gigerenzer G, Krüger L, Morgan MS (eds) The probabilistic revolution 1800-1930, vol 2. Probability in Modern Science, Cambridge/Mass, pp 355–369

Küppers B-O (1990) Information and the origin of life. MIT Press, Cambridge

Küppers B-O (1992) Understanding complexity. In: Beckermann A, Flohr H, Kim J (eds) Emergence or reduction. De Gruyter, Berlin, pp 241–256

Küppers B-O (1995) The context-dependence of biological information. In: Kornwachs K, Jacoby K (eds) Information. new questions to a multidisciplinary concept. De Gruyter, Berlin, pp 135–145

Küppers B-O (2000) The world of biological complexity: origin and evolution of life. In: Dick St. J (ed) Many worlds. Templeton Foundation Press, Pennsylvania, pp 31–43

Küppers B-O (2010) Information and communication in living matter. In: Davies P, Gregersen NH (eds) Information and the nature of reality. Cambridge University Press, Cambridge, pp 170–184

Küppers B-O (2013) Elements of a semantic code. In: Küppers B-O, Artmann S, Hahn U (eds) Evolution of semantic systems. Springer, Berlin, pp 67–85

Küppers B-O, Sumper M (1975) Minimal requirements for template recognition by bacteriophage Qß-replicase: approach to general RNA-dependent RNA synthesis. Proc Natl Acad Sci USA 72:2640–2643

Mills DR, Peterson RL, Spiegelman S (1967) An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. Proc Natl Acad Sci USA 58:217–224

Schrödinger E (1944) What is life? Cambridge University Press, Cambridge

Spiegelman S (1971) An approach to the experimental analysis of precellular evolution. Q Rev Biophys 4:213–253

Volkenstein MV (1994) Physical approaches to biological evolution. Springer, Berlin