Physico-chemical Constraints Connected with the Coding Properties of the Genetic System

JEAN LEHMANN*

Institut de Physique de la Matière Condensée, Université de Lausanne, 1015 Lausanne, Switzerland

(Received on 14 October 1998, Accepted in revised form on 5 November 1999)

New insights on the origin of the genetic code, based on the analysis of the physico-chemical properties of its molecular constituents (RNA and amino acids), are reported in this paper. We point out a symmetry in the genetic code table and show that it can be explained by the nature of the anticodon–codon interaction. The importance of the strength of this interaction is examined and a correlation is found between the free-energy change ($\Delta G_0$) of anticodon–codon association and the volume of the corresponding amino acids. This correlation is investigated in conjunction with the well-known one linking the hydrophobicity of the anticodons with that of the amino acids. We show that they can be considered separately and that the energy vs. volume correlation may be explained by the process implicating the peptide bond formation between two successive amino acids during translation. This interpretation is supported by a statistical pattern of bases (purines or pyrimidines), observed in present coding genes, and by considerations involving the availability of the different kinds of amino acids. Finally, we try to explain the hydrophobicity correlation when reconstructing the events at the time of the so-called “RNA World”. The whole of our investigation shows that the genetic code might be sufficiently robust to exist without the participation of pre-existing proteins, and that this robustness is a consequence of the physico-chemical properties of the four bases of the genetic system.

1. Introduction

Since the complete establishment of the genetic code map by 1965 (see Hayes, 1998), many attempts have been made to understand how such a complex system has arisen (for a review, see Di Giulio, 1997). Essentially, two different (and complementary) approaches of the problem have been undertaken. One is centred on the importance of the historical context, which is studied notably by the comparison between tRNA sequences (Eigen & Winkler-Oswatisch, 1981; Saks & Sampson, 1994; Di Giulio, 1999) and/or tRNA synthetase sequences (Nagel & Doolittle, 1995; Nicholas & McClain, 1995), which permit the construction of phylogenetic trees. The principal aim of these works is to connect the different groups of codons, and therefore to establish an expected coherence of the organization of the code based on evolutionary considerations. The second approach, centred on the physico-chemical properties of the molecules implied in the process of coding, is closely related to our investigations. Up to now, hydrophobicity has been the parameter that has aroused much interest (see Section 2).
Here, we show the results of new analysis on organization of the genetic code. The main observation reported here is the discovery of a correlation of the genetic code which connects the strength of the anticodon–codon association with the volume of the corresponding amino acids (Section 4). With some historical insights given by results of pre-biotic experiments (Section 5), the interpretation of this correlation (Section 6) allows us to put forward explanations of other observations on the code, and especially those concerning hydrophobicity (Section 7).

These considerations lead us to propose what could have been the general process of coding at the beginning of life’s organization, and permit us to point out that the four bases of the genetic system display complementarity.

2. Hydrophobicity and the Genetic Code

Observations on hydrophobicity of the molecular compounds of the genetic coding system have long suggested the non-random organization of the genetic code (Woese, 1965; Weber & Lacey, 1978; Jungck, 1978; Lacey & Mullins, 1983; Blalock & Smith, 1984; Taylor & Coates, 1989; Lacey et al., 1992, 1993). This parameter was measured for the four nucleotides and the 20 coded amino acids and a strong correlation was found between average ranking measures of the anticodon doublet (i.e. the first two bases, from 3’ to 5’) and the corresponding amino acids, with a few exceptions (Weber & Lacey, 1978; Lacey & Mullins, 1983) (see Fig. 1). Moreover, the four nucleotides can be ordered as [U, C, G, A], from the most hydrophilic (U) to the most hydrophobic (A) (Weber & Lacey, 1978; Lacey & Mullins, 1983). It was thus proposed to set the genetic code table in this order (Jungck, 1978). The second position of the anticodon shows the best correlation property (Blalock & Smith, 1984; Taylor & Coates, 1989): the most hydrophilic amino acids are coded by A (anticodon U) and the most hydrophobic by U (anticodon A), which are at opposite ends of the list.

3. The Universal Genetic Code and its Symmetry

In most biological textbooks, the genetic code is represented in a table with three entries, each corresponding to one position of the codon. Some order is immediately apparent in the pattern of degeneracy families in the table depending on the succession in which the four bases are written. The usual one is [U, C, A, G] (see e.g. Levin, 1994), which groups the codons of the two-fold degenerate families.

Given these considerations, one might ask if there is an optimal way to write the table. Examining the results of the permutations in the succession of the bases, the symmetry which is apparent with the order [U, C, G, A] renders this solution attractive [Fig. 2(a)]. Indeed, it helps to establish the parameters of the degeneracy, as demonstrated in Fig. 2(b). These are the number of hydrogen bonds created by the two first bases of the anticodon when it interacts with the codon, and the chemical family of the base at the second position of the anticodon: purine (A or G) or pyrimidine (C or U), abbreviated respectively to R or Y. Surprisingly, the order is the same as that established using hydrophobicity considerations (Section 2). As far as we know, the importance accorded to this degeneracy symmetry is new, although the prediction of the degeneracy of the codon family based on these parameters has
Fig. 2. Symmetrical representations of the genetic code. (a) The universal genetic code table written with the succession [U, C, G, A], which exhibits the degeneracy symmetry shown by the dashed line: four-fold degenerate codon families are indicated by grey rectangles and two-fold ones by squares. The mitochondrial code is more symmetrical, because UGR and AUR are also two-fold codon families (Jukes & Osawa, 1991). (b) Schematic representation of the degeneracy (mitochondrial code). Each small arrow indicates whether the considered codon belongs to a two-fold (below the inclined plane) or a four-fold (above the inclined plane) degeneracy family. The energy types of 1st and 2nd codonic positions are indicated by W (weak: A or U) and S (strong: G or C). The vertical axis indicates the R/Y type of the 2nd codonic position.

already been investigated (Ninio, 1973; Lagerkvist, 1978; Jayaram, 1997).

The origin of the degeneracy becomes evident in the mammal mitochondrial genetic system, where one tRNA (whose 3rd base is a U) can bind with any of the members of a four-codon family (Levin, 1994), the other codons obeying the simpler wobbling rules. The mitochondrial code is also interesting because it is more symmetrical than the universal code: UGR is the code for tryptophane and AUR for methionine (Jukes & Osawa, 1991). As a consequence, the symmetry is also valid for the two-fold degeneracy. These observations suggest that the mitochondrial genetic system is more subject to basic physical constraints than the nuclear one.

Bearing in mind the importance of energy [as shown in Fig. 2(b)], it is appropriate to wonder whether some properties of the amino acids might be correlated to the strength of the anticodon–codon interaction. Using free-energy change predictions of RNA duplex association (Freier et al., 1986; Turner et al., 1987; Turner & Bevilacqua, 1993) to make more accurate estimations for the different pairs, a very good correlation was found with the van der Waals volume of the corresponding amino acids (Fig. 3): small amino acids belong to the strongest codons and larger ones are coded by weaker codons. This suggests that a physical process is responsible for at least part of the organization of the genetic code. There are, however, some values (corresponding to asparagine, arginine and tryptophane) which do not correlate well.

Another interesting fact is that it is possible to fix a threshold $\Delta G_0$ above which all codon families are constructed with A or U at the mid-base position (see Fig. 3): apart from the miscorrelated points, this corresponds to the largest amino acids.

### 4. The Two Correlations in the Genetic Code

At this stage of the proceedings, it is important to emphasize that the amino acids coded by the genetic system follow two correlations. The first concerns the hydrophobicity (Fig. 1) and the second an energy dependence with respect to the
FIG. 3. Correlation between van der Waals volume of the amino acids (values from Darby & Creighton, 1993) and anticodon–codon free-energy-change interactions ($\Delta G_0$) in the universal genetic code, but with UGR as tryptophan and AUR as methionine for simplicity (case of the mitochondrial code, see comment in text). Each small circle indicates one interaction value (anticodon–codon with only Watson–Crick base pairing), while large dots correspond to average values over 4 circles (for a four-fold degenerate family) or 2 circles (for a two-fold degenerate family). Estimations of $\Delta G_0$ for each particular anticodon–codon interaction are made by applying the nearest-neighbor method described in Turner et al. (1987), which is normally valid for RNA duplex association. Additional positive free energy for association was not taken into account, because of the probable underestimation of the negative free-energy changes of anticodon–codon interaction (see Yoon et al., 1975; Turner et al., 1987). The dashed line shows the $\Delta G_0$ threshold above (below) which all codon families are constructed with A or U (G or C) in the second position.

volume (Fig. 3), which is a new observation on the genetic code. For simplicity, we will refer below to the hydrophobicity and the volume correlation, respectively. One might ask if there is a relationship between the two. Figure 4 shows the dependence of the intrinsic parameters of amino acids and anticodons: it does not indicate any simple rule for predicting hydrophobicity from van der Waals volume (amino acids) or $\Delta G_0$ (anticodons). This observation suggests that the assignment is the result of two distinct processes.

As mentioned in the introduction, much of our interest is concentrated on the physico-chemical aspects implied by translation, which we would like to depict as a particular case of a polymerization process. The question is now the following: are there physico-chemical reasons why a code emerges from such a process? We believe this to be the case. This conviction is based on the existence of the two correlations mentioned, which are plausible reflections of physico-chemical constraints, notably those connected with the polymerization of the amino acids based on an RNA template. Our main hypothesis is therefore that only amino acids satisfying the requirements expressed by the two correlations should enter into the genetic code table, in which their hydrophobicity and a volume parameter determine their respective positions.

One could ask if these correlations are strong enough to be taken into consideration. The possible interpretation of why some amino acids do not really fit the main sequence of the correlations (notably asparagine, arginine and tryptophane in the case of the volume correlation) will be discussed below (Section 6). Concerning the other values of the graphs (Figs 1 and 3), it must be emphasized that hydrophobicity is expressed by average ranking values (Lacey et al., 1983), $\Delta G_0$ values are estimations (see Fig. 3 for method), and van der Waals volume is a discrete variable (it depends in particular on the number of atoms in each amino acid), so very high correlations cannot really be expected. A good indication of a close link existing between the last two parameters ($\Delta G_0$ and volume) is the grouping of the codon families belonging to the same amino acid (three cases: serine, leucine and arginine), also observed when they are not situated in the main sequence (arginine).

5. Prebiotic Experiments and Coding Sequences

Before analysing the correlations, one might wonder if some aspects of the coding phenomena might be predicted or observed from the hypothesis stated in the previous section.

The physico-chemical constraints connected with the parameters (hydrophobicity, volume and $\Delta G_0$) are assumed to be most important at the beginning of life’s history. Indeed, if these constraints are sufficient to explain the origin of the code, no pre-existing functional proteins
(such as the synthetases) are required for the translation process to work. The system can therefore be identified with the “RNA World” (Gesteland & Atkins, 1993); it contains RNA and free amino acids; polypeptides with enzymatic activities are not necessarily present.

In reconstructing the events of this period, it is important to consider the effect on the translation process of the lack or rarity of any of the amino acids. Because we asserted that a particular amino acid cannot be coded by any codon family, the occurrence of gaps in the genetic code table (which are equivalent to stop codons) can be predicted.

We considered the prebiotic experiments in which amino acids are produced (Miller, 1987), and recorded the percentages of the different kinds of amino acids in a simplified genetic code table (Table 1), to establish which particular codon families might have been affected. The first and second codonic positions show an astonishing difference in the [amino acids (R)/amino acids (Y)] ratio (the 3rd position is of lesser importance in this analysis): amino acids coded by Y in the first position are only present in traces, while the second position displays a well-balanced proportion of amino acids (R) and amino acids (Y). It would seem obvious, but is nevertheless important to note, that the amino acids produced in these experiments (13 in all) are naturally the simplest ones: the more complex of the 20 coded amino acids can only be synthesized by means of enzymes which occur in contemporary organisms. Moreover, among the amino acids produced, the simplest ones are the most abundant (Miller, 1987).

Accordingly, only RNA sequences with a great excess of R in the first position should have been translated to any extent at that time: these sequences can therefore be schematized by RNN (N standing for any nucleotide).

It is instructive to consider these observations in parallel with some results of statistical investigations on present-day coding sequences. Many

---

**Table 1**

Proportion of amino acids produced in prebiotic experiments shown in an R/Y universal genetic code map

<table>
<thead>
<tr>
<th>First position of codon</th>
<th>Second position of codon</th>
<th>Y</th>
<th>R</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>0.6%</td>
<td>0%</td>
<td>0%</td>
<td>0.6%</td>
</tr>
<tr>
<td>R</td>
<td>50.4%</td>
<td>49.0%</td>
<td>99.4%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>51.0%</td>
<td>49.0%</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

Note: The percentages are calculated from the total amount of coded amino acids produced in two kinds of experiments (Miller, 1987, Tables 2 and 3).
TABLE 2
Frequency of purines and pyrimidines in the three codonic positions of prokaryotic coding sequences

<table>
<thead>
<tr>
<th>Kind of base</th>
<th>1st pos.</th>
<th>2nd pos.</th>
<th>3rd pos.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>37.0%</td>
<td>51.3%</td>
<td>53.6%</td>
</tr>
<tr>
<td>R</td>
<td>63.0%</td>
<td>48.7%</td>
<td>46.4%</td>
</tr>
</tbody>
</table>

Note: Nucleotide frequencies for the three codonic positions obtained from the analysis of the 44th release of the EMBL database for prokaryotes, by treating 11 748 coding sequences containing about 12 007 000 nucleotides (only complete coding sequences beginning with ATG were retained). Results for Eukaryotes were very similar.

Attempts have been made to measure Shannon’s Information (Shannon, 1948) contained in genes. Up to now, however, results have tended to show only a very low level of correlation between nucleotides (Lio et al., 1996; Tsonis et al., 1997), the most remarkable fact being that the frequencies of the four different kinds of nucleotides depend on their codonic position (Shepherd, 1981; Eigen & Winkler-Oswatish, 1981; Jukes, 1996) (see Table 2). An important excess of R exists at the first position [it should be mentioned that, as this position displays an R-Y degeneracy symmetry, Fig. 2(a), a well-balanced ratio could be expected], the other two positions remaining statistically “normal” in this general R/Y analysis. As a consequence, present-day coding sequences can also be schematized by RNN (this means that complex amino acids occur less often than the simpler ones).

These observations seem to confirm our basic hypothesis, because random assignment of amino acids to codons (with respect to their physicochemical properties) would not be expected to affect the general pattern of coding sequences.

6. The Volume/Energy Correlation

In view of the previous observations, we have to find a plausible explanation of why misassignment of one amino acid to a particular primordial tRNA should not normally occur or, if it does, why translation is improbable. These two alternative (or complementary) phenomena may be expected if the hypothesis stated in Section 4 is valid.

When reconstructing events at the time of the origin of the genetic code, an important fact should be noted: while both correlations might be reflected in events during a primordial assignment process, it is impossible that the two could have occurred simultaneously. Indeed, the hydrophobicity correlation does suggest that the amino acids interacted with the anticodon nucleotides during some stage of the assignment process (Woese, 1965; Weber & Lacey, 1978; Jungeck, 1978; Kuhn & Waser, 1981; Lacey & Mullins, 1983; Blalock & Smith, 1984; Lacey et al., 1992, 1993). The volume correlation could not have been expressed at that moment, because it implies anticodon–codon interaction. This observation indicates, as mentioned above for other reasons, that the whole process occurred in two stages.

Any explanation of the volume correlation must lie in a process that would exhibit special behavior in some cases (asparagine, arginine and tryptophane; see Fig. 3). To start with, if we try to find a direct relationship between the parameters implied, many contradictions appear. Anticipating the model developed below, we can imagine a primordial tRNA hairpin structure consisting of an anticodon loop stabilized by a stem of a certain length, which can be considered as the simplest RNA molecule with coding properties potentially able to catalyse the polymerization of amino acids (Kuhn & Waser, 1981; Di Giulio, 1994) (see Fig. 5). Indeed, it is often considered as the ancestor of present-day tRNAs (Di Giulio, 1999). The simplicity of this kind of molecule is very encouraging, from the thermodynamic point of view, if we want to describe a process which allows a minimum of predictions. Supposing that an amino acid binds covalently at the 3’ end of the stem (as mentioned, it cannot interact with the anticodon at this moment), we immediately notice that the volume of the [primordial tRNA + amino acid] structure is very close to that of the primordial tRNA alone. In particular, exchanging one kind of amino acid for another will not affect the global properties of the structure because of the size of the RNA, even if it consists of the smallest possible hairpin. However, examining the parameters independently, the $\Delta G_0$ parameter has a natural connection with the characteristic time $\langle \Delta t \rangle$ during which a tRNA structure binds to its cognate
FIG. 5. The translation process at the moment when a primordial tRNA carrying an amino acid has been brought by diffusion into contact with the subsequent codon of a hypothetical RNA coding strand. In the case of anticodon-codon complementarity (according to the wobbling rules), the translation becomes effective if the evolving amino acid (with lateral chain R) is able to bind covalently with the peptide’s last amino acid (with lateral chain R). For this to occur, the anticodon-codon interaction (ΔG₀) involving the codon must be high enough so that the inequality Δt allotted ≥ Δt necessary is verified (explanations in text). The area shaded light grey indicates schematically the area of conformations of the backbone of the amino acid. For clarity, no ribosome is shown here (see comments in text).

codon. Since the association rate constants have been found to vary only moderately for the different anticodon-codon pairs (cf. Grosjean et al., 1978), each association equilibrium constant is approximately inversely proportional to the corresponding dissociation rate constant (Grosjean et al., 1978). Therefore, the dependence between ⟨Δt⟩ and ΔG₀ can be evaluated by

\[ ⟨Δt⟩ \approx A \exp(- ΔG₀/RT), \]

where R is the perfect gas constant, T is the temperature and A is a constant.

An interpretation of the correlation is that small amino acids stay a long time in a place where potentially they can bind to a hypothetical growing peptide, whereas larger ones stay a shorter time in such a place.

For the amino acid volume, it is more difficult to find an explanatory parameter which can be linked to the correlation by convincing arguments. It is a curious fact that the correlation concerns one of the simplest properties of the amino acids: it is not directly related to any chemical or other special physical properties of the amino acids such as the polarity or the hydrophobicity [Fig. 4(a)].

The question is the following: are there physico-chemical reasons why the time necessary for the formation of the covalent bond between the incoming amino acid and the growing peptide should be longer for small amino acids compared with larger ones, as the volume correlation suggests? The problem seems theoretically difficult, but fortunately, such a dynamical aspect has already been studied experimentally. We found evidence in the paper by Milstien & Cohen (1970) entitled “Rate acceleration by stereopopulation control: models for enzyme action”. As indicated by the title, they studied parameters connected with the chemical reactivity of biological molecules, and hence their importance for enzymes.

Their findings show that intramolecular reaction rates may be enhanced by the adjunction of atoms on the molecules. This adjunction produces conformational restrictions on the reactive group and thus increase the population of most productive conformers (considering a statistical ensemble of molecules), hence the term “stereopopulation control”. In particular, Milstien & Cohen examined the lactonization of hydrocoumaric acid and alkyl derivatives (Fig. 6) and found that the increase in the reaction rate of the most alkylated molecule compared with the simplest form is of the order of 10¹¹.

There is a striking analogy between this kind of reaction and the process of peptide bond formation during translation (Fig. 5), which involves the 20 different kinds of amino acids.

Firstly, the respective positions of the reactive groups (NH₂ in the case of the amino acids and COOH for the hydrocoumaric acids) are very similar: they are both situated at the free ends of the molecules, which are the regions most likely to be influenced by conformational properties. Secondly, the ribose which binds the amino acid to the primordial tRNA (or any tRNA of a
present-day organism) should play a very similar role to that of the aromatic part of the hydrocoumaric acids. Indeed, the conformational restrictions are the result of the mutual interactions between the atoms of this ring and those of the fluttering part of the molecule, which is constituted by the amino acid in the case we examine.

Comparison of the different reaction rates (Fig. 6) reveals significant insights: on the addition of only two groups of methyl (CH$_3$) to the fluttering carbon chain, the relative reaction rate increases from 1 (compound I) to 4400 (compound III), whereas such a difference does not exist between compounds I and II, where atoms are only added to the aromatic ring. Because the position of the lateral chain of any amino acid (R$_j$ in Fig. 5) is very similar to that in the mentioned methyl groups, such difference in reaction rates for peptide bond formation should be expected between two amino acids such as glycine (R$_j$ = H) and alanine (R$_j$ = CH$_3$), and by extension to all the other amino acids: the greater the number of atoms, the higher the reaction rate should be.

These considerations lead us to define the variable RF$_j$: rate factor of the amino acid aa$_j$, characterized by the lateral chain R$_j$.

As suggested by the observations above, this variable is assumed to be very closely dependent on the size of the conformational space of the backbone of the amino acid cs$_j$, defined when the amino acid aa$_j$ is bound to the 3' end of a tRNA, as depicted in Fig. 5. It is therefore connected with the number of atoms of the lateral chain of the amino acid, and hence with its volume $v_j$, which limits the size of this space.

To summarize the whole of our argumentation and to show the expected consequences on the translation process of the different values taken by $\Delta G_0$ (anticodon–codon) and volume (amino acids), two characteristic times have to be considered: the time necessary for the formation of the covalent bond ($\langle \Delta t \rangle$ necessary), depending on RF$_j$, and the time allotted for such a reaction ($\langle \Delta t \rangle$ allotted), depending on the strength of the anticodon–codon interaction ($\Delta G_0$).

Let us consider the translation process at the moment when a charged primordial tRNA has been brought by diffusion to the codon (cd$_j$) situated next to the one (cd$_i$) which binds the previous primordial tRNA carrying the peptide corresponding to the already translated part of the RNA (Fig. 5). We assume that the anticodon of the arriving primordial tRNA is complementary (according to the wobbling rules) to the cd$_j$ codon. For the translation to be effective, the amino acid aa$_j$ (whose lateral chain R$_j$ is of any kind a priori) has to form a covalent bond with the peptide's last amino acid (with lateral chain R$_i$).

The requirement for this to occur can be stated as

$$\langle \Delta t \rangle \text{ allotted} \geq \langle \Delta t \rangle \text{ necessary}$$

where

$$\langle \Delta t \rangle \text{ allotted} \approx A \exp(-\Delta G_0/RT)$$

refers to the anticodon–codon interaction involving the cd$_j$ codon, and

$$\langle \Delta t \rangle \text{ necessary} = \theta(\text{RF}_j)$$

refers to the peptide bond formation between the amino acids aa$_i$ and aa$_j$. Under standard conditions, $\theta$ is assumed to decrease monotonically as RF$_j$ increases.

Here, $\theta$ is shown to be dependent on RF$_j$ rather than directly on $v_j$. Indeed, as mentioned above, not all the amino acids are in correspondence with the main sequence of the volume.
correlation (Fig. 3). A possible interpretation is that the lateral chain of special amino acids (namely asparagine, arginine and tryptophane) may interact in a particular manner with the atoms of the ribose, and possibly with other atoms situated in the vicinity of the reaction site. The consequence of such interactions would be that conformations corresponding to high probability of reaction are favored (asparagine) or disfavored (arginine and tryptophane) in some cases, resulting, respectively, in an enhancement or a decrease of the reaction rates. These considerations lead us to propose a more precise definition for $RF_j$:

$$RF_j = \int_{c_j} f(c) p(c) \, dc \div \int_{c_j} dc,$$

where $f(c)$ is a factor of reactivity and $p(c)$ the probability for the amino acid to be in the $c$ conformation. This probability depends on the interactions of the lateral chain $R_j$ with the surrounding atoms.

However, the volume correlation does suggest that a simple dependence between $RF_j$ and $v_j$ should be valid for almost all the 20 amino acids. To simplify the discussion in the following, the size (or the volume) of an amino acid may be taken for its $RF_j$.

The explanation of the volume correlation leading to the last inequality is in agreement with the observations discussed in Section 5, because it shows that a primitive genetic system which does not contain large [complex] amino acids is only able to translate sequences with codonic pattern RNN (see also the discussion in Section 9). In this paper, $a[b]$ is used to express the idea that the two properties $a$ and $b$ are correlated but not equivalent. However, the inequality implies that it is possible for large amino acids to be coded just as well by weak codons (to a lower limit which depends on the size of the amino acid) as by strong codons. This coding ambiguity should be reduced by the following considerations.

As mentioned just before the inequality, no assumption was made about the kind of amino acid being carried by a particular primordial tRNA (with respect to its anticodon). In Section 7, we will see that a sorting with respect to hydrophobicity should take place at the moment when an amino acid binds to the 3' end of a primordial tRNA (considering a hypothetical process existing at a time before the arrival of the synthetases). However, as this probability of binding (no matter how it happens) is by hypothesis independent of the size of the different amino acids, the occupancy of a population of different primordial tRNAs by these amino acids should reflect their respective abundance in solution. As mentioned in Section 5, at first approximation this abundance is inversely proportional to the complexity [size] of the amino acids in the prebiotic experiments—this is also the case in the cells of contemporary living organisms (data not shown here). Because translation implies a statistical process (the codon on the point of being translated is checked by many tRNAs before the arrival of the correct one), the relative abundance of the small amino acids will make them occur for the translation of the majority of the strong codons. Thus, this kind of competition depending on concentration should reduce the ambiguity of coding. The quantitative aspect of these statistics will not be developed in this article.

These considerations lead us to discuss the constraints affecting this part of the coding process in a general way (Fig. 7). Even in contemporary organisms, the statistical aspect of the coding phenomenon is of great importance: it implies that anticodon–codon mismatch is equivalent to $\Delta G_0$ being sufficiently unfavorable as to prevent the translation from occurring before the arrival of the correct tRNA. Our explanation enables the constraints to be extended to those due to the differential reactivity of the amino acids, expressed by the various values of the $RF_j$.

Examining the different kinds of interactions involved, it seems appropriate to wonder whether ribosomes are necessarily required. As we are concerned with the first steps of the organization of the genetic system, this question is of prime interest.

The necessity for the primordial tRNA carrying the peptide to be momentarily blocked on the $cd_j$ codon (Fig. 5) reveals that such a translation system cannot take place without the participation of molecules able to proceed to this stabilization. The presence of ribosomes (or a primitive form or
Fig. 7. General considerations on the constraints responsible for part of the genetic coding (second level of sorting). (a) Required conditions for the translation to occur at the moment when a charged primordial tRNA has been brought by diffusion into contact with the last untranslated codon of a hypothetical RNA coding strand (see Fig. 5). When anticodon–codon complementarity does not exist (mismatch), \( \Delta G_0 \) anticodon–codon is unfavorable (left axis, where \( \Delta G_0 \) is indicated), and this prevents the translation in all cases (dashed area). This means that the primordial tRNA leaves the codon while retaining its amino acid. When there is anticodon–codon complementarity (according to the wobbling rules), translation occurs only if the \( \Delta G_0 \) anticodon–codon corresponds to a translation state for the amino acid carried, the number of these states depending on the \( RF_j \) value of the amino acid. Each column represents the possible translation state(s) for a particular amino acid, ranging from the smallest (left) to the largest. The different translation states represent average \( \Delta G_0 \) anticodon–codon interactions, taken over four values (for a four-fold degenerate codon family, in black) or two values (for a two-fold degenerate codon family, in grey) (see Fig. 3). (b) Details of the set of states ringed in part (a), showing the mixing of the energy levels of the particular codons belonging to neighboring codon families (see text for comments). This schematic diagram does not contain all the amino acids and codons of the genetic code.

The First Level of Sorting: Hydrophobicity

For a more complete view of the primitive coding process, it now remains to find an explanation of the hydrophobicity correlation, which is the image of a process that evidently occurs just before the one discussed in the previous section. Hydrophobicity sorting can be considered to be effective at the moment when an amino acid binds covalently to the 3′ end of primordial tRNA (see the discussion at the beginning of Section 6).

Many authors (Woese, 1965; Weber & Lacey, 1978; Jungck, 1978; Kuhn & Waser, 1981; Lacey & Mullins, 1983; Blalock & Smith, 1984; Lacey et al., 1992, 1993; Di Giulio, 1994) have already worked on the significance of this correlation, for which it is not obvious to find precise molecular arguments. The present interpretation shown in this section is thus perhaps the most speculative part of this work, not least because it concerns catalytic properties of RNA molecules, which are very difficult if not impossible to predict theoretically (Gesteland & Atkins, 1993). However, recent experimental investigations in this domain have produced remarkable results (Ekland et al., 1995), giving us insights which
suggest the hypothetical process linked to this correlation.

Our assumption is that this process is mediated by the primordial tRNA itself, and that it occurred without the participation of external molecules at this stage of evolution, notably because of the probable absence of synthetases (Wetzel, 1995) and also because an important property of the amino acids (their size) could have been coded independent of the presence of these hypothetical molecules, as we have just shown. It is thus important to point out some physico-chemical properties of RNA hairpin structures that might be responsible for this coding capacity.

Firstly, the estimation of the increase in $\Delta G_0$ due to the formation of an RNA loop is a minimum when there are seven nucleotides in the loop (Turner et al., 1987), which is the size of an anticodon: this size is thus thermodynamically preferred to others. Secondly, the stability (equilibrium constant) or a hairpin depending on the length of the stem, an appropriate size would promote the regular (statistical) change between two conformations of the RNA: the close hairpin structure and the non-apparative RNA strand. These features should give catalytic properties to the hairpin we want to show.

Based on these observations, we consider an RNA hairpin structure with six bases paired in the stem and, like contemporary tRNA, four free nucleotides at the $3'$ end of the molecule. The global $\Delta G_0$ is then about $-6.3$ kcal mol$^{-1}$ (see Fig. 8, top left), which confers some stability on the hairpin.

Our intention is not to discuss the exact probability of such a molecule being present in an RNA World [thermodynamical arguments can be found in Turner & Bevilacqua (1993) and hypotheses on their function as primordial tRNA are exposed in Di Giulio (1999)], nor to be sure that the proposed number of nucleotides is at the optimum to allow association/dissociation to take place in the best way, but merely to show that this kind of RNA would have the capacity to bind an amino acid situated at the origin on the anticodon covalently at the $3'$ end. Concerning the amino-acid–anticodon interaction hypothesis (see Lacey et al., 1992 and references therein), it should be mentioned that, because most other nucleotides of the primordial tRNA are associated in base pairing, the anticodon probably constitutes the most specific target for a free biomolecule. Also, the unpaired nucleotides at the $3'$ end of the stem are probably less able to localize an amino acid in a specific place, as the anticodon should do, which is of importance for a process implying the formation of a covalent bond. We therefore interpret the hydrophobicity correlation in the following way (Fig. 8). Diffusing in a space filled by water (which could be delimited by a membrane), different kinds of amino acids encounter anticodon nucleotides and an intermolecular binding is created which is weak, but of higher magnitude when there is hydrophobicity correspondence. As the hairpin obeys the thermodynamic laws, dissociation takes place in a statistical manner, which allows the $3'$ part of the strand to associate with the nucleotides situated at the border of the anticodon, and thus to be close to the amino acid. To experience binding, the amino acid must, however, be oriented in such a way as to present its carboxylic group near the alcohol group of the ribose. Most of the specific lateral group of the amino acid is then necessarily situated close to the nucleotides of the anticodon (Fig. 8, magnification 2). It is probably at this moment that the hydrophobicity correspondence between amino acid and anticodon is most important, because incompatibility would prevent the correct positioning of the amino acid, and therefore decrease the probability of the reaction. Whether “occupied” by the amino acid or not (depending on the binding event), the RNA then returns to the initial configuration, which is more stable.

Some observations already mentioned seem to corroborate the mechanism leading to the attachment of the amino acid to the primordial tRNA. In particular, the fact that most of the hydrophobicity correlation concerns the second anticodonic position (Blalock & Smith, 1984; Taylor & Coates, 1989) instead of the first one is in agreement with the model, because spatial constraints would prevent strong interaction of the amino acid with the first position (see Fig. 8, magnification 2).

This theoretical process, leading to the selective binding of amino acids onto the $3'$ end of primordial tRNA, is an expected result of the thermodynamic instability of such small RNA
FIG. 8. Hypothetical process representing the first level of sorting (amino-acid hydrophobicity). The first part of the diagram (top left) indicates the thermodynamic parameters of the primordial tRNA (according to Turner et al., 1987). The succession of events is as follows. A diffusing amino acid encounters the anticodon of a primordial tRNA and intermolecular association takes place, but with higher magnitude when there is hydrophobicity correspondence. Magnification 1 shows a particular amino acid (distinguishable by its lateral chain R) interacting with the anticodonic nucleotides. As the RNA hairpin is thermodynamically unstable, dissociation occurs and part of the 3' end is then able to associate with the nucleotides situated at the border of the anticodon. This configuration is favourable for the formation of a covalent bond between the 3' end of the RNA and the amino acid. For this to occur, the amino acid must be oriented in such a way that most of its lateral chain (R) interacts with the 2nd position of the anticodon (magnification 2), which stabilizes the amino acid (and therefore enhances the probability of the formation of the bond) if there is a sufficiently high hydrophobicity correspondence. The RNA then returns to the initial configuration (bottom center), which is more stable (drawn here in the case of a binding event).

8. Complementarity in the Group of the Four Bases of the Genetic Alphabet

On examination of the two successive processes leading a free amino acid to be incorporated into a coded peptide (Fig. 9), astonishing facts appear. A first remark concerns the two properties of the amino acids subject to sorting. As visible in Fig. 4(a), they can be considered as approximately mutually orthogonal: hydrophobicity cannot be deduced from the volume and reciprocally. It is however, possible to qualify this affirmation, in the sense that an amino acid with a large lateral chain generally displays the strongest water-affinity dependence, either hydrophobic or hydrophilic.

Thus, the ambiguity of the assignment based on these two parameters is minimal, because they allow an optimal characterization of the different kinds of amino acids, which are therefore well separable by the two processes of sorting [Fig. 9(b)]. Another fortunate consequence of this orthogonality concerns the problem due to the mixing of the energy levels pointed out at the end of Section 6, which can be at least partially removed. Indeed, even if the hydrophobicity sorting does not completely differentiate all the different kinds of amino acids, its effect is notably to separate amino acids having the same approximate volume, which indirectly allow codon families to be distinguished whose respective codons are mixed on the $\Delta G_0$ scale [see Fig. 7(b)].
short, the two different levels of sorting perform mutual refinement.

One could then ask if the four different kinds of nucleotides have to exhibit some special molecular properties to perform this two-step separation in an optimal way.

As indicated by the dashed line in Fig. 3, it is possible to separate the different codon families into two distinct sets: the codon families constructed with A or U at the second position of the anticodon (above the dashed line, weak codons), and those constructed with G or C (under the dashed line, strong codons). Remarkably, A and U are, respectively, the most hydrophobic and most hydrophilic nucleotides (see Section 2), but they also belong to the weak interaction set [indicated as W in Fig. 2(b)]. The same reasoning holds for G and C: they are, respectively, below and above a mid-value on the hydrophobicity scale (although the difference is less pronounced than for A and U), and both belong to the strong interaction set [S in Fig. 2(b)]. It should also be remembered that hydrophobicity selectivity is best performed precisely by the second (anti) codonic position.

This observation shows that the four bases of the genetic alphabet display complementarity: The [U, C, G, A] succession ranks the nucleotides from the most hydrophilic to the most hydrophobic, but also with energy types [W, S, S, W]. This is another argument for writing the genetic code table according to this succession [Fig. 2(a)]. It shows that the two parameters (hydrophobicity and energy type) always permit any one of the four bases to be distinguished from the others (Fig. 10).

This complementarity property might explain why the incorporation into the genetic alphabet of other kinds of bases, which can notably be produced in laboratory experiments (see Piccirilli et al., 1990) did not happen.

In view of these observations, it seems possible to describe more completely the way in which information is coded in genes, in terms of both
the (well-known) hydrophobicity and the conformational properties of the amino acids. This could give new insights into the way the folding of a protein is coded in genes. The conformational space connected with the $RF_j$ (see Section 6 for its definition) must, however, not be confused with the one observed when the amino acid is included in a protein because the context is different.

9. Discussion

The main point this paper aims to make is that it is possible for the genetic code to exist without the participation of pre-existing functional proteins. Indeed, according to our interpretation, the emergence of such a coding system comes from minimization of the physico-chemical constraints connected with the polymerization of amino acids mediated by RNA. The consequence of this affirmation is that it is not a relevant question to ask whether the genetic code is optimized or not: the genetic code is the optimization of this polymerization process.

The problem of the accuracy of the primitive coding system is of prime interest and has been discussed at many stages. We have shown that important ambiguities can be removed while considering aspects of the coding phenomena that are connected, notably, the two parameters implied in the successive sorting processes, and the importance of the various concentrations of the different kinds of amino acids. Nevertheless, as it is subject to fluctuations, this last parameter shows that an important “wobbling” should remain in the assignment, the optimality of the polymerization (i.e. the code) being dependent on these concentrations (see Section 6). Starting from such a primitive system, the appearance of proteins with enzymatic activities should have notably enhanced the accuracy of the coding phenomenon.

Considering the present-day translation machinery, one could then ask if the basic physico-chemical parameters examined (hydrophobicity, volume and $\Delta G_0$) are still relevant.

It is well known that the tRNA contains many modified nucleotides. The first base ($3'$) of the anticodon is very often I, Q or S, which modulates the strength of the anticodon–codon interaction (Grosjean et al., 1978). Moreover, the kind of base situated on the $3'$ and $5'$ side of the anticodon is often correlated with its base composition, and also changes the way the interaction is realized (Grosjean et al., 1978). Other features of the present-day tRNA would certainly affect specific interactions with the ribosomes and the mRNA, but a complete description, and especially that of the anticodon–codon interaction inside the ribosome, will probably be very difficult to obtain.

Nevertheless, considering the reaction leading to the peptide bond formation, it is hard to imagine how an elaborate ribosome could “standardize” the kinetics for the 20 amino acids: this would imply at least 20 conformations of the region around the reactive site to limit specifically the size of the conformational space given to the different amino acids, which is not without creating contradictions. Indeed, the existence of the volume correlation precisely suggests that these constraints should at least partially remain.

The evaluation of the importance of the hydrophobicity constraints in the present genetic system is more problematic, because of the action of the synthetases. The appearance of these molecules could have changed the rules, which would mean that the hydrophobicity correlation may simply be a vestige of one of the early translation mechanisms. As the modification of any assignment would imply the simultaneous appearance of untidiness in the two correlations, this could prevent drastic modifications of the “universal” code once established, and hence explain why it is nearly the same in all living organisms and very similar to those of the organelles. We have already expressed our reservations about the mechanism we proposed to explain the emergence of the hydrophobicity correlation, but we want to emphasize here that the most important related points, especially those discussed in Section 8, are not affected by this interpretation, whether it is true or not.

Evolutionary considerations were taken up in Section 5, where we raised the question of the availability of the different kinds of amino acids at the time of the “RNA World”. Our interpretation of the volume correlation shows that only small fragments of RNA could have been translated, because the primitive system does not contain complex amino acids. The reason is precisely
that the high values of the $RF_j$ correspond to the complex [large] amino acids. Thus, the low reactivity of the simple [small] amino acids would impose strong constraints on the RNA sequences susceptible to being translated.

An interesting consequence of this phenomenon is the importance of the reading frame. One could argue that there always exists a reading frame which is better translated than the other two. This is because they correspond to different amino acids. For example, the sequence UAC|UAC|UAC|UAC... codes for a peptide composed of tyrosine, which is not produced by prebiotic experiments. When read as CUA|CUA|CUA..., the same sequence could be translated to a polyleucine because this amino acid is present. To generalize, the observations made in Section 5 show that a sequence constructed with the motif RNN|RNN|RNN|RNN..., read in that frame, is the best template of polymerize “prebiotic” amino acids. Indeed, as no mechanism susceptible of indicating the beginning of a hypothetical primitive gene would exist at that time, this asymmetry is important to characterize (favour) one special reading frame. It follows as a corollary that a primitive genetic system (without start codon signalling) which already contains all the 20 amino acids cannot evolve because all three reading frames would be read with the same efficiency. This means that no specific sequence (frame) would be preferentially translated (anarchy).

Starting from a prebiotic system with about ten amino acids (see Miller, 1987), the genetic code must have gradually evolved towards the present one by the specific filling up of the “stop” codons (i.e. the codons without assignment) of the table with the new amino acids produced by the emergent enzymatic ways, in accordance with the rules dictated by the constraints. At the same time, the translation process would gradually become more efficient (because of the presence of these new amino acids, with higher values of $RF_j$), and translated sequences would increase in size and complexity. Considering the three stop codons of the current genetic code, one could say they are simply the last “unfilled” codons.

This scenario for evolution suggests an exciting analogy with that of the periodic table of the elements: starting only from hydrogen, helium and lithium at the first second of our Universe, this table was gradually filled by the elements produced in stars and supernovae. However, this idea should be examined with care and we will not develop it here.

I would like to thank F. Rothen and N. Rivier for support and helpful comments, and for the critical review of the manuscript. The analysis of the coding sequences was done with the help of A. J. Koch. I am also indebted to M. Cibils, D. Watson, D. Bonnaz, P. Vogel, A. Sassara and B. Hirt for helpful comments, and to D. Watson for the correction of the English version.

REFERENCES


