Emergence of a Code in the Polymerization of Amino Acids along RNA Templates

Jean Lehmann[1]*, Michel Cibils[2], Albert Libchaber[1]

1 Center for Studies in Physics and Biology, The Rockefeller University, New York, New York, United States of America, 2 Section de Mathématiques, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Abstract

The origin of the genetic code in the context of an RNA world is a major problem in the field of biophysical chemistry. In this paper, we describe how the polymerization of amino acids along RNA templates can be affected by the properties of both molecules. Considering a system without enzymes, in which the tRNAs (the translation adaptors) are *not* loaded selectively with amino acids, we show that an elementary translation governed by a Michaelis-Menten type of kinetics can follow different polymerization regimes: random polymerization, homopolymerization and coded polymerization. The regime under which the system is running is set by the relative concentrations of the amino acids and the kinetic constants involved. We point out that the coding regime can naturally occur under prebiotic conditions. It generates partially coded proteins through a mechanism which is remarkably robust against non-specific interactions (mismatches) between the adaptors and the RNA template. Features of the genetic code support the existence of this early translation system.

Introduction

A major issue about the origin of the genetic system is to understand how coding rules were generated before the appearance of a family of coded enzymes, the aminoacyl-tRNA synthetases. Each of these ∼20 different enzymes has a binding pocket specific for one of the 20 encoded amino acids, and also displays an affinity for a particular tRNA, the adaptor for translation [Fig. 1(a)]. These adaptors are characterized by their anticodons, a triplet of base located on a loop. The synthetases establish the code by attaching specific amino acids onto the 3′ ends of their corresponding tRNAs, a two-step process called aminoacylation [1]. The first step (activation) involves an ATP, and leads to the formation of a highly reactive intermediate, aa–AMP (aa = amino acid). The second step consists of the transfer of the amino acid from AMP onto the 3′ end of the tRNA. Those tRNAs can subsequently participate in the translation of RNA templates, during which codons about to be translated are tested by the anticodons of incoming tRNAs. When anticodon-codon complementarity occurs, an amino acid is added onto the nascent protein through the formation of a new peptide bond [2].

How could a translation system operate in the absence of the synthetases? Recent works have shown that particular RNA stem-loops of ∼25 bases can self-catalyze the covalent binding of amino acids onto their own 3′ ends [3,4]. These RNAs however require aa–AMP as a substrate because they cannot manage the activation step in their present form. In addition, they show little specificity for the amino acids, raising the question of how a code could be generated by them. Some answers will likely be provided by the activation step if possible to implement on these small RNAs. This issue is not examined in the present paper.

Based on an earlier investigation [5], the present analysis shows that the translation process itself can contribute to the establishment of coding rules. Consider an elementary translation system constituted by RNA templates made up of two types of codons {I, II}, tRNAs with anticodons complementary to these codons, and two types of amino acids {1, 2}. Suppose that the tRNAs are *not* selectively loaded with amino acids (i.e. the rates of loading only depend on the relative concentrations of the amino acids). Our analysis shows that it is possible to observe a coded polymerization. We calculate the probability of codon I being translated by amino acid 1 and the probability of codon II being translated by amino acid 2, the coding regime occurring when both probabilities are *simultaneously* higher than 0.5. These probabilities are functions of the anticodon-codon association and dissociation rate constants, the amino acids concentrations and their respective kinetic constants of peptide bond formation. One general configuration allows a coding regime to occur: the amino acid with the slow kinetics (i.e. the "slow" amino acid) is more concentrated in solution than the "fast" amino acid. Given two appropriate codons, the competition for the translation of the codon dissociating quickly from its cognate tRNA (i.e. the "weak" codon) is won by the fast amino acid. As for the "strong" codon, for which the amino acid kinetics are equal or higher than the anticodon-codon dissociation rate constant, the higher concentration of the slow amino acid makes it a better competitor in that case. Although other types of polymerization are possible, we show that this coding regime is favored under prebiotic conditions. It is furthermore remarkably robust against anticodon-codon mismatches. We conclude our analysis by showing that this model can naturally be implemented by a system of four codons and four amino acids thought to be a plausible original genetic code.
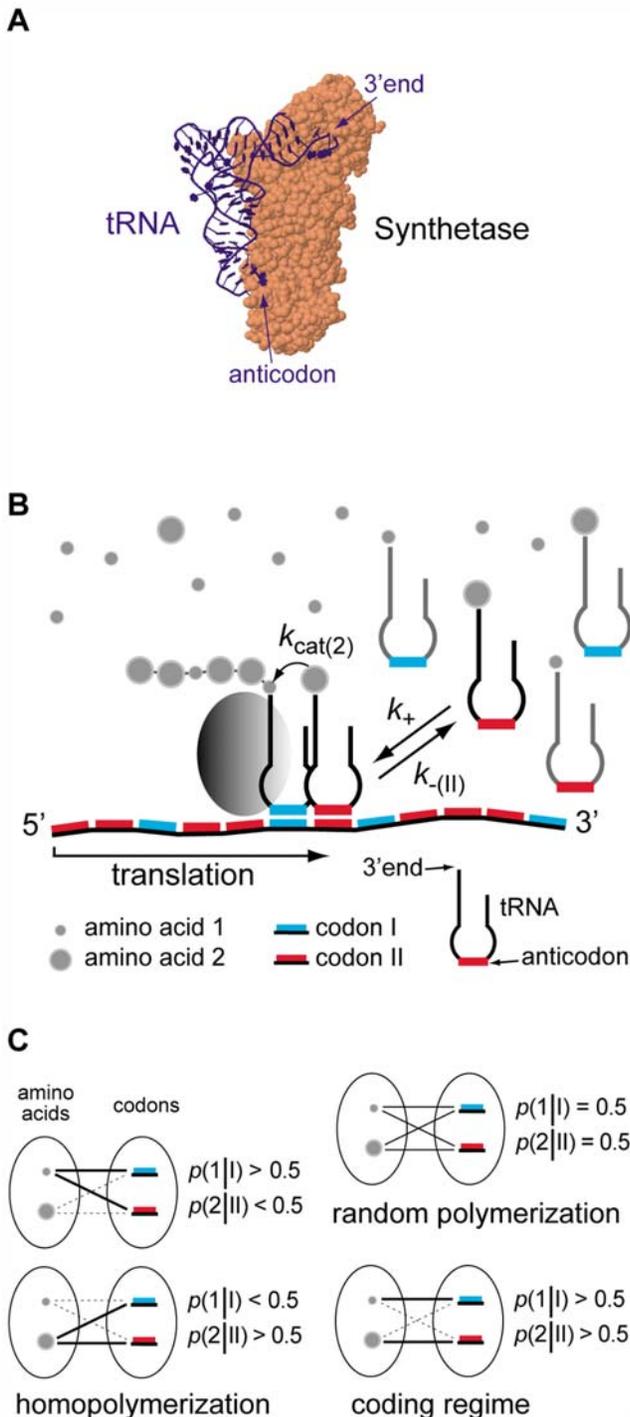
Figure 1. Model of the elementary translation system. (a) Synthetase with cognate tRNA (structure 1ZJW from *Protein Data Bank*). (b) Kinetic scheme of the elementary translation process. Two types of tRNAs (complementary to two template codons) are *unselectively* loaded with two types of amino acids (the rates of loading are only concentration-dependent). The grey rugby ball is a stabilizing cofactor (see text for explanations). (c) Characteristic regimes of the polymerization process. doi:10.1371/journal.pone.0005773.g001

Results

Model

Let us consider two small RNA stem-loops (hereafter called tRNAs) characterized by their anticodons and both capable of loading *as efficiently* two types of amino acids {1, 2} onto their 3' ends. The rates of aminoacylation will thus simply follow the relative concentrations of these amino acids in solution, [aa$_1$] and [aa$_2$]. These tRNAs are involved in the translation of an RNA template made up of two types of codons {I, II} complementary to these anticodons. Translation is governed by a Michaelis-Menten type of kinetics: the first step is characterized by anticodon-codon association rate constants (k_+) and dissociation rate constants (k_-), and the second (irreversible) catalytic step is characterized by a kinetic constant k_{cat} depending on the amino acids [Fig. 1(b)].

It is assumed that the association rate constants k_+ of all anticodon-codon couples are alike. This situation is expected since it already occurs in the context of anticodon-anticodon interactions [6], which is similar. In a first approximation, only complementary matchings are considered. We therefore have k_+ = constante, $k_{-(I)}$ for codon I and $k_{-(II)}$ for codon II. As for the kinetics of peptide bond formation (k_{cat}), an earlier work showed that this variable may strongly depend on the side-chains of the amino acids [5]. Accordingly, two constants ($k_{cat(1)}$ and $k_{cat(2)}$) are defined.

The model includes a ribosome-like cofactor capable of stabilizing the tRNA carrying the nascent protein on the template [Fig. 1(b)]. Although our analysis may not clarify the molecular origin of this cofactor, two of its properties can be specified, which are required to validate our conclusions:

1) The cofactor does not have a catalytic site for peptide bond formation which could minimize the side-chain effect mentioned above. Modern ribosomes have a catalytic site, the peptidyl-transferase center [7]. It is still not clear to us how this specialized part of the ribosome manages the different side-chains, although the problem has already been considered [8].

2) The cofactor does not have a decoding center. This evolved structure on the small subunit of modern ribosomes allows for an increase in the fidelity of anticodon-codon recognition [9]. This structure is inconsistent with the simplicity of the kinetic scheme described here [Fig. 1(b)]. Our model is therefore in agreement with the hypothesis that the original ribosome was only made up of the large subunit [10].

Overall, the cofactor considered here is *a priori* simple since it does not have specialized parts. Avoiding a break in the continuity of the evolutionary process implies that it was already made up of RNA.

Considering now the dynamic of the translation process, it is assumed for simplicity that the relative concentrations of the aminoacyl-tRNAs remain constant over time. This can be guaranteed by the reversibility of the aminoacylation process [1], which will prevent the accumulation of aminoacyl-tRNAs unfit for translation.

With the above hypotheses, let us define the probability $p(1|I)$ of codon I being translated by amino acid 1:

$$p(1|I) = [aa_1] \frac{k_{cat(1)}}{k_{-(I)} + k_{cat(1)}} \Big/ \left([aa_1] \frac{k_{cat(1)}}{k_{-(I)} + k_{cat(1)}} + [aa_2] \frac{k_{cat(2)}}{k_{-(I)} + k_{cat(2)}} \right). \tag{1}$$

One can similarly write $p(2|I)$, and verify that the normalization condition $p(1|I) + p(2|I) = 1$ is satisfied. It is convenient for the analysis to relate the different kinetic constants and concentrations

with ratios. We denote

$$a = k_{cat(2)}/k_{cat(1)}$$
$$b = k_{cat(1)}/k_{-(I)}$$
$$c = k_{-(I)}/k_{-(II)}$$
$$d = [aa_2]/[aa_1],$$

by defining the kinetic constants in order to have $a \geq 1$ and $c \leq 1$. Then, expression (1) becomes

$$p(1|I) = \frac{ab+1}{d(ab+a)+ab+1}. \qquad (2)$$

Similarly, the probability of codon II being translated by amino acid 2 is

$$p(2|II) = \frac{d(abc+a)}{d(abc+a)+abc+1}. \qquad (3)$$

Typical configurations of (2) and (3) are shown in Fig 1(c), which groups together the main possible outcomes of the polymerization process: random polymerization, homopolymerization, and coding regime. We are particularly interested in the case when this system displays some homogenous coding properties Σ, defined as

$$\Sigma = p(1|I) = p(2|II) > 1/2. \qquad (4)$$

It can be shown that there is only one solution to the coding problem (i.e. $p(2|I) = p(1|II) > 1/2$ cannot occur).

A few general statements can be made about relation (4). With a, b, c, $d > 0$ to be physically meaningful, the coding regime Σ cannot be observed if the amino acids are characterized by identical k_{cat} ($a = 1$) or if they are at the same relative concentrations ($d = 1$). In both cases, the only possibility is $p(1|I) = p(2|II) = 0.5$ and $a = d = 1$. If both codons I and II display identical k_- ($c = 1$), we also get $p(1|I) = p(2|II) = 0.5$, and $d = (ab+1)/a(b+1)$.

Amino acid requirements for the coding regime

An examination of expressions (2) and (3) when b varies within the interval $]0, \infty[$ shows that the coding regime (4) can be satisfied only if $c < 1$ and

$$1 > d > 1/a. \qquad (5)$$

Condition (5) signifies that *the amino acid with the highest k_{cat} must be less concentrated in solution than the other amino acid. This relative concentration must still be higher than $1/a$.*

To check whether condition (5) could be reasonably fulfilled at the origin of Life, let us consider some results of the well-known prebiotic synthesis experiments conducted by Miller [11], which revealed what amino acids of the genetic code are the easiest to be generated. The histogram of Fig. 2 shows that glycine and alanine display a similar abundance, and are about one order of magnitude more frequent than the next two amino acids, aspartic acid and valine. Considering now the chemical step, how may each of these amino acids affect the probability of peptide bond
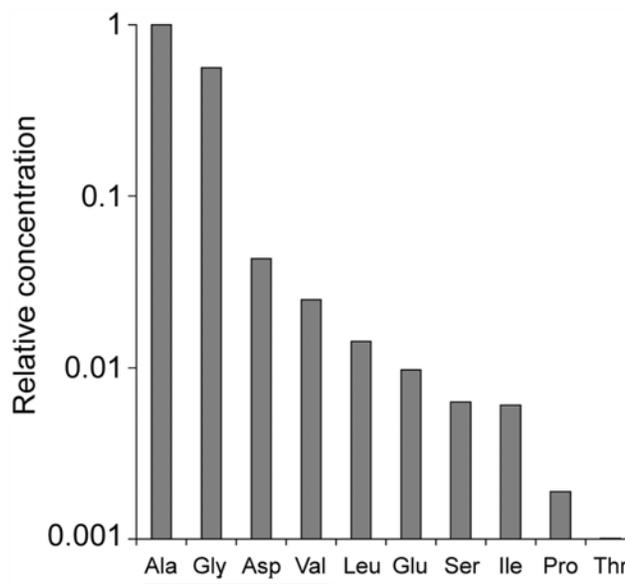

Figure 2. Relative abundance of primitive amino acids. Relative abundance of the 10 most frequent amino acids of the genetic code synthesized in an experiment thought to reproduce the conditions of the prebiotic Earth. Graph established from the data of Table 3 in ref. [11]. doi:10.1371/journal.pone.0005773.g002

formation (k_{cat})? Studies of intramolecular reactions [12–14] show that the size of the group(s) of atoms bound to the carbon in position 1 or 2 after a nucleophile is usually very critical for a reaction rate. When hydrogens are substituted with a dimethyl, the relative reaction rate (k_{rel}) may not appreciably change, although the result depends on the system and the position of the substitution [14]. When these substitutions involve bulkier groups (such as diethyl groups), a sudden jump of at least two orders of magnitude of k_{rel} is typically observed (Fig. 3). Bulky substituents restrict rotation around bonds, which contributes to the localization of the nucleophile (entropic effect) [13]. These data suggest that glycine and alanine may be characterized by similar k_{cat}, but that a much higher k_{cat} is expected for aspartic acid and valine (Fig. 4). Restricting our analysis to the four most abundant amino acids, the above considerations show that two categories may be established: {Ala, Gly} and {Val, Asp}. Each of these categories comprises amino acids that are similar with respect to k_{cat} and concentration. This degeneracy is examined further below. If we assume (as suggested by the above data) that these two categories are related by $a \sim 100$ and $d \sim 0.1$, we can conclude that condition (5) may be fulfilled in some primitive environment.

Codons for the coding regime

Let us now examine how well any two codons (related by $c = k_{-(I)}/k_{-(II)}$) can satisfy relation (4). To a low value of c corresponds a high value of Σ (see below). The ΔG_0 values of all complementary anticodon-codon interactions of the genetic code span approximately from -2 to -6 kcal mol^{-1} at T = 310 K [5]. These estimates refer to crude anticodon-codon interactions: they do neither include additional ΔG_0 contributions occurring within modern ribosomes [9], nor the effect of tRNA anticodon loop refinement taking place in the present-day genetic system [15–16]. With $K_{eq} = k_+/k_- = \exp(-\Delta G_0/RT)$, and with $k_+ =$ constant (see above), the lowest value of c lies somewhere between 10^{-2} and 10^{-3}.

A particular solution of Σ is obtained when $ac = 1$ which is physically relevant: for a given a, Σ increases relatively rapidly as c

k_{rel}

Bimolecular reaction

CH_3COO^- + $CH_3COOC_6H_5Br(p)$ 1.0

Intramolecular reactions

I   ~ 1000 M

II   ~ 3600 M

III   ~ 180,000 M

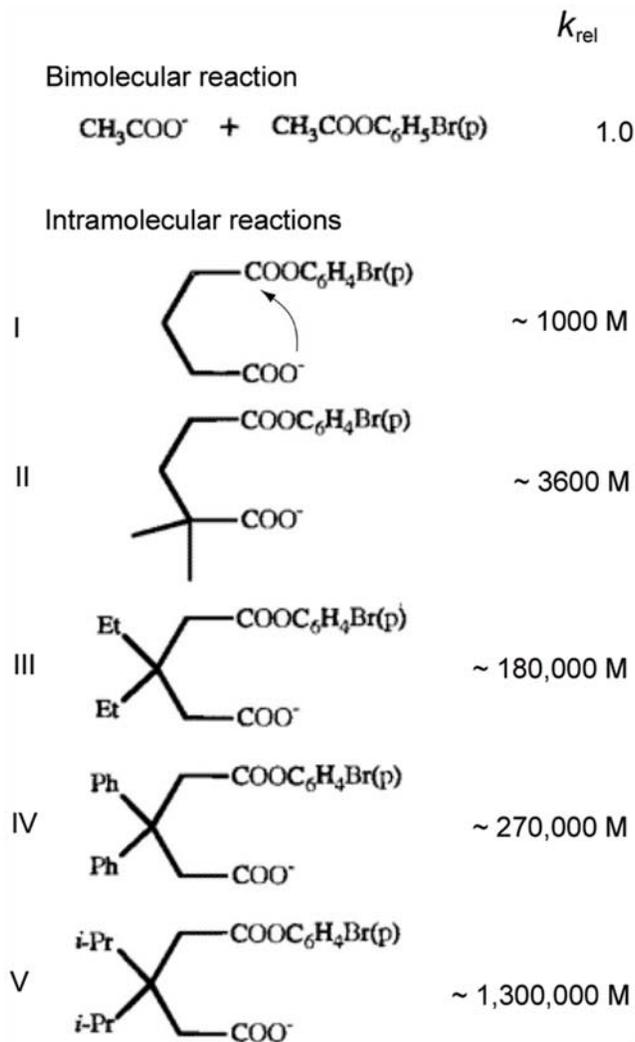IV   ~ 270,000 M

V   ~ 1,300,000 M

Figure 3. Effect of the local environment on a reaction rate.
Relative reaction rates (k_{rel}) of a bimolecular reaction and some corresponding intramolecular reactions. In the intramolecular systems, the nucleophilic attack (indicated by a small arrow in compound I) leads to the cyclization of the compounds. Adapted from ref. [12].
doi:10.1371/journal.pone.0005773.g003

decreases between 1 and $1/a$ [see Fig. 5(a–b)]. The turning point $c = 1/a$ sets up the maximal value of the "easy gains" for Σ, i.e. gains which do not require an unrealistic $\Delta\Delta G_0$ value between the two codons. When $c < 1/a$, Σ increases less and less, until it reaches a maximal value. Furthermore, a simplification in the algebraic treatment of relation (4) occurring when $ac = 1$ (implying $b = k_{\text{cat}(1)}/k_{-(I)} = k_{\text{cat}(2)}/k_{-(II)}$) allows us to determine that $\Sigma_{(ac=1)} = d(a+1)/[ad^2 + d(a+1) + 1]$ has a maximum with respect to d when $d = a^{-1/2}$. Since $b_{(ac=1)} = (1 - a^2 d^2)/a(d^2 - 1)$, this maximum is also characterized by $b = 1$. This result shows that the above estimates of a and d are well compatible with a polymerization process in a coding regime. Taking these values ($a = 10^2$; $d = 10^{-1}$) together with $c = 10^{-2}$, one gets $\Sigma = 101/121 \approx 0.835$.

Fig. 5(a) shows $p(1|I)$ and $p(2|II)$ as a function of b for three characteristic values of d, which allows to graphically check our results with the above numerical values. Remarkably, more realistic calculations including the effect of mismatches show that the coding regime is only marginally affected by them (see below), demonstrating the robustness of this coding system.

Connecting amino acids with codons

The b parameter makes a connection between two kinetic constants of different origins: $k_{\text{cat}(1)}$ is determined by an amino acid (bound to an RNA) while $k_{-(I)}$ is determined by RNA. Among the two constants, only $k_{-(I)}$ could be tuned by the translation machinery, through the selection of the tRNA anticodon loop. Anticodons of two bases (if structurally possible) are expected to have a high $k_{-(I)}$ because of the low anticodon-codon ΔG_0 and the importance of thermal fluctuations, while the opposite trend is anticipated for larger anticodons. An examination of Fig. 5(a) from low to high values of b reveals that a coding regime occurs at $b \sim 1$, in the transition between two types of homopolymerization. Below $b = 1$, the outcome of polymerization is controlled by k_{cat}; above this value, it is controlled by the relative concentrations. The transition region is sensitive to small differences in kinetic constants precisely because $k_{\text{cat}(1)} \approx k_{-(I)}$. The existence of a correlation in the genetic code reflecting a dependence between these two kinetic constants (Fig. 3 in ref. [5]) supports our analysis, and allows us to connects $b \sim 1$ with 3-nt anticodons, this size being structurally associated with 7-nt loops [17].

An elementary form of the genetic code

The above analysis shows that $\Sigma \sim 0.8$ is achieved with the two categories {Ala, Gly} and {Val, Asp} when $c = 10^{-2}$. Since any two codons of the existing genetic code can be characterized by c values much closer to 1 (implying a random polymerization), it can be concluded that an initial coding system with already 64 codons and 20 amino acid could not work. For several reasons, it has been proposed that the original set of anticodons and codons was limited to $^{5'}GNC^{3'}$, where N is U, C, G or A [18–20]. Remarkably, these four codons encode the above four amino acids, the two categories {Ala, Gly} and {Val, Asp} being associated resp. with {GCC, GGC} and {GUC, GAC}. Both codons in each category display identical anticodon-codon ΔG_0 estimates, and the $\Delta\Delta G_0$ between the two categories is ~1.9 kcal mol^{-1} [5], implying a c value of ~0.045.

Fig. 5(b) shows the level of coding Σ as a function of c with $a = 10^2$ and $d = 10^{-1}$. In this configuration, the maximal value that Σ can theoretically reach ($c \rightarrow 0$) is $da/(da+1) \approx 0.91$. The plot shows that Σ is sensitive to small variations of c within the region 1–0.01 (i.e. $ac \geq 1$), where it rapidly increases as c decreases. As for the "GNC system" ($c = 0.045$), Σ reaches a value as high as 0.77. Below 0.01, c rapidly falls to 0 as Σ tends to its maximal value.

The above analysis still leaves this coding system of four amino acids and four codons with an issue of degeneracy. Considering for instance {Val, Asp} and {GUC, GAC}, one should ask whether a mechanism could specifically assign Val to GUC and Asp to GAC, as it occurs in the modern genetic code. It can be noticed that in each of the two amino acids categories, one amino acid is hydrophobic (Val, Ala) while the other one is hydrophilic (Asp, Gly). This suggests the possibility of a discrimination during the loading of the amino acids on the tRNAs. This loading indeed necessarily implies an intermolecular association, which is usually strongly conditioned by that type of property [21]. The first two chemical steps (activation and aminoacylation) may thus contribute to a reduction of the mentioned degeneracy, a property which has been suggested earlier [5].

Perturbation of the coding regime by mismatches

This Section discusses briefly the issue of mismatches. Let us consider the possibility that codon I is "read" by tRNA II and codon II is "read" by tRNA I. In Fig. 1(b), this implies that a red
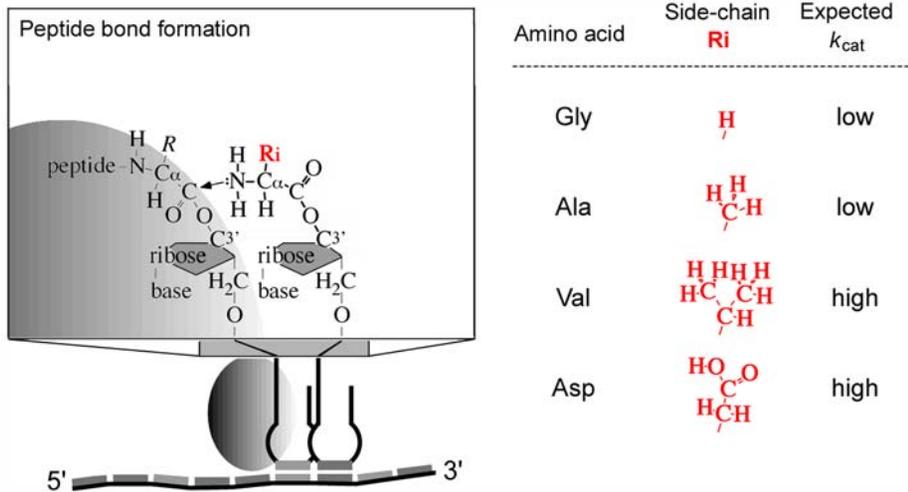
Figure 4. Amino acids side-chains and kinetics of peptide bond formation. Elementary translation and expected effect of the side-chains of alanine, glycine, aspartic acid and valine on the kinetics of peptide bond formation (k_{cat}) (discussed in the text).
doi:10.1371/journal.pone.0005773.g004

anticodon may also bind a blue codon and vice-versa. These interactions are called "mismatches".

Let us assume that the association rate constant (k_+) of these mismatches is identical to the one of complementary anticodon-codon interactions. The corresponding dissociation rate constant ($k_{-(\text{mm})}$) will usually be much higher than $k_{-(\text{II})}$ (the highest of the two dissociation rate constants of complementary interactions). Experimental data about the G·U "wobble" base-pair however suggest that these two constants could also be similar [22].

In addition to a, b, c, and d previously defined, let us define an additional parameter: $e = k_{-(\text{II})}/k_{-(\text{mm})}$. Let us now rewrite $p(1|\text{I})$ while including the additional terms due to mismatches:

$$p(1|\text{I})^* = \left(\left[tRNA_{(\text{I})}\text{aa}_1\right] \frac{k_{\text{cat}(1)}}{k_{-(\text{I})}+k_{\text{cat}(1)}} + \right.$$

$$\left. \left[tRNA_{(\text{II})}\text{aa}_1\right] \frac{k_{\text{cat}(1)}}{k_{-(\text{mm})}+k_{\text{cat}(1)}} \right) \frac{1}{\alpha}$$

$$\alpha = \left(\left[tRNA_{(\text{I})}\text{aa}_1\right] \frac{k_{\text{cat}(1)}}{k_{-(\text{I})}+k_{\text{cat}(1)}} + \right. \qquad (1')$$

$$\left[tRNA_{(\text{II})}\text{aa}_1\right] \frac{k_{\text{cat}(1)}}{k_{-(\text{mm})}+k_{\text{cat}(1)}} + \left[tRNA_{(\text{I})}\text{aa}_2\right] \frac{k_{\text{cat}(2)}}{k_{-(\text{I})}+k_{\text{cat}(2)}}$$

$$\left. + \left[tRNA_{(\text{II})}\text{aa}_2\right] \frac{k_{\text{cat}(2)}}{k_{-(\text{mm})}+k_{\text{cat}(2)}} \right)$$

Let us consider a system in which $[tRNA_{(\text{I})}] = [tRNA_{(\text{II})}]$ (i.e. the number of copies of the two tRNAs are identical). Following the hypothesis about the (non-selective) self-aminoacylation process described above, one has, therefore,

$$\frac{\left[tRNA_{(\text{II})}\text{aa}_1\right]}{\left[tRNA_{(\text{II})}\text{aa}_2\right]} = \frac{\left[tRNA_{(\text{I})}\text{aa}_1\right]}{\left[tRNA_{(\text{I})}\text{aa}_2\right]} = \frac{[\text{aa}_1]}{[\text{aa}_2]} = \frac{1}{d}$$

and

$$\left[tRNA_{(\text{I})}\text{aa}_1\right] = \left[tRNA_{(\text{II})}\text{aa}_1\right],$$
$$\left[tRNA_{(\text{I})}\text{aa}_2\right] = \left[tRNA_{(\text{II})}\text{aa}_2\right].$$

Rewriting expression (1') with the parameters a, b, c, d, e, one gets

$$p(1|\text{I})^* =$$
$$\left(\frac{b}{1+b} + \frac{bce}{1+bce} \right) \Big/ \left(\frac{b}{1+b} + \frac{bce}{1+bce} + \frac{abd}{1+ab} + \frac{abcde}{1+abce} \right) \quad (2')$$

Similarly, one gets for $p(2|\text{II})^*$:

$$p(2|\text{II})^* = \left(\frac{abcd}{1+abc} + \frac{abcde}{1+abce} \right) \Big/$$
$$\left(\frac{abcd}{1+abc} + \frac{abcde}{1+abce} + \frac{bc}{1+bc} + \frac{bce}{1+bce} \right) \quad (3')$$

It can be verified that when $e = 0$ (i.e. the dissociation rate constant of the mismatches has a very large value), one gets the initial relations (2) and (3).

Since relations (2') and (3') cannot be further simplified, it is not straightforward to analyze Σ^* with the new expressions $p(1|\text{I})^*$ and $p(2|\text{II})^*$. However, Σ^* can be numerically examined, and it turns out that the results are rather similar to those for Σ as long as $e \leq 1$.

Fig. 6 is similar to Fig. 5(a), but it includes de perturbation introduced by e. It can be seen that even when $k_{-(\text{mm})} = k_{-(\text{II})}$ (i.e. $e = 1$), Σ^* is not dramatically different from Σ, and when $k_{-(\text{mm})} = 10 \ k_{-(\text{II})}$ (i.e. $e = 0.1$), the effect of mismatches on Σ^* becomes totally negligible. Another way to examine the effect of the perturbation is to keep b set to 1 (with $ac = 1$), and establish the difference $|p(1|\text{I})* - p(2|\text{II})*|$ (Table 1).

Discussion

Although the molecular organization of genetic code is now known in detail, there is still no agreement on the reason(s) for which it has emerged. Early studies have shown that the codon table is highly structured with respect to amino acids hydrophobicity properties, suggesting that basic physico-chemical considerations could contain the solution to this problem [21,23–25]. More recent works have shown that this table is ordered with respect to
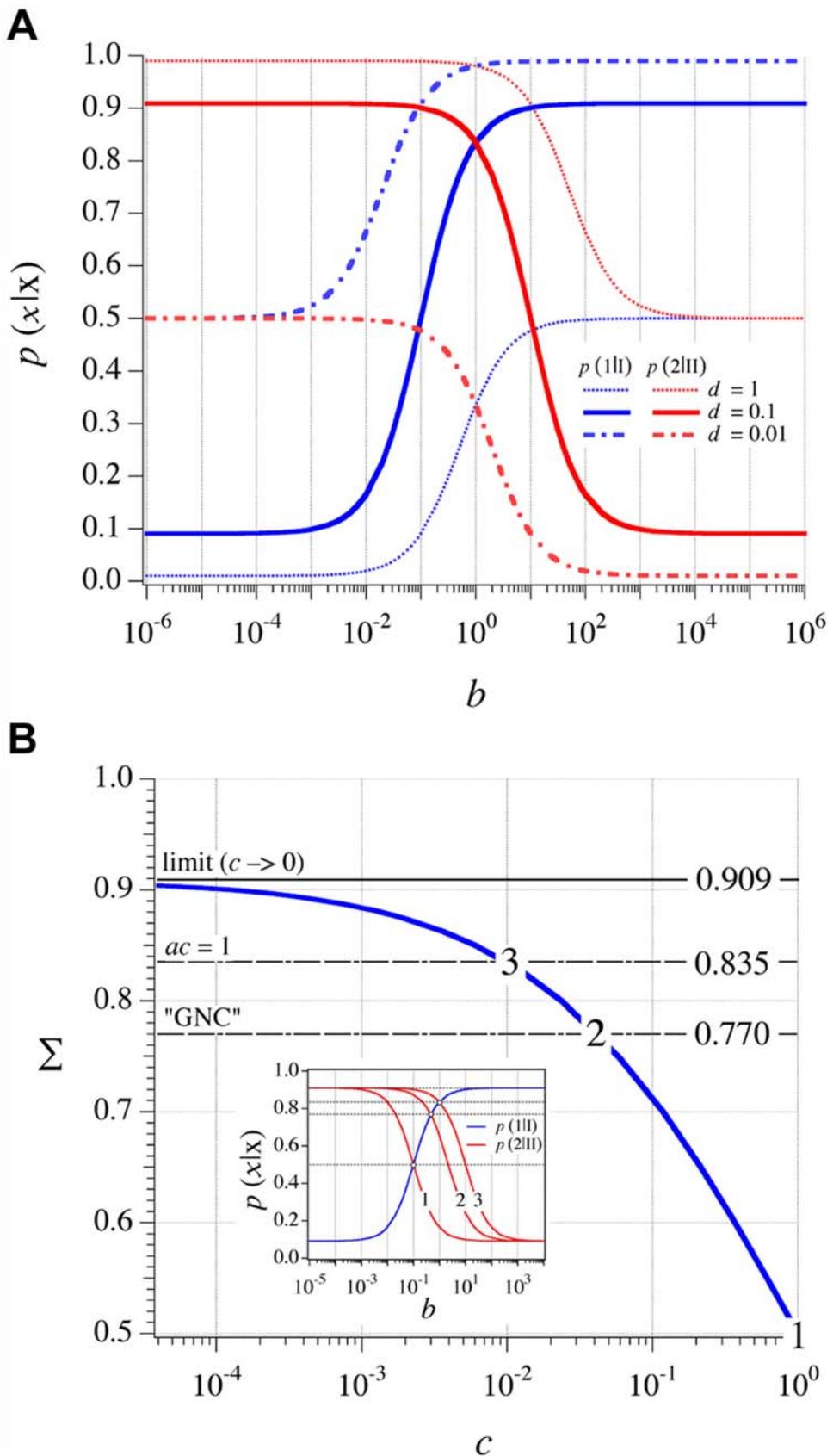
Figure 5. Regimes of the polymerization process. (a) Effect of the value of the ratio $k_{cat(1)}/k_{-(I)}$ on the polymerization process: $p(1|I)$ and $p(2|II)$ as a function of b for three significant values of d when $a = 10^2$ and $c = 10^{-2}$. When $d = 1$, either homopolymerization of aa_2 ($b < 1$), or random polymerization ($b \gg 1$) are observed. When $d = 0.1$, a coding regime $\Sigma = 0.835$ is observed at the transition between two types of homopolymerization,

which occurs at $b=1$. When $d=0.01$, either random polymerization ($b\ll 1$) or homopolymerization of aa_1 ($b>1$) are observed. (b) Level of coding Σ as a function of c when $a=10^2$ and $d=10^{-1}$. Three significant values of Σ are indicated. The inset shows a plot similar to (a) for three particular values of c (reported from the main graph).
doi:10.1371/journal.pone.0005773.g005

features of the aminoacyl-tRNA synthetases [26] and the tRNAs [27,28]. For instance, the mechanisms of aminoacylation as well as identity elements on the tRNAs are specific to certain groups of codons. Although these facts are fundamental, and have inspired scenarios for the evolution and the expansion of the code [26–28], evolutionary considerations may not, in essence, provide an answer to the origin of the code (since it is a prerequisite for biological evolution).

The present analysis shows that behind (the origin of) the code lies a problem of polymerization catalysis: how could different types of monomers (the amino acids) be involved in a same polymerization process? Whatever the exact operating mechanism(s), a single (or uniform) catalyst usually favors only one particular substrate. The four nucleotides A, G, C and U of RNA can generate conditions for the polymerization of different amino acids. In the elementary translation described here, these
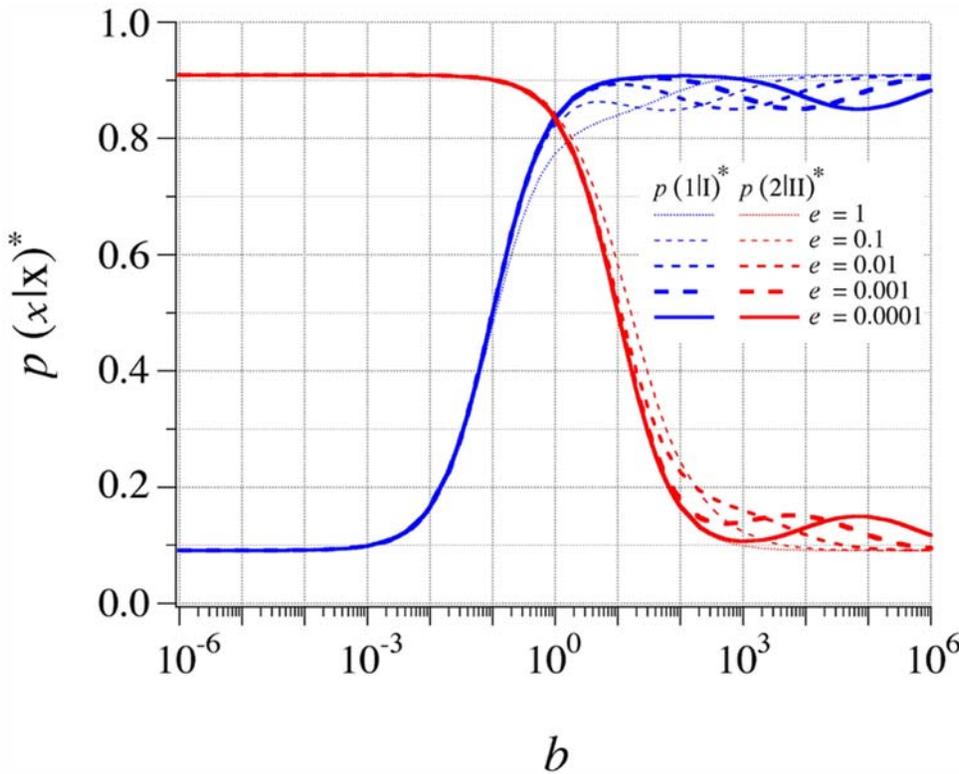

Figure 6. Effect of mismatches on the polymerization process. Probabilities $p(1|I)^*$ and $p(2|II)^*$ as a function of b for five values of $e \leq 1$ ($a=10^2$; $b=1$; $c=10^{-2}$; $d=0.1$). This diagram shows that the intersection between $p(1|I)^*$ and $p(2|II)^*$ (which defines the coding regime Σ^*) is only slightly affected by mismatches, as long as the dissociation constant of these mismatches ($k_{-(mm)}$) remains equal ($e=1$) or higher ($e<1$) than the highest dissociation constant of the complementary matches ($k_{-(II)}$).
doi:10.1371/journal.pone.0005773.g006

Table 1. Perturbation of the coding regime by mismatches.

e	$\Delta\Delta G_0$ (kcal mol^{-1})	$p(1|I)^*$	$p(2|II)^*$	$|p(1|I)^* - p(2|II)^*|$
1	0	0.7739	0.8347	0.061
0.7226	0.2	0.7825	0.8434	0.061
0.1	1.417	0.8225	0.8443	0.022
0.01	2.834	0.8334	0.8360	0.003
0	very large	0.8347	0.8347	0

Numerical results with $a=10^2$, $b=1$; $c=10^{-2}$, $d=10^{-1}$ (for which $\Sigma \approx 0.8347$). The $\Delta\Delta G_0$ indicated is the free-energy difference between the weakest complementary interaction (codon II) and the mismatch. A value of $e=0.7226$ implies a $\Delta\Delta G_0$ of 0.2 kcal mol^{-1}, which is the approximate difference between AU (Watson-Crick) and GU (Wobble) base-pairs [20].
doi:10.1371/journal.pone.0005773.t001

conditions are given by the kinetics of anticodon-codon interactions.

Among the identified requirements to set this polymerization process in the coding regime are the relative frequencies of the amino acids: the small amino acids (glycine and alanine) must be more abundant than the large ones. Although the result of Miller's experiment shown here (Fig. 2) is only indicative, it likely reflects a robust general trend which originates form the fact that complex amino acids require more chemical steps (and more energy) to be synthesized than simple ones. They are therefore expected to be less abundant, whatever the exact conditions of the environment. Our analysis thus integrates a frequency distribution which appears to be rather fundamental.

One should consider the issue of the initiation of protein synthesis in the system described here. This step is critical since a small amino acid may only weakly stabilize the initial tRNA on the ribosome cofactor. Large hydrophobic amino acids such as Leucine or Isoleucine are possible candidates since they are found in prebiotic synthesis experiments (Fig. 2). Also, the ester bond connecting these amino acids to the 3′ end of the tRNA is less prone to hydrolysis as compare with other amino acids [29], which might be critical for initiation. Another possibility is that a dipeptide already present at the 3′ end of the first tRNA [3] may help initiate translation.

In conclusion, our results show that the properties of amino acids and RNA can naturally impose a partially coded polymerization along RNA templates. We also found that the associated coding mechanism is remarkably robust against mismatches. When supplied with "meaningful" RNA sequences, translation systems of this kind should be capable of generating pools of proteins a small fraction of which will be functional. The feed-back action of these proteins on the translation itself may further increase its efficiency, allowing more codons to be added to its repertoire. In this evolutionary perspective, it can be speculated that a critical effect of emerging synthetases will be to establish only the [amino acid – tRNA] configurations that are fit for translation, a "learning" action that RNA alone cannot logically achieve.

Author Contributions

References

1. Berg P, Bergmann FH, Ofengand EJ, Dieckmann M (1961) Enzymic synthesis of amino acyl derivatives of ribonucleic acid. J Biol Chem 236: 1726–1734.
2. Gromadski KB, Rodnina MV (2004) Kinetic determinants of high-fidelity tRNA discrimination on the ribosome. Mol Cell 13: 191–200.
3. Illangasekare M, Yarus M (1999) A tiny RNA that catalyzes both aminoacyl-RNA and peptidyl-RNA synthesis. RNA 5: 1482–1489.
4. Lehmann J, Reichel A, Buguin A, Libchaber A (2007) Efficiency of a self-aminoacylating ribozyme: Effect of the length and base-composition of its 3′ extension. RNA 13: 1191–1197.
5. Lehmann J (2000) Physico-chemical constraints connected with the coding properties of the genetic system. J theor Biol 202: 129–144.
6. Grosjean HJ, De Henau S, Crothers DM (1978) Physical basis for ambiguity in genetic coding interactions. Proc Natl Acad Sci U S A 75: 610–614.
7. Sievers A, Beringer M, Rodnina MV, Wolfenden R (2004) The ribosome as an entropy trap. Proc Natl Acad Sci U S A 101: 7897–7901.
8. Ledoux S, Uhlenbeck OC (2008) Different aa-tRNAs are selected uniformly on the ribosome. Mol Cell 31: 114–123.
9. Ogle JM, Brodersen DE, Clemons WM, Tarry MJ, Carter AP, et al. (2001) Recognition of cognate transfer RNA by the 30S ribosomal subunit. Science 292: 897–902.
10. Agmon F, Bashan A, Zarivach R, Yonath A (2005) Symmetry at the active site of the ribosome: structural and functional implications. Biol Chem 386: 833–844.
11. Miller SL (1987) Which organic compounds could have occurred on the prebiotic earth. Cold Spring Harbor Symposia on Quantitative Biology LII: 17–27.
12. Lightstone FC, Bruice TC (1996) Ground state conformations and entropic and enthalpic factors in the efficiency of intramolecular and enzymatic reactions. 1. Cyclic anhydride formation by substituted glutarates, succinate, and 3,6-endoxo-Δ^4-tetrahydrophthalate monophenyl esters. J Am Chem Soc 118: 2595–2605.
13. Armstrong AA, Amzel LM (2003) Role of entropy in increased rates of intramolecular reactions. J Am Chem Soc 125: 14596–14602.
14. Jung ME, Piizzi G (2005) Gem-Disubstituent effect: Theoretical basis and synthetic applications. Chem Rev 105: 1735–1766.
15. Agris PF (2004) Decoding the genome: a modified view. Nucl Ac Res 32: 223–238.
16. Olejniczak M, Dale T, Fahlman RP, Uhlenbeck OC (2005) Idiosyncratic tuning of tRNAs to achieve uniform ribosome binding. Nat Str Mol Biol 12: 788–793.
17. Auffinger P, Westhof E (2001) An extended structural signature for the tRNA anticodon loop. RNA 7: 334–342.
18. Eigen M, Schuster P (1978) Hypercycle – principle of natural self-organization. C. Realistic hypercycle. Naturwissenschaften 65: 341–369.
19. Ikehara K, Omori Y, Arai R, Hirose A (2002) A novel theory on the origin of the genetic code: A GNC-SNS hypothesis. J Mol Evol 54: 530–538.
20. Lehmann J, Riedo B, Dietler G (2004) Folding of small RNAs displaying the GNC base-pattern: implications for the self-organization of the genetic system. J theor Biol 227: 381–395.
21. Lacey JC, Wickramasinghe NSMD, Cook GW (1992) Experimental studies on the origin of the genetic code and the process of protein synthesis – a review update. Origins Life Evol Biosphere 22: 243–275.
22. Turner DH, Sugimoto N, Jaeger JA, Longfellow CE, Freier SM, et al. (1987) Improved parameters for prediction of RNA structure. Cold Spring Harbor Symposia on Quantitative Biology 52: 123–133.
23. Woese CR (1965) Order in the genetic code. Proc Natl Acad Sci U S A 54: 71–75.
24. Woese CR, Dugre DH, Saxinger WC, Dugre SA (1968) The molecular basis for the genetic code. Proc Natl Acad Sci U S A 55: 966–974.
25. Lacey JC, Mullins DW (1983) Experimental studies related to the origin of the genetic code and the process of protein synthesis – a review. Origins Life Evol Biosphere 13: 3–42.
26. Schimmel P, Giege R, Moras D, Yokoyama S (1993) An operational RNA code for amino acids and possible relation to genetic code. Proc Natl Acad Sci U S A 90: 8763–8768.
27. Delarue M (2007) An asymmetric underlying rule in the assignment of codons: possible clue to a quick early evolution of the genetic code via successive binary choices. RNA 13: 1–9.
28. Rodin SN, Rodin AS (2008) On the origin of the genetic code: signatures of its primordial complementarity in tRNAs and aminoacyltRNA synthetases. Heredity 100: 341–355.
29. Matthaei JH, Voigt HP, Heller G, Neth R, Schöch G, et al. (1966) Specific interactions of ribosomes in decoding. Cold Spring Harbor Symposia on Quantitative Biology 31: 25–38.