Guimarães RC, Moreira CHC, Farias ST 2008 Self-referential formation of the genetic system. In: *The Codes of Life – the rules of macroevolution*, Ed. Marcello Barbieri, Springer, Dordrecht, The Netherlands. Series *Biosemiotics*, V. 1, 436 pg, Chapter 5, pg. 69-110, ISBN 978-1-4020-6339-8.

# SELF-REFERENTIAL FORMATION OF THE GENETIC SYSTEM

*Romeu Cardoso Guimarães*[1] *, Carlos Henrique Costa Moreira*[2] *, Sávio Torres de Farias*[1]

[1] Dept. Biologia Geral, Inst. Ciências Biológicas, Univ. Federal de Minas Gerais, 31270.901 Belo Horizonte MG Brasil

[2] Dept. Matemática, Inst. Ciências Exatas.

**Correspondence**: RCG, Tel +55-31-3499.2608, Fax +55-31-3499.2570, e-mail romeucg@icb.ufmg.br

**Keywords**: Early proteins; Gene definition; Genetic code; Palindromic tRNAs; Self-reference.

**Abstract**

Formation of the genetic code is considered a part of the process of establishing precise nucleoprotein associations. The process is initiated by tRNA dimers paired through the perfect palindromic anticodons, which are at the same time codons for each other; the amino acid acceptor ends produce the transferase function, in a manner similar to the reaction occurring in ribosomes. The connections between nucleic acids and proteins are bidirectional, forming a self-feeding system. In one direction, proteins that are resistant to degradation and efficient RNA-binders stabilize the tRNAs that are specifically involved with their production; in the other direction, these tRNAs become fixed with the correspondences which are the amino acid codes. Replication of the stabilized tRNAs becomes elongational, forming poly-tRNAs, the precursors of the mRNA strings (genes), and of ribosomes. The linear order in the gene sequences follows the temporal succession of the encoding of tRNA pairs. The whole encoding process is oriented by the tRNA pairs. The core sequence of proteins shows the predominant aperiodic conformation and the anticodonic principal dinucleotides (pDiN) are composed of two purines or two pyrimidines: (1a) Gly / Pro; (1b) Ser / Ser; (2a) Asp, Glu / Leu; (2b) Asn, Lys / Phe. Members of the following pairs, with pDiN composed of a purine and a pyrimidine [(3a) Ala / Arg; (3b) Val / His, Gln; (3c) Thr / Cys, Trp; (4) Ile, Met / Tyr, and iMet / Stop], are added, respectively, to the mRNA heads / tails. It is indicated that: (a) The Last Universal Common Ancestor populations could, at some early stages, be composed of lineages bearing similar genetic codes, due to the simple and highly deterministic character of the process; (b) Genetic information was created during the process of formation of the coding/decoding subsystem, inside a proto-metabolic system already producing some amino acids and tRNA-like precursors; (c) Genes were defined by the proteins that stabilized the system, as memories for their production.

## TABLE OF CONTENTS

# 1 Introduction

Cellular physiology is didactically described in the top-down tradition, from genes to the phenotype (Alberts et al. 2002). This description is similar to that of man-made factories, with designers or programmers – the genes, and workers – the RNAs and proteins that build the phenotype. The Darwinian thought placed a directing agency in the interactions between organisms and the environmental contexts but the main focus remained in the genetic variants that allowed some phenotypes to survive the interactions while others were selected out. We remain with the task of explaining how such a genetic system first came into being. An apparent paradox arises, of how could the planners start their enterprise just gathering together into the factory some workers-to-be that never experienced their jobs, which would be similar to slave-hunting practices. A bottom-up perspective is taken is this work, devising a mechanism for the spontaneous origin of the biotic system, inside which the genetic processes arose. The spontaneity of processes may be categorized as self-organizational and genes are considered the memory part of the system. Sections 1-2 present some basic concepts about the living system components, relevant to understanding the genetic code. The model is presented in Sections 3-6, with some technical details of the process of formation of the coding/decoding system. The derived conceptual implications comprise Section 7.

# 2 The biotic world

## *2.1 Strings and folding*

The basic constituents of the living system are nucleic acids and proteins so that cells are called nucleoprotein systems. These ‘noblest’ components are strings, polymers where the monomers are respectively nucleotides and amino acids. The nobility in these strings comes from the specificity in the order of the monomers in the sequences, which bring about the main functions of the system. Amino acids are fundamentally twenty, at the lower range of the number of letters in the symbolic alphabets of written languages.

The chains may be very long and are intricately folded to acquire specific spatial configurations. The plasticity of the tri-dimensional arrangements is a consequence of the string constitution and is a main factor in the regulatory and adaptive behavior of the entities. The main influences in the acquisition of the spatial architectures are the sequences in the polymers – the primary structure – and their interactions with water (Alberts et al. 2002). The folding is obtained through weak associative bonds, guaranteeing the dynamic character of the structures. In nucleic acids, the main structuring rule is the base-pairing that builds the double helices; in RNA, single-stranded loops are frequent. The secondary structures are more varied in proteins, due to the small size and the diversity of amino acids. The complexity of the sequence organization required to guarantee the protein conformations is ordered from the simplest (aperiodic) – coils, where most of the interactions of the amino acids are directed towards the environment, to the turns, with short-range interactions among the neighbor amino acids, then to the helices, with more regular interactions along a segment of the chain, and to the strands that, in order to form sheets, have to produce compatible segments in distant portions of the chains.

## *2.2 Hydropathy and cohesiveness*

The reactivity of molecules towards water, the most abundant constituent of cells, is designated hydropathy. Hydrophilic parts of molecules mix well with water and the hydrophobic parts repel or are repelled by water (Kyte and Doolittle 1982). Nucleotides are

amphiphilic, containing the hydrophilic sugars and phosphates and the more hydrophobic nucleobases. Amino acids span the whole hydropathy range so that proteins depict complex organization. Protein aggregates tend to form globules with membrane-like surfaces, for which the contribution of lipids may be accessory. The same principles governing the internal organization and folding of the macromolecules apply to the formation of associations between proteins or between proteins and nucleic acids. The nucleoprotein system is tightly associated: proteins and nucleic acids are both very sticky and form aggregates, the components communicating with each other in almost contiguity.

### *2.3 Networks and stability*

The cellular components form a large communication network (Barabási and Oltvai 2004). The transformations involving both the macro- and micro-molecules are of the same kind, justifying the extension of the term metabolism to the whole network and the description of the cell as a metabolic system (Guimarães et al. 2007 and references therein). In spite of so many interesting attributes that support the proposition of metabolic networks as original in forming living systems (Kauffman 1993) there is an enormous difficulty in modeling or obtaining them experimentally. Polymers with replication abilities are considered a necessary pre-requisite for the origin of bio-systems (Orgel 2002).

Chemistry and especially biochemistry study mainly stable objects. Stability of a molecule means also its abundance and this is a driving force in chemical reactions, together with the energetic gradients. The constituents of cells, however, are not particularly stable. The synthesis of a peptide bond or of an ester bond is accompanied by the formation of water molecules but, in the aqueous environment, the polymers are continually being forced into the reverse reaction, of de-polymerization through hydrolysis. So, a characteristic of the biologic system is the requirement for continuous replacement of the damaged or decaying polymers. The consumption of components creates another driving force on metabolism, accentuating the mass gradient in the anabolic direction. The driving force of consumption of products is considered more important for the organization of metabolism than the availability of precursors or substrates, in the sense of being pathway-specific, and the latter are categorized mainly as prerequisites for the anabolic reactions.

Only the DNA molecule can be efficiently repaired (see Zenkin et al. 2006, for the transcriptional repair) and this is the reason for its widespread double-strandedness. RNA is less stable than DNA due to its partial base-pairing and to the presence of two hydroxyls in the sugar. When RNAs are more stable, this character is partially obtained through nucleotide modifications but mainly through the protection provided by associated proteins that may be intrinsically stable. The need for re-synthesis of sequences makes memory structures necessary; these are provided by the nucleic acids. A template RNA may serve the translation of various copies of proteins, and a segment of DNA or RNA the replication, transcription or reverse transcription of various copies of their complements. Cells could survive and evolve only when they acquired the capacity to stabilize the templates and to re-synthesize polymers in excess of the degradation rates.

### *2.4 The ribonucleoprotein (RNP) world and pre-biotic chemistry*

Studies on cellular origins became more complicated when it was realized that DNA probably was not an early molecule. Deoxyribonucleotides are derived from ribonucleotides and DNA is vastly less reactive than RNA, not able to participate efficiently in metabolic reactions (Orgel 2002). To the contrary, cellular RNAs are known that can catalyze metabolic transformations – ribozymes, the most remarkable being the ribosomal peptidyl transferase activity (Yusupov et al. 2001). There is *in vitro* evidence even for self-replicating RNAs (Hughes et al. 2004). The context outlined is of an early RNP living world, that later became

enriched with DNA, therefore acquiring a stable memory and separating this function from the strictly metabolic ones, but the hypothesis of the early existence of a pure RNA world does not seem plausible. At times when pre-biotic ribozymes may have been abundant so should have been the pre-biotic peptides and protein-type catalysts.

The pre-biotic chemical scenario points to the abundance of some amino acids and scarcity of nucleotides. The latter are complex molecules while amino acids are small. The most abundant amino acids recovered from syntheses under plausible early Earth conditions are among the simplest, Gly and Ala (Miller and Lazcano 2002). The easiest way to obtain peptidic and nucleotidic polymers so far advocated, albeit with short sizes and low productivity is based on mineral surfaces, especially from clays (Cairns-Smith 1982). The plausible scenario is that a mineral order can be transferred to the polymers and these could be more easily adjusted to one another in formation of aggregates since they were based on a common origin (Basiuk and Sainz-Rojas 2001). A problem remains with the pre-biotic metabolism, wherefrom a sufficient supply of monomers could be guaranteed. A variant of the 'genetic take-over' hypothesis is acceptable, stating that nucleic acids became later replicated through means independent from the minerals that first guided their polymerization (Ferris 2002). The necessary character of the pre-biotic metabolic system that survived in the form of cells is that it was apt to produce internally the genetic subsystem.

## 3 **The coded biotic world**

In cells, proteins are polymerized in amino acid sequences following the order of the nucleotides in a template RNA. Such strict correspondence could be easily understood if some kind of fit existed between nucleic acid surfaces or pockets and the amino acids. This stereo-chemical hypothesis is not presently in favor for the generation of the whole code but it could have been responsible for some of the correspondences (Yang 2005; Yarus et al. 2005). In fact, the correspondences are mediated by adaptor RNA molecules, the transfer RNAs (tRNA), each one carrying a specific amino acid and bearing a code, a triplet of nucleotides (an anticodon) that matches by base-pairing a triplet in the template mRNA (the codon) (Ibba et al. 2000; Barbieri 2003).

The matrices of codes and anticodes are shown in **Table 1**. The largest matrix of anticodes is the eukaryotic, containing 46 tRNAs. However, such complexity is not necessary for decoding the full set of 64 codons. There are simpler matrices that rely more on the wobbling abilities of 5' G and U than on the retention of all kinds of bases at the 5' position. The bacterial anticodes are smaller, the archaeal intermediate and the vertebrate mitochondrial is the most intensely reduced, to the total of 23 tRNAs, including the iMet (initiator; Osawa 1995). The maximum size of the anticode is due to the absence of base A in the 5´ position and of the three tRNAs corresponding to the Stop codons.

[ENTER **Table 1**]

### *3.1 Hypotheses of early translation*

Attempts at developing a plausible explanation for the formation of the translation system have not been satisfactory so far (Trevors and Abel 2004). A problem with the direct early translation models stems from their assumption that preexisting RNAs served as templates (early mRNAs) for the alignment of aminoacyl-tRNAs (ac-tRNA; or with some early form of the tRNAs) and that this system gradually acquired efficiency and functionality. Following

this premise, based on the traditional evolutionary genetics, the learning process would have been very slow, through trials, errors and some successes that became fixed (Poole et al. 1998; Tamura and Schimmel 2003).

The template could be efficiently translated by an early specific ac-tRNA only if it had a homogeneous or repetitive sequence. This is not an interesting start, but it could be supposed that useful variations would be superimposed on this, if they were coherent in the template and in the new tRNAs; the necessary coherence in the evolution is the main problem. It is more plausible that the early template was heterogeneous but it could be efficiently translated only by a non-specific set of tRNAs, producing the correspondent highly variable proteins. This system would evolve by selection of the interesting and functional proteins and mRNAs, again in coherence. Specific tRNAs would not be adequate for translation of heterogeneous mRNAs due to the frequent occurrence of non-sense segments in the template.

Various attempts have been made to envisage the constitution of an early small set of amino acids that would also fit a coherent region of the genetic code matrix. This set should also ideally correspond to some interesting protein property so that a functional system would be constructed, to become a seed for further evolution. The proposals tend generally to concentrate on and vary somewhat upon the set localized on the 3' C row of the anticodon matrix (Val, Ala, Gly, Asp and Glu; Lehman 2002; Klipcan and Safro 2004).

We will not detail all the argumentation involved in this choice, but most of the important parameters that have been proposed in support of it are also interesting for understanding the code organization and can be summarized as follows:

• The amino acids in the 3' C row of the anticodon matrix belong generally in the category of the simpler and smaller. This produces consistency with some of the most abundant and stable molecules obtained in pre-biotic syntheses.

• None of them are, in the more widely accepted biosynthesis routes of amino acids, derived from others belonging in the code. They are all originated directly, or with a low number of transformations, from the most basic of the metabolic processes, the routes of glycolysis, of the pentose shunt and of the Citrate Cycle. This line of thought states that the early attributions were those of amino acids originated from the basic routes of biosynthesis and that, with complexification of the routes, producing new amino acids in families of derivation, the codes attributed to them also formed families of similarity. So, there was a co-evolution in the derivation of both amino acids and their codes (Wong 2005).

• The amino acid set is heterogeneous enough to produce peptides with a variety of properties, thus opening the way for the ample functionality of proteins (Oba et al. 2005).

• The anticodon set is rich in G+C bases, guaranteeing higher thermal stability in the pairings with the codons, and homogeneous at the 3' C, making easier the mutual transformations inside the set, so that an earlier single starting code would have had less difficulty in producing the others.

• The set of present day catalysts of the formation of the aminoacyl-tRNAs, the aminoacyl-tRNA synthetases (aRS), is rich in class II enzymes, and these are suggested to have generally preceded class I. In our model, only the three last amino acids in this set (Gly, Asp, Glu) are fixed early.

## 4 **The self-referential model**

Our study of the formation of the genetic code (Guimarães et al. 2007, and earlier references therein) allowed the proposition of a procedure for obtaining synthesis of proteins that is considered plausible and consistent with the present day knowledge on cellular processes, and can dispose of the requirement for an early mRNA to be translated. It is called self-referential

to the tRNAs and based on a simple type of symmetry produced by the dimerization of tRNAs through the complementary anticodons (Grosjean et al. 1986; Grosjean and Houssier 1990). We will start with a presentation of the model and then give an account of its consistency with the main attributes that have been found adequate and necessary to fit the evolutionary paradigm, e. g., of going from simple to complex arrangements and of providing a driving force or a phenotype with fitness value that can be selected for.

The process is based on a small machine-like system (**Figure 1**). Its simplicity offers a fast and high probability mode of evolution, to fill the matrix in few steps. Such expediency is necessary in the light of the estimated short time span available for the origin of life on Earth (Poole et al. 1998). The slow part of the process is sent back to the origin of replication and of the tRNAs. The simplicity and the highly deterministic character of the process suggest that the Last Universal Common Ancestor populations could, at some early stages, be composed of lineages bearing similar genetic codes. Simulations of the probabilities of the different codes inside these populations are being conducted (Farias et al., in preparation).

[ENTER **Figure 1**]

*4.1 The pools of reactants: tRNAs and amino acids*

The main ingredients of the coding system are of three kinds – the tRNAs, the amino acids and the catalysts of their union, accomplishing the aRS function. This function was initially obtained with the participation of, e. g., the tRNAs and metals, the RNA component being possibly also ribozymic, or prebiotic peptides. The tRNA pool contains the full set of 64 anticodons; replication is able to generate them easily. The early amino acid pool is probably not full, according to the list of amino acids obtained by chemical synthesis under conditions imitating the supposed early Earth conditions (see Trifonov 2004; in the decreasing rank order: GALVDEIPS, plus marginally T) or considered not derived from other amino acids belonging in the code (see Wong 2005; only the rank order is different: GASDEVLI, plus marginally P and T).

*4.2 Stages in the formation of the coding system*

The process is divided into three stages (**Figure 1**): (1) dimer-directed protein synthesis, (2) primitive mRNA translation, and (3) maturation of the mRNA structure. A precise stepwise succession of pairs of tRNAs being recruited and fixed with correspondences, to compose and to be integrated in the system, is delineated.

*4.3* (1) *Protein synthesis directed by tRNA dimers*

Various kinds of dimers may be formed, including the ones not relying upon the pairing through the anticodons (e. g., Martinez-Giménes and Tabarés-Seisdedos 2002), and many of them could be producing oligo- or polypeptides concomitantly. Proteins produced by each type of dimer would be largely repetitive, composed by either one amino acid, such as the Ser, which is coded by complementary pDiN, or up to many, in the cases where each tRNA in the dimer would demonstrate greater affinity for a different amino acid and lower affinities for others.

Our model says that the dimers formed by pairing of tRNAs through the anticodons of the perfect palindromic kind were those demonstrating higher stability, therefore more apt for the task of performing the transferase activity (see Guimarães et al. 2007). If all of them were at work concomitantly upon the early set of amino acids, the model needs to invoke a self-

stimulatory molecular selection process mediated by the proteins, choosing which ones would be fixed at each moment of the formation of the code.

The properties of the proteins produced by the pool of dimers would be able to explore the whole space of possibilities opened by the types of amino acids available and the semi-repetitiveness of each of them. The earliest self-stimulatory process relevant for the start of the RNP-based protein synthesis would depend on a few of these properties, the main one being the ability of proteins to bind to the RNAs. When the RNA-binding protein is also stable against degradation, it would protect the RNA and form a stable RNP. The RNA component should continue being adequate for replication and for participating in the synthetase and transferase reactions.

The set of correspondences is this stage is the GPS group of amino acids, attributed to the NCC : NGG and the NGA : NCU sets of triplets. It is noteworthy that there should be no need for external compartmentalization to facilitate the nucleoprotein binding when the nascent proteins are composed of amino acids with strong RNA-binding properties; they are not released and stay bound to the tRNAs, with immediate cohesiveness.

### *4.4* (2) *Translation of primitive mRNA and the triplet coding*

This is the most complex stage, at the transition from the dimer-directed to the translational mode of protein synthesis, with separation of the anticodon and codon functions. The set of nine amino acids corresponding to the pDiN of the homogeneous sector form the core structure of proteins with predominantly aperiodic conformation: five of the amino acids are characteristic of coils and turns (GPSDN), three of α-helices (DLK). Ribosomes and mRNAs are derived from poly-tRNA strings, which may be elongated after the RNA stabilization is guaranteed. Evidence indicating the derivation of rRNA from tRNA has been presented by Bloch et al. (1984, 1989). The synthetase function of proteins is developed, accompanied by the establishment of the hydropathy correlation (Farias et al. 2007).

The change from dimer-directed to templated protein synthesis introduced modifications in the way the stability of the tRNA couples active in the transferase function is obtained and helped in the restriction of the coding to the anticodon triplets. In the paired anticodon configuration, the base pairs in the loops holding the tRNAs together should point to one surface with the anticodons in the middle and other bases flanking the triplets might be involved in the pairing (Grosjean and Houssier 1990). In the ribosomal side by side configuration, the anticodon : codon pairing becomes restricted to the three bases at the center of the loop, and the flanking bases could pair laterally with the neighbor tRNA (see Smith and Yarus 1989).

### *4.5* (3) *Maturation of the generalized mRNA structure*

The tRNAs belonging to the mixed pDiN sector are integrated into the system next. Dimers do not penetrate ribosomes but dictate that new attributions should follow the pairing of tRNAs carrying them. The NRY attributions are ligated to the 5’ ends of mRNAs and the NYR to the 3’ ends, forming the generalized mRNA structure 5’ (NRY) (NRR, NYY) (NYR) 3’ (**Tables 2 and 3**). It is indicated that the codes corresponding to the tRNAs of the homogeneous pDiN sector are ligated *in tandem*, following the succession of the stages, and that the staggered mode of addition of the codes corresponding to the tRNAs of the mixed pDiN sector is due to their belonging already in the mRNA or DNA mechanisms.

[ENTER **Table 2**]

The chronology proposed for the fixation of the genetic code correspondences obeys the rules that the tRNAs were recruited in palindromic pairs and that the tRNAs with homogeneous pDiN were fixed earlier than the tRNAs with mixed pDiN. It is possible that some kind of pre-coding dictated by direct interactions between amino acids and oligonucleotides, belonging in the RNA world, might be relevant to the process, but we cannot find any link between such proposals (e. g., Seligmann and Amzallag 2002) and the standard code. They are based mostly in thermodynamic considerations, possible stereochemical affinities, specific amino acid binding to some RNAs not tRNA-like but containing segments similar to their correspondent codons or anticodons (Yarus et al. 2005), or models for possible ribozyme activities that could have generated the amino acid transformations (Szathmáry 1999; Copley et al. 2005).

We also find it difficult to envisage an RNA-based mechanism that would have distinguished and directed the early fixations to the tRNAs with homogeneous pDiN and the late ones to the tRNAs with mixed pDiN. It is more plausible to admit, based mainly on the hydropathy correlation being established in dependence of protein properties (see below), that the correspondences were fixed as codes by the self-stimulatory effects of the protein binding and stabilization of the RNAs that were producing them. It is suggested that the early tRNAs of the homogeneous pDiN sector had more repetitive and simpler sequences than those of the mixed pDiN sector, and were then more adequate for being bound to the early repetitive and simpler proteins produced in the dimer-directed stage. The rationale based on the simplicity of the interacting partners is also pointed at by (a) the simple sequence character of the palindromic triplets (lateral bases repeated), as compared to the other types, and (b) the non-directionality that they allow for the interactions (the lateral bases may be equally the start or the endpoints in an interaction).

On the amino acid side, simplicity has always been a main theme in guiding the proposals for the early attributions of the code. These are usually chosen as subsets from the lists of amino acids obtained by chemical synthesis under conditions imitating the supposed early Earth conditions. Considering the first nine amino acids in the ranks shown above, we see that 2/3 of them belong in the sector of homogeneous pDiN, corresponding to the first three pairs of our model (GPSDEL). On the protein side, our prediction is that a core sequence should be composed by the nine amino acids of the homogeneous pDiN sector and should acquire predominantly the aperiodic conformation. This may be considered certified by the data of Sobolevsky and Trifonov (2005, 2006). The amino acid composition of the 21 most abundant types of conserved octamers belonging in the universal loop-and-lock structures of proteins presents the rank order G(DLT)SAP(KR), from which 2/3 belong in the homogeneous pDiN sector (GPSDLK).

[ENTER **Table 3**]

*4.6 The tRNA dimers orient the entire process*

In the next sections, data are presented to demonstrate that the description of the entire genetic code through the dimer-oriented rationale is physiologic. The consistency of the results with biochemical data and evolutionary prescriptions is surprising in the light of the constraints imposed by the necessity of attributions having to be inserted into the system obeying the tRNA pairs, which is considered an argument in favor of the model. Our main novelty concerning the role of tRNA dimers is the proposition of the palindromic dimer-directed protein synthesis in the first stage and that, in later stages, dimers do not enter the ribosomes but orient the entrance of new attributions. This rationale incorporates the earlier studies of

Miller et al. (1981) and Yamane et al. (1981), showing that dimers could be regulatory to protein synthesis, modulating the availability of tRNAs for translation, and of Smith and Yarus (1989), showing that the neighbor tRNAs in the P and A sites of ribosomes interact side by side via the bases lateral to the anticodons in the loop, which is considered a different kind of dimer. The non-ribosomal dimer-directed transferase activity could be experimentally tested, either utilizing present day tRNAs (possibly too large and rigid for facilitating the reaction) or the various kinds of mini-tRNAs that have been used as acceptors for the aRS function or for spontaneous aminoacylation (see Beuning and Musier-Forsyth 1999), but containing anticodon-like loops, able to dimerize.

*4.7 Processes forming the code*

Three mechanisms are discerned in the formation of the codes. Two are primary: one for the attribution of correspondences between triplets and amino acids, the other for the generation of the specific punctuation signs. The mechanisms forming the primary attributions are similar in the sense that they involved the fishing of a second attribution by the first, dictated by the triplet pairing rules. (a) For amino acid coding, anticodonic pairings directed the attribution to the second tRNA that dimerized with the first. (b) For the localization of the termination signs, their tRNAs formed pairs with the initiation codon and competed with the initiation anticodon for this codon, therefore being deleted (see **Table 4**). The specific punctuation was developed after all amino acid attributions were completed, which is consistent with the specific punctuation being the most complex of the translation mechanisms. The third mechanism (c) is called secondary, for the generation of the hexacodonic attributions, derived from expansion of the specificities of the class I ArgRS and LeuRS. The expansions of Arg into the YCU and of Leu into the YAA triplets were established after the formation of their respective primary tetracodonic attributions (NCG and NAG), so that Ser was originally octacodonic (NGA and NCU), and Phe was originally tetracodonic (NAA).

*4.8 Amino acid coding*

It is known that tRNAs can form dimers through the pairing of complementary anticodons. The thermal stability of the dimers is high, equivalent to the formation of about seven base-pairs in the common RNA helices (Grosjean et al. 1986; Grosjean and Houssier 1990). It is indicated that the type of mini-helix formed either has a peculiar stability by itself or receives additional support, either from special base modifications or from other bases in the anticodon loop, besides the anticodon triplets. Thermal stability of dimers of present day tRNAs is also not much influenced by the G+C contents of the triplets, but there are indications that early coding might have been influenced by this character (Ferreira and Cavalcanti 1997), its main feature being that all boxes at the core of the matrix are simple. Our model requires a perfect palindromic topology in the triplet pairs and obeys only partially the thermodynamic stability principle: in each of the sectors, the initial boxes are at the core and the last ones to be filled with attributions are at the tips.

The tRNA dimers are considered proto-ribosomes and proto-mRNAs. In ribosomes, the two tRNAs are front to back in the A and P sites, guided by the codons in the mRNA (**Figure 1**). In the dimers, the anticodon loops are associated head to head through a mini-helix and the anticodons are simultaneously codons for each other, making unnecessary the presence of an external template, and the tRNA acceptor ends hang to different sides. In both cases the two tRNAs are held together in a stable structure so that the acceptor ends are placed in contact and the transferase reaction facilitated. This reaction is driven towards peptide synthesis due to the peptide bond being covalent and the dimer association dynamic, dependent only on hydrogen bonds. Polymerization of peptides works as a sink, providing a suction force to the

system. This force works as long as there are dimers of ac-tRNAs. Dimers of an ac-tRNA and a non-acylated tRNA may be inhibitory when the concentration of the latter is higher. When only one synthetase is available, acylating one tRNA type, the system may be de-repressed when a second synthetase arises with specificity for the other tRNA in the pair. In this way, dimerization is also a driving force for the fixation of catalysts that can acylate the second member of a pair. In this double-catalyst situation, there are difficulties in deciding which came first since the concentrations of the members of the pair will fluctuate and tend to get equilibrated; any decision on which came first relies upon external factors. A limited complementary behavior of the amino acids being recruited into the code is seen, when the amino acids follow the hydropathies of the tRNAs in the pairs.

*4.9 The palindromic triplets and pairs*

The type of pairs found meaningful for initiating the coding process is configured as perfect palindromes, containing the same bases in the extremities, which we shall call the fishing triplets (see Guimarães et al. 2007). Among all types of triplet pairs, the palindromic configuration is the one guaranteeing full and long lasting single-strandedness which is a requisite for their ability to produce stable dimers through the formation of the mini-helices. Considering the restriction of accepting only the perfect standard base pairs, it is indicated that the palindromic coding was developed before the exclusion of A from the 5' position of the triplets. The exclusion of 5' A became necessary in the complex boxes due to its wide wobbling possibilities that would produce ambiguity in their decoding. In our scheme, the mechanism of 5' A exclusion may have been initiated as early as in the stage of incorporation of the two acidic amino acids, sharing the UC box. The remaining 5' G solved the ambiguity problem and was enough for decoding. The later extension of the 5' A exclusion to all boxes may have been due to its benefit to the regulatory mechanisms, when some perfect palindromes (ANA : UNU) yielded to imperfect ones (GNA : UNU), with the consequent acceptance of G : U pairings.

*4.10 Steps in the coding at each box*

After (a) the initial coding of the fishing triplets, the perfect palindromic triplets in each of the paired boxes, the (b) 5' degeneracy was developed and accepted by the synthetases and the ribosomal decoding mechanisms, integrating all triplets in a box (see **Table 5**). This function is accessory to the main one provided by the pDiN. The full wobbling provided by 5' U would be sufficient for decoding all kinds of codons with the same pDiN, as in the mitochondrial anticodes, but the extended anticode of eukaryotes utilized more extensive 5' details. (c) When complex boxes were developed, the first occupier of the box receded to 5' R and conceded 5' Y to the new occupier(s). These concessions follow a consistent rationale and indicate that all 5' Y attributions are late in relation to the 5' R in the respective boxes (see below, Variant codes).

*4.11 Proteins organized the code*

Peptides formed through the mechanism of tRNA dimerization are partially organized. The set of amino acids coming from a tRNA pair have properties correlated with those of the anticodons. The next set in the succession may have characteristics independent from the previous one and a collection of independent pairs will compose a large pool of repetitive peptides each containing one or more types of amino acids. Details of further organizational steps require examination of real proteins and of the genetic code structure, to model a succession of dimers that is relevant to physiology and to the formation of the code. The model also provides a mechanism for the entrance of templates and ribosomes into the system. Various imprints of protein structure and properties, and of the mechanism of protein

synthesis, were detected in the structure of the genetic code, indicating that a part of the general structure of the matrix was configured in dependence of protein properties.

*4.12 Stages indicated by the hydropathy correlation*

The long known correlated distribution of hydropathies of amino acids and of the types of triplets was the first evidence showing that the correspondence between amino acids and triplets is not entirely arbitrary. It was reexamined considering the hydropathies that the amino acids present as residues in proteins (Farias et al. 2007). Previous studies utilized the hydropathies of amino acid molecules in solution and could not offer a rule to understand how the correlation was established and how to explain the deviations (Lacey and Mullins 1983).

Our study showed that nineteen of the attributions conformed to a wide correlation area and four, belonging in the homogeneous pDiN sector, were identified as outliers from the correlation (Gly-CC : Pro-GG and Ser-GA : Ser-CU; the GPS group; **Table 3**). The steepness of the regression line angle of inclination grows from the other attributions in this sector (DELNKF; 43$^{o}$) to the mixed pDiN sector (64$^{o}$). Taking the GPS set as the first amino acids to become encoded (Stage 1), it is indicated that peptides with this simple constitution were not able to produce the correlation. This was established when the peptides had a richer amino acid complement, with the entrance of the other amino acids of the homogeneous sector (Stage 2). When amino acids of the mixed sector entered (Stage 3), completing the full set of encoded amino acids, the correlation became stronger, indicating greater sensitivity of the fitting mechanisms. Demarcation of Stage 4 (Ile, Met / Tyr) was not dependent on the hydropathy correlation data but highlights the installation of the specific punctuation system.

The palindromic pairs are the only types possible inside the set of outliers that can join the two Ser boxes and accommodate the Gly and Pro boxes without ambiguities; this was the first indication of the meaningfulness of the palindromic pairs. The outliers are presently charged by class II aRS (the class of enzymes typically acylating the 3’ OH of the terminal adenosine of tRNAs). The other amino acids in the homogeneous sector (Stage 2) are charged by one couple of aRS class I (the enzymes typically acylating the 2’ OH of the terminal adenosine of tRNAs; the pair of boxes with Glu and Leu, together with AspRS class II), and the couple of atypical synthetases (the pair of boxes with Lys and Phe, together with AsnRS class II). The end result of the aRS class distribution in the homogeneous sector is five class II, the two atypical and two class I, in contrast with the mixed sector, where there are three class II and eight class I. The more hydropathy-sensitive character of the mixed sector is related to the enrichment in class I enzymes and may derive from their mode of docking on the tRNA acceptor stem (see below, aRS class characterization).

The correlation was established by the protein catalysts but each sector of tRNAs produced a characteristic high affinity and specificity, whose superposition formed the wider correlation area. The best fit produced was with class I and the triplets of intermediate hydropathies (mixed pDiN); the correlation established by class II enzymes does not discriminate pDiN types. Both sectors exploited the full hydropathy range but the homogeneous more the hydrophilics than the hydrophobics; the mixed sector contains all the moderately hydrophobic and most of the hydrophobics. Hydroapathetics are all (GPSTH) typical class II. Hydrophobics are all acylated at the 2’ position of the ribose: Cys plus all at the central A column (class I and the atypical PheRS class II).

It is indicated that when the catalysts unite two substrates in a product, the substrates should be hydropathically compatible and coherent so that they can be adequately placed in contact thereby obtaining the facilitated reaction. This interpretation suggests that at the times of fixing the attributions, either the anticodons or some correlates of them in the acceptor arm (possibly akin to the operational codes; see Schimmel 1995) participated in the substrate contacts, which is reminiscent of the physico-chemical hypothesis. Direct tests of the first

possibility are problematic, since the anticodons are presently far from the aminoacyl-adenosine synthesis site and some of them do not even interact directly with the synthetases.

*4.13 Selection in the regionalization of attributes*

An alternative hypothesis would suggest evolutionary minimization of errors or optimization of the distances between the properties of amino acids and of triplets. The mechanism suggested by this hypothesis is that various distributions of attributions once existed, produced by point mutations that changed slightly the triplet character. When this change would code for an amino acid with properties very different from the original attribution, such coding would be strongly selected against. The end result of this process would be the observed regionalization of the attributions, so that similar amino acids would be attributed to similar codes and the changes produced by point mutations usually would not change drastically the amino acid character. The only tests possible of this hypothesis are simulations of the evolutionary process, and they do show that the present distribution of attributions is among the best for minimization of errors (Knight et al. 1999). Otherwise, the hypothesis does not refer to mechanisms of origins of the attributions.

We propose that both catalyst-driven and selective optimization hypotheses are complementary and refer to different aspects or moments of formation of the code. Our attempt is to maximize biochemical mechanistic explanations and to reduce the more vague propositions of natural selection, based only in the considerations that the present state of the system is optimized, and that other variants once existed but tended to produce worse products (phenotypes) and were selected out. The outlier attributions are considered the earliest to be fixed, based on the premise that the catalysts responsible for them had properties different from the ones that produced the correlated attributions.

*4.14 Protein structure and nucleic acid-binding*

The dynamics in the succession of stages of formation of the code follows a symmetry that builds the picture of a levorotatory windmill (**Figure 2**). The extreme stages correspond to starting with the single class II-only pair of boxes (Gly with Pro) and ending with the single class I-only pair (Ile and Met with Tyr). Stage 1 contains three small and hydroapathetic amino acids in the hydropathy correlation outlier attributions (GPS). Five of the six amino acids in the final stages (3c and 4) are hydrophobic (CWIMY), only Thr being hydroapathetic. In the intermediate stages, hydropathies are correlated with the anticodon complementarity. The main characters found relevant for formation of the code were the acquisition of metabolic stability by the proteins and their ability to bind to RNA, wherefrom an RNP system could be developed.

[ENTER **Figure 2**]

*4.15 Protein stability and non-specific punctuation*

The fundamental property of protein metabolic stability (half-life; Varshavsky 1996), which depends strongly on the amino acid residing at the N-ends (the head), was found compatible with the frequency of amino acids residing on the N-ends of most proteins (Berezovsky et al. 1997, 1999). Accordingly, the amino acids that destabilize the proteins, when residing in the N-ends, were found to be concentrated in the C-ends (the tail). The net result of these studies is that present day proteins demonstrating higher stability show this polar distribution of the amino acids. These properties correlate with and support the staging proposed for the genetic code (**Table 3**). Such relationships indicate that the property of metabolic stabilization of proteins is primordial and intrinsic to the amino acids. Therefore, their polar distribution in

proteins and locations in the code matrix were dictated by this property, in the same way as the protein degradation mechanisms were subsequently adjusted to the amino acid properties.

The model is valid both for the chronological order of amino acid encoding and for the generation of the polar organization of protein sequences. When such order is inscribed in the code, it may be considered another punctuation system: start the sequences with stabilizing amino acids and direct the destabilizers to the tails. This is called a non-specific punctuation system with respect to the variety of amino acids satisfying the rules and in contrast with the traditional system, which is specific to one tRNA at initiation and to three codons at termination. It is indicated that the non-specific preceded the specific punctuation system and that this was superposed on the former: Met is a strong stabilizer and the Stop codons belong in boxes containing amino acids that are strong destabilizers or preferred in the C-ends.

The non-specific system can be identified in the homogeneous sector alone. The chronologic succession of entry of the nine amino acids in this sector corresponds to the construction of peptides demonstrating the polar organization of the sequences, which is the character of fully structured small proteins. It can be read from **Table 3** that the three amino acids in Stage 1 contribute to form stable protein heads and that the amino acids of Stage 2b should not be incorporated earlier, to the cost of destabilizing the peptides due to the properties of Lys and Phe; the amino acids of Stage 2a show heterogeneous properties. In the mixed sector, the other protein-stabilizing amino acids are concentrated in the NRY quadrant (MVTA) and the other destabilizers in the NYR quadrant (HRWY). The chronology of entry of the attributions belonging in the homogeneous sector corresponds to their order in a string but those of the mixed sector were added to the primordial string of the homogeneous sector in a different way, the NRY being placed in the N-ends and the NYR in the C-ends. A modification of one of the NRY (iMet) was responsible for starting the formation of the specific punctuation system.

*4.16 Specific punctuation*

The puzzle set forth by observing that the two pDiN, the CAU of Met, utilized for elongation, and the CAU of iMet, utilized for initiation, coincide complementarily or directly with the pDiN of the boxes where the Stop signs reside (respectively, UA and CA) inspired a more detailed search for the links between the tRNAs involved in the whole punctuation system (**Table 4**). The recognition of the $tRNA^{iMet}$ by the initiation system, with the wobble position in the 3' extremity, is indicated by the observation that various Start codons may be accepted, with variation in the 5´ position (NUG). The second codons shown were selected according to the criteria of having 5' R and of corresponding to amino acids that are strong stabilizers of the N-ends of proteins against degradation. It is shown that the initiation system is built upon a configuration of the two first codon : anticodon pairs where the pDiN are contiguous, forming a tetra-nucleotide without the possibility of interruption by a wobble pair (codons NUGRNN).

In the anticodon triplets corresponding to the Stop codons there is a constant 3' A, identical to the central A of the initiation anticodon, both forming a standard base pair with the central U of the initiation codon. This is indicated to be the main source of the competition between the initiation anticodon and the anticodons corresponding to the Stop codons. These conflicts led to the exclusion of the latter tRNAs and their substitution by the protein Release Factors. The 3' A is preceded by two Y. The central Y forms a pair with the 3' G of the initiation codon and the 5' Y forms a pair with the 5' R of the second codon. This mode of coding of termination is indicated to be a necessary consequence of the installation of the initiation mechanism based on the slipped pDiN; when the code matrix is full, conflicts between anticodons competing in the initiation process will automatically arise. Other characters of the

Trp tRNA or of its recognition by the termination system were developed so that it could be retained with avoidance of the conflicts, which is consistent with data in Rodin et al. (1993).

[ENTER **Table 4**]

*4.17 Nucleic acid-binding*

The compilation of the amino acids which are preferred in the conserved sites of nucleic acid-binding motifs (see Guimarães et al. 2007) relative to the ones showing up in the non-conserved sites in the same motifs (**Table 3**), was able to identify the homogeneous sector of the code with the RNA-binding ability (GPSLKF, plus V and M of the mixed sector) and the mixed sector with the DNA-binding (AHTC, plus E of the homogeneous sector) or with the ability for binding both kinds of nucleic acids (RQWIY). Most of the RNA-binding motifs are highly enriched in Gly, some of them being rich in Pro. Among the nine amino acids in the homogeneous sector, six are preferred in RNA-binding motifs; only Glu is preferred in DNA-binding motifs, and Asp and Asn do not show up in the conserved sites of nucleic acid-binding sequences. Among the eleven amino acids of the mixed sector, only Val and Met are added to the list of the preferred in RNA-binding sequences. It can be said that the homogeneous sector belongs in the RNP-world functions and the mixed sector in the fully developed nucleoprotein world.

*4.18 Protein conformations*

An order of increasing complexity of protein conformations can be also correlated to the succession of stages of amino acid entry into the code (**Table 3**). The full set of amino acids preferred in coils and turns (GPSDN) is completed in the homogeneous sector, which is poor in the amino acids preferred in the strands that form the β-sheets (FVTCWIY). Amino acids preferred in α-helices (DLKARHQM) are also more frequent in the mixed sector than in the homogeneous sector.

*4.19 Amino acid biosynthesis and possible pre-codes at the core of the matrix*

The criterion of obedience to the routes of biosynthesis of amino acids has been advocated by various authors as a guide for defining the succession of their entry into the code (see Davis 1999; Wong 2005). Nonetheless, the precise precursor-derived relationships are not entirely consensual. We simplified this rationale, saying that the restrictions should refer only to the most basic of the rules of derivation: amino acids which are consensually recognized as derived from others belonging in the code should not precede the precursors. Some of the non-derived amino acids are considered precursors to biosynthesis families: the S family (GCW), the D family (NKTIMR) and the E family (QPRK); F is precursor to Y or derived earlier in the same route forming Y; H may be considered non-derived, coming directly from modification of Ribulose-5-P, or derived from E or Q. Accordingly, S cannot be preceded by G, C or W; D by N, T or M; T by I; E by Q, P or H; F by Y; R and K cannot be preceded by either one of D or E. Other amino acids not belonging in these families are derived directly from the glycolysis pathway (VLA) or are the most abundant pre-biotically (GA).

Our staging is entirely compatible with the biosynthesis derivation rules, with the single exception of having Pro in Stage 1 while its precursor Glu is placed in Stage 2a. Fortunately, obedience to this requirement contributed positively to the staging, placing only Ser and Gly in Stage 1, and both coherently octacodonic, Pro (the final code) having substituted Gly (a pre-code) in the GG box at Stage 2a. This restricted set of Stage 1 amino acids is also backed by their biosynthetic relatedness, Ser being the biosynthetic precursor to Gly (Davis 1999). Ser is placed among the most interesting pre-biotic amino acids by Nanita and Cooks (2006)

due to its peculiar ability to form clusters with chiral selection. The last of the non-derived amino acids to enter the code are His and Val, in Stage 3b.

Another consideration derived from the biosynthesis rationale also contributed positively to the staging, namely the proposal (see Osawa 1995) that Arg can be considered an 'intruder', substituting for a previous amino acid in its codes. The first proposal for the predecessor to Arg was of Ornithine, which is in the biosynthesis routes for Arg, but various others have been added to a list of putative predecessors (see Jimenez-Montaño 1999). In the attempt to gather the most of biochemical precedents, adding the least of theoretical novelties, the list should be simplified to contain only amino acids already belonging in the code. We consider the calculations (see Osawa 1995) on codons that are more frequently used in proteins than are predicted from the codons available in the code (Lys > Asp > Glu > Ala; Lys is the most overused) and on the codons that are less used in proteins than predicted (Ser > Leu > His > Pro > Arg; Arg is the most underused). From these calculations, Lys would be a good candidate due to its basicity, but we propose it to have been Ala, in spite of the different biochemical character but based on the palindromic pairing mechanism. The early coding of Ala would have been octacodonic (N<u>GC</u> : N<u>CG</u>), in the same manner as the other pair of boxes in the core (N<u>CC</u> : N<u>GG</u>) are considered to have coded for Gly. Ala has also been considered a 'filler' in protein sequences (see Osawa 1995), relatively neutral due to its small side chain. Our proposal helps to explain the data on the overuse of Ala codons in present proteins: it had a greater number of codons earlier but had them reduced after the Arg intrusion. The substitution of Ala-<u>CG</u> (in the pre-code) by Arg (the final code) occurred in Stage 3a.

These suggested pre-codes could be tested in studies of variant codes (see below). There are reports of losses of Arg codons with still unknown destinations. They could also satisfy other authors proposing both Gly and Ala to have been early in the code, due to their being the most abundant amino acids in abiotic syntheses; their abundance might have forced their aminoacylation by tRNA dimers, resulting in the formation of the octacodonic attributions. Accordingly, the whole core of the matrix and the whole set of the hydropathy outliers would have passed though a fully octacodonic pre-code stage.

*4.20 Biosynthesis of Gly and Ser driven by Stage 1 protein synthesis*

The model concentrates on the proposition of a succession of steps that fill the matrix with the correspondences as they are in the standard code. Nonetheless, we may also consider other possible functions of the proteins being made. Among the whole population of dimers engaged in protein synthesis, some were being recruited and fixed into the code but others could be contributing in parallel to the enrichment of the metabolic system. The model cannot evaluate all these accessory components but it can suggest a new perspective for interpreting the proposition of the co-evolution between the formation of the codes and the development of amino acid biosynthesis routes, which is of protein synthesis working as a pulling force for the development of amino acid biosynthesis. When the transferase function works as a sink of amino acids, any developments able to regenerate the amino acids being consumed will be favored.

When the GPS group is being utilized for the first codings, it is indicated that the first biosynthetic routes to be more strongly pulled will be those generating them. Our specific proposition on this (Farias and Guimarães, in preparation) is for the biosynthesis of Gly directly from one-carbon units, the Gly synthase utilizing $CO_2$ for the carboxyl and another one-carbon unit, carried by the tetrahydrofolate pathway, for the α-carbon. The synthesis of Ser would come through the utilization of two Gly molecules and the Serine hydroxymethyltransferase. These routes are known to be active under conditions of scarcity of fermentable substrates (Sinclair and Dawes 1995). The pathways usually considered in the

studies of the co-evolution hypothesis (Davis 1999; Klipcan and Safro 1994) are based on the free availability of fermentable substrates, where the precursors to these amino acids are of three-carbon units as in the glycolysis pathway, Ser being derived from 3-phosphoglycerate and Gly being derived from Ser. Our model also indicates that Asp was recruited into the code before Glu, suggesting that the next biosynthesis routes to be fixed utilized two-carbon units (e. g., acetate derived from the catabolism of Gly and Ser) to generate Oxaloacetate directly, such as in the Glyoxylate Cycle or in the reverse Citrate Cycle. Only after the development of the full Citrate Cycle there would be free availability of α-Keto-Glutarate, from which Glu and then Pro are derived.

## 5 **The proteic synthetases**

The tRNAs are considered better guides for developing a model for the structure of the code on the basis of their forming an evolutionarily more conserved class of molecules when compared to proteins in general, that demonstrate enormous plasticity and high openness to regulatory modulation. On another side, the overall plasticity inherent to all kinds of molecular interactions, including those involving nucleic acids, only rarely approaching 100% specificity, leads to spreads that inevitably conduct to the formation of networks. The study of the genetic code will still need to consider other components of RNA plasticity, such as the role of the non-palindromic tRNA pairings, of the high anticodon degeneracy of eukaryotes, the richness of base modifications in tRNAs, and the differential usage of codons and of anticodons.

The synthetases are nowadays one per amino acid. They are grouped in two classes, each composing a homology family bearing a conserved domain for interaction with the acceptor arm of tRNAs and for their charging with the amino acid. The acceptor arms of tRNAs have some identity sites (operational codes; Schimmel 1995) but other identity sites may be spread along the tRNA sequences, at places other than the major code, the anticodon triplet. Besides the conserved domain characteristic of the aRS class, other domains are responsible for binding and recognition of the specific tRNAs. Class II docks on the acceptor arm of RNAs through the major (more external) groove of the double helix; class I docks through the minor groove, reaching more directly the bases. Through this double approaching mode, the identity sites on the acceptor arms may be fully explored (Pouplana and Schimmel 2001). The docking mechanism has a biochemical correspondence with the site of acylation on the hydroxyls of the ribose in the terminal adenosine of tRNAs: class I acylates typically the 2′ OH and class II typically the 3′ OH. A part of the network is formed when the two aRS classes share the tRNA set defined by one pDiN (**Table 5**). The topology of this network will form the core of a larger one, incorporating the variety of other functions and interactions of the aRS in the cellular network (Quevillon et al. 1997; Simos et al. 1998; Ibba et al. 2005).

Amino acids form groups of chemical or structural relatedness, partially correlated with the aRS classes and groups (see also Pouplana and Schimmel 2001). Class II enzyme pockets accept seven of the eight small or medium amino acids, only Cys being medium and class I. The twelve large amino acids are typical of class I (nine of them) and of the mixed sector (eight). The four large amino acids in the homogeneous sector are the couple of Glu and Leu (class I in the UC : AG pair) and the largest, Lys and Phe, correspond to the atypical acylation systems (in the UU : AA pair).

The anticodonic 5’ Y are typical of class I, irrespective of the aRS class at 5’ R, and of the specific punctuation. The 5’ Y location is the result of concessions from the first occupiers of the boxes: (a) From class I to class I or punctuation: Ile conceded to Met and this to iMet; Cys to Trp and this to X; Tyr to X; (b) From class II to class I: His to Gln; Asp to Glu; Asn to the

Lys class I of some organisms (many Archaea and a few Bacteria); (c) From class II to the class I expansions: Phe to [Leu]; Ser-CU to [Arg].

[ENTER **Table 5**)

*5.1 The atypical acylation systems*

Two acylation systems, for Phe and Lys, are atypical, each in a different way. It is indicated that the atypical character is consequent to the large size of Phe and Lys relative to the class II enzyme pocket and to the amino acids being at the extremes of hydropathies (**Table 3**), and that the development of the atypical behavior was consequent to a historical event of class I duplications being not available at the times of fixation of these two attributions while class II enzymes were available and adopted the large amino acids. The atypical couple was fished by the last tRNA pair of the homogeneous sector. The PheRS is class II but acylates the 2' OH of the terminal adenosine of the tRNA, which is the class I mode of activation. The high hydrophobicity of Phe required a conformational change in the enzyme, to achieve its peculiar mode of acylation. The LysRS is class II in some organisms (Eucarya and most Bacteria) and class I in others (Ibba et al. 1997), where a class I duplication was available at the time of the incorporation of Lys to the system. The class I LysRS fulfills the class I or punctuation homogeneity of all attributions to 5' Y of complex boxes; the 2' acylation, typical of class I enzymes, fulfills the homogeneity of all attributions to the central A column. LysRS class II is the only enzyme of the class to adopt the 5' Y triplets in complex boxes. It is concluded that both the class II PheRS and LysRS should be considered atypical.

The Selenocysteine (Sec) and Pyrrolysine (Pyl) attributions are punctual additions to the amino acid repertoire, called recoding (Baranov et al. 2003). Stop codons occurring internally in some mRNAs are decoded via suppressor tRNAs and utilize specific charging systems. They are also cases of atypical location of aRS class II on 5' Y anticodons. The AGU Stop codon is decoded by a $tRNA^{Sec}$ and charged by the normal SerRS class II, and the Ser-$tRNA^{Sec}$ is transformed into the Sec-$tRNA^{Sec}$. The GAU Stop codon is decoded by a $tRNA^{Pyl}$ and charged by either a PylRS or a ternary complex formed with LysRS class I and class II (Polycarpo et al. 2004).

*5.2 Regionalization and plasticity of the synthetases*

The regularities detected in the distribution of the aRS classes in the matrix are shown in **Table 3**. The combinations with the least deviations are: the central A plus the YR quadrant, typical of class I, deviants being only the HisRS class II and the atypical PheRS; the central G plus the YY quadrant, typical of class II, deviants being only the GluRS class I, the LysRS class I of some organisms and the ArgRS expansion. The contribution of synthetase classes to the building of an architecturally integrated network derives also from their specificities for the central purines, which do not distinguish the sectors: class II unites all central G boxes and class I the central A boxes. Further contributions derive from their spreads which were mostly due to the central Y ambiguity. The spreads become the norm rather than errors or deviations.

A large gap is highlighted between what can be proposed for the constitution of the primeval protein modules, peptides with the predominant aperiodic conformation, composed by the amino acids of the homogeneous sector, and the complex organization of the aRS, that will be difficult to fill. Various changes in the composition and genomic organization of the synthetase sets are being discovered, the majority occurring in the Archaea, which may help in tracing earlier states of the code. An intriguing feature of the collected examples is the high number of occurrences involving members of the families of amino acids derived biosynthetically from the acidics of the NUC box: Glu (Gln and Pro) and Asp (Asn and Lys).

Arg also enters the list due to being derived from either one of the acidic amino acids and to being supposed to have had a predecessor.

Some of the intermediate steps may be called expansions of the aRS specificities. In some organisms, synthetases may accept tRNAs which, in the standard code, are charged by a different enzyme. The paradigmatic cases are of the aRS for the two acidic amino acids: AspRS may accept the tRNAs for Asn to form Asp-tRNA$^{Asn}$ and this will later be transformed into Asn-tRNA$^{Asn}$ by an amidation enzyme; a similar pathway is followed for the formation of Gln-tRNA$^{Gln}$ from Glu-tRNA$^{Gln}$. There are many bacterial lineages that still keep the Glu-tRNA$^{Gln}$ pathway for obtaining Gln, and it has been proposed that a separate GlnRS arose first in the eukaryotic lineage, later being transfected to some of the bacterial groups (Skouloubris et al. 2003). Another instance of a synthetase with expanded specificity but that remained fixed as such in the standard code is the MetRS, that also charges the tRNA$^{iMet}$. Some archaea lack a separate CysRS and the charging of the tRNA$^{Cys}$ is achieved by a class II ProRS, which is ambiguous or bi-functional (ProCysRS; Stathopoulos et al. 2000; Yarus 2000). Another form of bi-functionality is the fusion of the ProRS and GluRS into a single polypeptide, in most eukaryotes (Berthonneau and Mirande 2000).

### *5.3 Specificity and timing the entrance of synthetases*

The scenario displayed by our staging model indicates a faster encoding by class II duplications, which predominate strongly in the homogeneous sector. Class I enzymes predominate in the mixed sector but their numbers only equilibrate with the class II numbers in the last stage. The wide-range asynchrony of the two aRS classes has a counterpoint in the short-range concerted duplications inside each of the classes, indicating the occurrence of coupled historical inductions possibly related to the entrance of the tRNAs in pairs.

The specificity and spread of the synthetase classes is indicated to run strictly through the pairs of columns. Class II occupies fully the central G and central C columns, including the proposed early occupier of the CG box (Ala) and the first occupiers of the CA and CU boxes (presently receded to the 5' G triplets), respectively Cys and Ser, the first through the dual-specificity ProCysRS. Class I occupies fully the central A and central U columns. The homogeneity in the central A column includes the atypical PheRS, refering to its 2' mode of acylation. In the complex central U boxes, class I corresponds to the second occupiers, in the 5' Y triplets.

The spread due to class II central Y ambiguity was to the first occupiers of the central U column: UG (His), UC (Asp) and UU (Asn), plus the atypical pair UU (Lys) : AA (Phe). The dispersion of class I replaced various early class II attributions, such as the dicodonic expansions of Arg and Leu, and the homogeneous sector ended up with a greater mixture of characters than the mixed sector. Nonetheless, the homogeneous sector maintained a neat regularity in the attribution of enzymes of the same class to all tRNA pairs: the hydropathy outliers occupy two pairs with class II; the other two have either the couple of class I (EL) or the couple of atypical acylation systems (KF). We take these regularities, clearer in the homogenous than in the mixed sector, to indicate that characters of the paired tRNAs may have guided the fishing of aRS of the same class. Such regularity was partially eroded in the mixed sector. Couples of aRS class I can be seen only in the UG (Gln) : AC (Val) and in the AU (Ile, Met) : UA (Tyr) pairs, while the CA (Cys, Trp) : GU (Thr) and the CG (Arg) : GC (Ala) pairs became class-discordant after the class I displacement of the previous class II occupiers of the CA (ProCysRS) and of the CG (AlaRS) boxes. The latter two substitutions were the main symmetry-breaking events to the configuration of the code.

The mechanism of formation of the code by the fishing of complementary anticodons is entirely consistent with the aRS class specificity for the complementary central bases, only adding a further specificity, namely that the complementary triplets are of the perfect

palindromic kind, united diagonally in the matrix. This regularity is not immediately apparent from the plain observation of the overall distribution of synthetases in the matrix. Only two diagonally-paired boxes are unambiguous, with synthetases of the same class: class II in the NGG : NCC pair (Stage I), class I in the NAU : NUA pair (Stage 4).

## 6 Evolutionary code variants and the hierarchy of codes

Variant codes are alterations affecting all proteins of the organisms or organelles, with redistribution of the pre-existing degeneracy. No case is known of full loss of any of the attributions, so that the standard configuration of the code is preserved, and our model survives also this test. The general mechanism is of a decoding system developing expanded capability (most frequently due to post-transcriptional modification of tRNAs, that become able to decode new codons, plus their acceptance by the synthetases, or ribosomal changes), substituting the previous meaning of one or more codons (the donor codons). The variants are considered to have developed after the standard code was formed, due to being dependent on changes in various components of the decoding system, which require complex genomes, containing sets of tRNA-modifying enzymes, besides ribosomes and the aRS sets. In each type or occurrence of a change, the expansion of the decoding system may be preceded or followed by the loss of the former meaning of the donor codon. When it is preceded by the loss of a codon (e. g., in genomes with strongly biased base compositions, such as in the mitochondria and in the firmicute bacteria) or of its meaning (e. g., loss of tRNA-modifying enzymes), it may be said that the expansion is an event of compensation for the loss. Genomic minimization or simplification is also indicated to be a causal mechanism, evidenced in mitochondria and in the firmicute-derived mycoplasmas.

It is indicated that the changes observed (see Guimarães et al. 2007) were those that could be tolerated, occurring upon attributions that are more expendable and less crucial to physiology. Attributions not tolerating losses, that became essential to all organisms and organelles, are in three of the complex boxes (Phe, [Leu]; His, Gln; Asp, Glu) and in the five simple boxes of the non-hexacodonic attributions (Pro, Gly, Val, Ala, Thr). The high prevalence of changes in the punctuation boxes is possibly due to the complexity and plasticity of the punctuation mechanisms, involving the protein factors. The high frequency of changes in the box containing the dicodonic components of the hexacodonics Ser and [Arg] indicates that these attributions are less crucial to physiology and more expendable than other codes.

We indicate that the functional hierarchy corresponds to a temporal hierarchy: attributions fixed earlier became more tightly integrated to other components of the physiological network (as hubs), and therefore more difficult to change. The later attributions would be more loosely coupled to the network, therefore more expendable. The more widely connected earlier attributions would be also more apt for adopting extra codes. Evidences for the physiological and temporal hierarchy are: (a) The vast majority of the donors are the 5’ R codonic attributions; there are only two losses of 5’ Y codons (AGY Ser). Such instability of the 5′ codons (or 5′ Y anticodons) may be among the forces resulting in the maintenance of the two types of 5′ bases in the cellular anticodes. (b) Stop codons changed most frequently to amino acid attributions, there being only two occurrences of the reverse path (AGR [Arg] or UCA Ser changing to Stop). (c) Codons with 5’ R changed to the 5’ Y meanings (UGA Stop to Cys, UAA Stop to Tyr, AGR [Arg] to Ser, AAA Lys to Asn), there being no example of the reverse path. (d) There are also changes between the 5’ Y (UGA Stop to Trp, the AUA exchanges between Ile and Met). (e) Other examples of changes between different boxes are also from late fixations conceding to earlier ones (AGR [Arg] to Gly, AGY Ser to Gly, CUG

Leu to Ser), there being one example of the reverse path (CUN Leu to Thr). The general panorama of the changes can be interpreted as loss of complexity of the matrix or reversal to simpler or earlier states.

## 7 **Discussion**

### *7.1 The systemic concept of the gene*

Nucleic acid molecules may be taken as mere physical entities, very interesting in themselves, especially due to their ability for the templating of replicas, but this property is not exclusive to them. Being a gene is a very different attribute, derived from the embedment of nucleic acid molecules in a biotic system where, besides replication, they can template for the synthesis of proteins through a code.

If this were the only process involved in their participation in the bio-world, the outcome would be dispersive: functions of the proteins would conduct mostly to interactions with the outer environment. Our investigation of the process of formation of the code is rooted on the perspective that protein functions should also feed back positively upon the nucleic acids, thereby providing for a double link between genes and proteins, and both links being of equal weight. When RNAs were demonstrated to be better candidates for the primeval nucleic acids, instead of the intrinsically more stable DNA molecules, the definition of the principal function of early proteins could be pinpointed to stabilization of the RNA.

These processes involving molecular affinities, replication and differential physical stability should be considered in the realm of the self-organization of systems (see the categorization of systemic approaches in Di Giulio 2005). The extension of the Darwinian concept of natural selection to the molecular world in the origins of life only adds confusion; it should be reserved for the definitely biologic process of differential reproduction of cells and organisms. In the present context, it is sufficient to say that self-stabilizing and self-constructive systems grow faster and remain longer, or the term ‘molecular selection’ should be adequately quoted.

In fact, we succeeded in showing that the genetic code organization presents clear signs of the relevance of the mechanism of RNA stabilization by proteins for fixation of the early attributions. This process established a stable nucleus (fixation of the GPS group of amino acids) around which the whole system could have developed. The temporal order of the stages in the model indicates the order of fixation of the paired boxes considering their final configuration in the code. The perspective of starting with a core, which grows outwards in all directions, differs from many of the technological models, attempting to sequentially lengthen and branch the chains of reactions in the clean test-tube chemistry. The process obviously includes the minimization of deviations from chemical norms or specificities, possibly starting with less specific and advancing to more specific catalysts, and may be modeled through the engineering optimization principles.

We would like to consider that the process of formation of the RNP genetic system as modeled here may be called self-cognitive, the term describing the reflexive and stimulatory association of the protein products to exactly the same RNAs involved with their production. The term cognition is derived from human communication affairs, among themselves or between humans and the environment, and spread into the realm of zoology. However, a correlate of it is frequently used by biochemists, relative to the molecular specificities of interaction, when saying that a substrate with high affinity to an enzyme, and this, are cognate to each other. When the binding of a stable protein occurred to a non-cognate RNA – which did not participate in its synthesis – a self-maintaining system would not be formed. So,

cognition is indicated to be at the basis of the productive and positive consequence or result of an interaction.

When we said (Pardini and Guimarães 1992) that 'the system defines the gene', still based on observation of the large amount of evidence of the ambiguity of genetic sequences, we were duly asked (J. D. Watson 1995; personal communication) to clarify the workings of the system. We can now offer some molecular details. The genes are defined by the two-way processes that construct the genetic system: they produce proteins that are meaningful to them, that is, when the proteins help them to become stable and to be integrated in a system. The nucleoprotein system is constructed by the circular or reflexive association, through the coding (digital, letter by letter) and through the stabilizing connections (analogic, based on sequence patterns), and both are important to the same degree.

It happens as if the nucleic acids were, surprised, telling to the proteins – 'you are my life!' – at the same time when the proteins were, equally surprised, telling to the nucleic acids – 'you are my genes!'. The definition of the producer by the product, in the present case, of the genotype (memory with code) by the phenotype (meaningful product), is just one specific case of a general process. For instance, in the building of some sentences in human languages, words are added sequentially but the precise meaning of the sentence is only reached when the last words are known; the last words are needed so that the first may be adequately chosen and correctly understood. The connections between the initial and final segments are multiple, simultaneous and entangled in the circular configuration.

The former protein-first 'or' nucleic acid-first hypotheses are now changed into proteins 'and' genes together. When asking about the formation of systems, the question of which component came first is not relevant. The intelligent question is how the components became associated and integrated in a system. A piece of nucleic acid may be a gene for the systems that are able to accept it productively, but not for those where that piece is not productive, due to deficiencies in any of the processes of the circular associations. So, the definition of the gene is relative to its belonging in a system, not absolute to a piece of nucleic acid.

### *7.2 Stability, abundance and strings as driving forces*

We appreciate the description of the forces of nature by the vacuum or the topologic metaphor, as if matter is drained or sucked down towards the eye of a sink or falls down along the slope of a hill, thereby reaching a more stable state. There is no need to discriminate some specific more important pushing or pulling forces, the best being to indicate the existence of a difference or a gradient from a less to a more stable state, and both types of forces are important. The argument allows a rationalization of the polymerization of proteins or of nucleic acids being effective for the maintenance of metabolic cycles or networks.

In the origins of the genetic code, the consideration is clear that formation of peptide bonds works as a sink of aminoacyl-tRNAs, against their being maintained as tRNA dimers. Another instance of the fundamental participation of the force of stability or abundance in driving the evolution of the system is at the origin of the template strings for translation. The coding system passed from the stage where, in the tRNA dimers, codons and anticodons resided in the same structures while, after the stabilized poly-tRNAs could work as or produce the template strings (mRNAs), the codon function became separated from that of anticodons.

In the maintenance of metabolic networks, the forces propelling the system may be considered to converge on the environmental sink of diluted products. The metabolic fluxes are kept on going only when the products are being continuously transformed so that they do not accumulate and would force the reactions in the reverse direction; they are either stored in a different form or structure, or are secreted/excreted from the biotic compartment.

The formation of long RNA strings is not only a possibility opened by the stabilization of RNAs by proteins. Their production is enforced by the greater efficiency of elongational

replication and of templated protein synthesis. Stable strings work as sinks or suction forces that consume tRNAs and amino acids. Their elongation drives the fixation of the dispersed components of the network into mRNAs, ribosomes and proteins, with dissipation of the whole space of codes to fill the matrix. The formation of strings also introduced other forces contributing to the structuring and coordination of the system. When tRNAs are replicated separately and independently from each other, the formation of the dimers in solution is dependent on mutual affinities and on the facilitation by external compartmentalization but also subject to fluctuations in concentrations, that may be inhibitory if not equilibrated. At elongational replication, the complementary tRNAs are linked *in tandem*, lowering the requirements on external compartmentalization, and are synthesized coordinately, with lower probability of generating conflictive concentrations.

### *7.3 Origins of the genetic system and of cells*

In considering other possible functions of the proteins being made during the process of formation of the code, which should contribute to the enrichment of the metabolic system, it is justified here to concentrate on the role of protein synthesis as a force driving the fixation of the pathways of amino acid synthesis. This leads to a reinterpretation of the proposition of the co-evolution between the formation of the codes and the development of biosynthesis routes (Wong 2005). Instead of saying that the availability of amino acids drove the chronology of their incorporation into the code, it is indicated that: (a) amino acids were recruited into the code according to their participation in protein composition, structure and functions in building the nucleoprotein system; (b) development of amino acid biosynthesis routes was driven by the consumption of the amino acids at protein synthesis.

The self-referential model for the origin of the genetic code is at the same time a model for the formation of the integral genetic system and of cells, and could also be called ‘the nucleoprotein aggregate’ or the ‘ribosomal’ model. It starts with the tRNA-directed protein synthesis. Proteins associate with the tRNAs and lead to the formation poly-tRNAs from which ribosomes and mRNAs (protein-coding genes) are derived. Metabolic routes are developed, driven by the consumption of amino acids at their incorporation into proteins. The protein aggregates build a growing globule, whose surface properties are improved by the aggregation of lipids.

### *7.4 Memories for self-production*

The relative autonomy of bio-systems with respect to environmental fluctuations, allowing them to maintain stable phenotypes and identity along the passage of time and of the generations, requires the possession of memory structures – structures providing for the regenerative abilities. We would like to consider that both types of memories discussed in this study of the genetic code formation are of the same basic kind of cyclic structures – *memory in cycles*. Cyclical routes inside the metabolic networks, besides being important due to the re-cycling and self-activation properties are also similar to attractors, due to being activated at high frequency by the network.

In the origin of the coding system, the stabilizing role of the proteins upon the nucleic acids composes a cycle of self-stimulation and the best term for the role of the nucleic acids is that of a memory structure – *memory in strings*. Mechanisms of this kind could be interesting for robotic engineering: the outcomes would be programmed with binding elements that could help them to attach to the producers and cognitively reinforce their connections, so that a string memory is created and maintained. This is different from the memories in Artificial Neural Networks. It will be interesting to look for counter-examples, of systems demonstrating self-referring properties that lack some kind of memory.

When the role of genes is established as a type of memory, produced inside and by the system, it becomes also clear that the distinction between genotype and phenotype is only didactic and descriptive of the top-down approach to physiology. It could be considered that genes are an integral part of the metabolic system and that this is the basic unit of the living; the phenotype is the whole system, including its memory structures, be they in the 'classic' part of metabolism or in the 'molecular biology' part.

### *7.5 What is life*

The definition of life proposed earlier (see Guimarães et al. 2007) can now be made clearer: Life is the process of stabilization and self-construction depicted by individualized metabolic systems. The material and architectural aspect resulting from the self-constructive process is highlighted. Cells or organisms with suspended metabolism, frozen or dried, or dead and fossilized, can be recognized on morphological bases alone. It is proposed that any 'self' process depends on the formation of memory-systems and become irrevocably bound to them, and to their evolution. Memories are one step further from simple stability properties, allowing for repetitive processes.

The string memory, when built in the form of nucleic acids, and more stably in the form of DNA, acquired also a striking expansive property. Stabilized elongation of chains, mainly through the process of incorporation of duplicates, with variations, is among the most important evolutionary processes, leading to increased genome sizes and to the large diversity. Functionality comes about through the coding machinery and the products feeding back upon the producers. Further expansion is obtained through the processes of multiple coding by the genetic sequences and of the multiple and plastic functions of the products in the metabolic networks.

The two expansive components, the genetic and the proteic, each have some attributes exclusive to them and with some autonomy of their mechanisms and dynamics, in spite of being kept irrevocably interdependent and integrated. The differences between the two components are clear when their community-forming processes are compared. Parts of the genetic memories may be interchanged and recombined between organisms, from the segmental horizontal transfers, now plainly utilized for production of modified genomes, to the sexual transfer of whole genomes. The workings of the DNA memory compartment are widely blind to the quality or source of pieces coming in. The networks and phenotypes communicate more plastically and variably, mostly through chemical means – the prionic-type transfers being possibly limited – and behaviors, but also up to the linguistic symbols. Networks and phenotypes are the places where expressions of the memory are put to tests, at their integration with the other components of the networks and at the interactions of the phenotypes with the environment. The expansive character of the living systems may be compared to the Big-Bang of cosmology. It is indicated that the biotic DNP world, mainly composed of unicellular forms, occupied the whole Earth surface in a very fast and almost explosive Big-Bio-Bang.

### *7.6 Information*

It seems impossible to refrain from using the term information in the biologic realm, and advisable to clarify and propose how it could be best understood and applied. It is utilized in two complementary contexts, referring to the internal workings of the cell and to the interactions between the cells or organisms and the environment. The term is taken from human affairs (as with the term cognition), but also in different contexts, when we communicate with others or interact with the environment, and is extendable to all interactions of living beings with others or with the environment.

A general definition of information can be taken as the property of a pattern that can be distinguished from noise. This is equivalent to saying that information is a difference (between a clear pattern and a noisy one) that makes a difference, and making a difference implies that the observer is able to discern what matters and what does not matter. The distinguisher is an observing or interpreting system, so that information is the relational quality of an object, defined by the system and relative to it. The pattern has some meaning to the system because it has receptors and processors that invoke reactions in the system.

The adequacy in the connection between the workings of the system and the inputs comes from the possibilities opened by the system's receptors. The induction of a reaction (meaning) by an interaction (information) is only possible when there is some kind of preparedness of the receptor to the input – the receptor discriminates what is and what is not informational. This is equivalent to saying that the input has some previously defined – evolutionarily adjusted – meaning to the receptor. In biochemical terms, it is said that the receptor has some kind of affinity to the input, or that there is some kind of cognitive interaction. Objects display a variety of properties, distinguishing them from other objects or from noise, but only a subset of these is utilized in each kind of interaction, which is the informational subset of inputs for that particular interaction. Some interactions are not productive because there is no fitting between the inter-agents, the inputs having no meaning to the receptors they reached. The same inputs may fit other receptors or the same ones under different contexts and then induce a reaction, indicating that the information is not an absolute property of the inputs or of the receptors, but a property of their mutual relations. So, information is a concept *a posteriori* to an interaction and is linked to the possible outcomes of it. A contact is informational when some reaction results.

The term 'transfer of information' may be misleading when implying that something energetic or material is being transferred, which is not always the case. After some meaningful contacts, the effector, agonist or messenger molecule may remain integrally preserved, indicating that the informational interactions involved only some minimal and transient exchanges of influences (Ricard 2004). The technological treatment of the process of transferring signals from sources to receptors, through channels that may introduce noise (Shannon 1948), is compatible with the molecular concept. When effectors and receptors interact, their contacts are mediated by weak bonds, very sensitive to modulation by a variety of physicochemical interferences (noise), from the electronic and thermal motions to the intermingling of small molecules (water, ions, etc.), inside the nanometric channel that separates the interacting molecules. However, the Shannonian concept of information explicitly leaves aside the question of meaning.

The application of the term is quite clear in the human communication context, through languages, but there are some gradations and quantitative possibilities in the outcomes of the events, that need to be qualified for clearness. Some prescriptive or normative instructions are almost dictatorial with respect to the responses invoked, as in a master-slave relationship or in digital computing. Others may be more democratic and respectful of the receptors, which may have some freedom to interpret the instructions and react accordingly. Some prefer to reserve the term informational to the interactions of the latter kind, the former being too mechanicist for their taste.

In the cellular context, there are also both kinds of reactivity. The almost dictatorial type is seen in the flow of string information, from genes to the primary sequence of proteins, through the genetic decoding machine. The order of nucleotides or codons in a template becomes translated almost automatically into that of amino acids in the protein. The more democratic type is seen in all other components of the plastic metabolic network, from gene regulation to the interactions between the cell and the environment (Markos 2002). This is a job of mainly the proteins, when they mediate the interactions of the system with the

environment, gene activation or repression, the processing of the transcripts, the timing and quantity of translation, so configuring the bulk of the phenotype.

In our model for the process of formation of the genetic code, string information arises during the process, when protein synthesis passes from the tRNA dimer-directed to the template-directed or translational mode, and this occurs in dependence of the protein properties that helped molding and structuring the code. In this context, protein properties define the character of the working system and the genes are just the memories, adequate for the maintenance of the system. The genetic memory is a store of string information for repetitive production of the primary structure of proteins.

**Acknowledgments**: Financial support from FAPEMIG and CNPq to RCG, and from PET/MEC to CHCM; doctoral fellowship from CAPES to STF.

**References**

- Alberts, B.; Jonhson, A.; Lewis, J. et al. (2002). *Molecular Biology of the Cell*. Garland Publishing, New York, USA.
- Barabási, A. L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* 5: 101-113.

- Baranov, P. V.; Gurvich, O. L.; Hammer, A. W. et al. (2003). Recode 2003. *Nucleic Acids Research* 31: 87-89.
- Barbieri, M. (2003). *The Organic Codes – An Introduction to Semantic Biology*. Cambridge University Press, Cambridge, UK.

- Baziuk, V. A. and Sainz-Rojas, J. (2001). Catalysis of peptide formation by inorganic oxides: high efficiency of alumina under mild condition on the earth-like planets. *Advances in Space Research* 2: 225-230.

- Berezovsky, I. N.; Kilosanidze, G.; Tumanyan, V. G. et al. (1997). COOH-terminal decamers in proteins are non-random, *FEBS Letters* 404: 140-142.

- Berezovsky, I. N.; Kilosanidze, G.; Tumanyan, V. G. et al. (1999). Amino acid composition of protein termini are biased in different manners. *Protein Engineering* 12: 23-30.

- Berthonneau, E. and Mirande, M. (2000). A gene fusion event in the evolution of aminoacyl-tRNA synthetases. *FEBS Letters* 470: 300-304.
- Beuning, P. J. and Musier-Forsyth, K. (1999). Transfer RNA recognition by aminoacyl-tRNA synthetases. *Biopolymers* 52: 1-28.
- Bloch, D. P.; McArthur, B.; Widdowson, R. et al. (1984). tRNA-rRNA sequence homologies: a model for the generation of a common ancestral molecule and prospects for its reconstruction. *Origins of Life and Evolution of Biospheres* 14: 571-578.
- Bloch, D. P.; McArthur, B.; Guimarães, R. C. et al. (1989). tRNA-rRNA sequence matches from inter- and intraspecies comparisons suggest common origins for the two RNAs. *Brazilian Journal of Medical and Biological Research* 22: 931-944.
- Cairns-Smith, A. G. (1982). *Genetic Takeover and the Mineral Origins of Life*. Cambridge University Press, Cambridge, UK.
- Copley, S. D.; Smith, E. and Morowitz, H. J. (2005). A mechanism for association of amino acids with their codons and the origin of the genetic code. *Proceedings of the National Academy of Sciences USA* 102: 4442-4447.

- Creighton, T. E. (1993). *Proteins: Structures and Molecular Properties*. Freeman, New York.
- Davis, B. K. (1999). Evolution of the genetic code. *Progress in Biophysics and Molecular Biology* 72: 157-243.
- Di Giulio, M. (2005). The origin of the genetic code: theories and their relationships, a review. *BioSystems* 80: 175-184.
- Farias, S. T.; Moreira, C. H. C. and Guimarães, R. C. (2007). Structure of the genetic code suggested by the hydropathy correlation between anticodons and amino acid residues. *Origins of Life and Evolution of Biospheres* 37: 83-103.
- Ferreira, R. and Cavalcanti, A. R. O. (1997). Vestiges of early molecular processes leading to the genetic code. *Origins of Life and Evolution of Biospheres* 27: 397-403.
- Ferris, J. P. (2002). From building blocks to the polymers of life. In: *Life's origin: The Beginnings of Biological Evolution*, Ed. Willian Schopf, University of California Press, Ltd. Los Angeles, pp 113-139.
- Grantham, R. (1974). Amino acid difference formula to help explain protein evolution, *Science* 185: 862-864.
- Grosjean, H. and Houssier, C. (1990). Codon recognition: evaluation of the effects of modified bases in the anticodon loop of tRNA using the temperature-jump relaxation method. In: C. W. Genhrke and K. C. T. Kuo, Eds., *Chromatography and Modification of Nucleotides*, Elsevier, Amsterdam, pp A255-A295.
- Grosjean, H.; Houssier, C. and Cedergren, R. (1986). Anticodon-anticodon interactions and tRNA sequence comparison: approaches to codon recognition. In: P. H. Knippenberg and C. W. Hilbers, Eds., *Structure and Dynamics of RNA*, Plenum, New York, pp 161-174.
- Guimarães, R. C.; Moreira, C. H. C. and Farias, S. T. (2007). A self-referential model for the formation of the genetic code. *Theory in Biosciences* **(in press)**.
- Hughes, R. A.; Robertson, M. P., Ellington A. D. et al. (2004). The importance of prebiotic chemistry in the RNA world. *Current Opinion in Chemical Biology* 8: 629-633.
- Ibba, M.; Becker, H. D.; Stathopoulos, C. et al. (2000). The adaptador hypothesis revisited. *Trends in Biochemical Sciences* 25: 311-316.
- Ibba, M.; Morgan, S.; Curnow, A. W. et al. (1997). A euryarchaeal lysyl-tRNA synthetase: resemblance to class I synthetases. *Science* 278: 1119-1122.
- Ibba, M. P.; Rogers, T. E.; Samson, R. et al. (2005). Association between archaeal Prolyl- and Leucyl-tRNA synthetases enhances $tRNA^{Pro}$ aminoacylation. *Journal of Biological Chemistry* 280: 26099-26104.
- Jiménez-Montaño, M. A. (1999). Protein evolution drives the evolution of the genetic code and vice-versa. *Biosystems* 54: 47-64.
- Kauffman, S. A. (1993). *The Origins of Order – Self-Organization and Selection in Evolution*. Oxford University Press, New York NY.
- Klipcan, L. and Safro, M. (2004). Amino acids biogenesis, evolution of the genetic code and aminoacyl-tRNA synthetase. *Journal of Theoretical Biology* 228: 389-396.
- Knight, R. D.; Freeland, S. J. and Landweber, L. F. (1999). Selection, history and chemistry: the three faces of the genetic code. *Trends in Biochemical Sciences* 24: 241-247.
- Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydrophatic character of protein. *Journal of Molecular Biology* 157: 105-132.
- Lacey, J. C. Jr. and Mullins, D. W. Jr. (1983). Experimental studies related to the origin of the genetic code and the process of protein synthesis - A review. *Origins of Life and Evolution of Biospheres* 13: 3-42.
- Lehmann, J. (2002). Amplification of the sequences displaying the pattern RNY in the RNA world: The translation → translation/replication hypothesis. *Journal of Theoretical Biology* 219: 521-537.

- Markos, A. (2002). *Readers of the Book of Life – Contextualizing Developmental Evolutionary Biology.* Oxford University Press, Oxford, UK.
- Martinez-Giménez, J. A. and Tabarés-Seisdedos, R. (2002). On the dimerization of the primitive tRNAs: implications in the origin of the genetic code. *Journal of Theoretical Biology* 217: 493-498.
- Miller, S. L. and Lazcano, A. (2002). Formation of the building blocks of life. In: *Life's Origin: The Beginnings of Biological Evolution*, Ed. Willian Schopf, University of California Press, Ltd. Los Angeles, pp 78-109.
- Miller, D. L.; Yamane, T. and Hopfield, J. J. (1981). Effect of tRNA dimer formation on polyphenylalanine biosynthesis, *Biochemistry* 20: 5457-5461.
- Nanita, S. C. and Cooks, R. G. (2006). Serine octamers: cluster formation, reactions, and implications for biomolecule homochirality. *Angewandte Chemie International Edition* 45: 554-569.
- Oba, T.; Fukushima, J.; Maruyama, M. et al. (2005). Catalytic activities of [GADV]-protein world for the emergence of life. *Origins of Life and Evolution of Biospheres* 35: 447-460.
- Orgel, L. E. (2002). The origin of biological information. In: *Life's Origin: The Beginnings of Biological Evolution*, Ed. William Schopf, University of California Press, Los Angeles, pp 140-155.
- Osawa, S. (1995). *Evolution of the Genetic Code*. Oxford University Press Inc, New York, USA.
- Pardini, M. I. M. C. and Guimarães, R. C. (1992). A systemic concept of the gene. *Genetics and Molecular Biology* 15: 713-721.
- Polycarpo, C.; Ambrogelly, A.; Bérubé, A. et al. (2004). An aminoacyl-tRNA synthetase that specifically activates pyrrolysine. *Proceedings of the National Academy of Sciences USA* 101: 12450-12454.
- Poole, A. M.; Jeffares, D. C. and Penny, D. (1998). The path from the RNA world. *Journal of Molecular Evolution* 46: 1-17.
- Pouplana, L. R. and Schimmel, P. (2001). Two classes of tRNA synthetases suggested by sterically compatible dockings on tRNA acceptor stem. *Cell* 104: 191-193.
- Quevillon, S.; Agou, F.; Robinson, J. C. et al. (1997). The p43 component of the mammalian multi-synthetase complex is likely to be the precursor of the endothelial monocyte-activating polypeptide II cytokine. *Journal of Biological Chemistry* 272: 32573-32579.
- Ricard, J. (2004). Reduction, integration and emergence in biochemical networks. *Biology of the Cell* 96: 719-725.
- Rodin, S. N.; Ohno, S. and Rodin, A. (1993). Transfer RNAs with complementary anticodons: could they reflect early evolution of discriminative genetic code adaptors?, *Proceedings of the National Academy of Sciences USA* 90: 4723-4727.
- Schimmel, P. (1995). An operational RNA code for amino acids and variations in critical nucleotide sequences in evolution. *Journal of Molecular Evolution* 40: 531-536.
- Seligmann, H. and Amzallag, G. N. (2002). Chemical interactions between amino acid and RNA: multiplicity of the levels of specificity explains origin of the genetic code. *Naturwissenschaften* 89: 542-551.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal* 27: 379-423, 623-656.
- Simos, G.; Sauer, A.; Fasiolo, F. et al. (1998). A conserved domain within Arc1p delivers tRNA to aminoacyl-tRNA synthetases. *Molecular Cell* 1: 235-242.
- Sinclair, D. A. and Dawes, I. W. (1995). Genetics of the synthesis of serine from glycine and the utilization of glycine as sole nitrogen source by *Saccaromyces cerevisiae. Genetics* 140: 1213-1222.

- Skouloubris, S., Ribas de Pouplana, L., Reuse, H. et al. (2003). A noncognate aminoacyl-tRNA synthetase that may resolve a missing link in protein evolution. *Proceedings of the National Academy of Sciences USA* 100: 11296-11302.
- Smith, D. and Yarus, M. (1989). tRNA-tRNA interactions within cellular ribosomes. *Proceedings of the National Academy of Sciences USA* 86: 4397-4401.
- Sobolevsky, Y. and Trifonov, E. N. (2005). Conserved sequences of prokaryotic proteomes and their compositional age. *Journal of Molecular Evolution* 61: 1-7.
- Sobolevsky, Y. and Trifonov, E. N. (2006). Protein modules conserved since LUCA. *Journal of Molecular Evolution* 63: 622-634.
- Stathopoulos, C.; Li, T.; Longman, R. et al. (2000). One polypeptide with two aminoacyl-tRNA synthetase activities. *Science* 287: 479-482.
- Szathmáry, E. (1999). The origin of the genetic code: amino acids as cofactors in an RNA world. *Trends in Genetics* 15: 223-229.
- Tamura, K. and Schimmel, P. (2003). Peptide synthesis with a template-like RNA guide and aminoacyl phosphate adaptor. *Proceedings of the National Academy of Sciences USA* 100: 8666-8669.
- Trevors, J. T. and Abel, D. L. (2004). Chance and necessity do not explain the origin of life. *Cell Biology International* 28: 729-739.
- Trifonov, E. N. (2004). The triplet code from first principles. *Journal of Biomolecular Structure and Dynamics* 22: 1-11.
- Varshavsky, A. (1996). The N-end rule: Functions, mysteries, uses. *Biochemistry* 93: 12142-12149.
- Wong, J. T. F. (2005). Coevolution theory of the genetic code at age thirty. *BioEssays* 27: 416-425.
- Yamane, T.; Miller, D. L. and Hopfield, J. J. (1981). Interaction of elongation factor Tu with the aminoacyl-tRNA dimer Phe-tRNA:Glu-tRNA. *Biochemistry* 20: 449-452.
- Yang, C. M. (2005). On the structural regularity in nucleobases and amino acids and relationship to the origin and evolution of the genetic code. *Origins of Life and Evolution of Biospheres* 35: 275 –295.
- Yarus, M. (2000). Perspectives: Protein synthesis - unraveling the riddle of ProCys tRNA synthetase. *Science* 287: 440-441.
- Yarus, M.; Caporaso, J. G. and Knight, R. (2005). Origins of the genetic code: the escaped triplet theory. *Annual Review of Biochemistry* 74: 179-198.
- Yusupov, M. M.; Yusupova, G. Z.; Baucom, A. et al. (2001). Crystal structure of the ribossome at 5.5Å resolution. *Science* 292: 883-896.
- Zenkin, N.; Yuzenkova, Y. and Severinov, K. (2006). Transcript-assisted transcriptional proofreading. *Science* 313: 518-520.

**Legends to the Figures and Tables**

**Figure 1** Production of genetic strings in the process of protein synthesis and of formation of the code.
(**1**) *Protein synthesis directed by tRNA dimers*. The system is self-referent, to the tRNAs (carriers). The amino acids (letters) are fished in couples through tRNA dimerization (fishing). Catalytic activities are in boldface. After each transferase reaction with synthesis of a peptide bond, an uncharged tRNA goes back to the pool and the peptidyl-tRNA is elongated.
(**2**) *Binding of proteins to the carriers and formation of elongated tRNA strings*. For replication the tRNA molecule is extended and duplicated. Since loops are labile (thin, boxed), without stabilization by proteins they break, originating copies of the original molecule. When the loops are stabilized (thick, dashed box) replication becomes elongational and chains of carriers (poly-tRNAs) are produced. The poly-tRNAs may acquire different configurations: some become the ribosomal RNAs and others form aligned anticodons, whose copies are codon strings (mRNA). The meaningful (to the system being formed) association is selective and specific (cognitive): proteins should be intrinsically stable and efficient binders, and should bind the same carriers that were involved in their production.
(**3**) *Ribosomal translation*. Protein synthesis is directed by strings and the carriers become the decoding system. Different types of cytosolic dimers of tRNAs (both uncharged, one of them charged, both charged), and their relative concentrations, may have regulatory functions.

**Figure 2** The symmetry in the succession of integration of pairs into the system builds the dynamic picture of a levorotatory windmill.
The drawing follows the matrix in **Table 1B**. The hydropathy outliers (flaps [1a] Pro : Gly; [1b] Ser : Ser [Arg]) conserve the central G and C, and the remaining pairs of the homogeneous pDiN sector are central A and U (flaps [2a] Asp, Glu : Leu; [2b] Asn, Lys : Phe [Leu]). Pairs 3a and 3b conserve the 3' G and C (flaps [3a] Ala : Arg; [3b] His, Gln : Val), and the remaining pairs of the mixed pDiN sector are the 3' A and U (flaps [3c] Thr : Cys, Trp; [4] Ile, Met : Tyr).

**Table 1** The matrices of the genetic coding system.
(**A**) The standard genetic code matrix. The order of bases follows the increasing hydrophobicity: U<C<G<A. The 3' base of the triplets may be generically designated by W (wobble) or N. In the main initiation codon (iMet) and the other less frequent initiators (not shown), the wobble position is the 5' base. In square brackets, the dicodonic components of the hexacodonics Arg and Leu. The principal dinucleotides (<u>pDiN</u>) of the triplets exclude the W position and identify the boxes. These are grouped in quadrants and sectors: pDiN with two R or two Y are called homogeneous; those with an R and a Y are called mixed.
(**B**) The standard genetic anticode matrix. The number of triplets is taken from eukaryotes but these have 5' inosine instead of guanosine in a half of the boxes. The order of bases follows the increasing hydrophilicity: A<G<C<U. The 5' base A of the triplets is rarely present. In the initiation anticodon, the wobble position is the 3' base. In parenthesis, the absent Stop anticodons.
(**C**) Symmetries in the genetic anticode principal dinucleotide matrix. Anticodons are simplified to the sixteen pDiN, which form eight complementary pairs. The axes are formed by the core (with G- and C-only pDiN; bold) and the tips (A- and U-only pDiN; underlined) boxes. Simple boxes, corresponding to one attribution only, are read as the core and the non-axial with central R. Complex boxes, corresponding to more than one attribution, are the tips and the non-axial with central Y.

**Table 2** The necktie model for the structure of primeval protein domains.
The sequence of the primeval mRNA module combines the chronology of incorporation of attributions to the genetic code, according to the stages (1a - 4) of the self-referential model, and the proposition of the universal loop-and-lock domains of proteins (Sobolevsky and Trifonov 2005, 2006). It is indicated that the codes corresponding to the tRNA pairs of the homogeneous pDiN sector are ligated *in tandem* and compose a loop, while those of the mixed pDiN sector are added in a staggered mode, the NRY to the N-ends and the NYR to the C-ends, originating a mini-loop-and-lock structure. The sequence is presented in the linear form in **Table 3**.

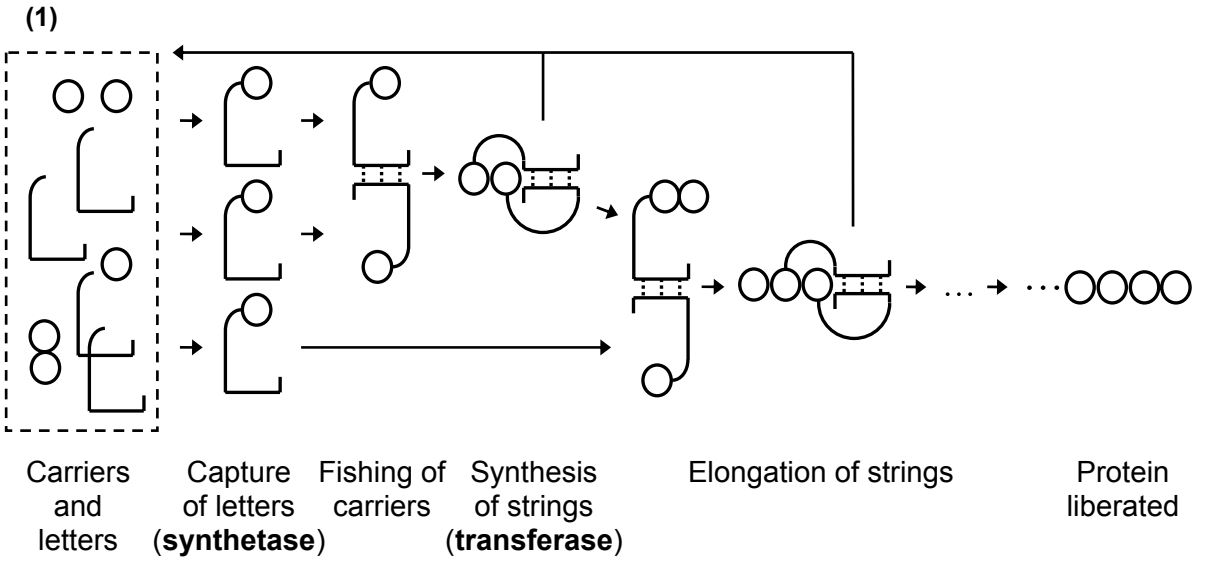
**Table 3** A model for the generalized structure of proteins based on the properties of amino acids in the quadrants of the genetic anticode matrix.
The stages of the stepwise model (second row) are arranged according to a linear protein chain. The core sequence contains the attributions in the successive pairs of the homogeneous pDiN sector (Stages 1a to 2b). The attributions of the RY pDiN are added to the N-ends (Stages 4 to 3a), and the attributions of the YR pDiN are added to the C-ends (Stages 3a to 4). The amino acid characters are summarized below.
(**A**) Characters related to aRS class specificity.
(**1**) Amino acid hydropathy (from Farias et al. 2006). Hydrophobics (< 0.111, FIMLVC) are acylated in the class I mode, including the atypical PheRS. Hydroapathetics (0.423 – 0.677, GPSTH) are acylated by aRS class II. The moderately hydrophobic (0.258 – 0.361, AWY) and the hydrophilic (> 0.809, NQEDRK) are heterogeneous with respect to the aRS classes. The outliers belong to aRS class II and to the homogeneous pDiN sector. The other attributions in this sector conform to a regression with $r^2 = 0.99$. Attributions of the mixed pDiN sector build a regression line ($r^2 = 0.85$) with steeper inclination.
(**2**) Amino acid size (in Å, from Grantham 1974). Small (GASP, < 32.5) and medium (DCNT, 54-61) are characteristic of aRS class II, including the Cys, charged by the bifunctional ProCysRS class II of some organisms. The twelve large amino acids (> 83) are characteristic of aRS class I, except for His and for the two atypical acylation systems: for Lys (class II in some organisms and atypically accepting the anticodonic 5' Y triplets of complex boxes) and Phe (acylating the ribose 2' OH, which is typical of class I enzymes). These two amino acids are also the largest of the homogeneous pDiN sector and of class II, and the extremes of hydropathy.
(**3**) pDiN hydropathy contrast. The lowest contrast between the pDiN of the homogeneous type is CC – GG = 0.732. The highest contrast between the pDiN of the mixed type is UG – AC = 0.407. It could be reasoned that it would be easier to start a process of hydropathy fitting utilizing the pDiN of the homogenous type, due to their high contrasts, and that the start with the pDiN of the mixed type would have been more difficult, requiring high discriminatory power. Nonetheless, the outliers reside in the homogeneous sector.
(**B**) Characters related to protein properties.
(**1**) Metabolic stabilization (half-life) of proteins dictated by the amino acid residing in their N-ends (adapted from Varshavsky 1996). Grade 1 is the strongest stabilizer, grade 9 the strongest destabilizer. Mixed pDiN sector: four of the five amino acids in the N-end set are strong stabilizers (grades 1 and 2); the C-end set has none of these and is rich in the destabilizers. In the middle set, of the homogeneous pDiN sector and the first stages of the construction of the code, the chronological sequence corresponds to starting with the strong stabilizers and ending with the strong destabilizers.
(**2**) Statistical preference of amino acids for being located in the N-ends (two first positions) or the C-ends (two last positions) of proteins (adapted from Berezovsky et al. 1997, 1999).

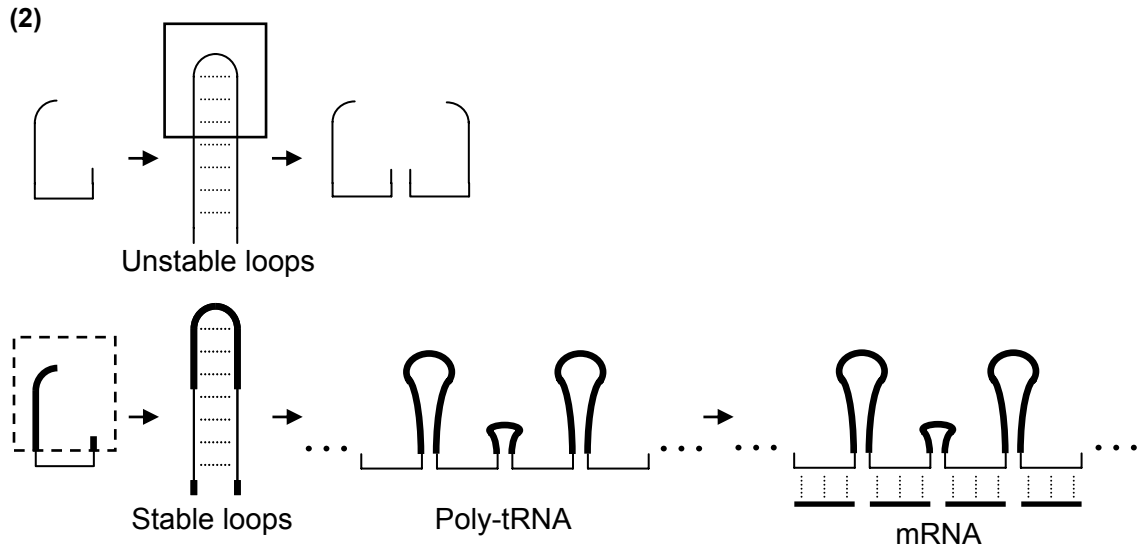
The statistical significance shows clear preferential locations at the protein ends only for ten amino acids but is entirely consistent with the metabolic stabilization grades. These data (B1, B2) are also in full agreement with the amino acids residing in the boxes where the specific punctuation triplets are. The overall consensus indicates that the polar distribution of the amino acids in proteins constitutes a non-specific punctuation system, based on the stabilization properties of the amino acids. This system is already shown by the middle set. With the addition of the mixed pDiN sector, the specific punctuation system was established, located on top of and in accordance with the non-specific system.
(**3**) Preferential participation in protein conformations (from Creighton, 1993). All amino acids characteristic of aperiodic strings (GPSDN) belong in the homogenous pDiN sector. This has only three of the eight characteristic of α-helices (ELKARHQM) and one of the seven (FVTCWIY) characteristic of β-strands.
(**4**) Preferential participation in conserved sites of nine RNA-binding and eight DNA-binding motifs or proteins. The monotonic basic motifs, rich in Lys and Arg, were not considered. Six of the eight amino acids, characteristic of RNA-binding motifs (GPSLKFVM) belong in the homogeneous pDiN sector. Nine of the ten amino acids, characteristic of DNA-binding motifs (EAHTC) or occurring equally in both types of motifs (RQWIY), belong in the mixed pDiN sector.

**Table 4** Conflicts between tRNAs directed the localization of the Stop signs.
The similarity between the triplet formed by the slipped pDiN of the initiation anticodon plus the first base of the second anticodon and the anticodons of the tRNAs corresponding to the Stop codons indicates that the latter competed with the former for the initiation codons. The conflict was solved with the exclusion of the tRNAs corresponding to the Stop codons. The second codons are listed according to the criteria of having 5' R and of corresponding to amino acids that are strong stabilizers of the N-ends of proteins against degradation.

**Table 5** The sharing of the tRNAs in a box by the different aRS classes generates a network system.
The sectors are similar in being filled from the core, with both boxes simple, to the tips, with both boxes complex, passing through the non-axial pairs, with one box simple and one complex, these sharing the two aRS classes. The sectors differ in: the homogeneous pDiN core is class II-only and the tips share the two classes; the mixed pDiN core has both aRS classes and the tips are class I-only. The full tetracodonic degeneracy of the first occupier of the Phe, His, Cys and Tyr boxes is not strictly necessary due to both the fishing and the final triplets being 5' R; it is necessary in the $Ser^{CU}$ , Asp, Asn and Ile boxes due to the fishing triplets being 5' Y and the final ones 5' R.

**Figure 1**


(1)
Carriers and letters
Capture of letters (**synthetase**)
Fishing of carriers
Synthesis of strings (**transferase**)
Elongation of strings
Protein liberated


(2)
Unstable loops
Stable loops
Poly-tRNA
mRNA

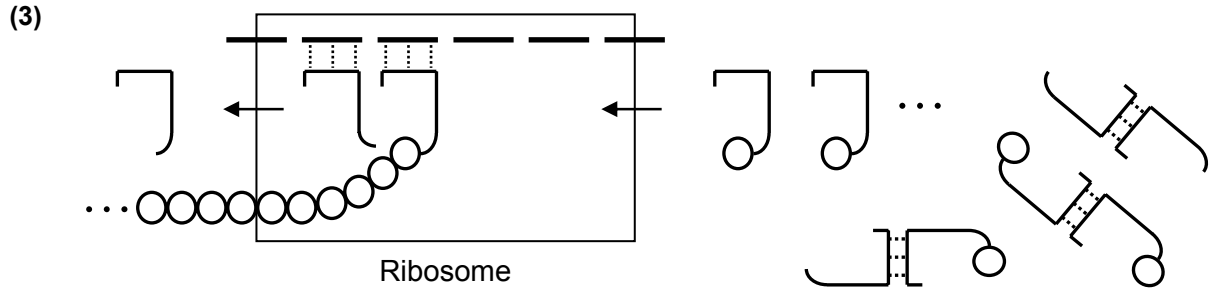

(3)
Ribosome

**Figure 2**

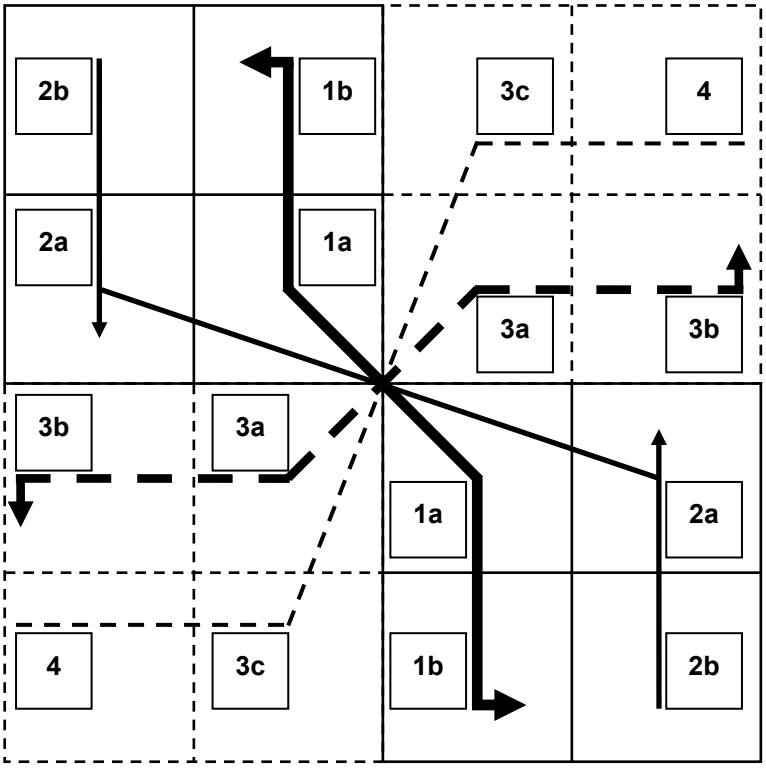


**Table 1**

**A**

| Central | U | | C | | G | | A | |
|---|---|---|---|---|---|---|---|---|
| | Quadrants | | | | Quadrants | | | |
| | Homogeneous pDiN YY | | | | Mixed pDiN YR | | | |
| 5′ U | UUU | Phe, F | UCU | Ser, S | UGU | Cys, C | UAU | Tyr, Y |
| | UUC | Phe, F | UCC | Ser, S | UGC | Cys, C | UAC | Tyr, Y |
| | UUG | [Leu] | UCG | Ser, S | UGG | Trp, W | UAG | Stop, X |
| | UUA | [Leu] | UCA | Ser, S | UGA | Stop, X | UAA | Stop, X |
| 5′ C | CUU | Leu, L | CCU | Pro, P | CGU | Arg, R | CAU | His, H |
| | CUC | Leu, L | CCC | Pro, P | CGC | Arg, R | CAC | His, H |
| | CUG | Leu, L | CCG | Pro, P | CGG | Arg, R | CAG | Gln, Q |
| | CUA | Leu, L | CCA | Pro, P | CGA | Arg, R | CAA | Gln, Q |
| 5′ G | GUU | Val, V | GCU | Ala, A | GGU | Gly, G | GAU | Asp, D |
| | GUC | Val, V | GCC | Ala, A | GGC | Gly, G | GAC | Asp, D |
| | GUG | Val, V | GCG | Ala, A | GGG | Gly, G | GAG | Glu, E |
| | GUA | Val, V | GCA | Ala, A | GGA | Gly, G | GAA | Glu, E |
| 5′ A | AUU | Ile, I | ACU | Thr, T | AGU | Ser, S | AAU | Asn, N |
| | AUC | Ile, I | ACC | Thr, T | AGC | Ser, S | AAC | Asn, N |
| | AUG | iMet | | | | | | |
| | AUG | Met, M | ACG | Thr, T | AGG | [Arg] | AAG | Lys, K |
| | AUA | Ile, I | ACA | Thr, T | AGA | [Arg] | AAA | Lys, K |
| | Mixed pDiN RY | | | | Homogeneous pDiN RR | | | |

**B**

| Central | A | | G | | C | | U | |
|---|---|---|---|---|---|---|---|---|
| | Homogeneous pDiN RR | | | | Mixed pDiN YR | | | |
| 3′ A | AAG | Phe | AGG | Ser | ACG | Cys | AUG | Tyr |
| | AAC | [Leu] | AGC | Ser | ACC | Trp | (AUC) | X |
| | AAU | [Leu] | AGU | Ser | (ACU) | X | (AUU) | X |
| 3′ G | GAG | Leu | GGG | Pro | GCG | Arg | GUG | His |
| | GAC | Leu | GGC | Pro | GCC | Arg | GUC | Gln |
| | GAU | Leu | GGU | Pro | GCU | Arg | GUU | Gln |
| 3′ C | CAG | Val | CGG | Ala | CCG | Gly | CUG | Asp |
| | CAC | Val | CGC | Ala | CCC | Gly | CUC | Glu |
| | CAU | Val | CGU | Ala | CCU | Gly | CUU | Glu |
| 3′ U | UAG | Ile | UGG | Thr | UCG | Ser | UUG | Asn |
| | UAC | iMet | | | | | | |
| | UAC | Met | UGC | Thr | UCC | [Arg] | UUC | Lys |
| | UAU | Ile | UGU | Thr | UCU | [Arg] | UUU | Lys |
| | Mixed pDiN RY | | | | Homogeneous pDiN YY | | | |

**C**

| Central | A | G | C | U |
|---|---|---|---|---|
| | Homogeneous pDiN RR | | Mixed pDiN YR | |
| 3′ A | AA | AG | AC | AU |
| 3′ G | GA | **GG** | **GC** | GU |
| 3′ C | CA | **CG** | **CC** | CU |
| 3′ U | UA | UG | UC | UU |
| | Mixed pDiN RY | | Homogeneous pDiN YY | |

**Table 2**

| | | | | | | |
|---|---|---|---|---|---|---|
| Homogeneous | | | 2a | 2a | | |
| pDiN sector | | | **Asp** | **Glu** | | |
| | 1b | **Ser** | | | **Leu** | 2a |
| 1b | **Ser** | | | | | **Asn** 2b |
| 1a | **Pro** | | | | | **Lys** 2b |
| | 1a | **Gly** | | | **Phe** | 2b |
| | | ‖ | | | ‖ | |
| Mixed | 3a | **Ala** | | | **Arg** | 3a |
| pDiN | 3b | **Val** | | | **His** | 3b |
| sector | | | | | **Gln** | 3b |
| | 3c | **Thr** | | | **Cys** | 3c |
| | | | | | **Trp** | 3c |
| | | | | | **X** | ---4 |
| | 4 | **Ile** | | | **Tyr** | 4 |
| | 4 | **Met** | | | | |
| N-end | ---4 | **iMet** | | | **X** | ---4 C-end |

**Table 3**

| Quadrant | *The NRY boxes of the mixed pDiN sector; the N-end set* | | | | | *The homogeneous pDiN sector; the middle set* | | | | | | | | | *The NYR boxes of the mixed pDiN sector; the C-end set* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Stage | 4 | 4 | 3c | 3b | 3a | 1a | 1a | 1b | 2a | 2a | 2a | 2b | 2b | 2b | 3a | 3b | 3b | 3c | 3c | 4 |
| Amino acid | Met | Ile | Thr | Val | Ala | Gly | Pro | Ser | Asp | Glu | Leu | Asn | Lys | Phe | Arg | His | Gln | Cys | Trp | Tyr |
| Hydropathy | Phob<br>0.059 | Phob<br>0.029 | Apath<br>0.438 | Phob<br>0.069 | Mod. phob<br>0.258 | Apath<br>0.423 | Apath<br>0.677 | Apath<br>0.508 | Phil<br>0.962 | Phil<br>0.935 | Phob<br>0.066 | Phil<br>0.809 | Phil<br>1.000 | Phob<br>0.000 | Phil<br>0.982 | Apath<br>0.573 | Phil<br>0.841 | Phob<br>0.111 | Mod. phob<br>0.325 | Mod. phob<br>0.361 |
| pDiN hydropathy | CAU<br>0.199 | RAU UAU<br>0.199 | NGU<br>0.433 | NAC<br>0.189 | NGC<br>0.403 | NCC<br>0.928 Outlier | NGG<br>0.196 Outlier | NGA: RCU<br>0.098:1 Outlier | RUC<br>0.931 | YUC<br>0.931 | NAG<br>0.059 | RUU<br>0.931 | YUU<br>0.931 | RAA<br>0 | NCG<br>0.564 | NUG<br>0.596 | NUG<br>0.596 | RCA<br>0.312 | CCA<br>0.312 | RUA<br>0.303 |
| Amino acid size | Large 105 | Large 111 | Med. 61 | Large 84 | Small 31 | Small 3 | Small 32.5 | Small 32 | Med. 54 | Large 83 | Large 111 | Med. 56 | Large 119 | Large 132 | Large 124 | Large 96 | Large 85 | Med. 55 | Large 170 | Large 136 |
| aRS class | I | I | II | I | II | II | II | II | II | I | I | II | I, II* | II* | I | II | I | I | I | I |
| Stabilization | 1 | 6 | 2 | 1 | 2 | 1 | 1 | 2 | 4 | 4 | 9 | 5 | 8 | 9 | 8 | 7 | 5 | 3 | 9 | 9 |
| Protein end | N | Both | N | N | Both | Both | N | Both | - | Both | Both | C | C | C | C | - | - | C | - | C |
| Protein conformation | α-helix | β-strand | β-strand | β-strand | α-helix | Aperiodic | Aperiodic | Aperiodic | Aperiodic | α-helix | α-helix | Aperiodic | α-helix | β-strand | α-helix | α-helix | α-helix | β-strand | β-strand | β-strand |
| Nucleic acid binding | R1 D0 | D2 R2 | D1 R0 | R4 D3 | D2 R1 | R7 D2 | R4 D2 | R1 D0 | - | D2 R1 | R5 D3 | - | R3 D2 | R3 D2 | D2 R2 | D2 R1 | D1 R1 | D1 R0 | D1 R1 | D2 R2 |

**Table 4**

| | Initiation ↓ | Elongation ↓ | |
|---|---|---|---|
| | First | Second | Third |
| | iMet | Val, Ala, Gly, Met, Thr, Ser$^{CU}$ | Any |
| Codons 5’ | W U G | R Y, G W | N N W |
| Anticodons 3’ | U A C | Y R, C W | N N W |
| Anticodons at the Stop hemiboxes 3’ A Y Y | A U U X from the Tyr box [ Deleted and substituted<br>A U C X from the Tyr box [ by the protein<br>A C U X from the Trp box [ Release factors<br><br>A C C Trp [ Retained | | |

**Table 5**

| Fishing triplet | Degeneracy | | Concession | aRS class | | Box type |
|---|---|---|---|---|---|---|
| Homogeneous pDiN sector | | | | | | |
| **1a** | | | | | | |
| Gly CCC | NCC | | | II | | Simple |
| Pro GGG | NGG | | | II | | Simple |
| **1b** | | | | | | |
| Ser AGA | NGA | | | II | | Simple |
| Ser UCU | NCU | RCU | | II | | Complex |
| | | YCU | [Arg] | | I | |
| **2a** | | | | | | |
| Asp CUC | NUC | RUC | | II | | Complex |
| | | YUC | Glu | | I | |
| Leu GAG | NAG | | | | I | Simple |
| **2b** | | | | | | |
| Asn UUU | NUU | RUU | | II | | Complex |
| | | YUU | Lys | II* | I | |
| Phe AAA | | RAA | | II[2'] | | Complex |
| | | YAA | [Leu] | | I | |
| Mixed pDiN sector | | | | | | |
| **3a** | | | | | | |
| Ala CGC | NGC | | | II | | Simple |
| Arg GCG | NCG | | | | I | Simple |
| **3b** | | | | | | |
| His GUG | | RUG | | II | | Complex |
| | | YUG | Gln | | I | |
| Val CAC | NAC | | | | I | Simple |
| **3c** | | | | | | |
| Thr UGU | NGU | | | II | | Simple |
| Cys ACA | | RCA | | II (ProCys) | I | Complex |
| | | YCA | Trp CCA | | I | |
| | | | Stop UCA | | | |
| **4** | | | | | | |
| Ile UAU | NAU | RAU | UAU | | I | Complex |
| | | YAU | Met CAU | | I | |
| | | | iMet CAU | | | |
| Tyr AUA | | RUA | | | I | Complex |
| | | YUA | Stop | | | |

SUBJECT INDEX

AUTHOR INDEX