Cogn Comput (2012) 4:115–139
DOI 10.1007/s12559-011-9116-1

Mutuality in Discrete and Compositional Information: Perspectives for Synthetic Genetic Codes

Romeu Cardoso Guimarães

Received: 19 March 2011 / Accepted: 14 October 2011 / Published online: 10 November 2011

Abstract The self-referential model for the formation of the genetic code proposes that protein synthesis was initiated by proto-tRNA dimers. Proto-tRNAs in the dimers recognize each other through anticodon pairing. The proteins produced recognize the producing dimers through binding, forming (proto)ribonucleoprotein (RNP) aggregates. Their functions were stimulated and specificities evolved through cycling. Such cycles would be among the first in the construction of living networks, and examples of processes that might be relevant for modeling cognitive networks. The protein synthesis process is considered a main drive for the living system′s specific attributes of anabolic and evolutionary semi-autonomy. Structures of the anticodon dimer networks are presented. Biological data point to the encoding having been installed on the modules of dimers formed by nonself-complementary triplets. Aminoacyl-tRNA adhesion interactions integrated the dimer networks into RNP networks. Specific questions are proposed for simulation and modeling that should help in designing experimental procedures aiming at testing the model and the development of synthetic genetic codes.

Keywords Genetic code · Origins · Evolution · tRNA dimers · Aminoacyl-tRNA synthetases · Networks · Self-reference · Self-organization · Simulation · Experimental · Pushing dynamics · Anabolic drive

Abbreviations

Amino acids	Are designated by three-letter abbreviations; groups of amino acids may be indicated by one-letter abbreviations
Anticodon	The code triplet of transfer RNA, the default notation of code triplets
aRS	Synthetase, aminoacyl-tRNA synthetase
Codon	The code triplet of messenger RNA
iMet	Initiator
MaRS	Multi-aRS complex
mRNA	Messenger RNA
pDiN	The principal dinucleotide of code triplets, excluding the wobble position
R	Purines
RF	Release factors
RNP	Ribonucleoprotein
rRNA	Ribosomal RNA
SRM	Self-referential model
Transferase	Peptidyl-transferase
tRNA	Transfer RNA
-, w, N (any base)	The wobble position of code triplets, anticodon 5′ or codon 3′
X	Stop or termination
Y	Pyrimidines

Introduction

The genetic code is a defining character of living systems, the quality that distinguishes them most clearly from the physicochemical realm. It is part of the central subsystem of cells—translation—where specific triplets in nucleic

R. C. Guimarães (✉)
Laboratório de Biodiversidade e Evolução Molecular,
Departamento de Biologia Geral, Instituto de Ciências
Biológicas, Universidade Federal de Minas Gerais,
31270-901 Belo Horizonte, MG, Brazil
e-mail: romeucardosoguimaraes@gmail.com

acid coding sequences of an mRNA correspond to specific amino acids (Fig. 1, Table 1). Its functioning is described as a circular structure where the characteristic components—nucleic acids and proteins—relate to each other in bidirectional mutuality, precisely and synergistically (Fig. 2). Recognizing that the puzzle of its origin refers to the establishment of a circular configuration of the interactions between very different components amplifies the complexity of the origins problem.

Outline

The first part of the text reviews the main approaches that have been pursued in attempts at deciphering the formation of the genetic code. Focus of the self-referential model (SRM) is in the encoding problem where the *compositions* of interactions between sites distributed in the sequences of the synthetases and of the tRNAs become correlated to the *discrete* correspondences between the anticodons and the amino acids. Contributions are given especially with respect to the anticodon dimerization mechanism and its chronological evolutionary development. The problem of enchaining triplets into mRNAs and genes is treated less extensively, while attention is given to identifying forces that gave rise to the coding system. The last part of the text is centered in the SRM, based on the tRNA dimer-directed protein synthesis [36–39], and suggests some experimental testing procedures. The dimers form network modules, some of which are amenable to propitiate encoding. Integration of the dimer modules is obtained from interactions between aminoacyl-tRNA synthetases. Ten of the 20 synthetases developed physical aggregation, implementing the integration of the system. Some questions to be addressed by simulation are briefly sketched.

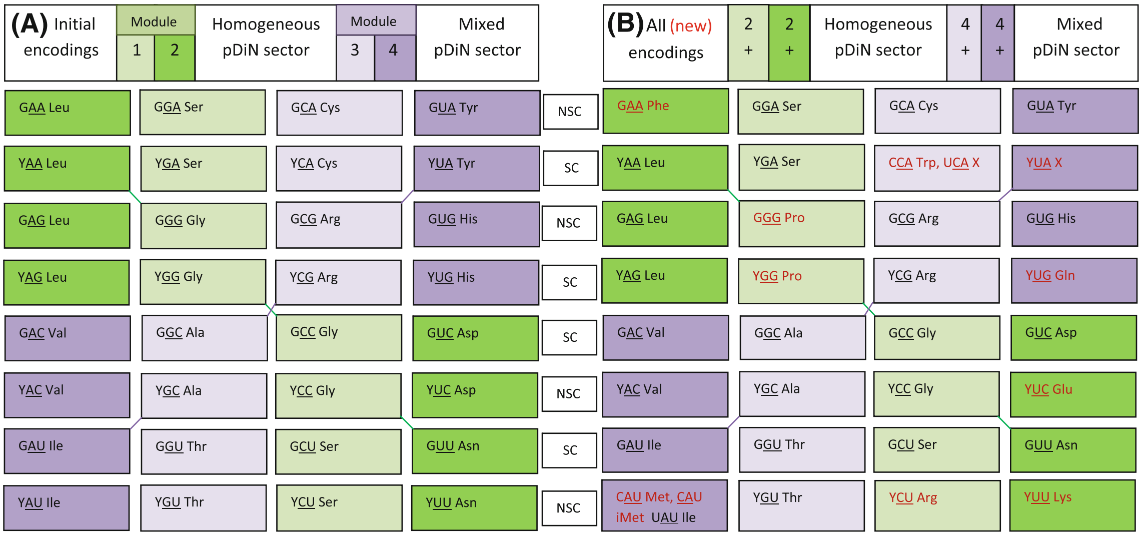
The Code is for Constructing Strings

The noblest components of cells are polymeric strings that may be called both together, irrespective of being templates for or products of the polymerization mechanisms, informational or genetic macromolecules. Monomers in proteins are relatively simple chemical compounds—amino acids—but in the coding sequences of genes or mRNAs, they are complex units—triplets of nucleotides. The polymers are characteristically non-monotonic and nonrandom sequences of the monomers, with a high degree of complexity and order.

Two Faces of the Genetic Code: Macromolecular Sequences and Protein-Dependent Encodings

Characters most relevant to biological specificity are well summarized in the genetic code, where the discrete correspondences between code triplets and amino acids and the compositional information in interacting polymers are joined in mutuality.

(a) Genetic information is considered the specific order in the sequences of codon triplets in the nucleic acid templates, which is transferred to protein sequences



(A)					(B)			
GAA Leu	GGA Ser	GCA Cys	GUA Tyr	NSC	GAA Phe	GGA Ser	GCA Cys	GUA Tyr
YAA Leu	YGA Ser	YCA Cys	YUA Tyr	SC	YAA Leu	YGA Ser	CCA Trp, UCA X	YUA X
GAG Leu	GGG Gly	GCG Arg	GUG His	NSC	GAG Leu	GGG Pro	GCG Arg	GUG His
YAG Leu	YGG Gly	YCG Arg	YUG His	SC	YAG Leu	YGG Pro	YCG Arg	YUG Gln
GAC Val	GGC Ala	GCC Gly	GUC Asp	SC	GAC Val	GGC Ala	GCC Gly	GUC Asp
YAC Val	YGC Ala	YCC Gly	YUC Asp	NSC	YAC Val	YGC Ala	YCC Gly	YUC Glu
GAU Ile	GGU Thr	GCU Ser	GUU Asn	SC	GAU Ile	GGU Thr	GCU Ser	GUU Asn
YAU Ile	YGU Thr	YCU Ser	YUU Asn	NSC	CAU Met, CAU iMet UAU Ile	YGU Thr	YCU Arg	YUU Lys

Fig. 1 The genetic anticode box structure and the chronology of encoding according to the self-referential model. The standard genetic anticode is shown in full in the *panel* on the *right*. Its formation reads from modules 1 and 2 (homogeneous pDiN sector, *left panel*, *upper left* and *lower right* quadrants, module 1 Gly, Ser and module 2 Leu, Asp, Asn) to the 2+ (new attributions in the *right panel* Glu, Pro, Lys, Phe, Arg), then going to the modules 3 and 4 (mixed pDiN sector, *left panel*, Val, Ile, Ala, Thr, Cys, Arg, Tyr, His) to finally reach the 4+ (new attributions in the *right panel* Met, iMet, X, Trp, Gln). An easier reading is shown in Table 1

Table 1 Chronology of encoding according to the self-referential model

Homogeneous pDiN sector					
Initial encodings (5 amino acids)	**Module 1**(a) Gly octacodonic -CC:-GG	(b) Ser octacodonic -GA:-CU	**Module 2** Leu octacodonic (a) -AG: (b) -AA:	(a) Asp tetracodonic -UC -----	(b) Asn tetracodonic ----- -UU
Full (5 new, 2+)	-CC Gly tetra -GG Pro tetra	-GA, RCU Ser hexa YCU Arg di	RAA Phe di YAA, -AG Leu hexa	RUC Asp di YUC Glu di	RUU Asn di YUU Lys di
Mixed pDiN sector					
Initial encodings (7 amino acids)	**Module 3**(a) -CG Arg tetra (hexa) -GC Ala tetra	(b) -GU Thr tetra -CA Cys tetra	**Module 4**(a) -AC Val tetra -UG His tetra	(b) -AU Ile tetra	(b) -UA Tyr tetra
Full (3 new plus punctuation, 4+)		RCA Cys di CCA Trp mono UCA X mono	RUG His di YUG Gln di	G, UAU Ile tri CAU Met mono CAU iMet mono	RUA Tyr di YUA X di

The process of encoding pairs of boxes is shown stepwise, going from simple to complex boxes and from higher to lower degeneracy. See legend to Fig. 1

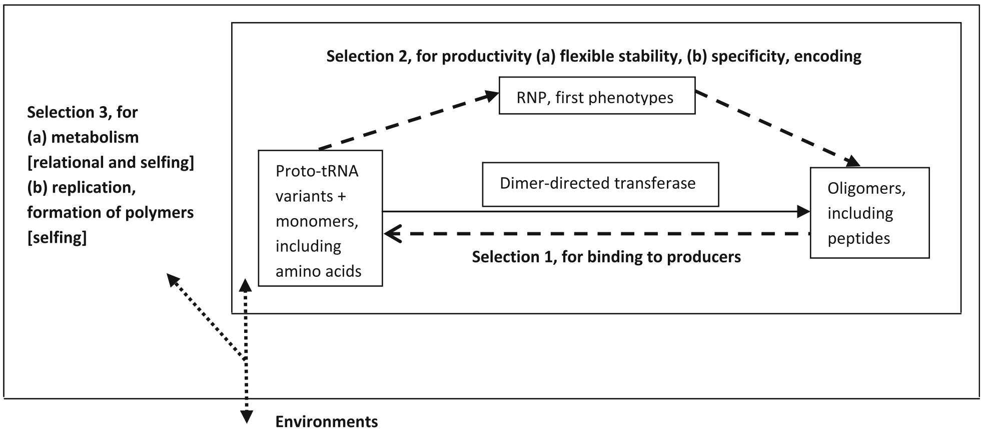

Selection 2, for productivity (a) flexible stability, (b) specificity, encoding
RNP, first phenotypes
Selection 3, for
(a) metabolism
[relational and selfing]
(b) replication,
formation of polymers
[selfing]
Proto-tRNA variants + monomers, including amino acids
Dimer-directed transferase
Oligomers, including peptides
Selection 1, for binding to producers
Environments

Fig. 2 The circular evolutionary process. The internal box shows the process of (proto)tRNA dimer-directed synthesis of oligomers that is fixed through selection (1) for the monomers able to produce the binding, which were the amino acids starting with Gly. The aggregate RNPs formed were the first phenotypes, maintaining the synthesis activity and whose accumulative power worked as anabolic drive and evolutionary pulling dynamics. Selection at this stage (2) was for productivity, reaching specificity between the reactants. Further selection (3) was for installation of sink-directed metabolic supply, originating the (*a*) network stability, and of nucleic acid replication, producing the (*b*) polymeric structures

through translation. It is maintained in cells through replication, transcription, and reverse transcription and in lineages through selection processes. Triplets and amino acids may be called the monomers or letters (the discrete components in the correspondences) of the macromolecular strings, these being analogous to words or sentences of common languages (the compositional wholes). While the sequences are linear (bi-dimensional), the working states of the long molecules are three-dimensional intricately folded dynamic configurations of the strings, very difficult to describe especially in the case of proteins.

(b) The encoding of triplets is obtained through protein-dependent mechanisms. The transfer RNAs (tRNA) are short, about 75 nucleotides, while the synthetase (aRS) enzymes are very long, in the order of about one thousand amino acids. Each synthetase recognizes a variety of molecular sites in a tRNA, which may not include the anticodon triplet and whose composition provides specificity for the charging with one amino acid [68]. The anticodons may be considered RNA guides (adaptors) that are utilized by the aRS and the ribosomes to obtain the synthesis of proteins directed by the codons, at translation of the templates. Release Factors are enzymes utilized by the ribosomes for obtaining termination of the synthesis of a protein, which substituted the tRNA guides while maintaining the directions given by the codons [51].

In accordance with the two faces of the code, there are also two levels of encodings. (c) Amino acid-triplet encoding refers to the development of specificity of the aRS, corresponding to one amino acid and a set of cognate tRNAs. The tRNA set may have more than one component, each one with a different anticodon, and an even higher number of complementary codons in mRNAs; the correspondence of one amino acid to more than one anticodon or codon is called degeneracy of coding (Table 2). (d) A higher level of encoding refers to the ordering of chains of triplets in mRNAs, corresponding to chains of amino acids in proteins. The anticodon triplets are pre-encoded as in (c) but the chains of codons in mRNAs receive meaning due to the initiation mechanisms of specific binding of ribosomes and mRNAs, from which translation can start.

Degeneracy

The codons or the anticodons are called codes in the sense that they are correlated with the collective of the specificity sites in the tRNAs and in the enzymes, which are the direct meaning-conveyers; the meaning of a code is an amino acid or the termination. There being 64 codons for 20 amino acids, degeneracy of coding is the rule. Main component of the degeneracy rule is the acceptance by the synthetases of variation in the 5′ position of anticodes, keeping constant the other two positions (the principal dinucleotide, pDiN), which define the 16 boxes of the triplet matrix.

Structural studies on ribosomes produced an adequate explanation for the wobbling in the 3′ position of codons at pairing with the 5′ position of anticodons and defined that the pDiN are most important for the decoding mechanism [66, 67]. Interactions of the ribosomal decoding site with the curvatures of the minihelix formed by the codon and anticodon showed an rRNA contiguous duplet (a homogeneous dinucleotide, AA) in tight and multiple contacts with the paired pDiN, while the contact with the wobble position is single and looser with a G nucleotide coming from a distant site. It is assumed that evolutionary development of the pDiN degeneracy resulted from a long interplay between (1) the aRS-dependent correlation between the (a) distributed interaction sites in them and in the tRNAs, and (b) the anticodon, plus (2) the bi-partite composition and the conformation of the ribosomal decoding site mini-helix.

Only two amino acids (Met, Trp) and the initiation of a protein sequence (iMet) are strictly monocodonic, the correspondence between the distributed specificity sites in tRNAs and in the synthetases being made with the entire triplet codes. Six other amino acids are monoanticodonic and dicodonic. For the other 12 amino acids and for termination, the degree of degeneracy ranges from dianticodonic to hexacodonic (Table 2). Only the hexacodonic (Ser, Leu, Arg) and the termination codes (X) correspond to proteins accepting variation in the pDiN besides the 5′ bases.

The aRS-pDiN Rule

The SRM adopts the rationale of developing complexity from initial simpler forms, utilizing the aRS-pDiN rule: initial encodings were fully 5′ degenerate for a pDiN (forming the tetracodonic boxes, *simple* with respect to

Table 2 Degeneracy in the genetic code

Codons 64	**Degeneracy**	**Hexa** 18		**Tetra** 20	**Tri** 6		**Di** 18		**Mono** 2
Anticodons 46		Penta 10	Tetra 4	Tri 15	Deleted	Di 2	Di 6	Mono 6	Mono 3
Synthetases 20		2	1	5	**Release Factors**	1	3	6	2
		One pDiN with full 5′ degeneracy			**pDiN with partial 5′ degeneracy**			**No 5′ degeneracy**	
One pDiN per synthetase		-	-	G, YCC Gly G, YGG Pro G, YAC Val G, YGC Ala G, YGU Thr	-	G, UAU Ile	YUG Gln YUC Glu YUU Lys	GAA Phe GCA Cys GUA Tyr GUG His GUC Asp GUU Asn	CAU Met CCA Trp
Two pDiN per synthetase or Release Factor **(a)** Punctuation		-	-	-	YUA X UCA codon degeneracy; pDiN: 3′ conservation, different central bases	-	-	-	CAU iMet codon degeneracy; pDiN slipped in same triplet
(b) Complementary		-	G, YGA Ser	-	-	-	-	GCU Ser	-
(c) 3′-related central base conserved		G, YAG Leu G, YCG Arg	-	-	-	-	YAA Leu YCU Arg	-	-

The multiplicity of codes reaches six codons or five anticodons per amino acid. Most of the degeneracy corresponds to acceptance of different bases in the wobble position of codes. Three synthetases and the Release Factors are highly plastic, accepting two different pDiN. The RFs correspond to different central bases, conserving the 3′ (anticodons UCA and YUA). The hexacodonic synthetases are distinct from each other: SerRS conserves the homogeneous pDiN sector with complementary pDiN (NGA:GCU), LeuRS conserves the homogeneous pDiN sector and the central base (NAG, YAA), ArgRS conserves the central base but traverses the homogeneous and the mixed pDiN sector (YCU, NCG) aRS specificity). When a new encoding is inserted in an already occupied box, the pDiN will be shared by different enzymes (forming a *complex* box), each utilizing a fraction of the 5′ variants, with a higher degree of specificity for some defined triplets containing the same pDiN. The consequent reduction in the degree of degeneracy is equivalent to increased informational attributes for specificity.

The Circular Evolutionary Process

The traditional mode of describing the code, following the flow of information encoded in strings in the direction from nucleic acids to proteins, considers only the nearly frozen and almost deterministic half of the loop. Chains of triplets in mRNA sequences direct the enchaining of amino acids in proteins so that mRNAs belong among the producing components in the system, proteins being the immediate products (Fig. 2). The other half describes the functions of proteins in constructing the phenotypes, but recognizes that some of them (e.g., synthetases, ribosomal proteins, translation factors) participate in the formation and evolution of the encoding–decoding system, therewith closing the loop. Such a circular structure of the interactions follows the same rationale of the biological evolutionary process where there are genetic variants producing phenotypes whose success is evaluated through some fitness measures that result in permanence, and consequent fixation in the pools, of the adequate configurations and progressive loss of the inadequate states.

Proteins are main components of phenotypes and when the producing components—nucleic acids—are informed on the outcomes—the proteins—via variations in their adequacy in formation of the nucleoprotein aggregate structures and in the functions of the aggregates, the producers may follow evolutionary changes in the direction of maintenance of a diversity of performances. According to this description, the living system and its central subsystem, the genetic code, belong in a class of performance-driven systems. The description follows top-down approaches where the performance of the whole influences the qualities of the components, among which are the genes and the genetic code. In the rationale describing the flow of information stored in string sequences, genes may be at the top, but in the system′s evolution rationale, phenotypes are the top and genes are edited according to the fitness of the integral system. In present-day organisms, the code is in an almost frozen or universal state. The known modifications from the standard form are few [3, 52] and do not drastically challenge the canonic rules. Otherwise, they indicate that the structure of the code followed evolutionary processes that can be investigated. Experiments in the area of the origins and evolution of the known coding system would also be able to set the bases for construction of synthetic organisms.

Formation of the Code

The large collection of studies on the formation of the code can be grouped under two wide categories of approaches: (a) the block structure of the matrix and the (b) stepwise addition of single codes [43, 53].

(a) The Block Structure Approach

Examination of the whole matrix of correspondences as a block structure poses a formidable problem in the attempts of understanding how the code was originated. It is supposed that there were at the beginning some long RNAs—mRNA-to-be—containing a given number of sites that would become codons when some kind of protein synthesis mechanism, involving anticodons, could utilize them as templates. Since the RNAs to be translated are external to the decoding system, this mode of evolution is called hetero-referential. The evolutionary rationale would be of selecting among the initial random sequences—in the sense of having structures not directed by systemic fitness values—some that would acquire functions during the process. The evolutionary focus on a certain fitness character would arise after some long strings of RNA meet other long strings of proteins, and the partners would be able to produce an aggregate depicting functionality. Various codes would be formed more or less concomitantly, some of them to become fixed more expediently than others in a process of mutual adjustments, until a reasonable number of codes could be set to become the core for further adjustments and increase in number. In such cases, a great problem would be the presence from the beginning of a large number of nonsense triplets on which the synthesis of proteins could not progress. A large portion of the evolutionary focus would be on producing correspondences for those triplets, to finally obtain reasonably long sense stretches (open reading frames) in the midst of only a few nonsense triplets. Further evolution of longer sequences would be on using the nonsense triplets for termination of the strings, as stop codes.

Optimization for Minimization of Errors

The hetero-referential supposition is also at the roots of the rationale following the trend of optimization of the distribution of codes among components of the matrix so that minimization of the consequences of errors would be obtained [43, 65]. The observational basis for this approach is that the distribution of amino acids is reasonably clustered with respect to similarities of their properties so that point mutational or translational simple changes would most frequently not result in drastic alteration of the character of the amino acids and of their effects on protein function. The optimization/minimization rationale should also be understood as nonabsolute (especially at the early times of formation of the code) since organisms are typically adapted to a variety of environmental contexts, always with fluctuations of varying degrees and usually not optimized to a specific environment, to the cost of losing adaptivity. In fact, statistical tests demonstrate only a near-optimality of the code, which is consistent with studies centered on the adaptive landscapes of organisms [50]. This same rationale applies to the mechanisms for development of the near-universality of the code, which is indicated to have been derived from the widespread occurrence of horizontal gene transfers especially at early times of formation of cells [95]. In an evolutionary network perspective, minimization/optimization is understood as the gradual construction of networks [2, 14], starting with the earlier codes that become more densely connected than the last codes. A drawback to the application of selective optimization protocols in attempts to decipher the origins of the code is that they cannot consider specific mechanisms of its formation and that they have obligatorily to choose among a multitude of characters of the system some that would seem most relevant for the tests of optimality, that is, ′what to optimize for?′

(b) The Stepwise Addition of Codes

When the process of building strings is seen stepwise, the main problem resides in finding the initial segment of Ariadne′s thread that could guide the way along the matrix of correspondences, now envisaged as a labyrinth metaphor for the chains of triplets in the genetic strings. The gradual construction should ideally try to delineate biochemical mechanisms for encoding that could be valid for all single-letter attributions and should also attempt to eventually come to the definition of an encoding–decoding system, with its own rules of integration and regulation. When systemic integration is obtained, this approach should reach the block structure from the bottom up. Some hints at regulatory mechanisms include the load of uncharged tRNAs in the pools [18] and possible effects of tRNA dimerization through pairing of the anticodons [61, 100]. The constructive biochemistry-centered approaches also incorporate selection of the adequate amino acids and triplets all along the process, including gradual adjustments for mutual adequacy of neighbors, therefore not questioning the optimization/minimization process.

The block structure and the stepwise addition approaches converge partially when the former attempts to define some small sets of triplets or amino acids as initial among the full sets [43]. Investigations on the sequential order of entries into the code tend to focus more strongly on the amino acids, considering aspects of the corresponding triplets secondarily. This line follows the concept that the meanings are more important than the codes and a prevailing tradition that the correspondences are arbitrary or ‘accidental’, assuming no chemical relatedness between triplets and amino acids; this is akin to the linguistic concepts where arbitrariness is the rule, except for only a few onomatopoeic or iconic correspondences. The biochemical support for the arbitrariness would be the distance from the acylation sites (the 3′ terminus of tRNAs, where amino acids are attached) to the anticodon positions in present-day tRNAs plus the lack of physical contacts between anticodons and some of the aRS. Otherwise, it should be considered that there are evidences for nonarbitrariness, such as the hydropathy correlation [27], indicating that the present day lack of direct contacts between codes and meanings would have been introduced after the initial encoding process; it is supposed that at encoding times, contacts or associations between acylation and anticodon-correlated sites would have existed, thereafter having been lost or modified [78].

Order of Triplet Encoding

Information on this side of the correspondences is rather difficult to obtain from mutational studies, these being widely position independent and nondirectional. Searches for patterns in the matrix of triplets follow the block structure approach and have explored a large variety of procedures, usually symmetries [45] and some sophisticated renormalization rules but, in spite of the appeal provided by some mathematical regularity [48], they have yet to prove relevant to biochemistry. There are prospects for applicability in string construction rules [24].

A question rarely discussed is whether there were to start with a complete set of triplets, the full space of 64 empty rooms (triplets) to receive (encode) the amino acid guests, or a limited one. In spite of not being necessary to the proposals, it seems that the block structure approach would be more akin to the idea of preexistence of large and diverse pools and the stepwise addition approaches more compatible with low diversity pools. A reasonable supposition would be that the early mechanisms of producing strings were highly nonspecific so that it would be expected that the pools would reach high diversity rapidly. In this case, problems of interference or competition between analogous or similar triplets might be envisaged. It would be more in accordance with experimental synthesis of strings that early sequences would be simpler and of a more monotonic kind, again highlighting the possibility of interference or of incompatibility between different kinds of monomers when they are utilized as substrates for some simple polymerization mechanisms [21, 22]. Low diversity would also be more compatible with the known difficulties in the chemical synthesis of nucleotide bases [74, 86].

Enchaining Encoded Triplets

Translation in cells utilizes mRNAs whose enchained triplets can be decoded after the adequate initiation binding to ribosomes. These harbor a duplet of pre-encoded tRNAs that enchain their attached amino acids (at initiation), or an amino acid attached to the incoming tRNA and a peptide chain attached to the previous tRNA (elongation). The tRNAs entering the ribosome are chosen for protein synthesis according to the quality of the minihelices their anticodons form with the codons of the mRNA. The rRNA contributes with the transferase activity and, via the decoding site, to the quality-checking process.

From studies on the origin of translation, the block structure rationale does not need to be involved with explanations on the enchaining of triplets, since it relies upon pre-existent long RNA chains to be translated. It is devoted to explain the overall structure of the decoding system, also not focusing on details of the origins of encoding the tRNAs and of the ribosomal machine. Most refined versions of the stepwise addition rationale should propose to start the genetic system only with encoded triplets belonging in oligomeric proto-tRNAs. Justifications for this include the observed difficulties in obtaining reasonably long RNA chains experimentally, even when starting with pre-made nucleotides in adequately activated forms [21, 22]. Production of chains, e.g., from catalysis directed by mineral surfaces, has been shown productive but rarely reaches oligomers longer than ~20mer. Prebiotic formation of amino acids is known for a long time and through a variety of procedures, as well as their spontaneous enchaining into peptides, from simple heating and drying experiments to mineral surface catalysis [11, 54].

Prospects for Early Poly-tRNAs

Starting with proto-tRNA oligomers seems necessary but faces the two problems of encoding and of derivation of long RNAs from the oligomers, including the rRNAs and mRNAs. The tRNAs have been considered possible early genes [20], genomes have been found largely punctuated by frequently clustered tRNA genes [83], which led to the 'genomic tag' hypothesis [56], and they participate in a variety of other functions [16]. The perspective of starting the construction of genomes from initial poly-tRNA has received some attention. Enthusiasm with the possibility came up from searches on tRNA-rRNA homology [9, 10], but the statistics backing the results was not considered definitive by the larger community. Otherwise, the idea of start building genomes through enchaining of small pieces bearing themselves some function (e.g., mini-exons, mini-introns, micro-RNAs etc.) is appealing and might still hold so that the proposition should deserve further testing. It follows the general principle of evolution by duplications [55] extended back to the origins.

External Pushing Dynamics for Encoding

The association of the 'first principles' of physical chemistry: (a) the overall chemical stability of the amino acids, substantiated by their abundance along geochemical environmental challenges, and (b) the thermodynamics of triplet pairs, was taken to support the hypothesis that the formation of the code would have followed such pushing dynamics, arising from natural spontaneity [93]. Encoding would be forced in consequence of the chemical law of mass action, that is, by the pre-existing abundance or concentration of amino acids. Enthusiasm with this reasoning is heightened by its relatedness to the chemical evolution principle, since the list of pre-biotic amino acids contains exemplars which are among the generally simpler in structure; these compounds would be easier to form due to, e.g., lower energy expenditure [44], and this character is usually combined with greater overall stability.

The Rows of the Matrix

Main results from this line of investigation converge in pinpointing an early set of pre-biotic amino acids headed by the most abundant Gly and Ala, followed by Val and Asp, and the $3'$ row of anticodes [43, 93]. This row would provide a set of triplets with high thermodynamic stability at forming codon-anticodon pairs due to the constancy of the $3'$C, especially when the $5'$G can be chosen from the triplets in the boxes, and even more so when the central base is C or G: Val-G$\underline{AC}$, Ala-G$\underline{GC}$, Gly G$\underline{CC}$ and Asp-G$\underline{UC}$. It is noted that triplets in the $3'$G row would offer the same thermodynamic possibilities of the $3'$C row, as well as the central G and central C columns, but the amino acids in these sets are not among the pre-biotic abundant.

Endogenous Components

Only some hints at thoughts about a significant role for the products of the code on the formation of the coding system can be gathered from the vast array of reports. It has been suggested that the first protein phenotypes would have been the unstructured coils and turns [49], composed mainly of amino acids extracted from the set belonging in the –YY quadrant of anticodes (GDESRNK). Other studies suggest early phenotypes to be composed of strands and sheets, due to their greater stability [28, 58]. Supporters of the pushing dynamics indicate that the large variety of properties of the most abundant pre-biotic amino acids (e.g., the 3′C row VAGD) would enable the production of correspondingly diverse phenotypes to be subjected to selection.

The Columns of the Matrix

An appealing observation is based on the distribution of the synthetase classes in the matrix. The aRS classes form distinct homology groups of enzyme sequences but both adopt the same chemical mechanisms for activation of the amino acids, except for the preferential ribose site to receive the amino acid: class I binds amino acids (LVIMCWRYQEK) to the 2′OH and class II (FSPATGHDNK) to the 3′OH. This is derived from their mirroring mode of docking the tRNA acceptor arm, from opposite sides [26]. They also depict unique specificities for the central purines of triplets: the entire central G column corresponds to class II and the central A column corresponds to class I, including the atypical PheRS whose sequence is class II but the acylation mode is of class I; the central Y columns show mixtures of aRS classes (see Figs. 1b, 4). LysRS is also atypical due to being class I in some organisms, class II in others, and the only class II in 5′Y anticodes of complex boxes, together with the recoded Selenocysteine (Sec) and Pyrolysine (Pyl) [71].

The hydropathy correlation is another attribute of the matrix that follows the column distribution: A is the most hydrophobic base and amino acids in the central A column are most hydrophobic, while U is the most hydrophilic base and amino acids in the central U column are most hydrophilic. The correlation is seen utilizing hydropathy data on free amino acid molecules in solution, but a revision of this correlation utilizing amino acid residue hydropathies showed neat improvements and pointed to the correlation being obtained through the action of the synthetases [27].

Rows Plus Columns

Attempts at unifying the two orthogonal modes of organization of the matrix have been pursued. The work of Higgs started with an energy-based investigation of the construction of amino acid molecules followed by application of the optimization rationale to define the early set to have been encoded [43, 44], leading to the proposal that the 3′C row attributions were the heads of the four columns. Trifonov's approach was of obtaining a consensus among the large variety of proposals on the early sets of amino acids, which again pointed to the list of pre-biotic abundant [93]. Addition of the thermodynamic rationale to the list allowed the pinpointing of the two most abundant amino acids and the two triplets that could form pairs with highest stability: codons Gly-C$\underline{\text{GG}}$: Ala-C$\underline{\text{CG}}$. These triplets would belong in the two strands of a coding mother-helix (following earlier suggestions [77]) and each of them would have become the head of a family of encodings [92, 93], the Gly and the Ala families.

Metabolic Supply

A new line of studies on the code origins was introduced by Wong [96, 97], noticing that the families of amino acid derivations—groups of amino acids where some require others as precursors in biosynthesis pathways, correspond to triplets that form groups of similarities, differing from each other by few changes, that is, there is mutual mapping between biosynthesis pathways and triplet structure. His model of metabolism/code co-evolution proposes that transformations of amino acids leading to such mutual mapping would derive from their occurrence upon amino acids bound to tRNA (aminoacyl-tRNAs) and that mutations on the anticodons would be distributed between the previous and the transformed amino acids.

The merit in this line of investigation (see also [15]) is in proposing links between the two 'first principles' of biology—metabolism and the genetic strings, but it was followed by much criticism and it is still looked at with skepticism [1, 79]. The metabolic order does not overlap well either of the two orthogonal orders, in spite of the tendency of the amino acid biosynthesis families to follow more extensively some rows and less extensively some columns. Examples of the mechanism of transformation of amino acids as parts of aminoacyl-tRNAs have been documented [16, 97] only for the origins of fMet (the formylated Met-tRNA^{i} in bacteria), Asn, Gln, and Sec.

Latest versions in this line either reinforce the 3′C row proposal [17], adding Ser to the set of five amino acids (GASDEV), based on the known metabolic inter-conversions between Gly and Ser, or add still others to complete a set of up to ten Phase 1 amino acids (added Leu and Ile, plus marginally Pro and Thr; [97]). The Phase 1 set would be possibly obtained from pre-biotic chemistry or in the heterotrophic scenario. This scenario for the early metabolic routes is centered on the dependence on the nutrient glucose, utilized through glycolysis, the Pentose Phosphate Shunt, and the Citrate Cycle, from whose metabolites the amino acids are derived. Metabolism would initially substitute the external pre-biotic and follow adding others, including the Phase 2 set, for which there are no indications of pre-biotic origin. The overlay between the lists of pre-biotic and of the co-evolution theory enabled the consensus obtained by Trifonov [93] and reinforced the pushing dynamics rationale for the formation of the code.

A main criticism to the co-evolution model, in the form proposed by Wong [96, 97] and Di Giulio [16, 17], is based on the observations that metabolic connections are mostly multiple and highly reticulated, forming networks, so that time-ordered successions based on specific choices of some linear pathways might not be reliable guides to infer precursor-product derivations that would have been relevant for the encodings [36]. The SRM follows the co-evolution idea but adopts less compromised and simpler assumptions: (a) any amino acid entering the code at a successive step should have its main family precursor already encoded in a previous step, not choosing a specific pathway; (b) the specific mode of amino acid derivation from precursor amino acids previously bound to tRNAs is not required.

The Synthetase Function

Aminoacyl-tRNAs are the only mixed molecules of the genetic system, containing an amino acid attached to an RNA. Specificity is obtained from complex interactions between enzymes and the substrates amino acids and tRNAs. There are two experimental routes investigating the origins of the specificity: (a) the RNA World hypothesis and (b) the inorganic catalyst precursors to the synthetase function.

The RNA World

Sophisticated research stemming from the RNA World hypothesis has advanced much on details of stereochemical specific binding of amino acids to pocket sites formed by RNA strands but the results are difficult to reconcile with the most common trends based on the order of amino acid encoding. There is in some cases consistency between the composition of the RNA-binding site and the code, but pointing sometimes to the codons other times the anticodons for the bound amino acids, and a large part of the data refer to complex amino acids, usually considered late entries [46, 101]. Some possibilities of metabolic derivation of amino acids from nucleosides have been studied [13], but the main problem yet to be solved is how the complex nucleotides would have arisen from geochemistry, in spite of some appealing advances in chemical syntheses [74]. There is always the possibility that the nucleic acid part of the code has been preceded by an RNA-mimic polymer [63] that would have been involved with the binding of amino acids and the installation of the primitive protein synthesis machinery. In this case, the RNAs were not original entities, but derived and dependent on proteins, having arisen inside and along the construction of an RNP system. Such questionings do not diminish the relevance of the participation of RNA in the early developments of metabolism and also do not question the precedence of RNA over DNA.

Acylation by Proto-Synthetases

It seems more fruitful the investment on some early amino acid acceptor that would look more like the known tRNAs—e.g., some form of mini-tRNA-like oligomers, proto-mini-tRNAs—where the amino acid is bound to a tail instead of a complex pocket of the RNA [8]. The chemistry involved in the acylation reaction is simple nucleophilic attack, provided that the steepest thermodynamic barrier has been previously overcome, which is the amino acid activation step [84, 88, 89]. In cellular protein synthesis, this is the first step of the aRS activity, forming an aminoacyl-phosphate bond at the expense of one ATP. Following reactions proceed thermodynamically downhill, through the intermediate aminoacyl-ribosyl (at the tRNA tail), which is an ester bond catalyzed by the aRS (the second and specific step), to the final amide or peptide bond, catalyzed by the ribosomal transferase. In studies on the origin of coding, the first aRS reaction and the transferase reaction are not usually considered since they are generic and nonspecific.

Before the advent of aRS proteins, the acylation reaction could be spontaneous or catalyzed by minerals (Ni is presently one of the favorite candidates; [12, 40]) and would be largely nonspecific, possibly directed by the availability of amino acids—the pushing dynamics—and by the not strongly selective chemical affinities. This could have given rise to some proto-codes [36], reserving the term code for the biologic, which is tRNA, ribosome, and protein dependent. It is still debatable how much continuity would be observed between proto-codes and the biologic code. An intermediate step would be obtained from early peptides, either pre-biotic or at the beginning of formation of RNPs, which would also start largely nonspecific and later develop specificity [11, 72]. Work with mini-tRNAs have concentrated on the mini-helices that mimic the acceptor stem plus the tail of tRNAs and found that this piece alone, lacking the other three arms, can be an adequate substrate for the acylation reaction, with preservation of the specificity [8]. This became known as the second genetic code, aside with the first that refers to the anticodons and codons, in spite of the evolutionary order most probably having been the reverse. Perspectives for work in this area should then concentrate on interactions between the acceptor segments of proto-tRNAs and the peptides that could propitiate the acylation reaction, together with the existence of other segments in the same proto-tRNAs that could acquire the anticodon function; furthermore, the two segments should demonstrate correlated structures and functions.

Pulling Dynamics, Anabolic Drive

The search for biochemical correlates to the long questioned and enigmatic vital force [6] was unfruitful and the concept has been discarded [82]. A substitute for the vital force became the set of auto (self) prefixes for a variety of the behaviors presented by living beings suggestive of automatisms, like the autocatalysis, self-stimulation, autopoiesis, etc. [80], inviting a clear biochemical characterization and possibly settling renewed discussions on the theme. Only a partial answer to these quests may be found in the usual description of the metabolic flow. In the formation and maintenance of transformation and production systems—generically, metabolic systems—a requirement is that an unimpeded flow should be guaranteed [70]. This would have to rely upon some kinds of adjustment mechanisms to the input sources, the intermediates in the production lines and the final outcomes, which are adaptation processes. When the final outcomes are expelled and diluted out, they stop contributing to the system; their function was of guaranteeing nonaccumulation, therewith avoiding blockades in the flux. Intermediate products in the transformation process should also not accumulate and some of them might even be toxic, having to be quickly processed.

The flux mechanisms would be at work in proto-metabolic systems but subjected mainly to external influences—the pushing dynamics. Would these gradients [11] be enough, the external sources effectively pressing the formation of an organized system for their consumption and dissipation? Nonspecificity in reactions would produce various directions, creating variety, but long-term tendencies would be toward equilibration. Chemical evolution potential of significant consequences would arise from the eventual appearance of self-stimulating cycles, which are typical of network structures [23, 91]. These might enter routes of auto-catalysis but these are prone to short duration due to exhaustion of substrates or accumulation of toxic components. Some taming of these would be provided for by the stabilizing properties of lateral routes of processing—functional redundancy—in the networks but the self-stimulating property will not be reinforced if these are directed away from the cyclic cores.

Physicochemical pushing dynamics may not be sufficient to satisfy the explanations looked for by biochemical research. It is widely accepted that life's origin was driven by pre-biotic geochemical gradients but it is hard to envisage how they would lead to metabolic pathways and it is indicated that other drives prevail in the biological realm. Such questioning has been enforced (e.g., [60]), but biochemical studies have not been clear in pinpointing where would the distinctness reside, besides describing the specificity of its typical nucleoprotein constitution. The pushing dynamics rationale ultimately points to entropic degradation, traversing the equilibrium processes, and nothing more than the four fundamental forces of physics is known. Otherwise, a main feature of biological networks is the dominance of feed-forward configurations [2, 14, 62]. They would be representative characters of the long known far from equilibrium or dissipative states of bio-systems but where are these structures pointing to, what are they feeding-forward to? No external guidance having been found, the answer has to be internal and endogenous.

Ribonucleoproteins in Protein Synthesis as Metabolic Sink

Productive self-stimulation depends on the products of a reaction being able to feed back upon the producers with some help in the production process. The seeds of a *bona fide* system arise when the self-feeding aggregate of producers and products reaches stabilization and unimpeded functions. In the biological context and with special interest to the origin of the genetic code, this rationale, obviously inspired on Eigen's hypercycle [19], says that the protein products—among a variety of possible products invented by a proto-metabolic system—were able to join their producers and start the formation of a system for their production. It is even possible that the producers were not exactly of the nucleic acid kind, having been driven toward these by the proteins, through mutual adjustments. The search of Noller [64] for a translation drive is RNA-centered and considers proteins amplifiers of the limited RNA functional abilities. Our proposal is systemic observing that the drive was installed at the proto-tRNA—protein associative and productive event, not choosing one of its components at the cost of the other; the drive is nucleoproteic ab initio. When the system reaches stability and maintenance of the productive cycle along the flow of time, it is said that it acquired memory properties. In the biological case, besides the dynamic memories of the productive cycles [87], a new kind was added, physically implemented through the template-dependent replication of nucleic acids.

The internal drive would reside in the collective of the metabolic pathways pointing to and feeding-forward in the direction of the protein synthesis process, a crucial point in gene expression. This involves RNPs, either as aggregates such as the ribosomes or as cytosolic components such as the tRNAs and the aRS. Search for origins should then concentrate on the RNPs. Proteins are majority in cell mass and in constructing phenotypes, overwhelmingly the main components of metabolism and of structures. RNAs are second in mass, the direct producers of proteins and active in much of the regulatory mechanisms, which are shared with the proteins even when the RNAs are ribozymic. DNA is the most important memory component of cells, from which RNAs are derived through transcription. The temporal order at the origins of the system, when RNAs would have accomplished the roles of memories, later transferred to DNA, is the reverse of the gene expression order. Other components, such as lipids and carbohydrates, are products of proteins and, even when polymeric, are not parts of the genetic macromolecules. The diversification and accumulative power of proteins are enormous, by themselves or as parts of RNPs or other aggregates; the nontoxic kinds were selected for; their stability would force reactions in the anabolic direction and this is further guaranteed by the irreversibility of translation, while replication and transcription may present reversible directions. Saturation or other challenges to the sink dynamics would be triggering mechanisms for installment of modifications of the plainly accumulative growth regime such as reproduction or formation of resistance states [18, 47].

The cellular system may be described as mainly devoted to the synthesis of proteins and these would be principal sinks of the anabolic flux. The consideration of living systems as sink-driven identifies a possible biochemical correlate to the long questioned vital force. This is an abstraction referring to the metabolic dynamics that creates suction potentials, which are analogs to other dissipative systems such as the eyes of cyclones. Departing from the psychological and nearly mystical connotations of vitalist propositions, the unique biological motive force can receive a naturalized definition in terms of the *anabolic drive*, pulled by the protein synthesis system. Manifestations of the suction dynamics at the frontiers of the system are found in the simplest form in the facilitated diffusion mechanisms of uptake of matter and energy. Proteins are very sensitive to environmental influences and developed some sophisticated sites working as receptors for materials of value to the system. As soon as some of these make contacts with the receptors, they are bound and become trapped inside via transformation into metabolites. Their concentrations in the immediate vicinity of cells are maintained at low levels therewith propitiating empty room—as if creating vacuum—to be replenished by diffusion.

Proteins Organized the Code

Being the genetic code a central character of the living as a main component of the translation machinery, it is necessary to discern in its structures and functions, characters that could be candidates for the self-centered properties. Such attributes are found when the code is considered a performance- or fitness-driven circular system. The origin of the code is envisaged as having been set inside a proto-metabolic geochemical system where production of a limited variety of oligomeric strings started, some of which serving the function of proto-tRNAs. Best candidates for this guiding role are mineral surfaces [21, 22, 42]. The process would be thermodynamically favorable in the sense of accomplishing reduction in the summed amounts of the free energies in the free floating monomers at oligomerization. Pre-biotic replication of some of the oligomer types would follow the same rationale, but is not strictly required in case the mineral-directed production would have been efficient. Associations between different kinds of oligomers would be occurring, again in accordance with thermodynamic principles, and some would have followed paths of mutual stimulation, when they were stable against degradation and did not impede the workings of the other components of the aggregates. Stabilization of the aggregate alone would result in greater productivity of some of the components. Development of specificities, such as the aRS functions, would be improvements in these same directions, followed by the introduction of replication and metabolic abilities (Fig. 2). The constitution of proto-tRNAs is not known and they could be only RNA-mimetic; it is considered that the complex structure of RNA is derived from enzymes, inside the RNP world.

When the protein-drive is recognized as a main 'force' in the formation of the living, the problem of formation of the code becomes centered on investigating which were the characters of proteins, among a multitude, which should be looked at preferentially as guides for walking safely along the entangled labyrinth formed by the enchained strings. The simplest of these properties should be, e.g., the stability of proteins against degradation and their ability to bind to RNAs (or proto-tRNAs; [37–39]). These two properties are considered minimal requirements for the formation of RNP (or proto-RNP) aggregates, which would be the seeds for the formation of the translation system. The rationale is that structures and functions of the proteins, dependent on specific amino acids and for the functions of forming the aggregates with the proto-tRNAs, created necessities (pulling dynamics) for those amino acids. Some of these would be chosen among the available pre-biotic pools. Metabolic sources substituted the pre-biotic sources and later added new kinds [36]. Accordingly, the metabolic pathways were fixed as responses to protein needs, especially for the formation of productive RNPs, and departing from the proposals that the pathways obeyed external influences thereafter propitiating the fixation of their products into proteins. Among a variety of amino acids offered by the primitive metabolic pathways, some were selected to take part on the protein stretches being formed, through criteria of fitting the functions of production of adequate RNP aggregates. It is considered that when a product is to be introduced in a system, it depends on the previous existence of consumption mechanisms, otherwise it would accumulate and block the production.

The Self-Referential Model

Fundamental indications of the SRM are: (a) protein synthesis started directed by dimers of charged proto-tRNAs, held together by the paired anticodons. The dimers are analogous to the mRNA-tRNA pairs at translation, one anticodon serving the function of codon to the other, and to ribosomes, structures where two tRNAs are hosted and whose tails can reach each other closely therewith facilitating the transferase reaction. This mechanism is clearly in the self-organization realm of processes [4]. (b) Cycles of dimer-directed peptide synthesis would be subjected to selection for the kinds of peptides that could maintain and, through stabilization, improve the productivity. This would be obtained only under maintenance of the self-referential condition. Anticodon complementariness is the first instance; in the dimers, tRNAs recognize themselves as a class of molecules. The second instance involves different classes of molecules but still requires selfing: the peptides would be able to bind to the dimers that produced them, the products recognizing their producers and the aggregate RNP remaining functionally apt.

The result is the selective increase in the RNP production. Were the process not self-referential, peptides not being able to bind and stabilize the same dimers that produced them, activities would follow dispersive routes, instead of the self-stimulating or convergent where producers and products recognize each other in mutuality to configure a self-feeding (proto)RNP system. Starting in the direction from producers to the peptides, self-reference would mean that the products would be recognizing the producers. Stabilization of RNPs would be able to lead to fixation of attributes of the partners that lead to maintenance of the productive cycling, that is, variants of the partners leading to productivity of the aggregate become progressively more abundant in the pools, to finally obtain specificity in the correspondences and improved productivity.

Self-reference is a complex term with different meanings according to the area of application, but it is being frequently utilized in different levels of biological organization [4, 69, 80]. It would belong among important factors responsible for evolutionary potentials, e.g., through adaptations via endogenization or internalization mechanisms. When external regularities are sensed by organisms, some of their internal mechanisms are adjusted, from repression of some routes to amplification of other routes, so that behaviors become ecologically adequate and may even lead to anticipatory behaviors, such as in the evolution of regulatory cascades. The encoding cycles would belong among the most basic kinds of biological self-reference. The nascent peptide with strong RNA-binding properties would stay held together with its producer, not being easily lost to the medium. When peptides are released, it is still possible that they could contribute to build self-stimulating cycles, and with the eventual benefit of population variety, but the process might run into problems of dilution of the effects, when some of the bindings would be directed to nonself-stimulating outcomes.

Pre-Biotic Continuity and Metabolic Autonomy

The main (a) pre-biotic to biotic amino acid substitution event identified by the SRM was of the simplest of all amino acids Gly, which stands up as the principal remnant evidence of continuity between the early peptide components: it is among the most abundant pre-biotic and the first to be synthesized by metabolism and encoded [36]. The second amino acid encoded according to the SRM is Ser, present albeit not abundant among the pre-biotic but definitely backed by metabolic evidence. The Gly-Ser anabolic pathway starts with the synthesis of the two-carbon (C2) amino acid Gly from $CO_2 + C1$-tetrahydrofolate; Gly receives another C1 to form Ser C3 and the pathway grows up to the formation of C4 acids, precursors to Asp and to components of the C4 side of the Citrate Cycle. Other C2 compounds can derive from the Acetyl-CoA Pathway of anabolism and the C2 Glyoxylate is also part of the Gly-Ser Cycle. This panorama for the early metabolism is inserted along with the proposals for the first cells being of the autotrophic-methylotrophic kind [57, 98] and indicates an early take up of metabolic autonomy in the system, departing from the pushing dynamics. In the mechanistic context (b), the SRM proposition of the proto-tRNA dimer-directed peptide synthesis should be general and applicable to both pre-biotic and biotic realms, which could be experimentally tested. A second component in the path toward acquisition of increased autonomy is (c) the advent of the fully double-stranded configuration of the nucleic acids in DNA.

Why, How, and Tests

The indications above propose answers to the questions of why the code was formed—because there were free energies to be dissipated—and how—through the dimer-directed protein synthesis mechanism. The propositions should be immediately amenable to experimentation. While the precise nature of proto-tRNAs is not known, it is proposed that tests could utilize mini-tRNAs of kinds similar to those already known, the acceptor stem and tail analogs [8], but receiving the addition of anticodon loops, to provide dimerization ability, and possibly some adjustment at the tails, for flexibility and facilitation of contacts to propitiate the transferase reaction. Test-tube compositions of mini-tRNAs bearing complementary loops could be designed and put to evolve spontaneously with the perspective of obtaining truly synthetic nucleoprotein systems. While longing for experiments, a collection of indirect tests of consistency is being pursued, between predictions of the SRM—some of them quite rigid—and empirical observations on protein structure and function. The work is mostly qualitative and generally independent from sophisticated statistics, since the number of components in the code matrix is low. The SRM propositions are fully consistent with empirical data and follow the general evolutionary principle of construction of complex forms from simple starting structures and functions.

Objectives

The following report concentrates in presenting new data on (a) the overall structure of the basic networks and subnetworks of triplet dimers, including some thermodynamic properties, and to propose mechanisms that could have led to the encoding processes. Details on protein and nucleic acid biochemistry that give support to the SRM are referred to other publications [36–39]. The other extreme of the network structure is also presented, referring (b) to the aRS aggregates (MaRS) found in the large eukaryotic cells, which is devoted to the function of integrating the dimer networks. It is expected that the challenges of deciphering the mechanisms of encoding and of integrating the system could be subjected to modeling and simulation, which would help clarifying the biologic problem, as in silico tests of some of the results and predictions of the model (see Section Questions to be addressed by bioinformatics). Such enlightenments would offer guidance for the proposed experimental tests that can be utilized for obtaining synthetic genetic codes.

Dimer Networks

Dimerization through pairing of anticodons leads to the formation of two large networks due to the possibility of the lateral positions forming pairs of the R:Y kind, which are analogs to the wobbling. In this configuration of the duplexes, only the central base pair remains with full capability of maintaining a Watson–Crick type of bonding, which is considered a prerequisite for stability of the dimer. In consequence of the restriction on the central base pair, the matrix is divided into two independent networks with identical topology: a central G:C and a central A:U network. The wide pairing abilities allowed to the lateral bases lead to each triplet participating in four dimers.

Subnetwork Modules

Each of the networks is composed of four subnetworks, again isolated from each other (Tables 3, 4). Formation of the subnetworks derives from combinations of: (a) the division of the triplets in each box into halves of different kinds: the nonself-complementary (NSC), with lateral bases both R or both Y, and the self-complementary (SC), with one lateral base R and the other Y; (b) the homogeneous or mixed character of the pDiN. In each network, there are two subnetworks containing pairs of NSC triplets: one with triplets of the homogeneous pDiN sector (R$\underline{\text{RR}}$:Y$\underline{\text{YY}}$; the pDiN bases are both R or both Y), the other with triplets of the mixed pDiN sector (R$\underline{\text{YR}}$:Y$\underline{\text{RY}}$; the pDiN have one R and one Y base). The other two subnetworks contain SC triplets: the distinction between these derives from the different combinations possible in a SC triplet, one kind with triplets 5′RNY3′ (R$\underline{\text{RY}}$:R$\underline{\text{YY}}$; SC-5′R) and the other triplets with 5′YNR3′ (Y$\underline{\text{RR}}$:Y$\underline{\text{YR}}$; SC-5′Y). Members of the NSC subnetworks are intra-sector while in the SC they are inter-sector, the latter providing integration inside each of the networks. The previously proposed constitution of the modules of the mixed pDiN sector [38, 39] is now revised, based on the strict composition of the subnetwork modules.

The distinction of the NSC and SC triplets is the basis for the splitting of all boxes into halves. The above plus the distinction of kinds of sectors are crucial to the SRM with respect to the triplet-related aspects of the mechanisms of encoding. Otherwise, aspects related to the aRS specificities follow the aRS-pDiN rule of degeneracy, which erase

Table 3 Modules of anticode dimers including the 5′A triplets

Triplets	CGG CAG	UGG UAG	GGG GAG	AGG AAG	CGA CAA	UGA UAA	GGA GAA	AGA AAA		CGC CAC	UGC UAC	GGC GAC	AGC AAC	CGU CAU	UGU UAU	GGU GAU	AGU AAU	Dimers	aRS, C central	aRS, U central
CCA CUA												-4.0 -2.7	-1.7 -1.0			-5.4 -3.1	-2.8 -1.4	0 0	Gly	Asp
CCG CUG												-6.3 -4.2	-4.0 -2.5			-5.4 -3.4	-3.1 -1.7	2 2	Gly	Asp
CCC CUC			-6.6 -4.5	-4.3 -2.8			-4.6 -3.4	-2.3 -1.8										2 2	Gly	Glu
CCU CUU			-4.8 -3.0	-2.5 -1.3			-5.7 -3.3	-3.4 -1.6										2 2	Gly	Glu
UCA UUA												-2.8 -1.6	-2.8 -1.2			-3.9 -2.0	-3.9 -1.6	0 0	Ser	Asn
UCG UUG												-5.1 -3.1	-5.1 -2.7			-4.2 -2.3	-4.2 -1.9	2 2	Ser	Asn
UCC UUC			-5.4 -3.4	-5.4 -3.0			-3.4 -2.3	-3.4 -1.9										2 2	Arg	Lys
UCU UUU			-3.6 -1.9	-3.6 -1.5			-4.5 -2.2	-4.5 -1.8										2 2	Arg	Lys
	Homogeneous pDiN; RRR : YYY; NSC triplets									Mixed : Homogeneous pDiN; RRY : YYR; SC triplets 5′ G										
	Homogeneous : Mixed pDiN; YRR : RYY; SC triplets 5′ Y									Mixed pDiN; YRY : RYR; NSC triplets										
GCA GUA										-3.5 -2.8	-2.5 -1.7			-4.6 -3.2	-3.6 -2.1			0 0	Arg	His
GCG GUG										-5.8 -4.3	-4.8 -3.2			-4.9 -3.5	-3.9 -2.4			4 4	Arg	His
GCC GUC	-5.3 -3.8	-4.3 -2.7			-3.3 -2.7	-2.3 -1.6												4 4	Arg	Gln
GCU GUU	-3.5 -2.3	-2.5 -1.2			-4.4 -2.6	-3.4 -1.5												4 4	Arg	Gln
ACA AUA										-3.1 -1.5	-3.2 -2.0			-4.2 -1.9	-4.3 -2.4			0 0	Cys	Tyr
ACG AUG										-5.4 -3.0	-5.5 -3.5			-4.5 -2.2	-4.6 -2.7			4 4	Cys	Tyr
ACC AUC	-4.9 -2.5	-5.0 -3.0			-2.9 -1.4	-3.0 -1.9												4 0	Trp	X
ACU AUU	-3.1 -1.0	-3.2 -1.5			-4.0 -1.3	-4.1 -1.8												0 0	X	X
Dimers	3 2	3 2	4 4	0 0	3 2	3 2	4 4	0 0		2 2	2 2	2 2	0 0	2 2	2 2	2 2	0 0	32 28		
aRSctG	Pro	Pro	Pro	Pro	Ser	Ser	Ser	Ser		Ala	Ala	Ala	Ala	Thr	Thr	Thr	Thr			
aRSctA	Leu	Leu	Leu	Leu	Leu	Leu	Phe	Phe		Val	Val	Val	Val	Met	Ile	Ile	Ile			

The matrix presents all dimers that can be formed by the 64 triplets, restricted by the central Watson–Crick base pair. Delta-G values are a measure for attributing weight to the dimers, calculated from the nearest-neighbor data on Watson–Crick [99], G:U [59] and A:C [30] base pairs; the 0.4 initiation penalty was included in the calculations for self-complementary dimers [99]. The right columns and bottom rows present the sum of dimers formed by each triplet and the synthetase attribution. The order of columns is 5′ C-U-G-A for grouping the Ile codes; the order of rows follows the standard (Fig. 1). Highlighted are all dimers in the NSC modules and the dimers formed with at least one 5′A triplet. The last were deleted from the standard anticode; therefore, their dimer counts are zero and they are not shown in Table 4. Dimers formed with SC triplets are not highlighted, except those having 5′A triplets; dimers containing the X anticodes were also deleted from the standard code and count zero, but were maintained in Table 4. Details on the NSC dimers are expanded in Table 5

Table 4 Types of modules of anticodone dimers in the central G:C and in the central A:U networks

a

Sub-networks		Homogeneous Central G							Mixed Central G					
		Non-self-complementary Module 1 (Ho:Ho)							Self-complementary 5′G (Mx:Ho)					
Homogeneous		Pro	Pro	Pro	Ser	Ser	Ser		Ala	Ala	Ala	Thr	Thr	Thr
Central C		CGG	UGG	GGG	CGA	UGA	GGA		CGC	UGC	GGC	CGU	UGU	GGU
Gly	CCG										-6.3			-5.4
Gly	CCC (a)			-6.6			-4.6							
Gly	CCU			-4.8			-5.7							
Ser	UCG										-5.1			-4.2
Arg	UCC (b)			-5.4			-3.4							
Arg	UCU			-3.6			-4.5							
Mixed Central C		Self-complementary 5′Y (Ho:Mx)							Non-self-complementary Module 3 (Mx:Mx)					
Arg	GCG (a)								-5.8	-4.8		-4.9	-3.9	
Arg	GCC	-5.3	-4.3		-3.3	-2.3								
Arg	GCU	-3.5	-2.5		-4.4	-3.4								
Cys	ACG (b)								-5.4	-5.5		-4.5	-4.6	
Trp	ACC	-4.9	-5.0		-2.9	-3.0								
X	ACU	-3.1	-3.2		-4.0	-4.1								

b

Sub-networks		Homogeneous Central A							Mixed Central A					
		Non-self-complementary Module 2 (Ho:Ho)							Self-complementary 5′G (Mx:Ho)					
Homogeneous		Leu	Leu	Leu	Leu	Leu	Phe		Val	Val	Val	Met	Ile	Ile
Central U		CAG	UAG	GAG	CAA	UAA	GAA		CAC	UAC	GAC	CAU	UAU	GAU
Asp	CUG										-4.2			-3.4
Glu	CUC (a)			-4.5			-3.4							
Glu	CUU			-3.0			-3.3							
Asn	UUG										-3.1			-2.3
Lys	UUC (b)			-3.4			-2.3							
Lys	UUU			-1.9			-2.2							
Mixed Central U		Self-complementary 5′Y (Ho:Mx)							Non-self-complementary Module 4 (Mx:Mx)					
His	GUG (a)								-4.3	-3.2		-3.5	-2.4	
Gln	GUC	-3.8	-2.7		-2.7	-1.6								
Gln	GUU	-2.3	-1.2		-2.6	-1.5								
Tyr	AUG (b)								-3.0	-3.5		-2.2	-2.7	
X	AUC	-2.5	-3.0		-1.4	-1.9								
X	AUU	-1.0	-1.5		-1.3	-1.8								

The central base pairs distinguish two large networks of identical topology, central G:C (a) and central A:U (b). Each of these is divided into four subnetworks. There are two subnetworks formed by the nonself-complementary triplets, inside the homogeneous and the mixed pDiN sector. The self-complementary triplets are of different kinds and form other two subnetworks: one combines the 5′G self-complementary triplets of both sectors, the other the 5′Y self-complementary triplets of both sectors. Highlighted (see Table 3) in the NSC modules are the first encoded (the two dimers at the *top left* corners) and the second encoded (the two dimers at the *bottom right* corners). The four other highlighted (*top right* and *bottom left* corners) are the competing NSC dimers that did not become encoded at this step. The SC dimers (not highlighted) also did not participate in the initial encoding mechanism the NSC/SC distinction; this is evident in the development of the simple box degeneracy where the role of the pDiN-correlates of specificity (which are distributed in the tRNA molecules) is dominant and the kind of 5′ base irrelevant.

All subnetworks depict a fully symmetric configuration with 16 pairs of triplets, since a triplet can form pairs with the other four triplets with complementary central bases. Such monotonic symmetry could, in principle, possibly offer a way for obtaining encodings propitiated by the thermodynamic stability of some pairs relative to others in the module. To the contrary, knowledge from the real anticode sets indicates that such physical principles were not enough or could even have been impeditive to obtain solutions through a reliable process. The encoding problem was solved with the help of strictly biological means, namely the elimination of the 5′A anticodons, which facilitated the initial encodings in the NSC subnetworks.

Elimination of 5′A Anticodons

The largest anticodon set (not counting the iMet) is the 45 of eukarya: 64 minus the three stop and one 5′A from each box [33]. This is considered the standard code, dominated by the 5′ G, C, and U constitution, and requiring the G:U base pair. In other kingdoms, there are further reductions. Modifications of 5′ bases are widespread, usually introducing restrictions (increased specificity) upon the wobbling range and limiting the experimental dimerization studies [34]; there are rare cases of specific formation of 5′A anticodons [68]. Splitting of boxes beyond the halving dictated by the NSC and SC triplets derives from the utilization of the mono-specificity of 5′C.

Biological explanations are rarely mono-factorial and, accordingly, the 5′A elimination can be understood as resulting from at least two factors acting in concert. Simple thermodynamic reasoning based on the weakness of A:U pairs, relative to the stronger G:C pairs, are not consistent with the maintenance of the even weaker 3′A:5′C or the 5′G:3′U pairs. (a) The maintenance of 5′A would result in decoding ambiguity in complex boxes (it wobble-pairs mostly with U, C and G [68]), so that mutations introducing 5′A would be continuously appearing and being eliminated, ever since the code was formed and up to the present, in a manner analogous to the termination suppressors [5]. (b) At formation of the code, 5′A elimination contributed with simplification and introduction of asymmetry in the NSC subnetworks, the symmetry-breaking helping to propitiate the encoding process. The SC-5′Y subnetworks remained untouched but the SC-5′R suffered a drastic reduction in size, while still remaining symmetric (Tables 3, 4). The 5′A elimination should have been triggered already at Module 2 + (see below), together with the entry of the first class I synthetases (GluRS/LeuRS) and the formation of the first complex box (Asp/Glu).

Encoding in Four Steps: Sectors and Modules

The clearest biochemical evidence for the SRM is the hydropathy correlation (see plot in [36]), which indicates a precise temporal succession of encodings corresponding to the NSC modules. Dimers formed with NSC triplets are more stable than those containing SC triplets due to assuming a definite configuration, while the SC allow for a variety of configurations [99]. The precise predictions with respect to early protein construction rules also posit challenges for tests of biological meaning. There are three correlation sets corresponding to the four NSC modules, each module containing two paired pDiN (boxes). These are distributed in the matrix (Fig. 1) according to sets of paired hemi-columns. This arrangement highlights the diagonal symmetry in the matrix, departing from the mere combination of rows and columns. Module 1 contains correspondences with no hydropathy correlation; the pDiN are of the homogeneous kind and highly hydrophobic (two R bases) or highly hydrophilic (two Y) but the amino acids are hydroapathetic: (1a) Gly-CC : Pro-GG and (1b) Ser-GA : Ser-CU, abbreviated as the Gly-Pro-Ser (GPS) set; Arg-$\underline{\text{CU}}$ is absent from this set. Other sets present significant correlations. Module 2 shows a moderate inclination of the regression line; both pDiN and amino acids are coherently hydrophobic or hydrophilic: (2a) Leu-AG : Asp, Glu-UC and (2b) Asn, Lys-UU : Phe, Leu-AA. The Arg-CU attribution clusters together with these Module 2 hydropathy-correlated attributions. The third set follows a steeper inclination of the regression line and is composed of two modules; the pDiN are of the mixed kind (one R and one Y base), consequently of intermediate hydropathy. Module 3: (3a) Arg-CG : Ala-GC and (3b) Thr-GU : Gys, Trp-CA; Module 4: (4a) Val-AC : His, Gln-UG and (4b) Ile, Met-AU : Tyr-UA.

It is indicated that the hydropathy correlation gathered strength stepwise, starting from the noncorrelated GPS set. The correlation could only be established when the protein constitution was rich enough to allow construction of the aRS functions, with addition of Module 2 amino acids, and became refined with the additions at Modules 3 and 4. The main punctuation boxes belong to the last pair (4b), and the third stop sign is added in the same 3′A row, these being derived from interactions with the slipped pDiN of initiation [38, 39]. The stepwise succession of modules indicated by the hydropathy correlation sets is plainly consistent with the amino acid metabolic pathways and shows that the division of pDiN (and triplets) into sectors of homogeneous versus mixed constitutions is physiologically significant: start encoding the triplets of the homogeneous pDiN sector and follow to the mixed pDiN sector.

Encoding the Four NSC Subnetworks

The encoding process combines thermodynamic stability of triplet pairs and the aRS-pDiN degeneracy and is applicable equally to the four modules (Table 5). (A1) In each NSC subnetwork, the pair with highest thermodynamic stability (GNG:CNC; at the 3′G and 3′C rows) has its two encodings propitiated. (A1′) Occupation of the triplets in this pair reduces the concentrations of the four other (conflictive or competing) pairs they would be able to form. (A2) The pairs remaining abundant in the module—of low thermodynamic stability—can then receive the second encodings (GNA:UNU; at the 3′A and 3′U rows). (B) In both cases, the aRS-pDiN degeneracy leads to full occupation—tetracodonic—of the triplets in a box with the initial encoding. All along the process of encoding, occupation of SC triplets is generally through expansion of the initial encoding of the NSC triplets, via the aRS degeneracy. The only direct entry of new amino acids into SC triplets is in the complex boxes of the YR quadrant (Trp, Gln), together with the X codes (Fig. 1).

Table 5 Encoding the nonself-complementary modules

(A) Encoding at the initial pairs of triplets			
(A1) First triplet pairs at the first pair of boxes in each module YNC:GNG			
Homogeneous pDiN sector		**Mixed pDiN sector**	
Module 1a	Module 2a	Module 3a	Module 4a
CCY Gly GGG Gly Pro	CUY Asp Glu GAG Leu	CGY Ala GCG Arg	CAY Val GUG His
(A2) First triplet pairs at the second pair of boxes in each module YNU:GNA			
Module 1b	Module 2b	Module 3b	Module 4b
UCY Ser Arg GGA Ser	UUY Asn Lys GAA Phe	UGY Thr GCA Cys	UAY Ile GUA Tyr
(B) Expansion of initial codes to all triplets in a box via the aRS-pDiN degeneracy			
(C) Concessions and substitutions			

Each module contains six triplets forming eight exchangeable or competing dimers. Modules of the homogeneous sector have two central R and four central Y triplets; in the mixed sector, four central R and two central Y triplets. (A) The delta-G weights direct the first encodings to the high stability pairs GRG:YYC homogeneous, GYG:YRC mixed. Other pairs these triplets would be involved with are reduced in concentration, allowing the other dimers to reach higher concentrations and become encoded: GRA:YYU homogeneous, GYA:YRU mixed. (B) Full-box encoding follows, dictated by the aRS-pDiN degeneracy. Formation of complex boxes results from (C) concession of some triplets to new occupiers. In the homogeneous sector, some of the initial encodings correspond to triplets that, in the standard code, belong to other amino acids (Pro, Glu, Arg, Lys), these entering at the completion of Module 2 (Fig. 1, Table 1, 2+)

Simple and Complex Boxes

The excess of complex boxes in the standard code (eight) over the four expected from the simple counts of amino acids (20) over boxes (16) points to more complicated constraints on their formation. Biochemical details accompanying the specific location of the simple versus complex boxes in the matrix are detailed elsewhere (Guimarães, in preparation). The thermodynamic factor is summarized in the description that complex boxes correspond to triplets forming pairs of low stability (boxes at the tips of the matrix, with A and U-only pDiN) plus the triplets forming pairs with intermediate stability and bearing a central Y (-CA, -UG, -UC, -CU). Conversely, simple boxes correspond to the triplets forming pairs with high thermodynamic stability (boxes at the core of the matrix, with G and C-only pDiN) plus the triplets forming pairs with intermediate stability and bearing a central R (-GU, -AC, -AG, -GA). We concentrate here on sketching some evolutionary forces (Fig. 1).

The process of enriching the encoded amino acid repertoire was (a) driven by the protein constitution and function necessities, inside a developing RNP system (see Section Proteins organized the code), but (b) had to rely upon the concomitant development of the amino acid biosynthesis and modification routes, up to the (c) halting point where the set of 20 was encoded. The fixation of the pathways of post-translational modification of the protein amino acids eliminated the pressures and the process of encoding reached completion. A main constraint invoked would be that, (d) at times of fixation of new codes, the available set of tRNAs (boxes) was limited, which required the division of a box between a previous and the new entrance, forming a complex box.

A main constraint influencing the distribution of simple and complex boxes is the frequency of amino acid usage in proteins. The degeneracy of codes per amino acid is generally correlated with the frequency of amino acid usage, with the exception of Arg, which presents excessive number of codons relative to its usage in proteins (see [68]). The mechanism is indicated that when the first occupier of a box did not maintain abundant utilization in proteins, it could concede some triplets to new occupiers and both would reach the adequate lower degeneracy and lower usage; the opposite result would be reached when an amino acid became of higher usage, which would trigger expansion of the degeneracy.

The homogeneous pDiN sector is occupied initially by five amino acids (GSDNL). At the completion of the central metabolic routes, in the transition from Module 2 to the mixed pDiN sector (Fig. 1, Table 1, 2+), the homogeneous sector receives another five amino acids (EPKFR); its four complex boxes are filled and two of the hexacodonics (Ser, Leu) are formed. Dominant features at this transition would have been (a) adaptations to thermophilic regimes, for which Glu and Lys contribute the most [25], and (b) the introduction (or massal amplification) of the ribose-phosphate backbone of nucleic acids, especially the introduction of DNA on the previous RNP system, for which the basic amino acids (Lys and Arg) are needed. It is not clear whether the transfer of the Gly-wGG codes to the new Pro-wGG, and of the Leu-RAA codes to Phe-RAA, would follow the same forces. The mixed pDiN sector is occupied initially by seven new amino acids, besides the Arg hexacodonic expansion (ATCVHIY). After all encodings of the elongation amino acids is completed (adding MWQ; Fig. 1, Table 1, 4+) with formation of three complex boxes, the punctuation system is introduced (iMet, X) and the fourth complex box of the mixed sector is formed. In this sector, it is clear the dominance of the installation of the punctuation system in the formation of the complex boxes; the addition of Gln might also be related to this, as indicated by some variant codes where X-Y$\underline{\text{UA}}$ are translated as Gln [52].

Pre-Biotic Continuity and the Nucleic Acid Autonomy

With respect to the amino acid pre-biotic requirements, the metabolic continuity is obtained with respect to Gly, pre-biotically abundant and the first amino acid in the Gly-Ser Cycle of anabolism [36]. Referring to the nucleic acid side, it is seen that plain thermodynamic directedness is consistent with and helpful in explaining the succession of encodings inside each NSC module but is inconsistent with the fixation of Module 2 (central A:U) before Module 3 (central G:C). The SRM proposes that the precedence of the full homogeneous sector over the mixed derived from the organizing influence of surfaces on the world of single-stranded proto-tRNAs (Modules 1 and 2). This is not an entirely de novo insight; structural studies on ribosomes have been pointing to planar configurations of active sites, which would be similar to ‘membrane’ surfaces [41].

In the prebiotic context, most abundant and stable surfaces would be mineral layers, among which there are many possibilities [42]. There are more complex hypotheses on the possible intervention of organic layers, either lipidic [76, 90] or peptidic [81]. It would be required from these surfaces that they expose regularly repetitive troughs and peaks, monotonic with respect to either the depths of the troughs or the heights of the peaks, to accommodate the R repeats or, alternatively, the Y repeats of the NSC triplets of the homogeneous sector. Binding and local increased concentration of these triplets would follow, therewith facilitating their dimerization in the watery immediate interlayer surrounds. In liquid medium, extended single-stranded oligo RNAs as well as the non-chiral oligo Gly are not organized, except for the stacking strength of oligo A, but it is expected that in aggregates, the two kinds of oligomers would organize and stabilize each other. The overwhelming influence of the external template surfaces would also be responsible for the absence of hydropathy correlation among components of Module 1. The correlation was established by the aRS function of proteins at Module 2+, aside with the influence of external organization upon the triplets of the homogeneous kind. Diversification of the enzymes spreads back into the Module 1 attributions, therewith erasing traces of the original mechanisms prevailing at that stage.

The NSC triplets of the mixed sector would require surfaces with regularly alternating troughs and peaks but with different depths and heights, non-monotonic and specifically adequate to accommodate a small Y between two large R, or the complements, a large R between two small Y; it is indicated that such specific kinds of surfaces were not available at the times and contexts where the code was formed. Otherwise, the SRM indicates that the mixed pDiN sector of attributions is typical of the double-stranded fully helical DNA World. Summarizing, the scenario for the early evolution follows the stages: (a) external surfaces were necessary for organization of the single-stranded proto-tRNA world; (b) formation of the early code is the realm of RNP, where the single-stranded RNAs and the proteins organize each other in mutuality, still in the homogeneous sector; (c) the mixed sector of the code reached the stage of nucleic acid self-organization where the double-stranded DNA could get freed from the organizing role of proteins or other surfaces, one strand serving this function for the other strand.

It is recognized that the most difficult problem remains with the origin of the chemical monomers of the proto-tRNA oligomers [86]. For their oligomerization, the SRM asks the help of catalysis by mineral surfaces, for continued production (fulfilling the replication function) and possibly for the transfer and continued maintenance of some mineral order to the oligomer sequences.

Building Strings of Codes

Acquisition of the replication ability, template-directed but realized by enzymes, should (a) follow the initial simple stabilization and proto-RNA-binding role of proteins so that the products of replication are protected from degradation as RNPs; this would also facilitate the (b) second step of producing longer stretches. It would also (c) follow closely the supporting metabolic developments, all this pointing to the formation of the code having been a slow process. The large magnitude of these difficulties would be among the reasons for maintenance of the encoding process upon only one kind of modules (the NSC), which would be more expedient, not requiring extensive innovations when going from an antecedent to the successive module.

Building strings of the triplet monomers should combine the activities of (a) replicational elongation, which is typical of the replication in the single-stranded kinds of RNA, utilizing the formation of hairpin loops at the extremities that generate the primers, and of (b) ligation of segments. The first should be more prevalent inside the modules and the second at the linking of a previous to a next module. The succession of the modules encoded should be reflected in the order in the strings, the chronology of encodings (time) becoming recorded in the two-dimensional space.

The SRM order corresponds to (other characters are described elsewhere [37–39]) the start with amino acids preferentially composing the non-periodic (coils and turns) protein structures, typical of the homogeneous pDiN sector, and following a path of increasing complexity, through the amino acids preferred in helices to end with those preferred in the strands and sheets, the latter being typical of the mixed sector. Strings corresponding to the homogeneous sector would be formed through elongational replication and ligation *in tandem,* and producing arrangements *in cis*, in the RNP realm. Reaching the DNP realm with the mixed sector, additions could utilize more extensively the *trans*-acting and combinatorial less restrictive properties of DNA genomes, to extend proteins in both directions: elongation of the head extends Module 1 backwards to the central R boxes of Module 3 (AT) and Module 4 (VIM) to reach iM; elongation of the tail extends Module 2 forward to the central Y boxes of Module 3 (RCW) and Module 4 (HQY) to reach X; this order is fully consistent with the N-end rule of protein metabolic stability [94]. Predictions from the model are that the poly-tRNAs (and the corresponding peptide stretches) possibly relevant to identify remnants of the encoding process should be mostly of the GPS set, then the DEL and the NKF. Such arrangements have been found (Sobolevsky, Guimarães and Trifonov, in preparation) but are very difficult to interpret especially due to the possibility of having arisen later and configuring instances of convergent evolution (see [85]).

Integration of the tRNA Networks into an RNP Network by Synthetase Aggregation

In the tRNA realm, all triplets participate in the formation of at least two dimers, and some integration is obtained with the formation of the eight modules (Tables 3, 4). In the nucleoprotein realm, the restricted specificity of the aRS for the pDiN or for the central bases starts the formation of a system integrated at higher levels albeit not completely. The first levels of integration are seen inside the NSC modules, when (a) the sets of triplets with the same pDiN are joined into two correspondences each, and (b) when the full $5'$ degeneracy is developed, forming a simple box from the joining of the SC to the NSC triplets in the box; this may be transient in the cases where complexity arises from concessions to new occupiers. Higher degrees show up when the aRS specificity is reduced to the central bases, which are the cases of the (c) class I hexacodonics (LeuRS, ArgRS) and of the central A column and of the (d) class II attributions in the central G column. The ArgRS specificity for the central C and the aRS class specificity for the central R produce a partial integration of the pDiN sectors. Both aRS classes show wide spreads along the central Y columns and these lead to the final 70% concordance of class II with the homogeneous pDiN sector and of class I with the mixed sector.

The upper level of integration shows up when the triplets belonging to one aRS participate in the formation of dimers with triplets belonging to other aRS (Table 6, Figs. 3, 4). It is suggested that the tRNA dimers propitiate physical contacts between the synthetases bound to them and these contacts are multiplied when a synthetase shares high numbers of dimers with other aRS. Most supportive of this rationale are the couples GlnRS-LeuRS and ArgRS-ProRS and/or -SerRS, which are highly connected and associating the mixed and the homogeneous pDiN sectors. Nonetheless, these interactions are limited to the modules with one specific type of central base pair, which is a constraint assumed by the model. The connectivity of $3'$R anticodons is double of the connectivity of $3'$Y anticodons (e.g., $P = 12$ dimers, $A = 6$ dimers; $Q = 8$, $E = 4$; $H = 4$, $D = 2$), and this contributes to the hub character of ProRS, ArgRS, and LeuRS. Another contribution to the formation of hubs is the high degeneracy of the hexacodonics, which also have at least the tetracodonic boxes in the $3'$R rows. The integrative power of SerRS is higher than the other hexacodonic, with respect to the number of aRS, due to the different central bases it accommodates; the integration promoted by ArgRS is higher with respect to the number of NSC dimers formed (eight, while others are ≤ 4). In consequence of these hubs, the central G:C attributions are fully integrated into one large network, while the central A:U remain divided into two subnetworks. The most isolated aRS belong in the central Y corners of the matrix; at present, we can only notice the relevant association in the mixed sector corner, of CysRS, TrpRS and GlnRS with the X positions, also backed by the finding of variant codes where X^{UA} can be translated as Gln [52].

Table 6 Integration of the networks of tRNA dimers by synthetase aggregation

Central G:C	Ser^S	Pro*	Gly	Arg*	Ala	Thr	Cys	Trp	X	Σ 32 +40
Ser^S –GA, GCU	-	-	2	2 +4	1	1	-	2	2	4 +10
Pro –GG *	-	-	2	2 +4	-	-	-	2	2	4 +8
Gly –CC	2	2	-	-	1	1	-	-	-	4 +2
Arg –CG, YCU *	2 +4	2 +4	-	-	2	2	-	-	-	8 +8
Ala –GC	1	-	1	2	-	-	2	-	-	4 +2
Thr –GU	1	-	1	2	-	-	2	-	-	4 +2
Cys GCA	-	-	-	-	2	2	-	-	-	4
Trp CCA	2	2	-	-	-	-	-	-	-	4
X UCA	2	2	-	-	-	-	-	-	-	4
Central A:U, Ho	Phe	Leu*	Glu*	Lys*	Gln*	X	Σ 16 +32			
Phe GAA	-	-	2	2	-	-	4			
Leu YAA, –AG *	-	-	2	2	8	8	4 +16			
Glu YUC *	2	2	-	-	-	-	4			
Lys YUU *	2	2	-	-	-	-	4			
Gln YUG *	-	8	-	-	-	-	8			
X YUA	-	8	-	-	-	-	8			
Central A:U, Mx	Val	Ile*	Met*	His	Tyr	Asp*	Asn	Σ 16 +8		
Val –AC	-	-	-	2	2	1	1	4 +2		
Ile G,UAU *	-	-	-	1	1	1	1	2 +2		
Met CAU *	-	-	-	1	1	-	-	2		
His GUG	2	1	1	-	-	-	-	4		
Tyr GUA	2	1	1	-	-	-	-	4		
Asp GUC *	1	1	-	-	-	-	-	2		
Asn GUU	1	1	-	-	-	-	-	2		

Numbers of dimers formed by the triplets belonging to an aRS, shared with other aRS. Ho, homogeneous pDiN sector; Mx, mixed pDiN sector; self-complementary connections, not underlined; nonself-complementary connections, underlined; self-aggregation, the SerRS^S; MaRS hetero-aggregation, *. Note the relative isolation of the X attributions, as well as the Trp-, Asn-, Gln, AspRS due to the absence of NSC connections, all at the corners of the central Y quadrants. The Gln- and AspRS were rescued to the RNP network via MaRS aggregation. The two central A:U subnetworks are one centered on the homogeneous sector dimers, to which the GlnRS and the X are joined, the other centered on the mixed sector dimers, to which the Asp- and AsnRS are joined. The summed data are presented in the graph format in Fig. 3

The wide range integration of the RNP system is mediated by the aRS–aRS adhesion interactions; these have been demonstrated experimentally [7, 32, 73, 75]. Integration of the two central base pair-specific large RNP-modules is accomplished and these communicate with various other components of the cellular machinery, working as multi-functional proteins. The integrative process has been slow, fully developed in the complex eukaryotic cells, due to a gradual acquisition by the aRS of segments demonstrating protein- or RNA-binding activities, or recruitment of such properties from auxiliary proteins that physically bridge and mediate the aRS contacts to

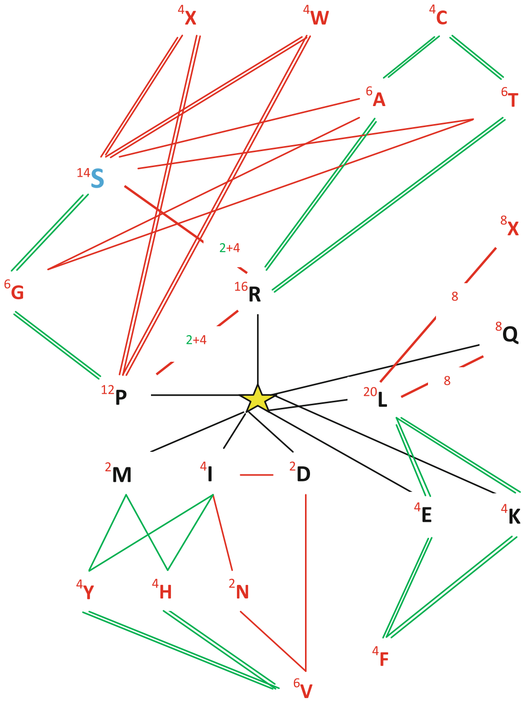


Fig. 3 Integration of the networks of tRNA dimers by synthetase aggregation. The central G:C dimer network is integrated due to the wide connectivity of the hexacodonic SerRS and ArgRS, which share dimers with a large variety of other aRS. The central A:U connections form two subnetworks: one for the central A of the homogeneous sector, the other for the central A of the mixed sector. The summed connections to each aRS are shown in the superscripts at the left of the nodes. The arches represent dimer connections; single and double (thin) or, when multiple, thicker and showing the numbers of shared dimers inside the lines. The self-associated SerRS (large S) is distinguished from the hetero-associated aRS of MaRS (united to the central star); the other ten aRS remain non-aggregated. Data originating the graph are presented as a matrix in Table 6

form multi-chain aggregates. Our analysis of the process of integration of the aRS-tRNA system assumes that it was operative in all kinds of cells, starting with a more diffusion-dependent organization, thereafter developing the physical adhesion mechanisms. The latter were beneficial especially to the large eukaryotic cells, where diffusion-driven contacts between reactants would be too slow and unreliable. The discussion should also not be impaired by the possibility that the process of acquisition and shuffling of adhesive properties, followed by selection, has not reached saturation or stability, which would only lead to the limitations of historical argumentation.

Ten of the 20 aRS Developed Physical Aggregation

Nine aRS are hetero-associated into a multi-synthetase complex (MaRS), well characterized in animal cells [75]; SerRS does not belong to MaRS but is found as an auto-associated dimeric enzyme [35]. Aggregation of LysRS class I and II was not added to our graph due to being utilized under highly specific contexts [71]. While an adequate explanation for the choice of the enzymes participating in the aggregates has not been reached, some descriptive remarks could be of help for guiding searches, distinguishing two types of factors. (a) A mostly nonspecific, quantitative driving factor would be the excessive number of dimers shared by some aRS. All five aRS presenting ≥ 8 dimers shared with other aRS are involved with the formation of aggregates. These include the three hexacodonic plus ProRS and GlnRS; saying otherwise, all attributions having a tetracodonic box in the 3′R rows plus GlnRS. (b) The other five aRS of MaRS have low number of dimers shared ($\leq$4) and all belong in the complex boxes of the 3′Y rows, which is indicative of specific choices still to be discerned: the aRS added to the MaRS complex are the two acidic, the LysRS—completing the set of basic amino acids and the amino acids in the initiation box; the set can also be described as all amino acids in complex boxes of the 3′Y rows, except Asn. In contrast to the hypotheses based on the pushing dynamics rationale for the early encodings, the distribution of aggregated aRS favors the 3′G and the 3′U rows (contributing with 4 aggregated/5 total amino acids and 5/7, respectively), while the aggregated aRS at the 3′A and the 3′C rows are quite scarce (respectively, 1/6 and 2/5). The list of apparently non-equilibrated distribution of characters among the aggregated aRS is extended with: excess of aRS class I over class II (7/4); excess of the homogeneous over the mixed

Phe 4	**Ser 14**	Cys 4	Tyr 4
Leu 20		Trp 4 (X 4)	(X 8)
	Pro 12	**Arg 16**	His 4
			Gln 8
Val 6	Ala 6	Gly 6	**Asp 2**
			Glu 4
Ile 4	Thr 6	**Ser**	Asn 2
Met 2 (iMet)		**Arg**	**Lys 4**

Fig. 4 Distribution of synthetases belonging in aggregates on the anticode box structure. A mostly nonspecific effect of the number of dimer connections per pDiN box may have directed the participation of aRS in aggregations. All aRS having tetracodonic boxes in the 3′R rows belong in aggregates while all aRS having tetracodonic boxes in the 3′Y rows do not. The specific component producing aggregation is highlighted by the distribution of aggregated aRS belonging in complex boxes. From the complex boxes in the 3′R rows, only GlnRS belongs in MaRS while from the complex boxes in the 3′Y rows, all belong in aggregates, except AsnRS. The 3′G and 3′U rows have high numbers of aRS in aggregates, contrasting with the low numbers in the 3′A and 3′C rows

pDiN sector (7/4); excess of central A and central U attributions over the central G and C (7/3); excess of Module 2 + over all other combined (6/4). Other characters indicate integrative equilibrium, when members of MaRS have at least one component from each of the NSC dimer modules, at least two members of each aRS subnetwork, and at least one central R and one central Y from each aRS subnetwork.

Physiologic Benefits: Regulatory Integration and Modularity

From the five aRS indicated by the quantitative factor (≥8 dimers shared) to be the best candidates for the development of aRS–aRS associations, only the SerRS was not observed to participate in MaRS. This absence is due to the SerRS being active as a self-dimer: the enzyme is self-sequestered and impeded from associating with other aRS. The SerRS kinetics is cooperatively stimulated by the binding of the first tRNA substrate to one component of the dimer [35]; this behavior should be related to the Ser anticodons being constituted by complementary pDiN. The SRM considers this peculiar behavior an example—the only remnant—of the original mechanism of dimer-directed protein synthesis. It demonstrates that the complementary proto-tRNAs could code for one amino acid only, therefore the absence of the hydropathy correlation, and explains the apparent 'disparateness' [29] of the different pDiN accepted by SerRS. Early fixation of the SerRS self-association would accomplish the role of freeing it from interferences from or upon other aRS, since SerRS is the most highly connected (to five other aRS) of all aRS. It could be indicated that preservation of other cases similar to the SerRS would have been dangerous to the system in the sense of producing excessive integration and loss of modularity. It is recognized that the dimer networks formed by present-day tRNAs would certainly acquire a different topology from the one shown here, probably reduced in various aspects due to the restrictions imposed by base modifications on the triplets [34].

There is for sure much information hidden inside the network structures, e.g., in relation to the variety of regulatory interactions. The two central G:C and central A:U dimer networks become integrated through protein adhesiveness to other proteins and to tRNA, these mechanisms complementing each other. The system accumulates benefits both from modularity—where components maintain some degree of independence from the whole, and from partial integration, not affecting all components and all modules to the same degree. The tRNA dimer lattices (Table 4) obey the central base pair restrictions and can only form modular structures, not being able to develop the functional integration, while proteins are free to follow their own rules, more related to the functionality of the ensemble—for instance, trapping tRNAs in close proximity to the aRS, which is needed in the large eukaryotic cells, and not obeying the SC versus NSC triplet distinction. The mechanism of cooperative self-stimulation observed with the SerRS self-dimers can be extended to processes occurring inside MaRS, now through cooperative hetero-stimulation and mutual integrative regulation. Biochemical details on the adhesive sites in SerRS and in MaRS are found elsewhere [7, 32, 73, 75].

Molecular Cognition and Synthetic Biology

Models of molecular systems may serve the role of inspirations upon which to model cognitive processes. Both seem to be based on self-organized networks of communicating components. The SRM is applicable to a part of the origin of life processes, at the foundation of truly synthetic biology projects namely the origin of the genetic code and of the nucleoprotein interdependency links. The model describes the formation of a cyclic structure of interactions, these being typical of network structures and of integrated organizations. It goes from the producers (triplet codes and amino acid meanings) to the synthesis of proteins (products) and from the products back to the producers through binding and stimulation of productivity. When such "molecular cognitive" closure is reached, the producers assume the role of memories or genes.

Another aspect of possibly even higher relevance to cognitive processes is (a) the ability of networks to grow constructively via incorporation of new components. A new agent in the system may be accepted if it can be inserted into one of the pathways through formation of at least two compatible connections, leading from the preceding element to the new and from this to the successive element. Inside the complexity of natural processes, this would accompanied with different gradations and in combination with the following: (b) the mere absence of compatible contacts, allowing for some aspects of demarcation from the exterior; (c) neutral participants, where contacts may only fill spaces inside the system; (d) the production of disruptions that may lead to selection (negative) effects on populations.

While experimentation in this area is invited, we are presently witnessing various advances following the line of manipulating the genetic memory component of biologic networks. The experiment of Gibson et al. [31] seems to crown one of these approaches, managing to substitute the full genome of a small bacterium with that from a close relative and containing some artificial changes. The outcome may be considered a new variant or species in the genus *Mycoplasma*. The really living component of the new cell lineage, which is the cytoplasmic metabolic and gene expression/regulation network, was not created de novo; it was obtained from an ancestor lineage and not disrupted. The whole set of genetic memories was derived from a related species, and the real novelty demonstrated was the very high level of plasticity of the network toward different memories or genes.

Questions to be Addressed by Bioinformatics

This work has progressed in a mostly qualitative mode of investigation utilizing punctually some elementary statistics but poses many questions that should gain from application of more sophisticated modeling and simulation. Some of the most evident questions are listed below, intending to stimulate interest, but certainly other questionings will come up from such studies, which could also help in the experimental planning. Fundamental to the model is the initial formation of self-stimulating nucleoprotein cycles that would be among the first in the construction of living networks and examples of processes that might be relevant for modeling cognitive networks.

The molecular recognition of producers—proto-tRNAs—and products—peptides—goes together with stabilization of the aggregate against degradation but excessive stabilization could result in reduction in the flexibility necessary for maintenance of the producing activity. How would the balance between stability and flexibility influence the number of cycles to allow for significant accumulation of the aggregates and for the development of specificity in the process? Such modeling should be important for planning on the composition of the first peptides. In the SRM, Gly is indicated to be most important among the amino acids involved with RNA binding but could also be too strong stabilizer. The second amino acid indicated is Ser, still RNA-binder but weaker, which would accomplish the role of modulator of the stabilization.

How would the encoding process benefit from repetitive utilization of the same mechanisms and the same kind of module, such as the NSC chosen by the SRM, in comparison with the utilization of combinations of different types of modules?

Was the anticodon 5′A elimination necessary to facilitate the encoding process? Tests are needed to demonstrate that the proposed symmetry-breaking resulting from the 5′A elimination from previously symmetric modules did facilitate the process. If the symmetry-breaking is necessary, could it have been installed on some other 5′ base? The known wobbling ambiguities are maximal with 5′U (any 3′ codon base), moderately high with 5′A (U, C, G > A), low with 5′G (U, C), and nonexistent with 5′C (G only) [68]. If the 5′A is formally not the best choice, the hypothesis that it was installed mostly in consequence of the ambiguity of decoding in complex boxes would be strengthened, in spite of not completely solving the complex problem. It is possible, for instance, that conservation of 5′U, despite its ample wobbling abilities, would have relied upon its great number of choices for chemical modification relative to other bases [33].

Our qualitative analysis of the aRS–aRS networks combined with the dimer-sharing by the aRS is surely partial and should be implemented through other procedures, e.g., by adding other dimer properties (such as thermal stability and NSC versus SC types, obtaining weighed graphs) or aRS characters. Application of the procedures to the real tRNA sets of different organisms, including the base modifications on the anticodons and on the neighboring bases, should make the model physiologically relevant.

How long would an algorithm be to describe the paths taken by the SRM to fill the code? Would there be other simpler rules to be followed, which would at the same time respect the physiological constraints? How would other possible models compare with the SRM, with respect to algorithm lengths?

Acknowledgments The continued collaboration of the colleagues in the Self-organization Group of the Center for Logics and Epistemology of UNICAMP, especially Alfredo Pereira Júnior, and of Carlos Henrique Costa Moreira and Sávio Torres de Farias are deeply appreciated, as well as the fruitful discussions with Arkadiusz Chworos.

References

1. Amirnovin R. An analysis of the metabolic theory of the origin of the genetic code. J Mol Evol. 1997;44:473–6.
2. Barabási AL, Oltvai ZN. Network biology: understanding the cell’s functional organization. Nat Rev Genet. 2004;5:101–13.
3. Baranov PV, Gurvich OL, Hammer AW, Gesteland RF, Atkins JF. Recode 2003. Nucleic Acids Res. 2003;31:87–9.
4. Batten D, Salthe S, Boschetti F. Visions of evolution: self-organization proposes what natural selection disposes. Biol Theory. 2008;3:17–29.
5. Beier H, Grimm M. Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs. Nucleic Acids Res. 2001;29:4767–82.
6. Bergson H. Creative Evolution. New York: Henry Holt; 1911.
7. Berthonneau E, Mirande M. A gene fusion event in the evolution of aminoacyl-tRNA synthetases. FEBS Lett. 2000;470: 300–4.
8. Beuning PJ, Musier-Forsyth K. Transfer RNA recognition by aminoacyl-tRNA synthetases. Biopolymers. 1999;52:1–28.
9. Bloch DP, McArthur B, Guimarães RC, Smith J, Staves MP. tRNA-rRNA sequence matches from inter- and intraspecies comparisons suggest common origins for the two RNAs. Braz J Med Biol Res. 1989;22:931–44.

10. Bloch DP, McArthur B, Widdowson R, Spector D, Guimarães RC, Smith J. tRNA-rRNA sequence homologies: a model for the generation of a common ancestral molecule and prospects for its reconstruction. Orig Life Evol Biosph. 1984;14:571–8.
11. Cleaves HJ, Aubrey AD, Bada JL. An evaluation of the critical parameters for abiotic peptide synthesis in submarine hydrothermal systems. Orig Life Orig Biosph. 2009;39:109–26.
12. Cody GD. Geochemical connections to primitive metabolism. Elements. 2005;1:139–43.
13. Copley SD, Smith E, Morowitz HJ. A mechanism for the association of amino acids with their codons and the origin of the genetic code. Proc Natl Acad Sci USA. 2005;102:4442–7.
14. Csermely P. Weak Links–Stabilizers of Complex Systems from Proteins to Social Networks. Berlin, Germany: Springer; 2006.
15. Davis BK. Comments on the search for the source of the genetic code. In: Messenger RNA Research Perspectives. Ed. Takeyama T. New York USA: Nova Science; 2008. pp 35–79.
16. Di Giulio M. The origin of the genetic code: theories and their relationships, a review. Biosystems. 2005;80:175–84.
17. Di Giulio M. An extension of the coevolution theory of the origin of the genetic code. Biol Direct. 2008;3:37.
18. Durfee T, Hansen AM, Zhi H, Blattner FR, Jin DJ. Transcription profiling of the stringent response in Escherichia coli. J Bacteriol. 2008;190:1084–96.
19. Eigen M, Schuster P. The hypercycle: a principle of natural self-organization. Berlin: Springer; 1979.
20. Eigen M, Winkler-Oswatitsch R. Transfer RNA, an early gene? Naturwissenschaften. 1981;68:282–92.
21. Ertem G. Montmorillonite, oligonucleotides, RNA and origin of life. Orig Life Evol Biosph. 2004;34:549–70.
22. Ertem G, O'Brien AMS, Ertem MC, Rogoff DA, Dworkin JP, Johnston MV, Hazen RM. Abiotic formation of RNA-like oligomers by montmorillonite catalysis: part II. Int J Astrobiol. 2008;7:1–7.
23. Eschenmoser A. Kinetic control. In: Stano P, Luisi PL, editors. Basic questions about the origins of life. Orig Life Evol Biosph 2007;37:309–314.
24. Faria LCB, Rocha ASL, Kleinschmidt JH, Palazzo R Jr, Silva-Filho MC. DNA sequences generated by BCH over GF(4). Electron Lett. 2010;46:202–3.
25. Farias ST, Bonato MCM. Preferred amino acids and thermostability. Genet Mol Res. 2003;2:383–93.
26. Farias ST, Guimarães RC. Aminoacyl-tRNA synthetase classes and groups in prokaryotes. J Theor Biol. 2007;250:221–9.
27. Farias ST, Moreira CHC, Guimarães RC. Structure of the genetic code suggested by the hydropathy correlation between anticodons and amino acid residues. Orig Life Evol Biosph. 2007;37:83–103.
28. Fishkis M. Steps towards the formation of a protocell: the possible role of short peptides. Orig Life Evol Biosph. 2007;37:537–53.
29. Fournier GP, Gogarten JP. Rooting the ribosomal tree of life. Mol Biol Evol. 2010;27:1792–801.
30. Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruthers MH, Neilson T, Turner DH. Improved free-energy parameters for predictions of RNA duplex stability. Proc Natl Acad Sci USA. 1986;83:9373–7.
31. Gibson DG, Glass JI, Lartigue C, et al. Creation of a bacterial cell controlled by a chemically synthesized genome. Science. 2010;329:52–6.
32. Golinelli-Cohen MP, Mirande M. Arc1p is required for cytoplasmic confinement of synthetases and tRNA. Mol Cell Biochem. 2007;300:47–59.
33. Grosjean H, Crécy-Lagard V, Marck C. Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes. FEBS Letters 2009. doi: 10.1016/j.febslet.2009.11.052.
34. Grosjean H, Houssier C. Codon recognition: evaluation of the effects of modified bases in the anticodon loop of tRNA using the temperature jump-relaxation method. In: Gehrke CW, Kuo KCT, editors. Chromatography and modification of nucleotides. Amsterdam: Elsevier; 1990. p. A255–A295.
35. Gruic-Sovulj I, Landeka I, Söll D, Weygand-Durasevic I. tRNA-dependent amino acid discrimination by yeast seryl-tRNA synthetase. Eur J Biochem. 2002;269:5271–9.
36. Guimarães RC. Metabolic basis for the self-referential genetic code. Orig Life Evol Biosph. 2011;41:357–71.
37. Guimarães RC, Moreira CHC. Genetic code–a self-referential and functional model. In: Pályi G, Zucchi C, Caglioti L, editors. Progress in biological chirality. Oxford: Elsevier; 2004. p. 83–118.
38. Guimarães RC, Moreira CHC, Farias ST. A self-referential model for the formation of the genetic code. Theory Biosci. 2008;127:249–70.
39. Guimarães RC, Moreira CHC, Farias ST. Self-referential formation of the genetic system. In: Barbieri M, editor. The codes of life: the rules of macroevolution. Dordrecht, NL: Springer; 2008. p. 68–110.
40. Guzman MI, Martin ST. Prebiotic metabolism: production by mineral photoelectrochemistry of alpha-ketocarboxylic acids in the reductive tricarboxylic acid cycle. Astrobiology. 2009;9:833–42.
41. Hartman H, Smith TF. GTPases and the origin of the ribosome. Biol Direct. 2010;5:36.
42. Hazen RM, Sverjensky DA. Mineral surfaces, geochemical complexities, and the origin of life. Cold Spring Harb Perspect Biol. 2010;2:a002162.
43. Higgs PG. A four-column theory for the origin of the genetic code: tracing the evolutionary pathways that gave rise to an optimized code. Biol Direct. 2009;4:16.
44. Higgs PG, Pudritz RE. A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. Astrobiology. 2009;9:483–90.
45. Hornos JEM, Braggion L, Mazini M, Forger M. Symmetry preservation in the evolution of the genetic code. IUBMB Life. 2004;56:125–30.
46. Illangasekare M, Yarus M. A tiny RNA that catalyzes both aminoacyl-RNA and peptidyl-RNA synthesis. RNA. 1999;5:1482–9.
47. Jones H, Cockell CS, Goodson C, Price N, Simpson A, Thomas B. Experiments on mixotrophic protists and catastrophic darkness. Astrobiology. 2009;9:563–71.
48. Jose MV, Morgado ER, Govezensky T. Genetic hotels for the standard genetic code: evolutionary analysis based upon novel three-dimensional algebraic models. Bulletin of Mathematical Biology 2010. doi:10.1007/s11538-010-9571-y.
49. Jurka J, Smith TF. β-turn-driven early evolution: the genetic code and biosynthetic pathways. J Mol Evol. 1987;25:15–9.
50. Kauffman SA. The Origins of Order–Self-organization and Selection in Evolution. New York USA: Oxford University Press; 1993.
51. Kisselev LL. Class I translation termination factors are functional analogs of aminoacyl-tRNAs. Mol Biol. 2003;37:791–802.
52. Knight RD, Freeland SJ, Landweber LF. Rewiring the keyboard: evolvability of the genetic code. NatRev Genet. 2001;2:49–58.
53. Koonin EV, Novozhilov AS. Origin and evolution of the genetic code: the universal enigma. IUBMB. 2009;61:99–111.
54. Lambert JF. Adsorption and polymerization of amino acids on mineral surfaces: a review. Orig Life Evol Biosph. 2008;38:211–42.
55. Li WH. Molecular evolution. Sunderland USA: Sinauer; 1997.

56. Maizels N, Weiner AM. The genomic tag hypothesis: what molecular fossils tell us about the evolution of tRNA. In: Gesteland RF, Cech TR, Atkins JF, editors. The RNA World. New York USA: Cold Spring Harbor Laboratory Press; 1999. p. 79–111.
57. Martin W, Russell MJ. On the origins of cells: a hypothesis for the evolutionary transition from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. Philos Transact R Soc Lond B. 2003;358:59–85.
58. Maury CPJ. Self-propagating beta-sheet polypeptide structures as prebiotic informational molecular entities: the amyloid world. Orig Life Evol Biosph. 2009;39:141–50.
59. Mathews DH, Sabina J, Zucker M, Turner M. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. J Mol Evol. 1999;288:911–40.
60. Mayr E. What Makes Biology Unique? New York USA: Cambridge University Press; 2004.
61. Miller DL, Yamane T, Hopfield JJ. Effect of tRNA dimer formation on polyphenylalanine biosynthesis. Biochemistry. 1981;20:5457–61.
62. Milo R, Shen-Orr S, Itzkowitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. Science. 2002;298:824–7.
63. Nielsen PE. Peptide nucleic acids and the origin of homochirality of life. Orig Life Evol Biosph. 2007;37:323–8.
64. Noller HF. The driving force for molecular evolution of translation. RNA. 2004;10:1833–7.
65. Novozhilov AS, Koonin EV. Exceptional error minimization in putative primordial genetic codes. Biology Direct. 2009;4:44.
66. Ogle JM, Brodersen DE, Clemens WM Jr, Tarry MJ, Carter AP, Ramakrishnan CV. Recognition of cognate transfer RNA by the 30S ribosomal subunit. Science. 2001;293:897–902.
67. Ogle JM, Carter AP, Ramakrishnan CV. Insights into the decoding mechanism from recent ribosome structures. Trends Biochem Sci. 2003;28:259–66.
68. Osawa S. Evolution of the genetic code. New York, USA: Oxford University Press; 1995.
69. Petrusz SC, Turvey MT. On the distinctive features of ecological laws. Ecol Psychol. 2010;22:44–68.
70. Plutynski A. Explaining how and why: developmental and evolutionary explanations of dominance. Biol Philos. 2008;23: 363–81.
71. Polycarpo C, Ambrogelly A, Berube A, Winbush SM, McCloskey JA, Grain PF, Wood JL, Söll D. An aminoacyl-tRNA synthetase that specifically activates pyrrolysine. Proc Natl Acad Sci USA. 2004;101:12450–4.
72. Poole AM, Jeffares DC, Penny D. The path from the RNA world. J Mol Evol. 1998;46:1–17.
73. Praetorius-Ibba M, Hausmann CD, Paras M, Rogers TE, Ibba M. Functional association between three archaeal aminoacyl-tRNA synthetases. J Biol Chem. 2007;282:3680–7.
74. Powner MW, Gerland B, Sutherland JD. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. Nature. 2009;459:239–42.
75. Quevillon S, Agou F, Robinson JC, Mirande M. The p43 component of the mammalian multi-synthetase complex is likely to be the precursor of the endothelial monocyte-activating polypeptide II cytokine. J Biol Chem. 1997;272:32573–9.
76. Rajamani S, Vlasov A, Benner S, Coombs A, Olasagasti F, Deamer D. Lipid-assisted synthesis of RNA-like polymers from mononucleotides. Orig Life Evol Biosph. 2008;38:57–74.
77. Rodin S, Ohno S, Rodin A. Transfer RNAs with complementary anticodons: Could they reflect early evolution of discriminative genetic code adaptors? Proc Natl Acad Sci USA. 1993;90: 4723–7.
78. Rodin S, Rodin A, Ohno S. The presence of codon-anticodon pairs in the acceptor stem of tRNAs. Proc Natl Acad Sci USA. 1996;93:4537–42.
79. Ronneberg TA, Landweber LF, Freeland SJ. Testing a biosynthetic theory of the genetic code: fact or artifact? Proc Natl Acad Sci USA. 2000;97:13690–5.
80. Rosslenbroich B. The theory of increasing autonomy in evolution: a proposal for understanding macroevolutionary innovations. Biol Philos. 2009;24:623–44.
81. Santoso S, Hwang W, Hartman H, Zhang SG. Self-assembly of surfactant-like peptides with variable glycine tails to form nanotubes and nanovesicles. Nano Lett. 2002;2:687–91.
82. Schwartz AW. Arsenate DNA: Evidence for a vital force? Orig Life Evol Biosph 2011;41. doi:10.1007/s11084-010-9231-0.
83. Seligmann H, Krishnan NM, Rao BJ. Possible multiple origins of replication in primate mitochondria: alternative role of tRNA sequences. J Theor Biol. 2006;291:321–32.
84. Serrano A, Perez-Castineira JR, Baltscheffsky M, Baltscheffsky H. H^+-PPases: yesterday, today and tomorrow. IUBMB Life. 2007;59:76–83.
85. Sobolevsky Y, Trifonov EN. Protein modules conserved since LUCA. J Mol Evol. 2006;63:622–34.
86. Szostak JW. Systems chemistry on early earth. Nature. 2009;459:171–2.
87. Tafforeau M, Verdus MC, Norris V, Ripoll C, Thellier M. Memory processes in the response of plants to environmental signals. Plant Signal Behav. 2006;1:9–14.
88. Tamura K. Origin of amino acid homochirality: relationships with the RNA world and origin of tRNA aminoacylation. BioSystems. 2008;92:91–8.
89. Tamura K, Schimmel PR. Chiral-selective aminoacylation of an RNA minihelix: mechanistic features and chiral suppression. Proc Natl Acad Sci USA. 2006;103:13750–2.
90. Tenera M. Life began when evolution began: a lipidic vesicle-based scenario. Orig Life Evol Biosph. 2009;39:559–64.
91. Trewavas A. A brief history of systems biology. Plant Cell. 2006;18:2420–30.
92. Trifonov EN. Glycine clock: eubacteria first, archaea next, protoctista, fungi, planta and animalia at last. Gene Ther Mol Biol. 1999;4:313–22.
93. Trifonov EN. The triplet code from first principles. J Biomol Struct Dyn. 2004;22:1–11.
94. Varschavsky A. The N-end rule: functions, mysteries, uses. Proc Natl Acad Sci USA. 1996;93:12142–9.
95. Vetsigian K, Woese C, Goldenfeld N. Collective evolution and the genetic code. Proc Natl Acad Sci USA. 2006;103: 10696–701.
96. Wong JTF. A co-evolution theory of the genetic code. Proc Natl Acad Sci USA. 1975;72:1909–12.
97. Wong JTF. Coevolution theory of the genetic code at age thirty. BioEssays. 2005;27:416–25.
98. Wood AP, Aurikko JP, Kelly DP. A challenge for the 21st century molecular biology and biochemistry: what are the causes of obligate autotropohy and methanotrophy? FEMS Microbiol Rev. 2004;28:335–52.
99. Xia T, SantaLucia J Jr, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry. 1998;37: 14719–35.
100. Yamane T, Miller DL, Hopfield JJ. Interaction of elongation factor Tu with the aminoacyl-tRNA dimer Phe-tRNA:Glu-tRNA. Biochemistry. 1981;20:449–52.
101. Yarus M, Widmann JJ, Knight R. RNA-amino acid binding: a stereochemical era for the genetic code. J Mol Evol. 2009; 69:406–29.