Formation of the genetic code

Romeu Cardoso 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 Brasil

Tel-Fax +55-31-3274.4988

Email: romeucardosoguimaraes@gmail.com

Formation of the genetic code

Romeu Cardoso Guimarães[1]

Abstract

Formation of the genetic code was investigated by integration of data on amino acid composition of protein functional sites and on development of amino acid biosynthesis pathways, with the novel proposition that encodings were directed by proteins synthesized by dimers of tRNAs paired through anticodons. Elimination of anticodonic 5'A facilitated the encoding process. The succession of encodings begins with Glycine and Serine, indicating that early metabolism was centered on assimilation of one-carbon units. Early amino acids (GS,LDN,EPKFR) compose peptides predominantly stabilizing the amino-termini of proteins against degradation, and building RNA-binding motifs with mostly non-periodic conformations. Compartmentalization of amino acids in modules of tRNA dimers results in minimization of the consequences of errors along the coding/decoding processes. Termination anticodons were deleted due to their interference with the initiation mechanism. The system of tRNA dimers is integrated through development of associations between the synthetases.

Keywords Genetic code; Origins; tRNA dimers; Aminoacyl-tRNA synthetases; Networks; Metabolic sink; Error minimization; Self-reference.

[1] Laboratório de Biodiversidade e Evolução Molecular, Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, 31270.901 Belo Horizonte MG Brasil.

TelFax +55-31-3274.4988, romeucardosoguimaraes@gmail.com

Introduction

The core components of living beings, the informational macromolecules – nucleic acids and proteins – are bi-directionally interdependent. They compose a system of interactive polymeric strings forming a multi-cycle structure with mutual stimulation activities (Figure 1). In one direction, various RNAs, including rRNAs, tRNAs, and mRNAs, are main participants in the system of protein synthesis. The mRNA sequences have consecutive codon triplets that direct the sequences of proteins through decoding via the genetic code, the set of correspondences between triplets and their meanings – amino acids or termination. The decoding process is largely deterministic. Correspondences are strict and between discrete units, the triplets and the meanings. Decoding is realized through adaptor molecules, the tRNAs, and the Release Factors (proteins; RF). Each tRNA carries both members of the code in different segments of the molecule: an anticodon triplet and the amino acid. Transfer RNAs are complex molecules, about 76 nucleotides long, substrates for both the synthetases and the ribosomes. The peptidyl-transferase reaction is catalized by the rRNA (Nissen et al. 2000).

In the other direction, the products (proteins) participate in various aspects of the process of their own synthesis. For termination of translation, the tRNAs are substituted by the RFs that recognize an X codon and cleave the peptidyl-tRNA bond, not forming a peptide bond and releasing the nascent peptide. The aminoacyl-tRNA synthetase (aRS) reaction accomplishes the amino acid/tRNA correspondence. Proteins interact with the nucleic acids through aggregative binding, including the aRS/tRNA cognate interactions, and participate in nucleic acid synthesis, especially as polymerizing enzymes. The binding interaction at formation of nucleoprotein aggregates is compositional in two concomitant aspects: (i) it depends on binding sites in both partners, which are short segments of the sequences that may present variable nucleotide or amino acid composition, frequently forming consensual motifs; (ii) sites are multiple and distributed in different positions of the long sequences of the interacting molecules (Beuning and Musier-Forsyth 1999). Combinations of sites define the strength and plasticity of the binding interactions. The function may be dependent mainly on the spatial structure of the interacting sites, which can be obtained with different sequences (Bashan and Yonath 2008). There is still a wide gap in knowledge on the organization of the system of encoding and decoding, and how it originated.

Outline

The self-referential model (SRM) for the formation of the coding system develops two fronts of investigation, (i) on the mechanistic aspects of the establishment of codes (Guimarães et al. 2008a, b) and (ii) on the development of metabolic support (Guimarães 2011). The first says that the original attribute is the formation of nucleoprotein associations (such as the aRS/tRNA) and that the code was shaped by the interplay of properties of these two components of the circular structures – dimers of tRNAs and the peptides produced by them, without participation of external strings to be translated. In the proto-biotic realm, nucleic acid-like oligomers would work as proto-tRNAs, apt to receive ligands of different kinds including amino acids, dimerize through base-pairing at some segments (proto-anticodons), and synthesize products through polymerization of the attached ligands, transferring them from one proto-tRNA to the other.

Kinetic reasons for the dimer-based code formation are (a) the higher transferase productivity in dimers, relative to that of acylated proto-tRNA monomers meeting each other freely in solution (for covalently ligated ribozymes, see Kuwabara et al. 1999), combined with (b) the turnover property facilitated by the hydrogen-bonded iterative dimerizations within the pool of acylated proto-tRNA monomers, not requiring the complex conformational changes of dual-specificity ribozymes (Lee et al. 2000). The acylation reaction is simple (Yarus 2011; Tamura 2011) and the turnover is propitiated when mediated by the dimerization through the ‘kissing’ interactions (Horiya et al. 2003) between proto-anticodons, typical of tRNA dimers (Moras et al. 1986).

Among the variety of products, peptides were selected for, on the basis of the ability to interact adequately with the proto-tRNAs, forming proto-ribonucleoprotein (RNP) aggregates. The ensemble acquired functionality, initially from stabilization (Figure 2), and its evolution led to development of the codes when some of the aRS/tRNA associations became fixed. Characters of both partners were adjusted in mutuality, their co-evolution in the self-stimulating cycles reaching the code specificities; proto-tRNAs evolved into the tRNAs and peptides into the synthetases, among other early protein specificities.

Possible configurations of proto-tRNA dimers accommodating a nascent peptide are sketched in Figure 3. Peptides produced recognize the producing dimers through binding. The joining of producer and product in the proto-RNP aggregate configures the closure of a self-referential functional loop. Activities of the dimers are stabilized in the RNPs by the bound peptides, stabilization being equivalent to stimulation of the production function. Specificity evolves through iteration of the production cycles with selection of variants in the producers, the products, and the RNPs, driven by productivity (Figure 2). Such cycles are among the first in the construction of the translation system. Formation of the cyclic structure of interactions is akin to the biological evolution paradigm: variants of components produce populations of phenotypes whose success is evaluated through fitness measures, resulting in permanence of the adequate configurations; these are fixed in the pools through selection dynamics with progressive dilution of the inadequate states. When the producing components (nucleic acids) and the products (proteins) bind to each other, variants are selected by criteria of mutual adequacy for formation of the nucleoprotein aggregate structures and for the functions of the aggregates (fitness). The ensemble (phenotypes) may follow evolutionary changes in the direction of maintenance or improvement of performances and the producers acquire the function of memories in the system.

Development of the second front of investigation stems from the observation that the initial encodings (Gly, Ser) do not coincide with the most abundant in pre-biotic syntheses (Gly, Ala). A search for metabolic sources of the former pair was adequately satisfied (Guimarães 2011); the Glycine-Serine Cycle (GSC), typical of facultative methylotrophs, is the best extant representative of such pathways. Systems chemistry reasoning indicates that the pre-biotic path leading to fixation of the proto-RNPs was of establishing a detoxification flux, driven (pushed) by concentrations of reactants. Early environments were rich in toxic – reactive – compounds such as organic acids and aldehydes, whose amination into amino acids was a simple mode of detoxification. The flux was established by polymerization of the amino acids and the consumption of the peptides in the proto-RNP aggregates, configuring a dynamic sink mechanism. After development of early proto-codes, where the proto-RNP constitution was dominated by pushing forces, selfing mechanisms arose when proteins acquired enzymatic capabilities and could start building biosynthesis pathways that substituted the pre-biotic sources. In the biological context, the protein synthesis system continued working as a sink, pulling the fixation of the biosynthetic pathways. The metabolic rationale says that the standard configuration of the code was established in concomitance with the formation of pathways for amino acid production, starting with precursors to the GSC.

The SRM predicts that the matrix of triplet codes (Box 1) should be organized according to the formation of dimers and that the succession (non-dated chronology) of encodings should correspondingly follow the dimer-directed structure. This approach was developed on the basis of correlations between biological data and the dimer organization, resulting in the proposition of a specific chronology (Guimarães et al. 2008a, b). This is now revised and updated (see also Guimarães 2012) with the separation of six steps of encodings, four of them corresponding to the modules of dimers (Table 1). The chronology is of interest if it is considered a record of some aspects of the cellular construction process, therewith presenting indications that can be evaluated in light of functional and evolutionary criteria.

The model integrates about a dozen attributes (albeit not all of them fully independent from each other) into a comprehensive scheme. Details are offered on *(i)* the relationship between the constitution of some of the main protein sites and the fixation of metabolic pathways (Guimarães 2011) that supply amino acids or their precursors for encoding, on *(ii)* the precedence of the RNP over the DNP domains of cellular processes (respectively the homogeneous and the mixed sectors; Box 1), and on *(iii)* the formation and distribution of simple and complex boxes (Box 2). Formation of codes in the RNP realm is more detailed than in the DNP sector. *(iv)* A rationale relating the initiation and termination codes is developed (Guimarães 2001), which leads to functional closure of the system. This is coherent with the preferential location of amino acids in the N- and C-ends of proteins (Berezovsky et al. 1999) and the N-end rule of protein stability against catabolism, which is dependent on the amino acids located at the N-termini (Varschavsky 1996; Meinnel et al. 2006; Wang et al. 2008). Realization that the code was shaped under the influence of protein structures and functions, as integral components of the early RNP aggregates, leads to reconsideration of previous study traditions, which are centered on external pressures (pushing forces – concentration gradients of large sets of available amino acids, derived initially from pre-biotic sources) and then followed by metabolic pathways (Wong 2005; Di Giulio 2008). The SRM scheme says that putative proto-codes might have obeyed the pushing forces but the standard code was configured as a result of the self-stimulating cycles, with mutual adjustment of the structures of the nucleic acid and protein components for the construction of functional and stable nucleoprotein aggregates. Production of the aggregates became a sink that organized the flux of consumption including the fixation of the supporting metabolic pathways (Guimarães 2012). Sink-driven organization is akin to the pulling dynamics inherent to the traditional 'facilitated diffusion' mechanisms. *(v)* The basal components of the translation system – synthetases and tRNAs – are physically integrated via synthetase aggregation. In consequence of the encoding process being *(vi)* directed by the construction of protein functional sites and motifs, combined with the allocation of codes to modules of dimmers, the resulting structure immediately presents the property of minimized consequences of mutational or translational errors. Proto-codes are considered next and some perspectives for further investigation are indicated.

Results

Anticodon dimer networks. The 64 triplets are divided into 16 boxes, each corresponding to a principal dinucleotide (pDiN; Box 1). Boxes of triplets are split into halves of the non-self-complementary (NSC; RNR, YNY; lateral bases of the same kind) and self-complementary kinds (SC; RNY, YNR; lateral bases of the complementary kinds) (Figure 4). Dimers formed by

NSC triplets identify the homogeneous (Ho; RRR:YYY) and the mixed (Mx; RYR:YRY) sectors, discriminated by the central base being of the same or of a different kind with respect to the lateral bases. Dimers formed by SC triplets (RNY:RNY, YNR:YNR) integrate the sectors and the columns (Figure 4). Two large networks are formed, one for the central G:C and the other for the central A:U triplets, each containing four subnetworks (Guimarães 2012).

Kinds of triplets and the principal dinucleotide. Mononucleotides are fundamental for mutational studies, but dinucleotides are the basic units of thermal behavior of nucleic acid sequences (Xia et al. 1998), and the minimum repeat size at interactions between nucleic acids or of nucleic acids with other molecules (Guimarães and Erdmann 1992), including the involvement of anticodon triplets with synthetase specificity. Triplets and dimers are composed of two overlapping dinucleotides (Figure 5a), with no functional distinction between them; four kinds of dimers are obtained, corresponding to the four kinds of sub-networks (Figure 4). Distinction between NSC and SC triplets is the basis for the splitting of boxes into halves and for the definition of the homogeneous and mixed sectors of the matrix. The distinction is modified when one of the dinucleotides acquired the pDiN function, the remaining base becoming the wobble position. This character is seen in all components of translation (Figure 5b): the codon:anticodon pairing, the synthetase specificity, and the rRNA site checking the quality of the codon:anticodon mini-helix. Possible consequences of the difference between the Ho and Mx pDiN (Figure 5b1) for the mini-helix conformation have not been investigated; experiments (Ogle et al. 2001) were conducted on pairs of the NSC homogeneous kind, codon UUU:anticodon GAA.

NSC triplets have simple sequences, composed of dinucleotide repeats, compared to the complex SC triplets that show no repetition. NSC triplets of the Ho sector are all-R or all-Y, repeats of one dinucleotide kind (RR + RR : YY + YY); NSC triplets of the Mx sector are inverted repeats (RY + YR : YR + RY), of intermediate complexity. NSC triplets are symmetry centers in the anticodon loop, more neatly defined in the case of the mixed sector where the central base is of a kind different from the lateral bases. Best indicators of the high complexity of SC triplets are the combination of one Ho and one Mx dinucleotide and the absence of the symmetry center inside the triplet. The first character explains their role of integrators of the four sub-networks of dimers corresponding to a specific type of central base pair into one network (Figure 4). The second would be involved with their exclusion from the initial encoding process, which were directed to the simpler repetitive structures of the NSC triplets. In the SRM chronology, all initial and most of the second encodings were installed upon NSC triplets of each box; the SC triplets were occupied via development of the simple box degeneracy. Only the second occupiers of complex boxes of the wYR quadrant (Gln, Trp, X) were encoded directly into the SC triplets.

Constitution of sequences in the RNP realm (the homogeneous sector) may be hinted at by the chronology. Both kinds of strings would not be richly heterogeneous. Proteins are composed by ten amino acids forming mainly non-periodic conformations. There being no records on the early RNA sequences besides the encoded triplets, it can only be suggested that these would represent general characters of the sequences. RNAs would be rich in homogenous tracts (oligo-R and oligo-Y) as indicated by the encoding of the NSC triplets of the Ho sector; these would be most adequate for composing hairpin loops presenting the predominant extended conformation which facilitates dimerization. The non-structured segments of peptides and RNAs would organize each other at formation of the RNPs. A problem with the homogeneous tracts (called segments with translational symmetry in thermodynamics of duplex formation; Xia et al. 1998) is that they would allow slippages in interactions and introduce positional interference with other interactions in nearby sites (Guimarães and Erdmann 1992). In hairpins, the slippage problem is reflected in dynamic variability in the sizes of loops and stems. Slippage problems are reduced at encoding the Mx sector NSC triplets (Modules 3 and 4), where translational symmetry is substituted by the rotational symmetry. This is typical of the duplex palindromes interacting with complex protein folds, in the DNP realm.

Sub-network symmetry-breaking. The peculiarity of the standard anticode of lacking 5′A triplets evidences the crucial symmetry-breaking event that generated distinct topologies in the subnetworks (Guimarães 2012) (Figure 4). This character is continuously implemented through elimination of mutants introducing the 5′A triplets, in a process analogous to the exclusion of termination suppressors (Beier and Grimm 2001), or through post-transcriptional modification of the 5’A.

The four asymmetrical NSC sub-networks

The same topology in maintained in both sectors, with eight (2 × 4) intra-sector dimers each:

Module 1: (Ho)GGR:YCY(Ho), Module 2: (Ho)GAR:YUY(Ho)

Module 3: (Mx)GCR:YGY(Mx), Module 4: (Mx)GUR:YAY(Mx)

The four symmetrical SC sub-networks

Triplets of the two sectors are combined. Subnetworks are of two kinds:

5′Y triplets with sixteen dimers each (4 × 4): (Ho)YGR:YCR(Mx) and (Ho)YAR:YUR(Mx)

5′G triplets with four dimers each (2 × 2): (Ho)GCY:GGY(Mx) and (Ho)GUY:GAY(Mx)

Chronology of encodings. Initial encodings were installed in the NSC modules covering the entire range of 16 boxes through the same mechanism. Thermodynamic data indicate the greater stability of NSC duplexes with respect to the SC, the latter assuming a variety of configurations (higher entropy; Xia et al. 1998) which would delay the process of choosing the most adequate for developing the nucleoprotein interactions. Hairpin structures containing a loop centered on NSC triplets plus lateral SC segments, as in present day tRNAs, seem ideal both for facilitation of initiation of duplex formation, via the lateral segments, and for obtaining stability at loop pairing, via the NSC central triplets.

Thermal stability of triplet mini-helices (Guimarães 2012) was the driving force for the order of encodings inside each module: start with the high-ΔG dimers (Figure 6, pairs a of Modules 1-4), then follow to the low-ΔG dimers (b). After 5′A elimination, the asymmetric configuration of the NSC modules facilitated the process. Creation of asymmetry could have utilized other bases, but it is indicated that the choice for the 5′A depended on concomitant additional benefits: (i) at wobble-decoding (Agris et al. 2007) of complex boxes, the 5’A would cause misreading; (ii) the less ambiguous 5′ bases (C, G) were preserved and (iii) the more ambiguous U could be subjected to a large variety of post-transcriptional modifications that restrict the ambiguity (Grosjean et al. 2010). Ordering the four NSC modules to establish the chronology relied upon amino acids, protein, and nucleoprotein properties; thermodynamic data on mini-helices were not directive in this context.

General protein characters. Initial peptide synthesis would have been largely non-specific with respect to the system being built but specific choices were introduced at protein construction driven by its consumption in formation of RNP aggregates, therewith forming the metabolic amino acid sink. Some physiological and evolutionary correlations are indicated by the chronology enlightening on the natural protein construction rules.

Formation of protein functional sites. Protein conformations, functional motifs, and binding sites are oligomeric stretches of sequences constructed as clusters of amino acids that share properties adequate for the respective functions of the stretches. Protein sequences were built following the drive originated at the formation of such functional stretches: (a) when they bind to the proto-tRNAs, the constitution of one member of the complex drives the constitution of the other; and (b) when the complex tends toward acquisition of thermodynamically stable configurations. The SRM chronology was tested for compatibility with some aspects of this clustering principle: *(i)* the temporal succession of amino acids should correspond to the formation of structural and functional clusters; *(ii)* the amino acids selected for encoding drove the fixation of biosynthesis pathways supplying them; *(iii)* the regionalization of codes and meanings in the matrix would correspond to *(iii.a)* functional clusters in proteins and to *(iii.b)* the modular organization of the dimers of triplets. The latter *(iii)* would be the basic reason for the minimization of the local differences between codes and meanings, and of consequences of point mutations or of translational errors (Vetsigian et al. 2006).

Four characters of proteins were examined. The hydropathy correlation (Farias et al. 2007) (Figure 7) indicates that early peptides were composed of Gly, the non-chiral amino acid, and

Ser, introducing the chiral character (Ho sector, Module 1). By the end of the Ho sector (Step 3), enzyme specificity is obtained, indicated by establishment of the correlation. The Ho sector contains all amino acids that are preferred in non-periodic protein conformations (coils and turns (Creighton 1993) (Figure 8) and three-fourths of the amino acids that preferentially participate in the constant sites of RNA-binding motifs (Guimarães and Moreira 2004) (Figure 9). Complementarily, the Mx sector attributions contain ≥80% of the amino acids that are preferred in the β-strand conformations (amino acids that are preferred in the α-helices are distributed evenly between the sectors), and of the amino acids that preferentially participate in the constant sites of DNA-binding motifs or that participate in both kinds of nucleic acid-binding motifs. It is clear that the RNP system (the Ho sector) preceded the DNP system (the Mx sector). In the former, the mainly non-periodic protein conformations would be most adequate for binding to the RNAs, rich in single-stranded loops of hairpins; the two non-structured stretches organize and stabilize each other in the RNPs. The DNP aggregates, richer in the complex rotationally symmetric structures, require the more complex protein folds built with major contributions from β-strands and sheets, together with the fully structured α-helices.

Enzyme specificity and the hydropathy correlation. Evolution of subsets of the correlation between hydropathies of the pDiN of anticodons and the correspondent amino acid residues (Farias et al. 2007) establishes the chronology between sectors and between modules of the Ho sector (Figure 7). Correspondences of the mature Module 1 (GPS) do not build a correlation: amino acids are hydroapathetic and anticodons are of highly divergent hydropathies. Both Gly and Ser correspond to the two members of a dimer (respectively, WGG:WCC and WGA:WCU), a condition that would have prevailed in the pre-biotic / early biotic transition. With the advent of the synthetases, the former was modified, substituting the WGG-Gly with Pro, while the SerRS conserved partially the dimer specificity (WGA and GCU).

Other correspondences were established through enzyme specificity forming two correlated sets. One is built with proteins composed by ten amino acids, at the end of the Ho sector (Step 3); the regression line built by the Module 2 plus the Step 3 amino acids (less Pro, of Module 1;

eight attributions of seven amino acids) presents a moderate inclination. The inclination of the regression line built by the amino acids of the Mx sector is steeper, meaning that in the presence of the whole set of attributions the synthetases were able to finely tune the hydropathies.

Protein metabolic stability. Consistency between the N-end rule (determination of protein half-life by the kind of amino acid present at the N-terminus (Varschavsky 1996; Meinnel et al. 2006; Wang et al. 2008), the preferential localization of amino acids in the N- and C-terminal segments of proteins (Berezovsky et al. 1999) (Figure 10), and the SRM chronology indicates that the N-end rule participated in the processes directing the construction of sequences. The encoding process started with amino acids that contribute to lengthening protein half-life and these constructed the N-terminal segments; then, the amino acids that may shorten protein half-life were added, forwarded to the C-terminal segments. The rationale applies straightforwardly to the order of the modules of the RNP realm, indicating tandem ligation of codes and of amino acids from Modules 1 to 2. The distribution of codes of the DNP sector follows a different mechanism: the wRY quadrant, composed entirely of stabilizing amino acids, is placed en bloc in the upward extension of the N-end segments, while the wYR quadrant, entirely composed of destabilizing amino acids, is placed en bloc in the downward extension of the C-end segments. Components of the Mx sector dimers are dissociated, in accordance with the typical trans-acting character of the DNP realm, and sorted in a staggered mode while retaining the order of the Modules 3 to 4. The last elongation codes, Module 4b, are also the loci of the main punctuation codes (treated further down).

Minimization of the consequences of errors. The regionalized distribution of the correspondences in the matrix has been considered a second aspect of the regularities in the code structure (Vetsigian et al. 2006). The regionalization is usually interpreted as consequent to evolutionary adjustments leading to a near-optimization of the distribution of codes or, through a reverse wording, to minimization of the functional consequences of mutational or translational errors. Our data (Figure 11) indicate that the reduced impact of errors is a necessary consequence of the process of encoding that combined (a) the construction of protein segments devoted to fulfill specific functions and (b) their allocation to specific modules of dimers. The hotel room/guest metaphor (José et al. 2011) would say that the modules are locations available for utilization by the amino acid guests; these were chosen through selective processes focused on the functionality at the construction of protein motifs adequate for the nucleoprotein associations.

Salient aspects in the chronology and in the structure of the coding system are as follows: Module 1 amino acids are fully homogenous through all characters; the set of amino acids preferred in the construction of non-periodic protein conformations is completed in the Ho sector; all amino acids contributing to the construction of RNA-binding motifs and to stabilization of protein N-terminal segments belong to the Ho sector and to the N-end extension. It is expected that tests of the extent of minimization utilizing these biological parameters should yield improved results when compared to the tests utilizing the amino acid polarity data that have been utilized so far (Higgs 2009).

Synthetase complexity. Evolution of synthetase active sites from simple to complex substrate kinds is supported by the various characters examined in the chronology. The route from Ho to Mx triplets was treated above and amino acid characters are now examined.

Amino acid size and hydropathy. A trend from small to large amino acid (Grantham 1974) substrates is demonstrated in the RNP sector after whose end a plateau of large average sizes is established (Figure 12). The hydropathy general trend (Farias et al. 2007) (Figure 13) is from hydroapathetic (Module 1) to hydrophobic (Module 4). Intermediate stages show a wide dispersion of sizes and hydropathy, the entire range being explored from Modules 2 to 3, which reflect the hydropathy correlation. These trends are also correlated with the early wider utilization of aRS class II, showing mixtures of classes in the intermediate stages and increasing class I in later stages. RNP structures are more hydrophilic than the DNP, in accordance with the substitution of U by T and that of ribose by deoxyribose. The concomitant introduction of hydrophilic and hydrophobic amino acids starting from Module 2 indicates the formation of organized protein structures, possibly including peptide bilayers (Santoso et al. 2002) at some early stages.

Simple and complex boxes. The functional distinction of the pDiN in triplets is considered an ancestral character of synthetase specificity. It is indicated that (a) synthetases established correspondences initially to the pDiN, with fully 5′ degenerate initial codes (simple boxes); (b) boxes containing more than one meaning (complex boxes) evolved from sophistication of synthetases with specificity extended to the whole triplets, involving specific subsets of the wobble position; (c) at formation of a complex box, the initial attribution receded to a subset of the triplets in the box and conceded other subset(s) to the new attribution(s). The rationale of starting with simple boxes, dominated by the pDiN specificity, is enforced by the similar arrangement in the ribosomal decoding site (Figure 5b3). This interacts with the curvatures of the mini-helix formed by the codon and anticodon through an rRNA contiguous AA duplet in tight and multiple contacts with the paired pDiN of mRNA and tRNA, while the contact with the wobble position is single and looser through a G nucleotide coming from a distant location (Ogle et al. 2001). There are no data to help in choosing a succession of events starting either from the synthetases or from the ribosome. It is safer to adopt the evolutionary principles of co-evolution between the components with gradual fixation of both characters after a long interplay.

Simple boxes were maintained in the standard code when the pDiN were composed only of G and C (Pro-WGG:Gly-WCC, Ala-WGC:Arg-WCG; with high ΔG (Guimarães 2012), the core of the matrix) and when the pDiN had one base, either G or C, and one A or U, but with a central R (Ser-WGA, Leu-WAG, Val-WAC, Thr-WGU) (Figure 14). Complex boxes were generated when the pDiN were composed only of A and U (Phe, Leu-WAA:Asn, Lys-WUU; Ile, Met-WAU:Tyr, X-WUA; with low ΔG, the tips of the matrix) and when the pDiN had one G or C and one A or U, but with a central Y (Cys, Trp, X-WCA; His, Gln-WUG; Asp, Glu-WUC; Ser, Arg-WCU). The ΔG_{max} at the non-axial pairs (WGA:WCU, WAG:WUC, WAC:WUG, WGU:WCA) are intermediate. When the lower stability of the intermediate ΔG pairs is associated with the instability of the central Y triplets, these became prone to develop complexity. Instability of pairs made by central Y triplets is indicated by the high frequency of variant codes involving the anticodon central Y quadrants (Table 2), and by the self-dimerization of central Y triplets. This occurs spontaneously with the $tRNA^{Asp}_{GUC}$ (Moras et al. 1986) and after cytosine deprotonation (Romby et al. 1986) with the $tRNA^{Gly}_{GCC}$, both cases involving a central Y:Y mismatch.

An interplaying between dimer intrinsic lower stability and the synthetase participation in its further stabilization is indicated, which would lead to the multiple encoding of the box. When dimer stability is intrinsically high, participation of the wobble bases in stabilization would be negligible; any kind is adequate, working mainly as space-filling. Synthetases were not required to participate in their stabilization and to develop specificity for a specific kind of wobble base, so that simple boxes are generated. When intrinsic dimer stability is low, the help of synthetases was necessary for ensuring stability but the state of this nucleoprotein system remained plastic – stability dictated by nucleic acid duplexes is stronger than when relying upon nucleoprotein aggregates. A simple box is generated initially but is still open to receive new encodings. At a new entrance, both this and the former occupier of the box are required to develop specificity for the base(s) in the wobble position. The symmetrical distribution of simple and complex boxes in the matrix (Figure 14c) indicated the mechanisms for their formation (above), and that they were fully utilized the code evolution traverse. Were the possibilities offered by the triplet pair stability rules not utilized to saturation, the symmetry and the rules would not be discerned.

Identification of the initial attribution in a complex box relies mainly upon (a) comparisons between characters of the different occupiers together with (b) examination of the direction of changes inside complex boxes in evolutionary variant codes (Knight et al. 2001) (Table 2). Changes in the meaning of codons go from a source (the standard) to a target (new meaning), reflecting both the evolutionary instability of the source and the tolerance of the system where the change was fixed to the substitution. In six of the eight complex boxes (the four belonging to the 3′Y boxes and two of the 3′R boxes containing termination codes) the initial occupier is identified with (a) the amino acid that is structurally simpler than the other amino acid in the box and/or than the punctuation mechanism, and that corresponds to the 5′G anticodes: the initial D conceded to (→) the new E; N → K; $S^{CU} \rightarrow R^{CU}$; I → M, iM; Y → X; C → W, X. Among the five complex boxes containing variants, one is not informative: the M ↔ I codes show inter-conversions and the iM is stable. Information from (b) the other four supports the rationale, indicating that the changes of meanings of triplets followed the direction from the late to the early occupier (K → N; $R^{CU} \rightarrow S^{CU}$; X → Y; W, X → C). These variants are interpreted as reversions from the late to the previous meaning of the triplet via loss of the mechanisms that were added at generation of box complexity, also saying that the mechanisms generating the early (simple box) meanings demonstrate higher evolutionary stability than those producing the new meaning(s) in the boxes. This rationale is consistent with Barabási and Oltvai’s (2004) network evolution principle – older nodes are more stable owing to development of greater number of links with other components of the network, and younger nodes are less stable owing to less dense connectivity. One component of the mechanisms would be the benefit for the ancient encoding of utilizing the lower wobbling variation of the 5′G, relative to the 5′U. It is noteworthy that there are no instances of changes of meanings inside complex boxes from the earlier encoded (presently 5′G) to the late 5′Y. The SRM proposes different explanations for the formation of the other two complex boxes belonging to the 3′R rows (F, L and H, Q; see below) but there are no evolutionary variants for checking.

Hexacodonics. Most complex synthetase specificities are the hexacodonics, all belonging to the RNP realm. When $\text{ArgRS}^{Y\underline{CU}}$ entered, at Step 3, SerRS receded to hexacodonic maintaining only the 5′G from the W$\underline{CU}$ box (W$\underline{GA}$ + G$\underline{CU}$; Module 1b). The other two synthetases show expansion of the specificity for the 3′ bases while conserving the central bases. LeuRS developed moderately wide 3′ specificity (3′R, W$\underline{AG}$ + Y$\underline{AA}$; Module 2), conceding G$\underline{AA}$ to PheRS. ArgRS developed wider 3′ specificity (G, U), at the transition from Steps 3 to 4 and bridging the sectors, adopting one Ho and one Mx pDiN (Y$\underline{CU}$ + W$\underline{CG}$).

Atypical systems. Formation of the two atypical aRS, both class II, can be unified by the dimer rationale. The 5′Y anticodes of complex boxes are occupied by class I enzymes or punctuation, the exception being the LysRS, which may be class I or II in different organisms (Cusack 1997); the class II is atypical. Enzymes of class II acylate the 3′-OH of tRNAs, the 2′-OH acylation being typical of class I enzymes, but PheRS behaves in the class I mode (Arnez and Moras 1997). These atypical systems belong to the dimer $\mathrm{G\underline{AA}}$ Phe:$\mathrm{Y\underline{UU}}$ Lys, the last encoded in the RNP sector (Module 2b). A suggestion of conformational strains being imposed upon the enzymes by the amino acids is indicated by their sizes – they are the only large amino acids of class II enzymes (Figure 12) – and by their extreme hydropathies (Figure 13).

Adaptation to higher temperature. Among the five amino acids in Step 3, there are the two most preferred in proteins of thermophilic organisms, Glu and Lys, plus the $\mathrm{Arg}^{\mathrm{Y\underline{CU}}}$ anticodons that are also preferred in thermophilics (Farias and Bonato 2003; Van der Linden and Farias 2006). Data are silent about Pro and Phe that also enter at this step. The concomitant incorporation of the two basic amino acids (Lys, Arg) would indicate the formation of the ribose-phosphate backbone of nucleic acids or of their massive accumulation, including the introduction of DNA in the system, which would also be related to the thermal challenges.

Phylogeny of enzymes and tRNAs. Revision of aRS sequences phylogeny through the ancestral sequence procedure (Farias and Guimarães 2007) applied to genomes of prokaryotes identified two groups in each of the classes. The picture is complex and indicative of varied adaptive features, but it is significant that, in both the bacterial and archaeal kingdoms, the cluster in class II enzymes containing the Module 1 attributions (GPS) is conserved. The conserved cluster also contains HisRS, the enzyme for another of the five hydroapathetic amino acids (Figure 15).

Tests are under design with the purpose of joining both enzymes and tRNAs in a conjoint tree, but the tRNA tree (Figure 16) is revealing in itself (Farias ST, Personal communication). Three basal sequences are identified: Gly, Leu, and Asp. The Asp group is large, branching further into three sections; two are headed by Ser and Pro, while the other is ambiguous. The five basal ancestors belong to the Ho sector; four of these were taken up by class II enzymes, including the Module I attributions; the fifth is Leu-tRNA of Module II, heading a class I lineage. The ambiguous branch derives from the two His-tRNA ancestors being unique in presenting distinct behaviors; both show similarities to the aromatic specificities but one clusters with Tyr and the other with Phe; the basal position of His^1-tRNAs may represent the exception to the rule.

Metabolic supply. ***Why amino acids.*** In the pre-biotic panorama, amino acids are among stable compounds of low reactivity or toxicity to chemical systems. Their high yields among other compounds obtained under conditions simulating early Earth environments (Zaia et al. 2008; Higgs and Pudritz 2009) attest to the stability as end-products under a variety of conditions, amination of reactive compounds including formation of amino acids being an easy mode of detoxification. A chemical systems perspective (Eschenmoser 2007) allows formulation of the rationale that polymerization would have been most effective with these stable and non-toxic compounds; such properties were extended to the peptides, which, by their turn, initiated an efficient flux of consumption at RNP formation.

Methylotrophy and autotrophy. A scenario on ‘nutritional’ conditions at times of formation of the early RNPs, which implement the flux of peptide consumption and obey the record written in the code chronology, points to utilization of amino acids constructed from C1 units, either in the most oxidized form CO_2 or in the various reduced states. Extant organisms following these regimes are oligotrophs, in the fuzzy borders between methylotrophy and autotrophy (Anthony 1982). The oligotrophic condition makes sense in the light of the nutrient scarcity expected to have prevailed on Earth during the origin of life. Some kind of heterotrophy based on more complex nutrients such as sugars formed by the formose reactions would have been transient, rapidly exhausted (Quayle and Ferenci 1978). Single-carbon compounds would have been the main reliable and constant sources for minimal heterotrophy; at conditions of heterotrophic crises, C1-utilization pathways would be safeguard restoration subsystems in the metabolic network. These would be among the reasons for their choice as starters in the code chronology.

Amino acids in this C1 realm are Gly C2 and Ser C3, the only amino acids that show up among compounds in the central metabolic pathways. They belong to the Glycine-Serine Cycle (GSC), typical of facultative methylotrophs (Guimarães 2011). It may be indicated that the prominence of the GSC among central pathways would have been higher during evolution of ancestral cellular forms (LUCA) than it is in extant systems. The GSC is the simplest among central pathways, containing C2-C4 compounds, the others reaching C5-C7; among its ten components, only two (2-P-Glycerate, P-Enolpyruvate) are phosphorilated and one is thio-activated (Malyl-CoA). An evolutionary route from the complex Ribulose-P pathway of obligate methylotrophy to the Ribulose-PP pathway of autotrophy has been advocated on the basis of enzymological characters (Quayle and Ferenci 1978). Evidence recorded in the code chronology extends and enriches this proposal that methylotrophy preceded and was a step towards autotrophy, adding that the route was installed at the simpler level of amino acid biosynthesis.

Biosynthesis. Biological fixation of synthesis routes for amino acids is not adequately rationalized in the absence of feasible mechanisms of consumption of the products of the pathways. Without consumption, they would accumulate to a point of provoking the reversal of the reactions with hydrolysis and other degradation modes to finally reach some equilibrium state. There are various alternatives for obtaining the anabolic direction via polymerization of the amino acids, either in solution or with guidance by surfaces (Lambert 2008). Evidence obtained from the code structure suggests that polymerization directed by dimers of proto-tRNAs was the route leading to cells.

The main aspect of the chronology is the full dependence of encodings on metabolic supply of amino acids or of their precursors in biosynthesis pathways. The early entrance of Gly and Ser disagrees with both traditions of studies on the code origins. The line based on the list of prebiotic abundant amino acids points to AGDV with some variation (Higgs 2009; Trifonov 2004). The metabolism co-evolution model is also based initially on these sets, with later addition of the biosynthesis-derived amino acids (Wong 2005; Di Giulio 2008). The beginning of encodings with Gly points in both directions: (a) of the prebiotic chemical systems, where it is among the abundant compounds, and (b) of the early metabolic pathways, such as the acetogenic Ljungdahl-Wood and the GSC (Madigan and Martinko 2006; Martin and Russell 2003), therewith suggesting a path of pre-biotic to biotic continuity. Glyoxylate is among the intermediates in oxidation of acetate, also a plausible pre-biotic compound (Nuevo 2010), highly toxic (Anderson 2004) and the immediate precursor to Gly by amination (Figure 17). Other main metabolic alternatives for obtaining Glyoxylate are the Glyoxylate Cycle plus the photorespiration and the Ethylmalonyl-CoA pathways (Guimarães 2011).

Glycine can also be obtained metabolically from C1-THF + methylamine (from CO_2 + NH_3) via the anabolic activity of the Glycine Decarboxylase Complex (GDC; Bauwe and Kolukisaoglu 2003). Methylamines are also substrates for methylotrophs. The next metabolic step in GSC is another detoxification event, where the reactive C1 (formate or formaldehyde) is carried as C1-THF and added to Gly, forming Ser via the Serine hydroxymethyl transferase, in a freely inter-convertible reaction. These amino acids configure a sub-cycle inside the GSC; it may be indicated that this sub-cycle would have been the seed for the complete GSC.

Modules 1 and 2 correspond to encoding of amino acids directly derived from the GSC (Figure 17). Step 3 is added by the end of the Ho sector (between Module 2 [Step 2] and Module 3 [Step 4]; Table 1), corresponding to elongation encodings dependent on metabolic sources of amino acids more complex than those belonging to the GSC. Encoding of these amino acids involved substitution of some of the earlier encodings, with reduction of their degeneracy. It is considered that the main central metabolic pathways reached maturation at Step 3, starting with gluconeogenesis derived from the C3 and C4 components of the GSC (respectively, the Pyruvate and the triose families, and the Oxaloacetate and the Asp families), followed by the sugar transformations and the Citrate Cycle. Amino acids entering at Step 3 belong to the C5 Glu family (Pro, Arg), plus Phe, whose synthesis requires a C3 and a tetrose; Lys may be derived from Asp or from 2-Oxoglutarate in different groups of organisms. After this step where the central pathways are mature, there are no metabolic constraints for the encoding of Module 3 and 4 (Steps 4 and 5) amino acids.

A remark is necessary on the chronological distance between the Glu to Gln modification (Step 3 to Step 5) with respect to the closeness between the Asp to Asn (Module 2a to 2b). The discrepancy may be related to a delayed evolutionary differentiation of the GlnRS, with prolonged maintenance of the mechanism of Glu-tRNA^{Gln} transformation into Gln-tRNA^{Gln}, which has experimental support (Skouloubris et al. 2003). Another remark refers to the early entry of the C6 Leu with respect to the simpler Pyruvate derivatives C3 Ala and C5 Val. In the absence of records on any other amino acid possibly preceding Leu in its location, it can only be suggested that if such a possibility ever existed, it was short-lived and left no traces; the choice for Leu would have been due to protein construction requirements, not satisfied by the other possibilities.

Metabolic interdependency. The basal metabolic role of the GSC is enforced by observing that its components can account not only for the amino acids in Modules 1 and 2 but also for most of the carbon skeletons of the nucleobases (Zrenner et al. 2006). From the five carbons in the purine ring, two come from one Gly molecule, two derive from C1-THF, which is a co-factor in the Gly-Ser inter-conversions in the GSC, and one comes from bicarbonate. From the four carbons in the pyrimidine ring, three come from Asp and one from the bicarbonate that formed Carbamoyl-P. In fact, some of the phylogenetic trees based on protein architecture place the pathways of Gly and Ser, and of the nucleobase metabolism (Caetano-Anollés et al. 2007) close together.

Overall data on the contribution of carbons of amino acids to the carbon skeleton of compounds in the entire mass of bacterial cells (Reitzer 2003) show that the GSC components account for a major portion of the carbon donation (~10% from Gly, reaching ~40% with Ser and ~70% with Asp; Figure 18). With the Step 3 entry of Glu (derived from the Citrate Cycle), 90% of the carbon donation is reached, the remaining 10% being distributed among other four amino acids of the Mx sector. The picture emerges that evolution of metabolic complexity relied upon mainly amino acids that through gluconeogenesis built the sugars from which the central sources of metabolites reached near completion. Considering the participation of amino acids in protein composition the baseline (Figure 18), the data on the number of amino acid post-translational modifications (Kyte 1995), which enlarge the protein alphabet, point to the important participation of Ser. Gly is not involved with these processes, while K and R (Step 3) and CHY (Mx sector) are other important contributors to these developments.

Punctuation. Organization of the code in eight pairs of boxes can only be hinted at through examination of the orthogonal matrix of correspondences by three indications. Two of them have already been examined: the SerRS with complementary pDiN and the two atypical aRS in the tips of the RNP sector. Punctuation is the third, at the tips of the DNP sector, but its generation does not come from pDiN complementariness.

The initiator function is accomplished by one specific tRNA^{iMet} that enters the ribosome and is transported directly to the empty P site, thereafter being able to transfer the amino acid to the aminoacyl-tRNA in the A site and to build the first peptide bond (Carter et al. 2001). The anticodon of iMet has the same sequence of the elongator tRNA^{Met} but is functionally different as seen from the ability to accept a small variety of codons (Kolitz et al. 2009; Etten and Janssen 1998), most frequently Val, sometimes Leu, and rarely others, resulting in the definition of a peculiar dislocation (slippage) of the wobble position: elongation reads Met codon AUG:anticodon UAC, but initiation reads iMet codons NUG:anticodon UAC. Complementariness between the elongation pDiN of the initiation box (WAU) and the main termination box (WUA) is suggestive of relatedness between mechanisms of initiation and termination. Otherwise, the inversion between the initiation-slipped pDiN (WAC) and the elongation pDiN (WCA) of the box where the secondary X codon resides is puzzling. A plausible mechanism could be developed for the relatedness between all four punctuation codes when the slippage process was taken into account and when the initiation process is considered a second-order reaction, involving the first two codons or amino acids to form the first peptide bond, instead of the simple introduction of the initiator tRNA in the translation system. The mechanism also suggests how the exclusion of the X anticodes was driven.

Initiation. Second amino acids indicated to compose the majority of initiation sequences (components of the ‘sequence context’ for initiation) are coded for by codons with 5′R (Val GUN, Ala GCN, Thr ACN); this set completes the entire quadrant where the Met box resides (Figure 10). The slipped pDiN of initiation becomes contiguous with the elongation pDiN of the second amino acid. This configuration (Table 3) would guarantee high dynamical stability to the initiation process; the tetra-nucleotide stretch excludes the possibility of interruption by variations due to wobbling in case the initiation pDiN were not slipped. Utilization of the slippage mechanism also indicates that this is secondary to the pre-existent elongation pDiN and adds to the concept that all elongation codes were present at times of installation of the punctuation system, including the tRNAs corresponding to the X codes (YCA for Trp, WUA for Tyr). According to this panorama, both processes were late: the utilization of the slipped iMet pDiN and the exclusion of the X tRNAs.

Termination. Searches were conducted looking for all possible combinations of triplet matches between initiation and termination codes and anticodes, either identity or complementariness, and with slipped or non-slipped pDiN. Matches considered relevant would have to account for all three real termination codes and indicate a consistent mechanism (Guimarães and Moreira 2004). The meaningful set of data showed evidence of competition derived from identity between anticodes of X tRNAs and of the iMet for the initiation codon, indicating the rationale that the X tRNAs were deleted in consequence of the competition conflicts (Table 4). The target identified was a combination of initiation codons forming the triplets AUG and GUN^{Val} or ACN^{Thr} (the slipped pDiN UG of the initiator codon plus the 5′R of the second codon). All X anticodes fit the competition rationale. They make a Watson-Crick base pair with the U and generalized R:Y base pairs with the other bases (AYY). Were a G in the first position (forming a U:G base pair) also significant for the competition, X anticodes would have spread to the Arg-GCY and the Gln-GUY. Data also indicate that competition pressures were less strong upon the Trp-CCA anticodes when the second codon is Thr-ACN, owing to the formation of the weak A:C base pair. It is considered that this set of matches is adequate to sustain the proposed explanation: deletion of X anticodes was a consequence of the installation of the initiation code based on the slippage mechanism.

Functional closure. Connections between the punctuation codes, the beginnings and the ends of genetic strings, are obtained via long-distance interactions, the X anticodes traversing tortuous channels to reach the initiation complexes. Since this process occurred after the entire space of elongation triplets has been dissipated, it is indicated that the encoding system reached completion with functional closure. A process related to this rationale is the closed-loop structure formed in eukaryotes when the poly A tail of mRNAs is bound via the poly A binding protein to the initiation factors (Walsh and Mohr 2011); in prokaryotes, analogous long-distance interactions are mediated by binding via the small ribosomal sub-units of polyribosomes (Arava 2009).

Building strings. Functional closure of the coding system is the result of a complex string-building process that can be displayed as a circular structure (Figures 8-10). The single-stranded RNP domain is dominated by hairpin replication: a strand is elongated through one end looping back upon itself, forming a replication primer. The strand becomes punctuated with plus and minus stretches that may form various hairpin structures, typical of RNAs (Bloch et al. 1984, 1989). This elongation mechanism would be dominant inside all modules. Intra-module sequences may be subjected to all possibilities of recombination and duplication; the order in Figure 6 refers only to the mechanism of encoding. The joining of Modules 1 and 2 requires ligation in tandem and obeys the order 1 → 2 but as soon as this is accomplished, the ensemble of all encodings may again be subjected to shuffling. The protein component in this realm is dominated by non-periodic conformations (coils, loops, and turns) plus some of the α-helix forming amino acids, and by amino acids contributing strongly to formation of RNA-binding motifs (Figure 9). The order of amino acids from Modules 1 to 2 and Step 3 follow the N-end rule of protein stabilization against catabolism, starting with the stabilizers and ending the strings with the destabilizers; adequacy of the whole protein for permanence is checked from the beginning, through the quality of their N-end segments. After guaranteed stability, long strings may be formed and the regulatory elements would be acting mostly in the cis configuration. The RNA strands contain a large fraction of loop segments widely open to interactions with distant segments of the same molecules, forming complex 3D arrangements, and with other molecules such as the proteins, which are also widely open owing to the predominance of the non-periodic conformations. Possible remnants of such clusters have been described (Sobolevsky and Trifonov 2006; Sobolevsky et al. 2012), most striking being the various configurations of the GPS combinations. It is not convenient to contrast the polar possibilities of their being due to ancestor conservation or to convergent evolution; it is more useful to think of both acting in concert, the latter reinforcing the former.

Addition of the Mx-sector amino acids is already in the DNP realm, with predominance of the amino acids forming DNA-binding motifs. The nucleic acid portion of the aggregates is self-organized through the major double-stranded configuration, and trans-acting regulatory mechanisms are typical of the DNP realm. Protein conformations privileged in the Mx sector are the β-strand- and sheet-building amino acids, which unite distant segments of proteins. The string-building process, as indicated by the N-end rule and the protein conformation preferences, follows the order of Modules 3 to 4 but, in accordance with the mostly informational interactions indicated by the long distances and the activities in trans, the members of the dimers are not joined one to the other in tandem as in the Ho sector segments. The WRY quadrant attributions (ATVIM) are all stabilizers of protein N-ends and added to the heads of the strings, elongating them to reach Met, after which the specific initiation system can be installed. Complementarily, and following the same trend already installed in the Ho sector, the WYR attributions (RCWHQY) are all destabilizers of proteins and added to the tails of the strings, elongating them to reach Tyr, after which termination is installed. The tips of the strings are then joined through the informational connections of the punctuation system and the circular structure is closed.

Halting. It is adequate to say that the standard code proved good enough to reach the nearly universal structure but not to mean that the system became frozen. It became stabilized to a degree that guarantees faithful construction of the cellular macromolecules and its conservation is a central stronghold, a common cellular core. Otherwise, the high complexity, diversity, and informational richness of living systems depended on the introduction of additions on top of the core while not subverting its main rules. Some of the additions are (a) the few instances of variant codes (Table 2), where some modifications of the standard code apply to all proteins in a lineage (Knight et al. 2001), (b) the many instances of recoding (Bekaert et al. 2010), where some codes may have meanings changed under special circumstances or contexts, and (c) the most frequent and a main character of organic evolution, the post-transcriptional (Grosjean et al. 2010) and post-translational (Kyte 1995) modifications of RNAs and proteins, highly enlarging the limited repertoire of the genetic code. Ingredients to the standard code are few and the enormity of the biodiversity would correlate better with the expanded alphabets of post-transcriptional base and post-translational amino acid modifications, each of these summing around 100 types and widely extending the combinatorial possibilities (Figure 18).

The latter (b, c) depend on the development of new sets of genes for pathways producing the modifications, which took a large role of halting the encoding process. The code became the core subsystem of a larger system and the requirements for its stability were the foundations and at the same time pressures for the development of other subsystems such as the post-transcriptional and post-translational modification pathways. The panorama is enforced that formation of the code was a slow process, immersed in an integral cellular system, interdependent with metabolic and gene duplication-modification and -divergence developments. To cope with mutational fluctuations and challenges to the stability of the coding system, and with pressures for adaptations to environmental changes, novelties in more complex protein and nucleic acid interactive sites were searched for among new enzymes to accomplish the post-translational modifications, instead of adding letters to the code.

Integration via synthetase aggregation. Present roles of tRNA dimers should be mostly regulatory, difficult to clarify on the basis of the complex structure of the networks. Experimental data show the potential of such interactions (Yamane et al. 1981; Miller et al. 1981) but they are constrained through specifications introduced by nucleotide modification in the triplets and in the neighboring bases. The effect of excesses of uncharged tRNAs in amplification of signals from amino acid starvation, which include the sequestration of remaining charged tRNAs into dimers with the excessive uncharged, is well known (Raskin et al. 2007; Altmann and Linden 2010). Special roles of the modules of SC dimers would be towards balancing equilibration, profiting from their symmetrical structures, but still limited to the formation of the central G:C and central A:U dimer networks. These networks were initially only dynamically integrated through the amino acid biosynthesis pathways starting with Pyruvate (central G:C Ala; central A:U Leu, Val), Oxaloacetate (central G:C Thr; central A:U Asp, Asn, Lys, Ile, Met) and 2-Oxoglutarate (central G:C Pro, Arg; central A:U Glu, Lys, Gln), but the main integrative development was the slow acquisition of sites producing physical aggregation of synthetases (Guimarães 2012). The basic encoders/decoders (aRS and tRNAs) form an almost fully integrated RNP system that acquired efficiency in locally trapping tRNAs, with mutual help among the enzymes.

Integration of the system of dimers through the synthetases started with their specificity restricted to the pDiN, erasing the distinction between the SC and NSC triplets at formation of simple boxes. Higher-level integration was introduced when ArgRS crossed the Ho/Mx-sector barrier. Integration of the sectors was expanded, albeit remaining partial, when some of the enzymes developed specificity for the central purines, forming the central G/class II and the central A/class I groups, the latter including the atypical PheRS. The mixtures of aRS classes in the columns led to the ~70% concordance of class II with the Ho sector (GPSDN plus the atypical F and K) and of class I with the Mx (VIMCWRYQ), the discordant being Ho/class I (LEKR) and Mx/class II (ATH).

A complex network of the synthetases connected through the numerous tRNA dimers is obtained in consequence of the latter being clumped into the eight modules and of the synthetases making contacts among themselves when they are bound to the tRNAs in the dimers (Figure 19). Each tRNA participates in formation of at least two dimers and each synthetase is connected to the network also through at least two dimers. The rule in the constitution of the nucleoprotein structures centered on the tRNA dimers is that each of these places in close proximity two different synthetases. It is suggested that contacts between the proteins are propitiated, which would lead to evolution of aggregative sites, based on the observation that all synthetases that are connected to the network through at least eight dimers (Q with L; P, R, S between themselves and with various other aRS) developed the aggregative property. The last four synthetases in this set are the main hubs of the dimer system, expectedly including the hexacodonics. The main hub role of ProRS was transient, reduced after elimination of the X tRNAs. The high integrative power of P, R, and S leads to formation of a large central G and central C sub-network, while the central A and central U attributions remained divided into two sub-networks. The central G:C sub-network acquired the ultrasmall-world (Barabási and Oltvai 2004) character with formation of four tri-node cycles with constant participation of SerRS. Some anticodes at the external corners of the central Y quadrants of both sectors present only SC dimer links but only two synthetases (Trp and Asn, as well as the deleted X anticodes) remain in this marginal condition; the two other synthetases in this set (Asp, Gln) were rescued to participate in the nucleoprotein network through protein aggregation. The isolated condition of X anticodes may have been among the mechanisms directing the location of the punctuation system.

Aggregation was a slow process of gradual incorporation of segments adequate for protein- or RNA-binding activities into the aRS sequences, or via recruitment into the Multiple aRS Complex (MaRS) of auxiliary proteins that mediate the aRS contacts to form the aggregates

(Guimarães 2012). Few enzymes are found aggregated in prokaryotes, the full MaRS is found in eukaryote cells. It is indicated that the physical cohesiveness was beneficial especially to the large cells, overcoming the unreliable and poorer dynamics of the diffusion-mediated contacts. Some synthetases participate in a variety of other organic functions (Guimarães 2012; Park et al. 2012), working as multi-functional proteins so that the tRNA charging system is only a portion of the large functional ensemble. The nine synthetases participating in MaRS are distributed among the sub-networks but more numerous among the central A:U dimer sub-networks. This may have been an integrative strategy since the dimer-mediated contacts were already more extensive inside the central G:C sub-network. The central G:C dimer network contributes to MaRS with the two hubs (ProRS, ArgRS); one of the central A:U sub-networks participates in MaRS with the LeuRS hub plus Glu-, Lys-, and GlnRS, and the other with Ile-, Met-, and AspRS. An inverse relationship is apparent between sub-network complexity (connectivity and the presence of minimal cycles) and the number of aRS incorporated into MaRS (Figure 19).

SerRS is the tenth enzyme demonstrating the aggregative property. Aggregation of LysRS class I and II is not considered owing to being utilized under specific contexts (Polycarpo et al. 2004). SerRS forms an auto-associated dimeric protein, differing from the MaRS hetero-associations, but the functional benefits of the dimerization suggest the possible reasons for the formation of MaRS. Kinetics of the sub-units of the SerRS dimer is co-operatively self-stimulatory since binding of one tRNA facilitates the binding of the second (Gruic-Sovulj et al. 2002), and similar hetero-stimulations may occur inside MaRS. SerRS is the only dimer hub (14 dimers shared with other enzymes) that does not belong to MaRS and is also the most connected to other enzymes (five links) through dimers. This apparently excessive connectivity may have been a reason for the isolation of SerRS from aggregation to other aRS through self-sequestration, preserving modularity and preventing excessive integration, which could include both mutual stimulation and interference. The possibility of the self-dimerization being related to the peculiar conservation of the complementary pDiN for SerRS is appealing but has not been investigated.

No adequate explanation can be offered as yet for the specific choice among the lowly connected synthetases to be included in MaRS. A quantitative driving factor seems to have been at work for propitiating aggregation in the main hubs (LPRS) plus Gln (all residing in the 3′R anticode rows), but the other five participants of MaRS have at most four dimer connections and are all in the 3′Y rows, curiously skipping the six-dimer-connected aRS. Other preliminary observations on the MaRS network are found in (Guimarães 2012) but details should wait for examination of the real cellular systems, where restrictions to dimer formation due to anticodon loop base modification should be decisive.

Discussion

A vast array of reports centred on the RNA World hypothesis attempt to picture what would have looked like an early RNA oligomer accomplishing the role of the proto-tRNA (Illangasekare and Yarus 1999; Lee et al. 2000; Tamura and Schimmel 2006; Yarus et al. 2009; Yarus 2011; Tamura 2011). A main drawback to the hypothesis is the lack of convincing evidence for the basic assumption – the pre-biotic existence of nucleotides and RNA (Powner et al. 2009; Szostak 2009). On these grounds, our stance is for an early realm of proto-tRNA oligomers, RNA-like, but not necessarily constituted by the complex nucleotides of the present RNA, containing generalized R and Y bases and backbones of unknown composition (Figure 3). Experiments on oligonucleotide synthesis directed by clay surfaces and utilizing activated monomers have been successful and attest to the predominance of sizes up to ~20mer and of sequences containing more of the homogeneous than of the mixed tracts. Furthermore, there are difficulties in obtaining homogeneous kinds of linkages between the sugars (Ertem 2004; Ertem et al. 2008).

Single strands of the synthetic Peptide Nucleic Acid (Egholm et al. 1992; Nielsen 2007) or of the Glycol Nucleic Acid (Zhang et al. 2005) could be considered among possible alternatives; a structure of the former kind is favored by our data that indicate the late encoding of basic amino acids. Such nucleotide-like oligomers would start making dimer-directed synthesis of oligomers built from their ligands. When this system iteratively reached the situation of producing peptide-assisted formation of proto-RNPs, enzymes evolved to yield the present-day nucleotides and RNAs. The SRM says that the first half of the genetic code was formed inside this RNP world. At some early stage in the development of tRNAs their structures might have looked like the mini-tRNAs that have been experimentally studied (Beuning and Musier-Forsyth 1999) or might have developed from other structures (Lee et al. 2000; Yarus 2011) but receiving the addition of an anticodon loop-like structure with the capacity of forming dimers.

In the half-century interval from the deciphering of the code correspondences, models for its formation have been almost invariably centered on the pushing dynamics rationale. Pre-biotic concentration gradients of amino acids would have driven the development of RNA acceptors or aptamers that evolved into the tRNAs. Long RNAs would pre-exist (Poole et al. 1998; Noller 2004), becoming genes (mRNAs) when a reasonable set of the tRNA adaptors would be able to translate them. This scenario is hetero(non-self)-referential since the mRNAs-to-be are external to the translation system under development. The magnitude of the adjustments between the reactants in this scenario would have been enormous. The complication in the senseless to sense transition is lower in the SRM that starts from simple oligomers to be encoded thereafter being ligated, building readily translatable strings. The flow of biological specificity (information) runs, at the origin, from the aRS/tRNA cognate couples to the long strings; in the full system, it runs from the DNA memories through the decoding system to the proteins.

Even the metabolism-based model of Wong (2005; also Di Giulio 2008) follows the pushing dynamics reasoning for the early assignments. Phase 1 encodings are taken from lists of amino acids that can be synthesized under pre-biotic conditions, phase 2 being of biotic origins. The mechanism of encoding is based on amino acid transformations being set upon aminoacyl-tRNAs; the derived amino acids would be assigned to variant tRNAs and both components of the system would form families of relatedness. The mechanism of aminoacyl-tRNA-based transformation has been documented for the Asp-Asn, Glu-Gln, Ser-Selenocysteine, and Met-iMet couples. In the SRM, the only metabolic constraints are as follows: the heads of amino acid families have to be encoded before the derived amino acids; the mechanism of aminoacyl-tRNA-based transformation is only one among various other possibilities.

The list of pre-biotic amino acids may reach ten, adding ESLIPT to the most abundant AGDV. Greatest attention has been given to the latter, all four residing in the 3′C row of anticodes, which would make it easy to obtain high ΔG duplexes when the 5′G is chosen. Recent reports in this direction come from distinct research lines, which may seem to be mutually reinforcing, but they lack an explicit mechanism for the encodings. The work of Higgs (2009) is derived from statistical calculations based on minimization of the consequences of changes of meanings of codes and proposes that the above four amino acids were the heads of the four columns of the code. The work of Trifonov (2004) is derived from a consensus among a large variety of proposals for the early encodings (of course, the SRM was excluded from the consensus, as an outlier together with a few others) and goes further in pinpointing the two most abundant among the pre-biotic, which may form a high ΔG duplex Gly-GCC : Ala-GGC (note the 5′G SC triplets). This couple of correspondences would be the heads of two families of amino acids and of triplets. The triplets would reside in the two strands of an original RNA, backing earlier ideas (Rodin et al. 1993, 1996). Trifonov was also able to show an evolutionary trend towards reduction of the Gly content of proteins (the Gly molecular clock; Trifonov 1999). The correlate Ala clock was not investigated owing to Ala not showing an intense decrease along protein sequence evolution and not being generally a frequent component of protein active sites.

Chemical systems considerations can be advocated against the rationale based solely on external pushing dynamics. Given long times, even in the presence of some polymerization mechanism, reactions would tend to equilibrate with their reverse (Eschenmoser 2007). Models based on these mechanisms (more extensively reviewed in Guimarães 2011, 2012) are considered proto-codes, which may have been important for, for example, pre-biotic evolutionary ‘learning’ periods, where peptides and proto-tRNAs might have co-existed and perhaps developed some mutual adjustments including the stereo-chemical affinities (Yarus et al. 2009). A proposition of the SRM is the intervention of an endogenous pulling dynamics established by the formation of a peptide-consuming process. This system can only be stabilized if it can propitiate the fixation of synthesis pathways for the amino acids composing the peptides. The SRM focuses on the simplest of all routes, based entirely on C1 compounds, departing from the heterotrophic tradition based on more complex nutrients.

The SRM has been gathering strength cautiously since the first preliminary report (Guimarães 1996), searching for empirical data produced by others and for other purposes that could fulfill its predictions, at the same time, refining the model that could only be published in full in 2008 (Guimarães et al. 2008a, b). Other components and characters are needed to be incorporated into the network since only about a dozen have been considered by now. Main predictions fulfilled were on the metabolism (Guimarães 2011) and on the tRNA dimer/synthetase network construction (Guimarães 2012), enabling the present stage of identifying the six-step chronology based on the four NSC modules of dimers. One main focus of investigation is centered on refining the theoretical details of the pulling dynamics, here identified with the construction of protein stretches – oligomeric sites and motifs – devoted to specific functions. It is expected that experimental tests on the proposed dimer-directed protein synthesis mechanism may progress into the building of synthetic RNP and coding systems.

A challenge pointed out by this work is the identification of the pre-biotic counterparts to the proto-cells (LUCA) metabolic pathways whose chemical mechanisms would be equivalent to the C1-utilizing methylotrophic/autotrophic routes. The biochemical task should not be easy, given the diversity of enzyme co-factors involved in extant organisms (Guimarães 2011; Madigan and Martinko 2006; Chistoserdova 2011). The traditional consideration of the GSC as integrator of the C1-C3 metabolism is now expanded, considering it a restoration reservoir of the metabolic network at periods of heterotrophic crises. This firm condition for stability would have contributed to the early fixation of the Gly and Ser encodings together with the GSC enzymes. For RNP-building, Gly contributed mostly with the RNA-binding property and Ser with the information-rich character. Biologic precedents qualify the methylotrophs and the acetogenic autotrophs as candidates for having composed early cellular communities. These oligotrophic organisms frequently form syntrophic networks which may also have exchanged horizontal gene transfers, indicated to have been frequent among early cells.

A main drive for open-ended evolution of the cellular system is indicated to be the sink dynamics exerted by the amino acid consumption in the protein synthesis subsystem that has to be kept active in maintaining the metabolic flux unimpeded. The diversifying and accumulative power of proteins is very large but not inexhaustible. A state is eventually reached where their accumulation saturates to the risk of triggering denaturation and degradation routes (Dill et al. 2011). It is considered that avoidance of such risks was obtained through mechanisms beyond the mere balance between proteolysis with re-utilization of amino acids. Cells would be losing or expelling out portions of their bodies, for example, through cytoplasm fission, gemulation, exosome formation, or other modes of excretion of cytoplasm chunks. The final solution was the development of the complex process of reproduction where excessive body growth was combined with genome duplication followed by the precise genome distribution into the more or less similar portions of the daughter cytoplasms. The filial cells (each cytoplasm portion containing a genome), especially those bearing a small cytoplasm – in the frequent cases of unequal cytoplasm partition (Nyström 2011), would have the protein synthesis activity and the anabolic drive restored. The rationale is reinforced by the discovery of shared components in both processes (Makarova et al. 2010; Deatherage and Cookson 2012) of exosome formation and of cell division.

References

Agris PF, Vendeix FA, Graham WD (2007) tRNA’s wobble decoding of the genome: 40 years of modification. J Mol Biol 366:1-13

Altmann M, Linden P (2010) Power of yeast for analysis of eukaryotic translation initiation. J Biol Chem 285:31907-31912

Anderson WB, Board PG, Anders MW (2004) Glutathione transferase zeta-catalyzed bioactivation of dichloroacetic acid: reaction of glyoxylate with amino acid nucleophiles. Chem Res Toxicol 17:650-662

Anthony C (1982) The biochemistry of methylotrophs. Academic, London

Arava Y (2009) Compaction of polyribosomal mRNA. RNA Biol 6:399-401

Arnez JG, Moras D (1997) Structural and functional considerations of the aminoacylation reaction. Trends Biochem Sci 22:211-216

Barabási AL, Oltvai ZN (2004) Network biology: understanding the cell’s functional organization. Nat Rev Genet 5:101-113

Bashan A, Yonath A (2008) Correlating ribosome function with high-resolution studies. Trends Microbiol 16:326-335

Bauwe H, Kolukisaoglu U (2003) Genetic manipulation of glycine decarboxylation. J Exp Bot 54:1523–1535

Beier H, Grimm M (2001) Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs. Nucl Acids Res 29:4767-4782

Bekaert M, Firth AE, Zhang Y, Gladyshev VN, Atkins JF, Baranov PV (2010) Recode-2: new design, new search tools, and many more genes. Nucl Acids Res 38:D69-74

Berezovsky IN, Kilosanidze GT, Tumanyan VG, Kisselev LL (1999) Amino acid composition of protein termini are biased in different manners. Prot Eng 12:23-30

Beuning PJ, Musier-Forsyth K (1999) Transfer RNA recognition by aminoacyl-tRNA synthetases. Biopolymers 52:1-28

Bloch DP, McArthur B, Widdowson R, Spector D, Guimarães RC, Smith J (1984) tRNA-rRNA sequence homologies: a model for the generation of a common ancestral molecule and prospects for its reconstruction. Orig Life Evol Biosph 14:571-578

Bloch DP, McArthur B, Guimarães RC, Smith J, Staves MP (1989) tRNA-rRNA sequence matches from inter- and intraspecies comparisons suggest common origins for the two RNAs. Brazil J Med Biol Res 22:931-944

Caetano-Anollés G, Kim HS, Mittenthal JE (2007) The origin of metabolic networks inferred from phylogenomic analysis of protein architecture. Proc Natl Acad Sci USA 104:9358-9363

Carter AP, Clemons WMJr, Brodersen DE, Warren RJM, Hartsch T, Wimberly BT, Ramakrishnan V (2001) Crystal structure of an initiation factor bound to the 30S ribosomal subunit. Science 291**:**498-501

Chistoserdova L (2011) Modularity of methylotrophy, revisited. Envir Microb 13:2603-2622

Creighton TE (1993) Proteins: structures and molecular properties. WH Freeman, New York

Cusack S (1997) Aminoacyl-tRNA synthetases. Curr Opin Struct Biol 7:881-889

Deatherage BL, Cookson BT (2012) Membrane vesicle release in bacteria, eukaryotes, and archaea: a conserved yet underappreciated aspect of microbial life. Infect Immun 80:1948-1957

Di Giulio M (2008) An extension of the coevolution theory of the origin of the genetic code. Biol Direct 3:37

Dill KA, Ghosh K, Schmit JD (2011) Physical limits of cells and proteomes. Proc Natl Acad Sci USA 108:17876-17882

Egholm M, Buchardt O, Nielsen PE, Berg RH (1992) Peptide nucleic acids (PNA). Oligonucleotide analogs with an achiral backbone. J Am Chem Soc 114:1895-1897

Ertem G (2004) Montmorillonite, oligonucleotides, RNA and origin of life. Orig Life Evol Biosph 34:549-570

Ertem G, O´Brien AMS, Ertem MC, Rogoff DA, Dworkin JP, Johnston MV, Hazen RM (2008) Abiotic formation of RNA-like oligomers by montmorillonite catalysis: part II. Int. J. Astrobiol. 7**:**1-7

Eschenmoser A (2007) Basic questions about the origins of life - kinetic control. Orig Life Evol Biosph 37:309-314

Etten WJV, Janssen GR (1998) An AUG initiation codon, not codon-anticodon complementarity, is required for the translation of unleadered mRNA in Escherichia coli. Molec Microbiol 27:987-1001

Farias ST, Bonato MCM (2003) Preferred amino acids and thermostability. Genet Molec Res 2:383-393

Farias ST, Guimarães RC (2007) Aminoacyl-tRNA synthetase classes and groups in prokaryotes. J Theor Biol 250:221-229

Farias ST, Moreira CHC, Guimarães RC (2007) Structure of the genetic code suggested by the hydropathy correlation between anticodons and amino acid residues. Orig Life Evol Biosph 37:83-103

Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185**:**862-864

Grosjean H, Crécy-Lagard V, Marck C (2010) Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes. FEBS Lett 584:252-264

Gruic-Sovulj I, Landeka I, Söll D, Weygand-Durasevic I (2002) tRNA-dependent amino acid discrimination by yeast seryl-tRNA synthetase. Eur J Biochem 269:5271-5279

Guimarães RC (1996) Anti-complementary order in the genetic coding system. Int. Conf. Orig. Life (ISSOL – The Astrobiology Society) 11**:**100

Guimarães RC (2001) Two punctuation systems in the genetic code. In: Chela-Flores J, Owen T, Raulin F (eds) First steps in the origin of life in the universe. Kluwer, Dordrecht, pp 121-124

Guimarães RC (2011) Metabolic basis for the self-referential genetic code. Orig Life Evol Biosph 41**:**357-371

Guimarães RC (2012) Mutuality in discrete and compositional information: perspectives for synthetic genetic codes. Cogn Comput 4:115-139

Guimarães RC, Erdmann VA (1992) Evolution of adenine clustering in 5S ribosomal RNA. Endocyt Cell Res 9:13-45

Guimarães RC, Moreira CHC (2004) Genetic code – a self-referential and functional model. In: Progress in biological chirality. Pályi G, Zucchi C, Caglioti L (eds) Elsevier, Oxford, pp 83-118

Guimarães RC, Moreira CHC, Farias ST (2008a) A self-referential model for the formation of the genetic code. Theory Biosci 127**:**249-270

Guimarães RC, Moreira CHC, Farias ST (2008b) Self-referential formation of the genetic system. In: Barbieri M (ed) The codes of life – the rules of macroevolution. Springer, Dordrecht, pp 68-110

Higgs PG (2009) A four-column theory for the origin of the genetic code: tracing the evolutionary pathways that gave rise to an optimized code. Biol Direct 4:16

Higgs PG, Pudritz RE (2009) A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. Astrobiology 9:483-490

Horiya S, Li X, Kawai G, Saito R, Katoh A, Kobayashi K, Harada K (2003) RNA LEGO: magnesium-dependent formation of specific RNA assemblies through kissing interactions. Chem Biol 10:645-654

Illangasekare M, Yarus M (1999) A tiny RNA that catalyzes both aminoacyl-RNA and peptidyl-RNA synthesis. RNA 5**:**1482-1489

José MV, Morgado ER, Govezensky T (2011) Genetic hotels for the standard genetic code: evolutionary analysis based upon novel three-dimensional algebraic models. Bull Math Biol 73:1443-1476

Knight RD, Freeland SJ, Landweber LF (2001) Rewiring the keyboard: evolvability of the genetic code. Nat Rev Genet 2:49-58

Kolitz SE, Takacs JE, Lorsch JR (2009) Kinetic and thermodynamic analysis of the role of start codon/anticodon base pairing during eukaryotic translation initiation. RNA 15**:**138-152

Kuwabara T, Warashina M, Nakayama A, Ohkawa J, Taira K (1999) tRNA^{Val} heterodimeric maxizymes with high potential as gene inactivating agents: simultaneous cleavage at two sites in HIV-1 tat mRNA in cultured cells. Proc Natl Acad Sci USA 96**:**1886-1891

Kyte J (1995) Structure in protein chemistry. Garland, New York

Lambert JF (2008) Adsorption and polymerization of amino acids on mineral surfaces: a review. Orig Life Evol Biosph 38:211-242

Lee N, Bessho Y, Wei K, Szostak JW, Suga H (2000) Ribozyme-catalyzed tRNA aminoacylation. Nat Struct Biol 7**:**28-33

Madigan MT, Martinko JM (2006) Brock biology of microorganisms. Prentice Hall, Upper Saddle River

Makarova KS, Yutin N, Bell SD, Koonin EV (2010) Evolution of diverse cell division and vesicle formation systems in archaea. Nat Rev Microbiol 8:731-741

Martin W, Russell MJ (2003) On the origins of cells: a hypothesis for the evolutionary transition from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. Phil Trans R Soc Lond B 358:59-85

Meinnel T, Sereno A, Giglione C (2006) Impact of the N-terminal amino acid on targeted protein degradation. Biol Chem 387:839-851

Miller DL, Yamane T, Hopfield JJ (1981) Effect of tRNA dimer formation on polyphenylalanine biosynthesis. Biochemistry 20**:**5457-5461

Moras D, Dock AC, Dumas P, Westhof E, Romby P, Ebel JP, Giegé R (1986) Anticodon-anticodon interaction induces conformation changes in tRNA: yeast tRNA^{Asp}, a model for tRNA-mRNA recognition. Proc Natl Acad Sci USA 83:932-936

Nielsen PE (2007) Peptide nucleic acids and the origin of homochirality of life. Orig Life Evol Biosph 37:323-328

Nissen P, Hansen J, Ban N, Moore PB, Steitz TA (2000) The structural basis of ribosome activity in peptide bond synthesis. Science 289**:**920-930

Noller HF (2004) The driving force for molecular evolution of translation. RNA 10**:**1833-1837

Nuevo M, Bredehöft JH, Meierhenrich UJ, D'Hendecourt L, Thiemann WHP (2010) Urea, glycolic acid, and glycerol in an organic residue produced by ultraviolet irradiation of interstellar pre-cometary ice analogs. Astrobiology 10**:**245–256

Nyström T (2011) Spatial protein quality control and the evolution of lineage-specific ageing. Phil Trans R Soc Lond B 366:71-75

Ogle JM, Brodersen DE, Clemens Jr WM, Tarry MJ, Carter AP, Ramakrishnan CV (2001) Recognition of cognate transfer RNA by the 30S ribosomal subunit. Science 293**:**897-902

Quayle JR, Ferenci T (1978) Evolutionary aspects of autotrophy. Microbiol Molec Biol Rev 42:251-273

Park MC, Kang T, Jin D, Han JM, Kim SB, Park YJ, Cho K, Park YW, Guo M, He W, Yang XL, Schimmel P, Kim S (2012) Secreted human glycyl-tRNA synthetase implicated in defense against ERK-activated tumorigenesis. Proc Natl Acad Sci USA E640-647

Polycarpo C, Ambrogelly A, Berube A, Winbush SM, McCloskey JA, Grain PF, Wood JL, Söll D (2004) An aminoacyl-tRNA synthetase that specifically activates pyrrolysine. Proc Natl Acad Sci USA 101:12450-12454

Poole AM, Jeffares DC, Penny D (1998) The path from the RNA world. J Mol Evol 46:1-17

Powner MW, Gerland B, Sutherland JD (2009) Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. Nature 459**:**239-242

Raskin DM, Judson N, Mekalanos JJ (2007) Regulation of the stringent response in the essential function of the conserved bacterial G protein CgtA in Vibrio cholera. Proc Natl Acad Sci USA 104**:**4636-4641

Reitzer L (2003) Nitrogen assimilation and global regulation in *Escherichia coli*. Ann Rev Microb 57:155-176

Rodin S, Ohno S, Rodin A (1993) Transfer RNAs with complementary anticodons: could they reflect early evolution of discriminative genetic code adaptors? Proc Natl Acad Sci USA 90:4723-4727

Rodin S, Rodin A, Ohno S (1996) The presence of codon-anticodon pairs in the acceptor stem of tRNAs. Proc Natl Acad Sci USA 93:4537-4542

Romby P, Westhof E, Moras D, Giegé, Houssier C, Grosjean H (1986) Studies on anticodon-anticodon interactions: hemi-protonation of cytosines induces self-pairing through the GCC anticodon of *E. coli* tRNA-Gly. J Biomol Struct Dyn 4:193-203

Santoso S, Hwang W, Hartman H, Zhang SG (2002) Self-assembly of surfactant-like peptides with variable glycine tails to form nanotubes and nanovesicles. NanoLetters 2**:**687-691

Skouloubris S, Pouplana LR, Reuse H, Hendrickson TL (2003) A noncognate aminoacyl-tRNA synthetase that may resolve a missing link in protein evolution. Proc Natl Acad Sci USA 100:11296-11302

Sobolevsky Y, Trifonov EN (2006) Protein modules conserved since LUCA. J Mol Evol 63:622-634

Sobolevsky Y, Guimarães RC, Trifonov EN (2012) Towards functional repertoire of the earliest proteins. J Biomol Struct Dyn (in press)

Szostak JW (2009) Systems chemistry on early earth. Nature 459**:**171-172

Tamura K (2011) Molecular basis for chiral selection in RNA aminoacylation. Int J Mol Sci 12**:**4745-4757

Tamura K, Schimmel PR (2006) Chiral-selective aminoacylation of an RNA minihelix: mechanistic features and chiral suppression. Proc Natl Acad Sci USA 103:13750-13752

Trifonov EN (1999) Glycine clock: eubacteria first, archaea next, protoctista, fungi, planta and animalia at last. Gene Ther Mol Biol 4:313-322

Trifonov EN (2004) The triplet code from first principles. J Biomol Struct Dyn 22:1-11

Van der Linden MG, Farias ST (2006) Correlation between codon usage and thermostability. Extremophiles 10**:**479-481

Varschavsky A (1996) The N-end rule: functions, mysteries, uses. Proc Natl Acad Sci USA 93:12142-12149

Vetsigian K, Woese C, Goldenfeld N (2006) Collective evolution and the genetic code. Proc Natl Acad Sci USA 103:10696-10701

Walsh D, Mohr I (2011) Viral subversion of the host protein synthetic machinery. Nat Rev Microb 9:860-875

Wang KH, Hernandez GR, Grant RA, Sauer RT, Baker TA (2008) The molecular basis of N-end rule recognition. Mol Cell 32:406-414

Wong JTF (2005) Coevolution theory of the genetic code at age thirty. BioEssays 27**:**416-425

Xia T, SantaLucia JJr, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry 37**:**14719-14735

Yamane T, Miller DL, Hopfield JJ (1981) Interaction of elongation factor Tu with the aminoacyl-tRNA dimer Phe-tRNA:Glu-tRNA. Biochemistry 20**:**449-452

Yarus M (2011) The meaning of a minuscule ribozyme. Phil Trans R Soc Lond B 366:2902-2909

Yarus M, Widmann JJ, Knight R (2009) RNA-amino acid binding: a stereochemical era for the genetic code. J Mol Evol 69:406-429

Zaia DAM, Zaia CTBV, Santana H (2008) Which amino acids should be used in prebiotic chemistry studies? Orig Life Evol Biosph 38:469-488

Zhang L, Peritz A, Meggers E (2005) A simple glycol nucleic acid. J Am Chem Soc 127:4174-4175

Zrenner R, Stitt M, Sonnewald U, Boldt R (2006) Pyrimidine and purine biosynthesis and degradation in plants. Ann Rev Plant Biol 57:805-836

Acknowledgments The continued collaboration of the colleagues in the Grupo de Auto-organização, Centro de Lógica e Epistemologia, UNICAMP, especially Alfredo Pereira Júnior, and of Carlos Henrique Costa Moreira, Edward N. Trifonov, and Sávio Torres de Farias are deeply appreciated, as well as the fruitful discussions with Marco V. José and Laurent Palka, and the efficient help of Sheila M. O. Reis with the illustrations.

Competing financial interests: The author declares no competing financial interests.

BOX 1 The orthogonal matrix of anticodon triplets

Boxes, principal dinucleotides and kinds of triplets

The matrix is built according to the increasing hydrophilicity order of bases AGCU, which also depicts the main symmetries. A box is identified by the principal dinucleotide (pDiN), the constant portion of the set of four codons (5'NNw3') or three anticodons (5'wNN3') that contain different bases in the wobble (w) position. Each box has a half where the lateral bases are of the same kind (both R or both Y; non-self-complementary, NSC), and a half where the lateral bases are of different kinds (one R, one Y; self-complementary, SC). Thermodynamic properties of the NSC and SC oligonucleotides are explored in Xia et al. (1998).

Sectors of kinds of triplets

The homogeneous (Ho) sector is identified by the principal dinucleotides that contain only purines (wRR quadrant) or only pyrimidines (wYY quadrant) and by the non-self-complementary triplets that extend the homogeneous character to the wobble position (RRR, YYY). The mixed (Mx) sector is identified by the principal dinucleotides that contain one R and one Y (wRY quadrant, wYR quadrant) and by their non-self-complementary triplets (YRY, RYR). The organization of the orthogonal matrix in sectors follows the diagonals. The NSC triplets are shaded.

Homogeneous sector wRR quadrant			Mixed sector wYR quadrant	
RAA	RGA		RCA	RUA
YAA	YGA		YCA	YUA
RAG	RGG		RCG	RUG
YAG	YGG		YCG	YUG
RAC	RGC		RCC	RUC
YAC	YGC		YCC	YUC
RAU	RGU		RCU	RUU
YAU	YGU		YCU	YUU
Mixed sector wRY quadrant			Homogeneous sector wYY quadrant	

BOX 2 Simple and complex boxes

A box may be simple or complex with respect to the number of attributions its set of triplets corresponds to. Simple boxes have all triplets in the set corresponding to one synthetase or amino acid (shaded). The wobble position is N and the attribution may be referred to the pDiN. The degree of code degeneracy for the amino acid in the box is tetracodonic or trianticodonic. Complex boxes have the set of triplets (or the pDiN) corresponding to more than one attribution, either two synthetases (or amino acids), or one or two amino acids and a punctuation code. The identity of the base in the wobble position is essential for the synthetase or the Release Factor specificity. In the standard code, eight boxes are simple and eight are complex. In six of the complex boxes the attributions correspond to the division of triplets into the self-complementary and non-self-complementary halves. In the other two, one half is further split to accommodate amino acids and the punctuation codes (YAU Ile, Met, iMet; YCA Trp, X). Hexacodonic attributions correspond to three synthetases (SerRS, LeuRS, ArgRS) with most complex specificity, combining one simple box and a portion of a complex box.

Homogeneous sector wRR quadrant			Mixed sector wYR quadrant	
RAA Phe YAA Leu	wGA Ser		RCA Cys CCA Trp UCA X	RUA Tyr YUA X
wAG Leu	wGG Pro		wCG Arg	RUG His YUG Gln
wAC Val	wGC Ala		wCC Gly	RUC Asp YUC Glu
R,UAU Ile CAU Met CAU iMet	wGU Thr		RCU Ser YCU Arg	RUU Asn YUU Lys
Mixed sector wRY quadrant			Homogeneous sector wYY quadrant	

Table 1. Six-step chronology of encodings. The succession of encodings follows the sectors and modules of dimers, with attributions dictated by the metabolic supply. Start with the Glycine-Serine Cycle (Modules 1 and 2, five amino acids) to the mature central metabolic pathways (Step 3, total ten amino acids). Physiologic correlations indicate formation of the RNP realm from the homogeneous sector attributions. At completion of this sector, Asp concedes YUC to Glu (derived from 2-Oxoglutarate of the Citrate Cycle) and recedes to GUC. Pro (derived from Glu) substitutes Gly-WGG. Asn concedes YUU to Lys (derived from Asp or from 2-Oxoglutarate) and recedes to GUU. Leu WAA concedes GAA to Phe (derived from Phosphoenolpyruvate and a tetrose) and recedes to YAA. The DNP realm (mixed sector) starts with the expansion of the ArgRS specificity (Arg is derived from Glu) from the YCU to the WCG. At completion of the elongation encodings, punctuation is installed utilizing a different pDiN from the same Met anticodon, which leads to deletion of the X tRNAs. These were substituted by the Release Factors, Trp YCA receding to CCA and Tyr WUA to GUA.

Homogeneous principal dinucleotide (pDiN) sector WRR:WYY anticodons						
Amino acids derived from the Glycine-Serine Cycle						
Step 1 (Module 1)	(1a) Gly WCC : Gly WGG		(1b) Ser WGA : Ser WCU	**Step 2 (Module 2)**	(2a) Leu WAG : Asp WUC	(2b) Leu WAA : Asn WUU
Concessions to amino acids derived from other central metabolic pathways						
Step 3 (2+)	Recede Concede	- Pro WGG	Ser GCU Arg YCU		Asp GUC Glu YUC	Leu YAA Asn GUU Phe GAA Lys YUU
Mixed pDiN sector WRY:WYR anticodons						
Step 4 (Module 3)	(3a) Ala WGC : Arg WCG		(3b) Thr WGU: Cys GCA, Trp YCA	**Step 5 (Module 4)**	(4a) Val WAC: His GUG, Gln YUG	(4b) Ile GAU, UAU, Met CAU :Tyr WUA
Step 6 (4+)	**Punctuation**					iMet CAU
	Recede Concede		Trp CCA X UCA			Tyr GUA X YUA

Table 2. Variant codes. Patterns discerned: (1) One half of the occurrences, including all residing in simple boxes, are in the hexacodonics. (2) Most frequent sources (the standard meaning of the triplet that suffered a change) are in X (seven) and Arg (six); there are only six changes involving central Y codons. Most frequent destinations (new meaning of the triplet; six) are unknown (?) or the donor codons are not used. (3) From an amino acid to another amino acid: one bidirectional exchange (Ile↔Met); six cases are unidirectional, one from early to late (Leu→Thr), five from late to early (Ser→Gly; Arg→Ser, Gly; Leu→Ser; Lys→Asn). (4) From an amino acid to X: one each from Gly and Ser. (5) From X to an amino acid: six. (6) Changes inside a complex box: not considering the bidirectional Ile↔Met, they are always directional from a late to an earlier attribution ($\mathrm{X}^{\underline{\mathrm{UG}}\mathrm{A}}$→Trp or Cys, $\mathrm{X}^{\underline{\mathrm{UA}}\mathrm{A}}$→Tyr, Lys→Asn, $\mathrm{Arg}^{\underline{\mathrm{AG}}\mathrm{G}}$→Ser). Chronological steps in parenthesis. Adapted from the list of >100 variations in Knight et al. (2001).

Attributions	**Complex boxes**		**Simple boxes**	
Hexacodonic	Ser (**1b**) AGY→	Gly (**1a**); ?	Ser (**1b**) UCA→	X (**6**)
	Arg (**3**) AGG→ AGR→ AGA→	Ser (**1b**) Gly (**1a**); X (**6**) ?	Arg (**3**) CGN→ CGG→	? ?
			Leu (**2a**) CUN→ CUG→	Thr (**4b**) Ser (**1b**)
Non-hexacodonic	Lys (**3**) AAA→	Asn (**2b**)	Total 22 types of variant codes	
	Ile (**5b**) AUA→	Met (**5b**); ?		
	Met (**5b**) AUG→	Ile (**5b**)		
	X (**6**) UAG→ UAA→ UAR→	Leu (**2a,b**); Ala (**4a**) Tyr (**5b**) Gln (**5a**)		
	X (**6**) UGA→	Trp (**4b**); Cys (**4b**); ?		

Table 3. Configuration of the initiation codes. The initiation codons show variation in the 5′ position, pairing with the constant iMet anticodon 3′U. The pDiN in the iMet is slipped with respect to the pDiN in the elongation Met, this maintaining the standard elongation pDiN pattern. It follows that this portion of the initiation context is a composite of the two first codon-anticodon pairs, adequately positioned to confer strong stability to the process via formation of a tetra-nucleotide run without a wobble position in between the two pDiN. The initiation function is not the mere starting the chain by the entrance of the iMet but the synthesis of the first peptide bond (≈) from iMet plus the second amino acid that is most frequently Val and Thr (Figure 10). The tetra-nucleotide shows a peculiar repeated constitution and complex rotational symmetries: direct repeats in the crossed strands in the case of Thr, inverted repeats in the same strands in the case of Val.

	Initiation			**Elongation**		
		↓		↓		↓
Codons	5′ N U G	.	A C N^w	.	N N N^w	.
Anticodons	3′ U A C	.	U G N^w	.	N N N^w	.
Amino acids	iMet	≈	Thr	~	3rd	~
Codons	5′ N U G	.	G U N^w	.	N N N^w	.
Anticodons	3′ U A C	.	C A N^w	.	N N N^w	.
Amino acids	iMet	≈	Val	~	3rd	~

Table 4. Localization of termination codes directed by the initiation code. The localization of the initiation code and of the main termination codes in the last pair of encoded boxes (Module 4b) indicates a unified mechanism for the punctuation system. This is dependent of protein factors but incorporates the guidance by RNA specificities. The suggested mechanism is derived from a series of tested pairings between various options of the code configurations, involving the initiation codons and the anticodons of the wYR quadrant also varying the slippage possibilities (Guimarães and Moreira 2004). Only the set of tests with positive results is shown. It is indicated that elongation anticodes in this quadrant would conflict via competition with the initiation anticodon for the initiation codon; the conflict is overcome by elimination of the interfering elongation anticodons which become substituted by the Release Factors. The gold test is the coincidence between triplet pairs and the real termination codes (yellow). Watson-Crick base pairs green, G:U blue, A:C not highlighted. (a) The initiation codon tested in frame against the anticodons of the wYR quadrant. (b) The first two positions of the initiation codon tested against the first two bases of the anticodons. (c, d) The last two bases of the initiation codon tested against the last two bases of the anticodons. Both options for the first base of the second codon [A of Thr (c), G of Val (d)], pairing with the first base of the anticodons, yield triplet pairs coinciding with the locations of the X anticodons (in the 5′Y of the boxes under test). It is suggested that the A:U Watson-Crick base pair (3′A from the Cys and Tyr boxes, central U from the initiation codon) is decisive for the guidance; were the non-Watson-Crick G:U base pair relevant for the mechanism, termination would also involve the Gln and some of the Arg triplets. The rationale has to incorporate other mechanisms for maintenance of the Trp anticode, which could not be discarded in favour of the X anticodes, at difference from the C<u>UA</u> Tyr anticode.

Patterns of the wYR quadrant anticodons	**Patterns of the initiation codon**			
	(a)	**(b)**	**(c)**	**(d)**
	A U G	A U G	A U G . A	A U G . G
Cys	A C G	A C **G**	**A** **C** G	**A** **C** G
Trp	A C **C**	A C C	**A** **C** C	**A** **C** **C**
X	A C **U**	A C U	**A** **C** **U**	**A** **C** **U**
Tyr	A U G	A **U** **G**	**A** **U** G	**A** **U** G
X	A U **C**	A **U** C	**A** **U** C	**A** **U** **C**
	A U **U**	A **U** U	**A** **U** **U**	**A** **U** **U**
Arg	G C G	G C **G**	**G** **C** G	**G** **C** G
	G C **C**	G C C	**G** **C** C	**G** **C** **C**
	G C **U**	G C U	**G** **C** **U**	**G** **C** **U**
His	G U G	G **U** **G**	**G** **U** G	**G** **U** G
Gln	G U **C**	G **U** C	**G** **U** C	**G** **U** **C**
	G U **U**	G **U** U	**G** **U** **U**	**G** **U** **U**

Figure legends

Figure 1. Nucleoprotein interdependency. One linear and three circular structures are shown. Translational movements (straight arrows) are of the mRNA (pointing to the right) and of the ribosome (pointing to the left) relative to each other. Decoding of codons (of mRNA) by the anticodons (of tRNAs) is shown by their pairing: anticodon 5′ base [thin pair] plus the principal dinucleotide [thick pairs]. The aminoacyl-tRNA enters the ribosomal A-site and receives the peptide that was attached to the tRNA in the ribosomal P-site via the transferase reaction (lower-middle white arrow), then moves to the P-site (upper-middle white arrow). The nascent peptide amino terminus (N-end) points out of the ribosome. After the transferase reaction and the intra-ribosomal movements, the uncharged tRNA moves to the E-site (exit) to be released and for possible re-utilization by the synthetases. The protein product may be a synthetase that binds to a cognate tRNA through the interacting (recognition) sites (yellow ellipses) and transfers an amino acid to the tRNA; this reaction is the decoding of a previously encoded tRNA/synthetase couple. Nucleic acid components, red; amino acids and proteins, green; ribonucleoproteins, violet.

Figure 2. Formation and evolution of the nucleoprotein system. Three levels in the process are shown. Dimers of proto-tRNAs are able to produce oligomers of their attached ligands. Iteration of the transferase activity with different ligands produces some oligomers that are able to bind the producing dimers and stabilize their functions, which were in the biological case the amino acid monomers and the peptide oligomers. Stability of the aggregates introduces a sink dynamics in the self-stimulated system, the proto-RNPs becoming robust consumption ‘forces’ for their components in the metabolic flux. Productivity (performance) under variable and fluctuating contexts drives the evolution of flexible (plastic) stability and of specificity, therewith evolving the tRNAs and the synthetases from the initial proto-RNPs, among the diversity of RNA and protein products. The third level is the evolution of selfing properties and partial autonomy, where the nucleic acids develop the role of template memories and the

proteins the roles of major catalytic components of metabolism and major constituents of biological structures, aside with the participation of RNAs.

Figure 3. Minimal structures of putative proto-tRNAs, dimers, and proto-RNPs. At some point in the evolution of tRNAs, a structure might have been like the 13mer hairpin shown. It would dimerize through (proto-anticodon) loop-pairing. The extended dimer is similar to extant tRNA dimers (Moras et al. 1986). A configuration that would be able to facilitate the transferase reaction is shown as a globular structure, the dimer surrounded by a peptide with binding propensity to the proto-tRNAs. Generic R bases, blue; Y, red; amino acids, green.

Figure 4. Modular structure of the dimer networks. The orthogonal matrix is separated into the central G:C and central A:U columns and the correspondent networks. The two networks are identical, each containing four sub-networks; networks and sub-networks do not overlap each other. Dimers formed by NSC triplets (bold) unite the columns diagonally; dimers formed by SC triplets unite the horizontal rows. 5′A anticodes are absent. The NSC modules contain eight dimers each (2 × 4), numbered according to the chronology of encodings (Table 1); connections are intra-sector. The 5′Y SC modules are untouched by the 5′A elimination and maintain the original 16 dimers each (4 × 4); the 5′G SC modules are drastically reduced to four dimers each (2 × 2); SC dimer connections are inter-sector. Homogeneous sector, green; mixed sector, pink. Integration of the dimer networks comes from the synthetase interactions (Figure 19).

Figure 5. Kinds of triplets and dimers. In dimers of anticodons (a), there is no distinction of the principal dinucleotide (pDiN), which is a functional property of the encoding/decoding mechanisms prevailing at the (b) synthetase, mRNA, and rRNA interactions with the tRNAs and anticodons. The dimers (a) are composed by two overlapping dinucleotides and form four kinds of structures. The NSC triplets and dimers of the homogeneous sector (green) are the simplest sequences, containing monotonic repeats and exposing planar surfaces, with translational symmetry but not rotational symmetry. The NSC dimers of the mixed sector (pink) form duplexes of intermediate complexity. Sequences have rugged surfaces and depict a distinct

symmetry centre; they are composed of inverted repeats in a strand but forming palindrome-like double-stranded configurations, with rotational but not translational symmetry. The SC structures are most complex, combining one homogeneous and one mixed dinucleotide and losing the symmetry centre (violet). At dimer formation, the SC triplets are divided into two forms, both with self-pairings: the 5′GNY and the 5′YNG. In the translation interactions (b), the distinction between NSC and SC triplets may be erased due to the differentiation of the pDiN. A possible difference between homogeneous and mixed pDiN at the mRNA:tRNA pairing (b1) and at the ribosomal quality-checking site (b3) has not yet been investigated. At the synthetase reaction (b2), the interactions may be idiosyncratic (non-degenerate) when the amino acid corresponds to a single tRNA or may reach high degeneracy levels (up to five tRNAs for Leu and Arg). The correspondences are set by complex mapping (compositional) of combinations of a variety of recognition sites (ellipses) in the synthetases and in the tRNAs to one (discrete) couple, the amino acid, and the anticodon triplet.

Figure 6. Encoding the non-self-complementary modules. A unified mechanism follows strictly the ΔG dictums. (Phase 1) Triplets forming the highest ΔG dimer GNG:CNC are occupied by the first encodings in the first pair (a) of boxes, and sequestered for these functions via adaptations of the synthetases and of the tRNAs. (Phase 2) Concentrations of the other four dimers that the GNG triplet was forming before encoding are reduced. (Phase 3) The synthetases develop the pDiN specificity with full degeneracy so that the UNC becomes also occupied and follows the cognate preferential dimerization with GNG. (Phase 4) The GNA:YNU dimers are at high concentrations and are apt for the encodings in the second pair (b) of boxes.

Figure 7. Correlation between hydropathies of anticodon principal dinucleotides and amino acid residues. The mature Module 1 set (GPS, black) is not correlated, indicated to have been fixed in the pre-synthetase stage. The mature Module 2 set (less Pro; blue) is correlated, with a 44^{o} inclination of the linear regression line ($r^2 = 0.991$, $y = 1.001 \times -0.002$). The mixed sector attributions (red) are correlated and with steeper inclination of the regression line (63^{o}, r^2

= 0.849, y = 2.172 × -0.443); attributions of Modules 3 and 4 are not distinguishable from each other. The atypical synthetases are indicated by asterisks. Principal dinucleotides are homogeneous hydrophobic AA, AG, GA, GG[#]; mixed intermediate AC[#], AU[#], UA, CA, GC, GU, CG, UG; and homogeneous hydrophilic CC, UC, UU, CU ([#], inter-sector overlap). Amino acid residues are hydrophobic FIMLVCAWY; hydroapathetic GTSHP and hydrophilic NQEDRK (data from Farias et al. 2007).

Figure 8. The homogeneous sector codes for the non-periodic protein secondary structures. All amino acids preferred in protein non-periodic segments (Np, coils, and turns) belong to the homogeneous sector (GNPSD); those preferred in α-helices (Hel) are distributed in the two sectors, namely, homogeneous (ELKR) and mixed (AMQRH); those preferred in β-strands (Str) are mostly in the mixed sector (VIYCWT), only one in the homogeneous sector (F). Numbers refer to the order of preference (1, first) in the list (data from Creighton 1993).

Figure 9. The homogeneous sector codes for the RNP realm. Amino acids preferred in conserved positions of RNA-binding motifs (abbrev. R) are 75% in the homogeneous sector (GPLKFS), with only VM in the mixed sector. Amino acids belonging to the DNP realm may be preferred exclusively in DNA-binding motifs (abbrev. D; 80% in the mixed sector: AHCT; with only Glu in the homogeneous sector), or preferred in both DNA- and RNA-binding motifs (abbrev. DR; IYRQW, Arg belongs to both sectors). Asp and Asn were not preferred in any nucleic acid-binding motifs. Numbers refer to the order of preference (1 first) in the list (Guimarães and Moreira 2004); highly basic motifs were excluded from the calculations owing to their non-specificity for bases, interacting mostly with the sugar-phosphate backbones.

Figure 10. Heads and tails of protein strings. Data from the N-end rule of protein stabilization against degradation (Varshavsky 1996) (a) and from statistical frequency of amino acids in the N- and C-terminal segments of proteins (Berezovsky et al. 1999) (b). The N-end rule classifies amino acids as strong stabilizers of proteins against catabolism (grades 1, 2; respectively GPMV, SAT) when the half-life of proteins bearing those amino acids at the N-terminus is

long; strong destabilizers (grades 8, 9; RK, LFWY) when the half-life is drastically shortened; and intermediate (grades 3-7; CDENQIH). Data on preferential location of amino acids at protein terminal segments suffer from database-dependency and from frequent ambiguities, when amino acids are preferred or avoided in both locations. H (head), significant statistical preference at the N-terminus (H1), at the second position (H2), or when the two first positions are summed as a dipeptide (H1 + 2); T (tail), significant statistical preference at the C-terminus (T1), at the second position (T2), when the preference is significant at both last positions individually (T1, 2), or when the two last positions are summed as a dipeptide (T1 + 2). There is overall consistency between the two modes of examining protein strings, indicating that properties of amino acids generating the N-end rule were utilized by cells to locate stabilizing amino acids at the heads and destabilizing at the tails, which may be considered a primitive punctuation system (Guimarães 2001).

Figure 11. Error-reduction structure. The property of the encoding/decoding system of reducing the consequences of errors is indicated to derive from (i) the clustering of attributions dedicated to the construction of protein motifs (ii) inside the modules of dimers. The characters examined are aligned with the modules, after transformation of the circular structures (Figures 8-10) into strings. (a, b) The core of strings starts being built by tandem ligation of codes of Module 1 (the N-ends of the cores) followed by Module 2 (the C-ends). Modules 3 and 4 are then added by staggered ligation, extending the N-ends upwards and the C-ends downwards. The attributions are shown (c) in the order (d) of encoding (see Table 1). Functional characters of the attributions are in e-h. The strict Module 1 attributions (GPS) form a conserved cluster through all characters: no hydropathy correlation (e), preferred in non-periodic conformations (f), in RNA-binding motifs (g) and contributing to protein stabilization (h). At the encoding of all attributions of the homogeneous sector (Steps 2 and 3), amino acids preferred in non-periodic conformations are completed and a half of the preferred in protein α-helices are added, together with some others preferred in RNA-binding motifs. Amino acids forming preferentially β-strands and DNA-binding motifs are typical of the mixed sector (Modules 3 and 4). The

components of the WRY quadrant of the mixed sector complete the sets of amino acids preferentially forming RNA-binding motifs and contributing to N-end stabilization. Amino acids that destabilize proteins when residing in the N-ends are all located in Module 2 and in the C-end extension.

Figure 12. Chronology and amino acid hydrodynamic size. A trend is shown, starting (Module 1) with small amino acids and increasing sizes steadily to the end of the homogeneous sector, thereafter maintaining large average sizes. Variability is high in intermediate stages and lower in the sets of Module 4. Synthetases class II (red) are typical of small and class I (blue) of large amino acids. The atypical synthetases are class II for the large amino acids: PheRS, acylating in the class I mode (2′), and LysRS (*), class I or II* in different organisms. ArgRS has codes in both sectors. Small GASP; Medium DCNTEVQH; Large MLIKRFYW (data from Grantham 1974); average per module, black.

Figure 13. Chronology and hydropathy of amino acid residues. The chronologic trend starts hydroapathetic (Module 1) and aRS class II (red), and ends hydrophobic (Module 4b) and aRS class I (blue). Intermediate stages explore the entire hydropathy range and utilize both aRS classes; the atypical aRS class II are at the extreme hydropathies.

Figure 14. Distribution of simple and complex boxes in the orthogonal matrix and along the chronology. The distribution is symmetrical and indicates combinations of properties of both the dimer thermal stability and of the synthetases in helping stabilization of the low (tips of axes (a) and of the intermediate (b) ΔG dimers. The dimers with highest ΔG in the pairs of boxes are shown (Guimarães 2012). The distribution follows strictly the guidance by the triplet constitution, not necessarily the aRS classes (Roman numerals in c). Presence of the symmetry in both sectors (c) indicates that the triplet guiding mechanism has been utilized in full, to near saturation, traversing the formation and the evolution of the code.

Figure 15. Phylogeny of synthetases through the ancestral sequence procedure. Groups examined had at least five adequately homogeneous sequences per synthetase specificity,

obtained from complete genomes (16 Archaea, 36 Bacteria). In consequence of the number of bacterial sequences being excessive with respect to computational requirements for examination en bloc, they were separated into two sections (S1, S2) based on 16S rRNA relatedness. PheRS is composed by two subunits in both kingdoms; GlyRS has two subunits in Bacteria. Adequate group sizes could not be obtained from AsnRS class II, GlnRS class I, and archaeal LysRS class I. Highlighted are the specificities belonging constantly to each of the two large branches of the trees. In aRS class II, the Module 1 specificities (GPS) are constantly clustered and occurring together with HisRS, the remaining hydroapathetic amino acid (ThrRS) being variable. In aRS class I, constant components of the large branch are homogeneous with respect to the triplet central A (LVIM). The constant members of the small branches are at the end of Step 3 (PheRS, ArgRS). Adapted from Farias and Guimarães (2007).

Figure 16. Phylogeny of tRNAs through the ancestral sequence procedure. The Selenocysteine is recoded (superscript R); sequences yielding two ancestors are shown with superscripts$^{1, 2}$. The location of the assignments in the six-step chronology is shown in parenthesis. Five heads of the branches belong to the homogeneous sector of attributions (GLDSP). The two His ancestors are ambiguous, grouping either with Tyr or with Phe. The cluster with Phe may be the exception to the rule demonstrated by the other heads of branches. From Farias ST (Personal communication).

Figure 17. Sketch of some central metabolic pathways relevant to the chronology of encoding. Glycine and Serine belong to a central pathway and were directly encoded; other amino acids derive from precursors in the central pathways. Construction of the network starts from C1 units to build C2 acids and amino acids, then going through C3 and C4 compounds to reach the complex sugars. (i) The Glycine-Serine Cycle (GSC, left panel, black). Glycine derives from C1 sources: (i1) amination of Glyoxylate, an oxidation product of Acetate coming from the autotrophic Acetyl-CoA Pathway; (i2) the Glycine Decarboxylase Complex in the anabolic direction; (i3) complex sources (e. g. Glyoxylate Cycle, Ethylmalonyl-CoA Pathway) not shown. Glyoxylate starts the GSC. Serine is formed by addition of a C1 unit to Glycine; this

reaction is reversible. C3 compounds derived from Serine are precursors to other amino acids: P-Enolpyruvate via Pyruvate forms Leu, Ala, Val; via Chorismate forms Phe, Tyr, Trp. C4 compounds from GSC: Oxaloacetate forms Asp; Malyl-CoA regenerates Acetyl-CoA and Glyoxylate. (ii) The Citrate Cycle (left panel, red font) starts from Oxaloacetate + Acetyl-CoA. Its C5 2-Oxoglutarate forms the Glu family of amino acids. (iii) The third source of amino acids comes from sugar pathways (left panel, green font), starting with gluconeogenesis, especially from C3 compounds. The glycolysis intermediate 3-P-Glycerate may be an important source of Ser (not shown). Ribulose-P (RuP) heads the obligate methylotrophic route, which regenerates the C3 Glyceraldehyde pool (Madigan and Martinko 2006). Ribulose-PP (RuPP) heads the photorespiration route, generating one Glyceraldehyde and feeding the GSC via Glycolate. Erythrose-P (ErP) is joined to 2× P-Enolpyruvate forming Chorismate. Biosynthesis of His comes from Phosphoribosyl-pyrophosphate (PRPP); PRPP participates together with Ser, in the biosynthesis of Trp. Amino acid biosynthesis families (right panel): Gly↔Ser; Ser→Cys, Trp; Asp→Asn, Lys, Thr, Met; Thr→Ile; Glu→Gln, Pro, Lys, Arg. Components of Modules 1 and 2 derived from the GSC are highlighted in colour; other components of these modules correspond to Step 3.

Figure 18. Metabolic contribution of amino acids to bacterial cell mass and to the protein alphabet along the chronology of encoding. (i) Participation of amino acids in total bacterial protein composition (black; taken as baseline for comparisons; from Reitzer 2003); (ii) contribution of amino acids with carbons to the biosynthesis of other components of bacterial cell mass (red; cumulative micromoles %; from Reitzer 2003); (iii) contribution of amino acids to the expansion of the protein alphabet via post-translational modifications (green; cumulative number of types of modifications %; from Kyte 1995). The GSC components (Gly, Ser) and its immediately derived (Asp) are highlighted as most important contributors of carbons to the skeletons of other components of bacterial cell mass, followed by Glu of Step 3. The highly informational amino acids, indicated by the participation in post-translational modifications are spread along the chronology [S (Step 1), KR (Step 3), and CHY (Steps 4-5)].

Figure 19. Integration of the networks of tRNA dimers by interactions between the synthetases. An RNP network is formed by aggregation of the aRS (nodes) that are also connected through binding to tRNA dimers (non-directional edges). There are three sub-networks: one integrating the central G:C tRNAs plus their aRS (Modules 1 + 3; eight amino acids plus the $X^{U\underline{CA}}$); one containing the central A:U Module 2 plus the Q and $X^{Y\underline{UA}}$ of Module 4; and one containing the central A:U Module 4 plus the D and N of Module 2. Superscripts are the total number of dimers (left) and the number of NSC dimers (right) connected to a node. Edges are black for the NSC dimers, red for the SC dimers. Highlighting: green, the self-aggregated SerRS; violet, the hetero-aggregated aRS of MaRS; yellow, the eliminated X anticodons; pink, the most isolated TrpRS and AsnRS (connected to cycles only through SC dimers and not belonging to MaRS). The large hubs are the hexacodonic attributions: SerRS and ArgRS in the central G:C sub-network, LeuRS in the central A:U Module 2 sub-network; in the central A:U Module 4 sub-network, the ValRS and IleRS hubs are moderately large. Basic structures in networks are the interwoven smaller cycles, formed by three (minimal) or four nodes. The central G:C sub-network is complex, the only one containing tri-node cycles and with five tetra-node cycles: a central Module 1 cycle (PGSR] is surrounded by four Module 3 cycles [the PWSX plus three centred on AlaRS and ThrRS (RACT, SACT, GACT)]. The central A:U Module 2 sub-network is of intermediate complexity, containing four tetra-node cycles. The central A:U Module 2 sub-network is simple, containing one cycle with two dangling nodes. The connectivity indices (average number of different nodes connected to a node; subscripts to the nodes) are indicators of complexity. They were calculated after elimination of the X, reducing the connectivity of some nodes [${}^{12}S^{4}_{5}$, ${}^{10}P^{4}_{3}$, ${}^{12}L^{4}_{3}$] and modifying the SWPX cycle into SWPR plus SWPG. Revised and updated from Guimarães (2012).

Fig.1

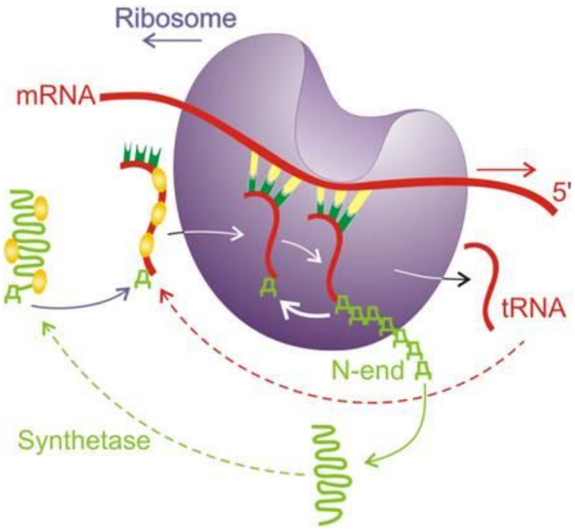


Fig.2

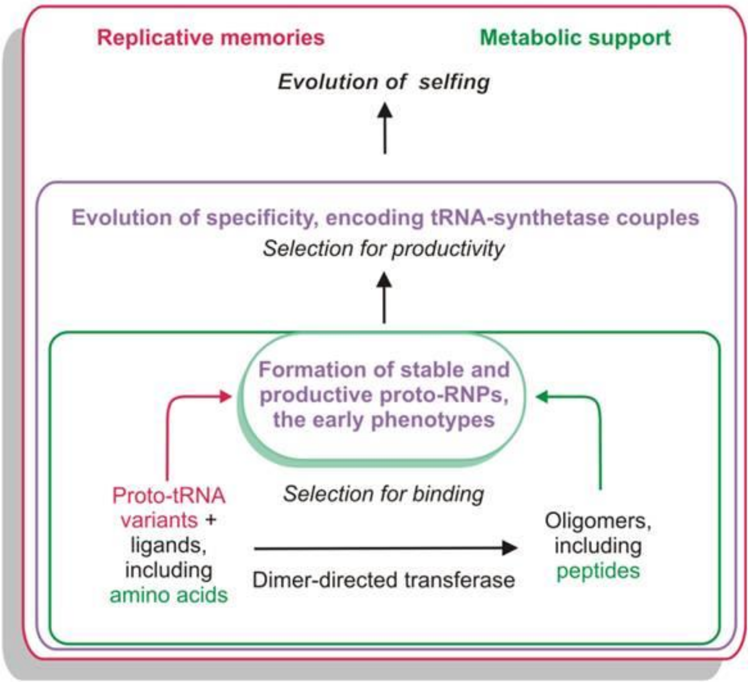
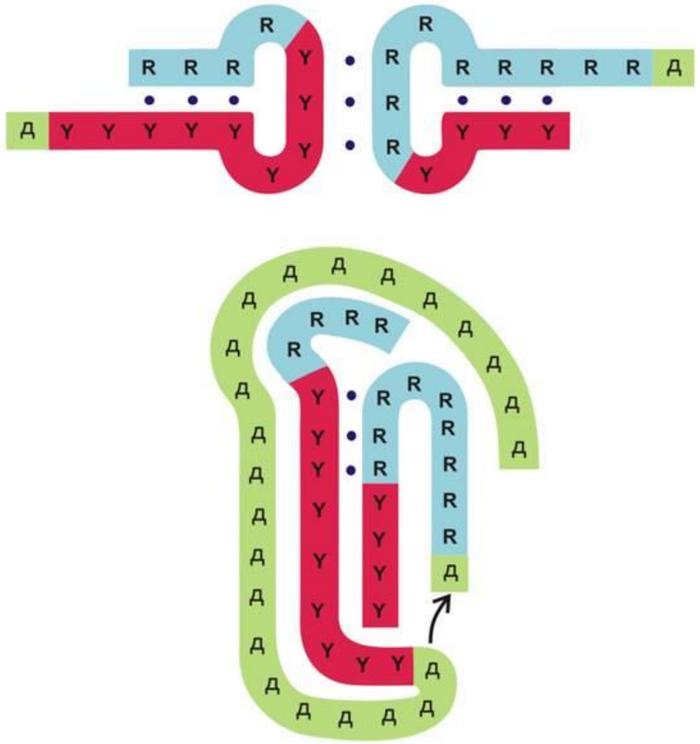

Replicative memories
Metabolic support
Evolution of selfing
Evolution of specificity, encoding tRNA-synthetase couples
Selection for productivity
Formation of stable and productive proto-RNPs, the early phenotypes
Proto-tRNA variants + ligands, including amino acids
Selection for binding
Dimer-directed transferase
Oligomers, including peptides

Fig.3



R
R R R
Y
Y
Y
Y
Д Y Y Y Y Y
R
R R R R R R Д
R
R
R
Y
Y Y Y
Д Д
R R R R
Y Y Y Y Y Y Y Y Y Y
R R R R R R R R R R R
Y Y Y Y
Д

Fig.4

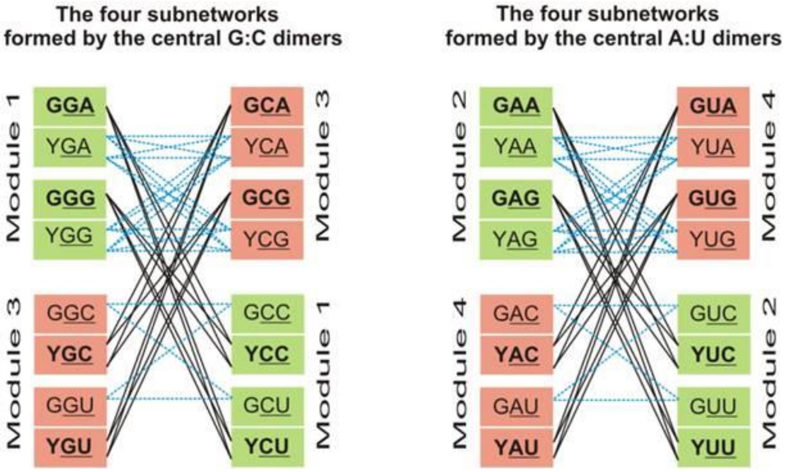

The four subnetworks
formed by the central G:C dimers
Module 1
GGA
YGA
GGG
YGG
Module 3
GGC
YGC
GGU
YGU
Module 3
GCA
YCA
GCG
YCG
Module 1
GCC
YCC
GCU
YCU
The four subnetworks
formed by the central A:U dimers
Module 2
GAA
YAA
GAG
YAG
Module 4
GAC
YAC
GAU
YAU
Module 4
GUA
YUA
GUG
YUG
Module 2
GUC
YUC
GUU
YUU

Fig.5

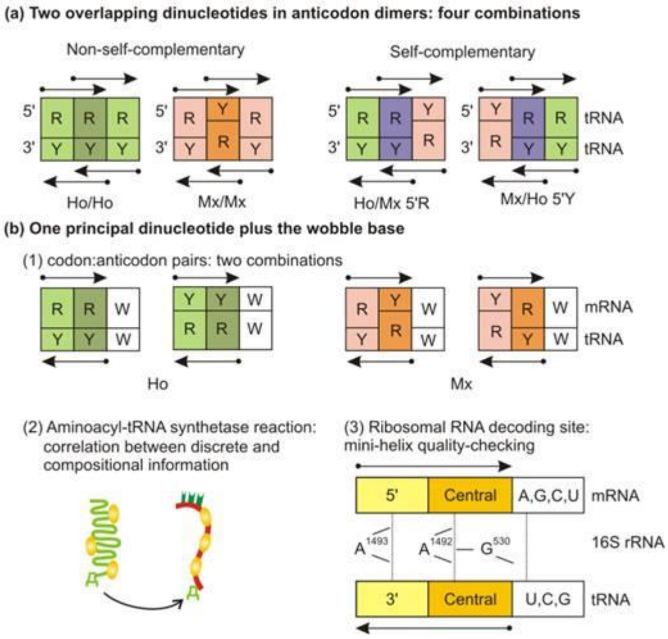

(a) Two overlapping dinucleotides in anticodon dimers: four combinations
Non-self-complementary
Self-complementary
5'
3'
R R R
Y Y Y
R Y R
Y R Y
R R Y
Y Y R
Y R R
R Y Y
tRNA
tRNA
Ho/Ho
Mx/Mx
Ho/Mx 5'R
Mx/Ho 5'Y
(b) One principal dinucleotide plus the wobble base
(1) codon:anticodon pairs: two combinations
R R W
Y Y W
Y Y W
R R W
R Y W
Y R W
Y R W
R Y W
mRNA
tRNA
Ho
Mx
(2) Aminoacyl-tRNA synthetase reaction: correlation between discrete and compositional information
(3) Ribosomal RNA decoding site: mini-helix quality-checking
5'
Central
A,G,C,U
mRNA
A1493
A1492
G530
16S rRNA
3'
Central
U,C,G
tRNA

Fig.6

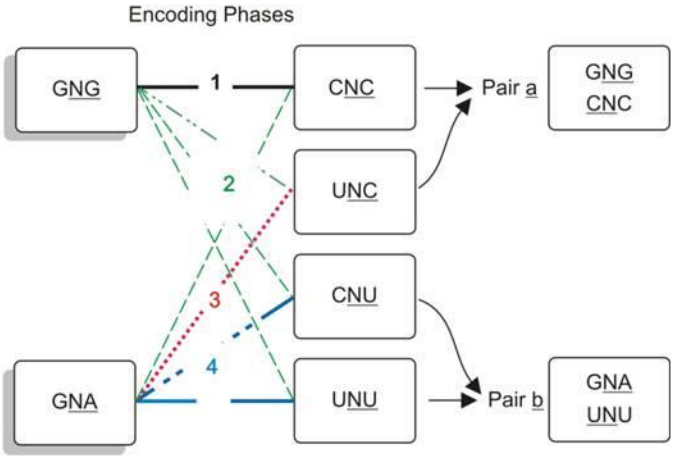
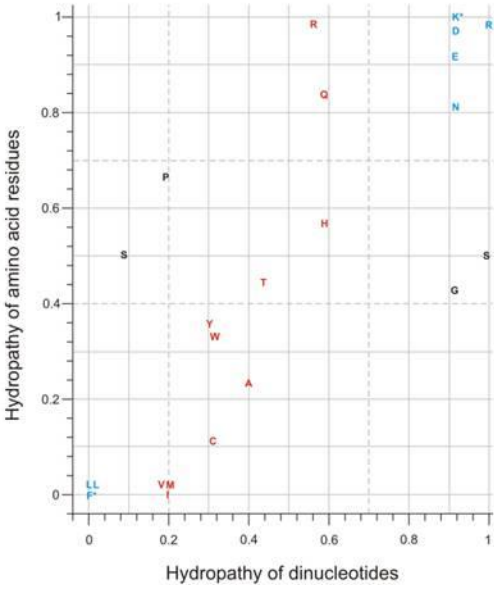

Encoding Phases
GNG
1
CNC
Pair a
GNG
CNC
2
UNC
CNU
3
4
GNA
UNU
Pair b
GNA
UNU

Fig.7



Hydropathy of amino acid residues
1
0.8
0.6
0.4
0.2
0
R
K*
D
R
E
Q
N
P
H
S
S
T
G
Y
W
A
C
LL
F*
VM
I
0
0.2
0.4
0.6
0.8
1
Hydropathy of dinucleotides

Fig.8

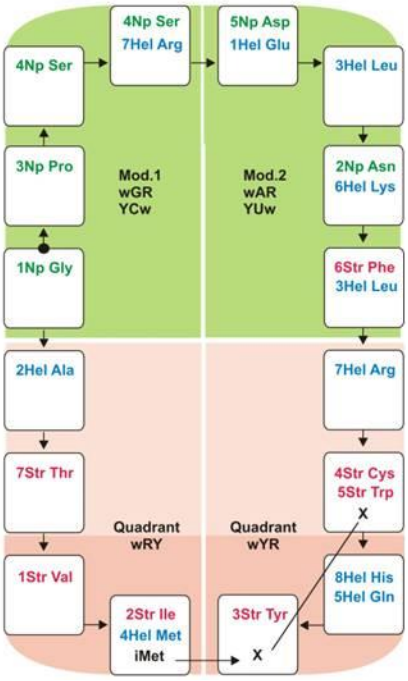

4Np Ser
7Hel Arg
5Np Asp
1Hel Glu
4Np Ser
3Hel Leu
3Np Pro
Mod.1
wGR
YCw
Mod.2
wAR
YUw
2Np Asn
6Hel Lys
1Np Gly
6Str Phe
3Hel Leu
2Hel Ala
7Hel Arg
7Str Thr
4Str Cys
5Str Trp
X
Quadrant
wRY
Quadrant
wYR
1Str Val
8Hel His
5Hel Gln
2Str Ile
4Hel Met
iMet
3Str Tyr
X

Fig.9

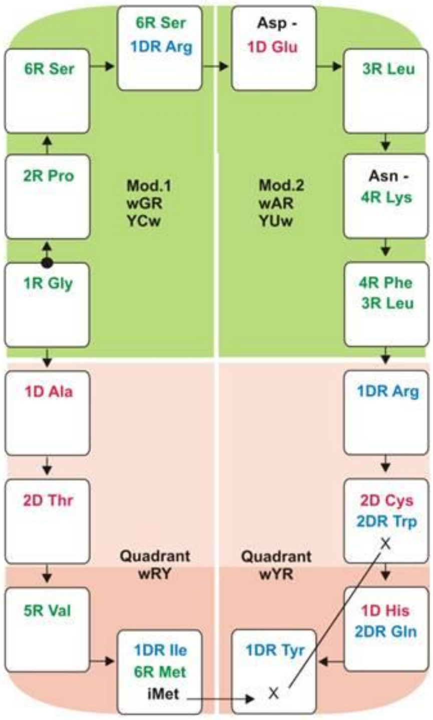

6R Ser
1DR Arg
Asp -
1D Glu
6R Ser
3R Leu
2R Pro
Mod.1
wGR
YCw
Mod.2
wAR
YUw
Asn -
4R Lys
1R Gly
4R Phe
3R Leu
1D Ala
1DR Arg
2D Thr
2D Cys
2DR Trp
X
Quadrant
wRY
Quadrant
wYR
5R Val
1D His
2DR Gln
1DR Ile
6R Met
iMet
1DR Tyr
X

Fig.10

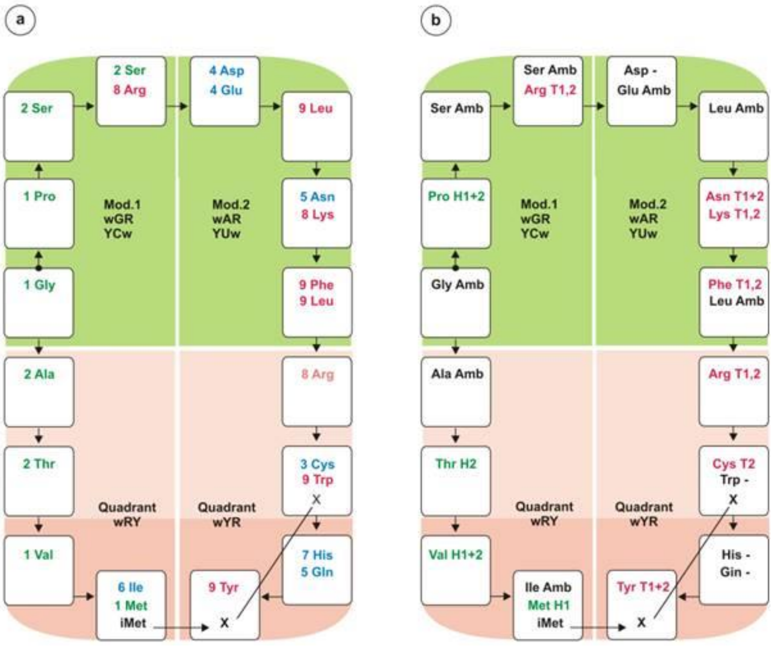

a
2 Ser
8 Arg
4 Asp
4 Glu
2 Ser
9 Leu
1 Pro
Mod.1
wGR
YCw
Mod.2
wAR
YUw
5 Asn
8 Lys
1 Gly
9 Phe
9 Leu
2 Ala
8 Arg
2 Thr
3 Cys
9 Trp
X
Quadrant
wRY
Quadrant
wYR
1 Val
7 His
5 Gln
6 Ile
1 Met
iMet
9 Tyr
X
b
Ser Amb
Arg T1,2
Asp -
Glu Amb
Ser Amb
Leu Amb
Pro H1+2
Mod.1
wGR
YCw
Mod.2
wAR
YUw
Asn T1+2
Lys T1,2
Gly Amb
Phe T1,2
Leu Amb
Ala Amb
Arg T1,2
Thr H2
Cys T2
Trp -
X
Quadrant
wRY
Quadrant
wYR
Val H1+2
His -
Gin -
Ile Amb
Met H1
iMet
Tyr T1+2
X

Fig.11

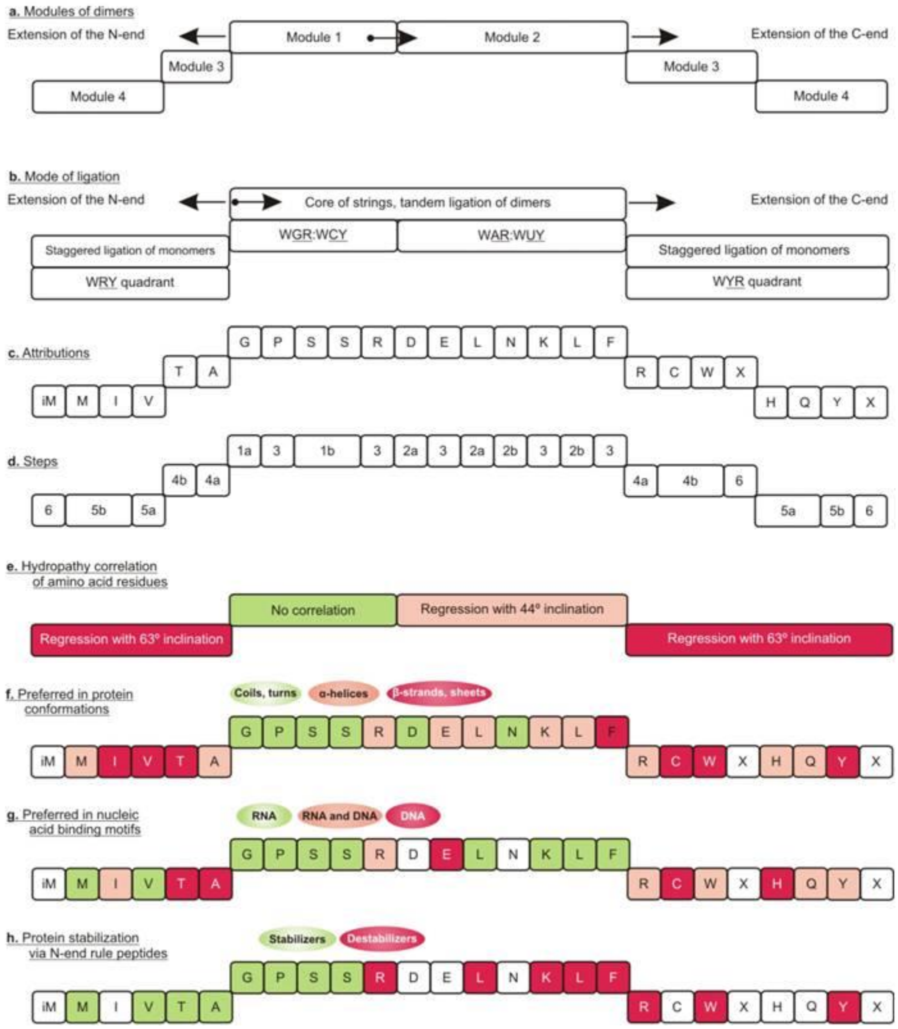

a. Modules of dimers
Extension of the N-end
Module 1
Module 2
Extension of the C-end
Module 3
Module 4
Module 3
Module 4
b. Mode of ligation
Extension of the N-end
Core of strings, tandem ligation of dimers
Extension of the C-end
WGR:WCY
WAR:WUY
Staggered ligation of monomers
Staggered ligation of monomers
WRY quadrant
WYR quadrant
c. Attributions
G P S S R D E L N K L F
T A
R C W X
iM M I V
H Q Y X
d. Steps
1a 3 1b 3 2a 3 2a 2b 3 2b 3
4b 4a
4a 4b 6
6 5b 5a
5a 5b 6
e. Hydropathy correlation of amino acid residues
No correlation
Regression with 44° inclination
Regression with 63° inclination
Regression with 63° inclination
f. Preferred in protein conformations
Coils, turns
α-helices
β-strands, sheets
G P S S R D E L N K L F
iM M I V T A
R C W X H Q Y X
g. Preferred in nucleic acid binding motifs
RNA
RNA and DNA
DNA
G P S S R D E L N K L F
iM M I V T A
R C W X H Q Y X
h. Protein stabilization via N-end rule peptides
Stabilizers
Destabilizers
G P S S R D E L N K L F
iM M I V T A
R C W X H Q Y X

Fig.12

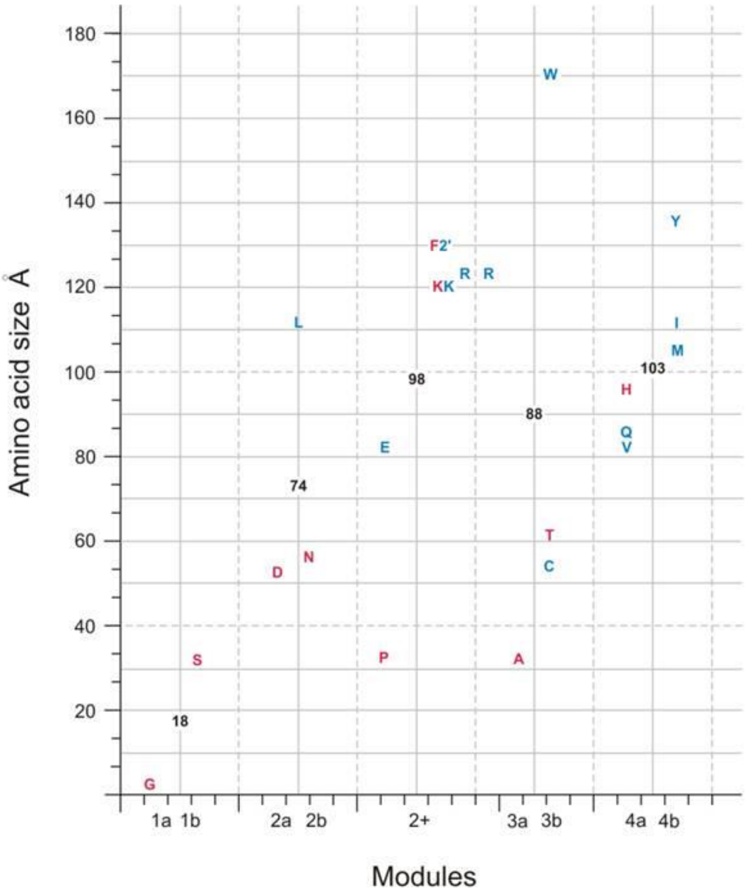
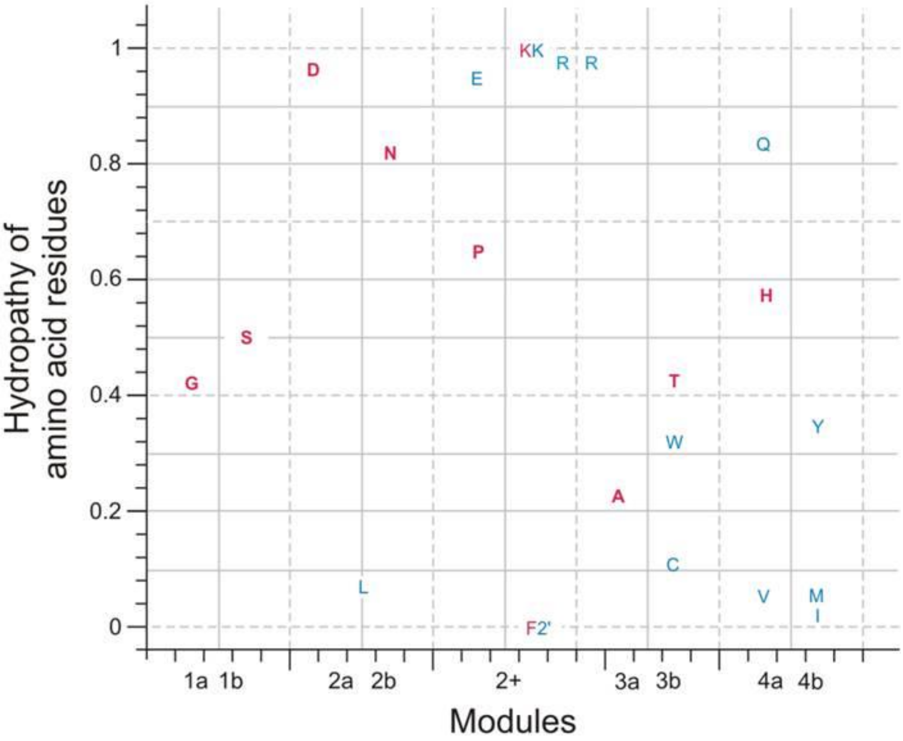
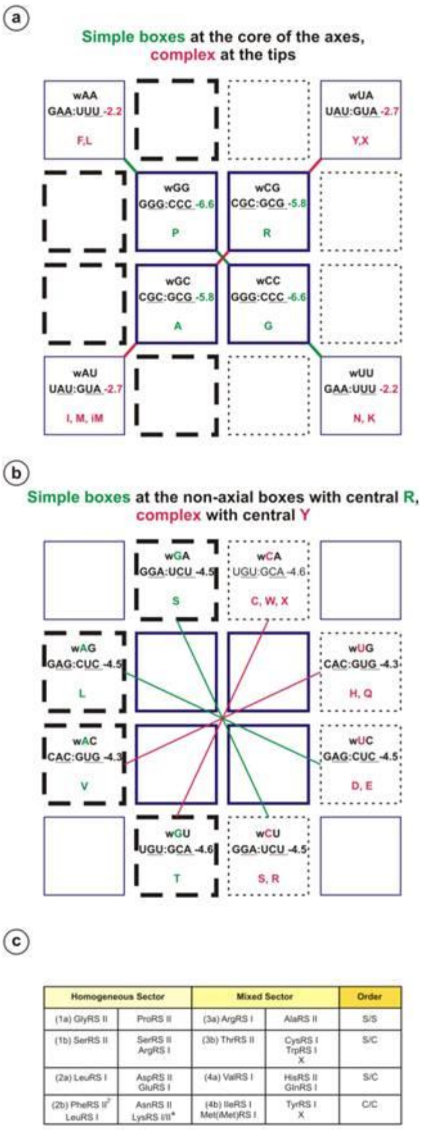

Amino acid size Å
180
160
140
120
100
80
60
40
20
W
Y
F2'
R
R
KK
L
I
M
103
98
H
88
Q
V
E
74
T
C
N
D
S
P
A
18
G
1a
1b
2a
2b
2+
3a
3b
4a
4b
Modules

Fig.13



Hydropathy of amino acid residues
1
0.8
0.6
0.4
0.2
0
D
E
KK
R
R
N
Q
P
H
S
G
T
Y
W
A
C
L
V
M
I
F2'
1a
1b
2a
2b
2+
3a
3b
4a
4b
Modules

Fig.14



Homogeneous Sector		Mixed Sector		Order
(1a) GlyRS II	ProRS II	(3a) ArgRS I	AlaRS II	S/S
(1b) SerRS II	SerRS II ArgRS I	(3b) ThrRS II	CysRS I TrpRS I X	S/C
(2a) LeuRS I	AspRS II GluRS I	(4a) ValRS I	HisRS II GlnRS I	S/C
(2b) PheRS II$^{?}$ LeuRS I	AsnRS II LysRS I/II*	(4b) IleRS I Met(iMet)RS I	TyrRS I X	C/C

-

Fig.15

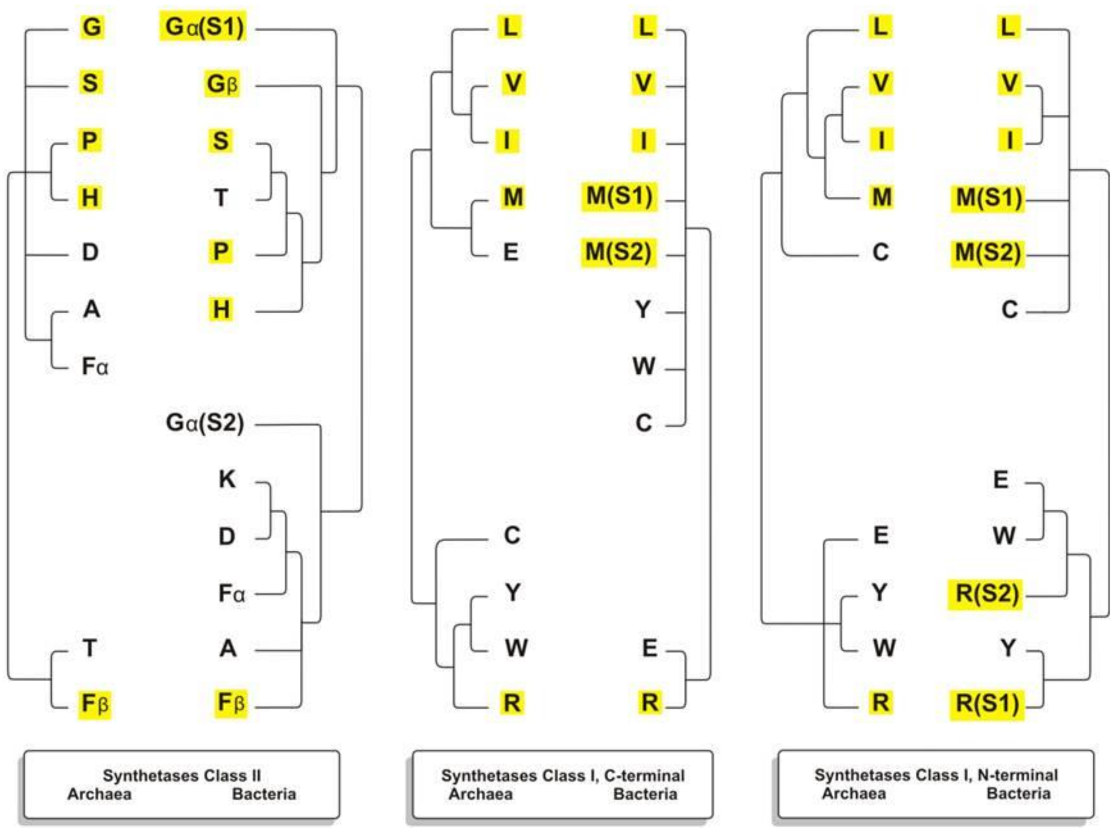

G
S
P
H
D
A
Fα
T
Fβ
Gα(S1)
Gβ
S
T
P
H
Gα(S2)
K
D
Fα
A
Fβ
Synthetases Class II
Archaea
Bacteria
L
V
I
M
E
C
Y
W
R
L
V
I
M(S1)
M(S2)
Y
W
C
E
R
Synthetases Class I, C-terminal
Archaea
Bacteria
L
V
I
M
C
E
Y
W
R
L
V
I
M(S1)
M(S2)
C
E
W
R(S2)
Y
R(S1)
Synthetases Class I, N-terminal
Archaea
Bacteria

Fig.16

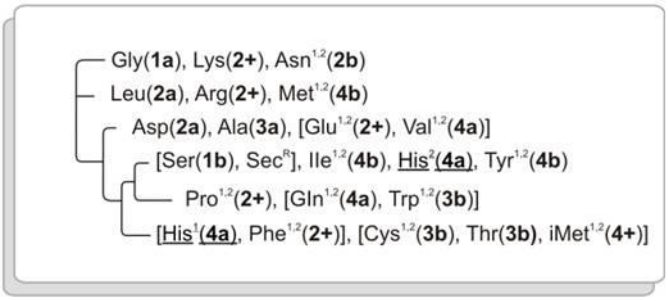
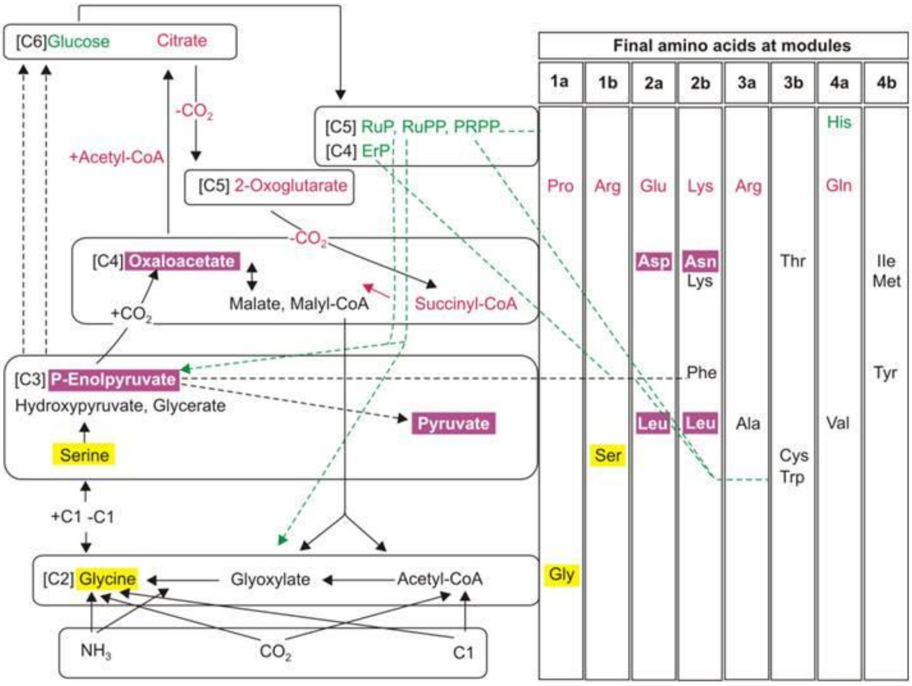
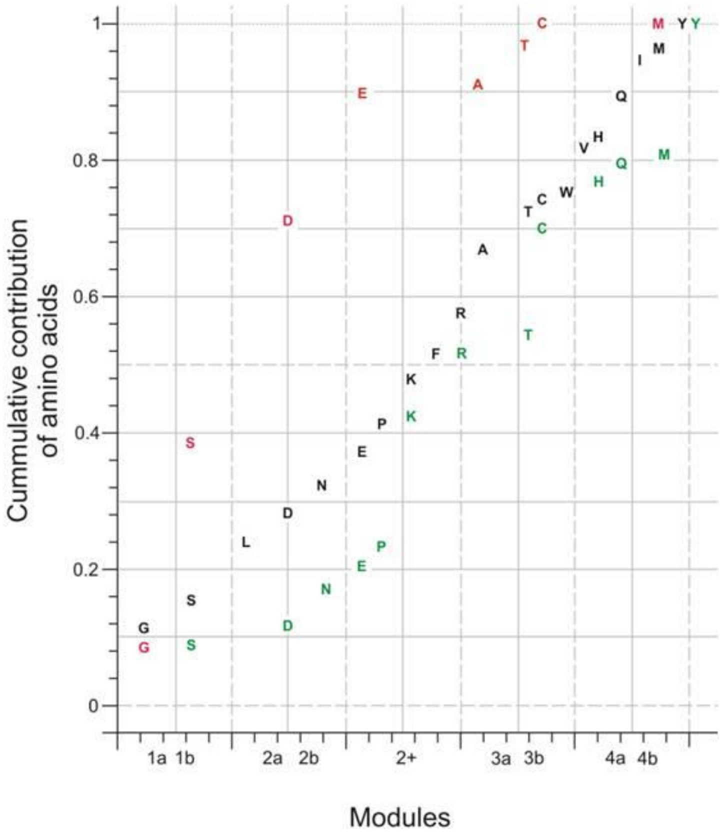

Gly(**1a**), Lys(**2+**), Asn^{1,2}(**2b**)
Leu(**2a**), Arg(**2+**), Met^{1,2}(**4b**)
Asp(**2a**), Ala(**3a**), [Glu^{1,2}(**2+**), Val^{1,2}(**4a**)]
[Ser(**1b**), Sec^{R}], Ile^{1,2}(**4b**), His^{2}(**4a**), Tyr^{1,2}(**4b**)
Pro^{1,2}(**2+**), [Gln^{1,2}(**4a**), Trp^{1,2}(**3b**)]
[His^{1}(**4a**), Phe^{1,2}(**2+**)], [Cys^{1,2}(**3b**), Thr(**3b**), iMet^{1,2}(**4+**)]

Fig.17



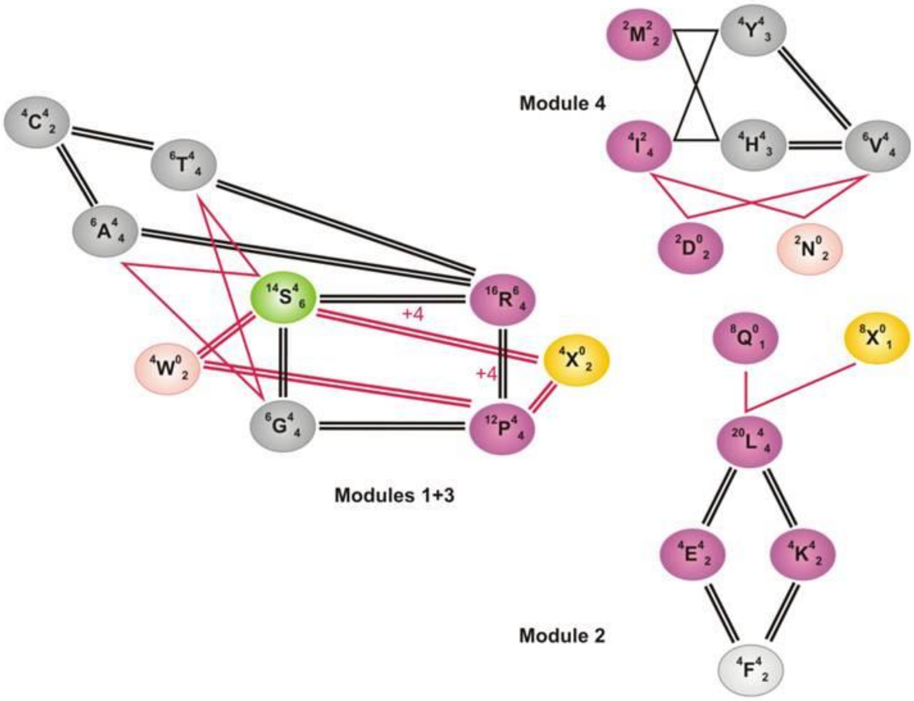
Final amino acids at modules							
1a	1b	2a	2b	3a	3b	4a	4b
						His	
Pro	Arg	Glu	Lys	Arg		Gln	
		Asp	Asn Lys		Thr		Ile Met
			Phe				Tyr
		Leu	Leu	Ala		Val	
	Ser				Cys Trp		
Gly							

Fig.18



Cummulative contribution of amino acids
Modules
0
0.2
0.4
0.6
0.8
1
1a
1b
2a
2b
2+
3a
3b
4a
4b
G
G
S
S
S
L
D
D
D
N
N
E
E
E
P
P
K
K
F
R
R
A
A
T
T
T
C
C
C
W
V
H
H
Q
Q
I
M
M
M
Y
Y

Fig.19



GLOBAL	CENTRAL G:C Modules 1+3	CENTRAL A:U Module 4	CENTRAL A:U Module 2
Three-node cycles	SGA, SGT, SRA, SRT		
Four-node cycles	PGSR, GACT, SACT, RACT , PWSX	VYIH, VYMH, MYIH, VNID	LEFK
Connectivity 58/20 = 2.9	28/8 = 3.5	20/7 = 2.86	10/5 = 2
9/20 aRS in MaRS = 0.45	2/8 = 0.25	3/7 = 0.43	4/5 = 0.8