Orig Life Evol Biosph (2011) 41:357–371
DOI 10.1007/s11084-010-9226-x

GENETIC CODE EVOLUTION

Metabolic Basis for the Self-Referential Genetic Code

Romeu Cardoso Guimarães

Received: 10 June 2010 / Accepted: 8 October 2010 /
Published online: 6 November 2010

Abstract An investigation of the biosynthesis pathways producing glycine and serine was necessary to clarify an apparent inconsistency between the self-referential model (SRM) for the formation of the genetic code and the model of coevolution of encodings and of amino acid biosynthesis routes. According to the SRM proposal, glycine was the first amino acid encoded, followed by serine. The coevolution model does not state precisely which the first encodings were, only presenting a list of about ten early assignments including the derivation of glycine from serine—this being derived from the glycolysis intermediate glycerate, which reverses the order proposed by the self-referential model. Our search identified the glycine-serine pathway of syntheses based on one-carbon sources, involving activities of the glycine decarboxylase complex and its associated serine hydroxymethyl-transferase, which is consistent with the order proposed by the self-referential model and supports its rationale for the origin of the genetic code: protein synthesis was developed inside an early metabolic system, serving the function of a sink of amino acids; the first peptides were glycine-rich and fit for the function of building the early ribonucleoproteins; glycine consumption in proteins drove the fixation of the glycine-serine pathway.

Keywords Genetic code · Glycine-serine pathway · Origin · Evolution · Self-reference

Introduction

The genetic code is a main character demarcating living beings from the physicochemical nonliving realm. It establishes the correspondences between triplets of nucleotides in nucleic acid coding sequences and the amino acids in proteins. The formation of the system of correspondences is a key problem for understanding the origins of life. Single triplets are found as anticodons in transfer RNAs (tRNAs) and chains of triplets, the codons of messenger RNAs (mRNAs), form the templates that are translated into protein sequences.

R. C. Guimarães (✉)
Lab. Biodiversidade e Evolução Molecular, Dept. Biologia Geral, Inst. Ciências Biológicas,
Univ. Federal de Minas Gerais, 31270.901 Belo Horizonte, MG, Brazil
e-mail: romeucg@icb.ufmg.br

Translation is obtained from ribosomes, structures that place in close contact one molecule of mRNA and a pair of tRNAs, each of these bearing an attached amino acid—aminoacyl-tRNA. Contiguous codons in the mRNA form base pairs with the anticodons of the tRNA couples and the amino acid that is bound to the tail of one tRNA is transferred to the other; the receptor tRNA becomes a peptidyl-tRNA, loaded with a peptide chain containing its previous plus the newly received amino acids. Successive reading of pairs of codons elongates the peptides.

Self-Referential Loops

The self-referential model (SRM; Guimarães et al. 2008a, b) for the formation of the genetic code proposes that dimers of proto-tRNAs composed the original protein synthesis machinery. Present day tRNAs are complex molecules about 72 nucleotides long and contain numerous modifications in the nucleosides, making it difficult to imagine the sequences of the original proto-tRNAs. The tRNAs present a structure suggestive of having originated from smaller segments that became duplicated (Bloch et al. 1989, and references therein; Randau et al. 2005). The proto-tRNAs correspond to some of these smaller segments, possibly hairpin structures with proto-anticodons in a loop and an amino acid-binding tail. There are at present no clues about the chemical composition of the proto-tRNAs, whose monomers might have been different from the nucleotides of current RNAs. Minimal requirements are that the proto-anticodons would have at least the triplet size and, in a population of proto-tRNAs, the proto-anticodons would be able to base-pair with formation of hydrogen-bonded dimers. The dimers are considered proto-mRNAs in the sense that one of the (proto)anticodons functions as a codon for the other anticodon. They are also considered proto-ribosomes, structures where pairs of tRNAs are bound together, catalyzing the formation of peptide bonds. The self-referential aspect of the model stems initially from the indication that the proto-tRNAs in a dimer refer to each other and that the peptides synthesized refer back to the dimers synthesizing them, establishing a positive feedback loop. The term also indicates the difference from other models that assume the code arising with the function of translating previously existent long RNAs, called hetero-referential.

Another key basis for the SRM is the reinforcement of the dimer-directed peptide synthesis activity by the peptides produced by the dimers, thereby inaugurating the formation of an auto-stimulated system. This would happen when the peptide composition is adequate for its own stability against degradation and for proto-tRNA binding, resulting in stabilization of a ribonucleoprotein (RNP) aggregate. The protein synthesis activity of the RNP complex should not be impaired by the protein binding and would be improved by the selective development of specificities, reaching the known correspondences of the code. This first evolutionary level, of single letter encodings, would be accompanied by the development of protein-assisted replication of the proto-tRNAs, to reach the stage of formation of stabilized long RNA chains, starting with poly-proto-tRNAs, from which mRNAs and ribosomal RNAs (rRNAs) were derived.

Hydropathy Correlation

A main evidence for the SRM was obtained (Farias et al. 2007, and earlier references therein) from a re-elaboration of the studies on the correlation between hydropathies of

nucleotide doublets—equivalent to the principal dinucleotides (pDiN) of the anticodon triplets, excluding the wobble (w) position—and of the amino acid residues in proteins, according to the code (Fig. 1). The novelty introduced—utilization of the hydropathies of amino acid residues in proteins, instead of the hydropathies of amino acid molecules in solution, enabled a refinement of the correlation data and the identification of clearly defined subsets. Such improvements indicated that the hydropathy correlation is a property of the RNP interactions (aminoacyl-tRNA synthetases with tRNAs), instead of reflecting prebiotic interactions between the proto-tRNAs and the amino acids. The correlation data were thought trustworthy due to the triplets in the four subsets forming clearly defined modules of self-contained complementary dimers, indicating an underlying structure to the triplet matrix.

The modules in the correlation could be arranged in a chronologic stepwise succession: start with the subset with no correlation and proceed to the subset presenting the correlation with a moderate inclination of the regression line to end with the subsets showing steep inclination. The pDiN of the two first subsets were composed of homogeneous types of

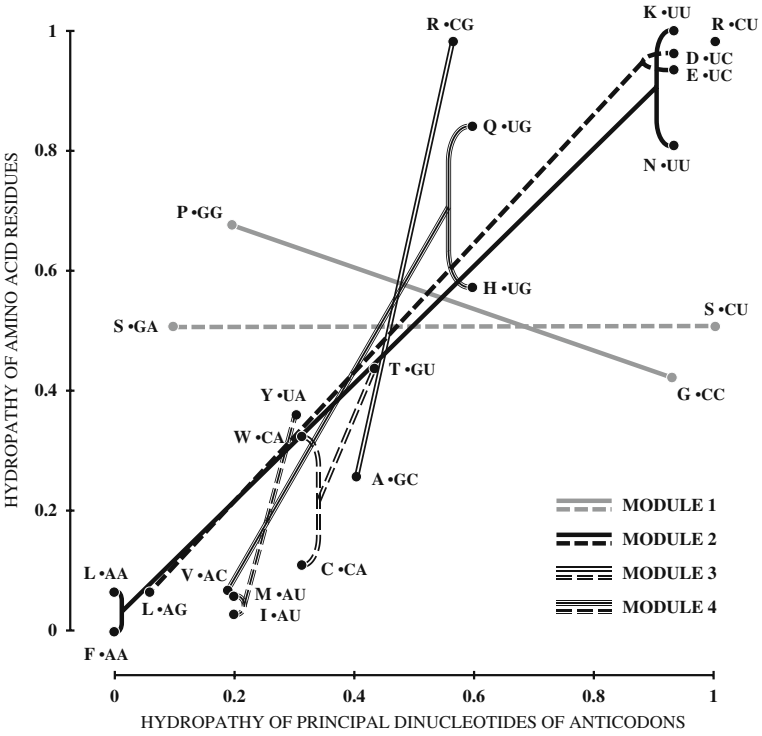

Fig. 1 Correlation between hydropathies of amino acid residues and principal dinucleotides of anticodons. Two sectors are distinguished by the principal dinucleotide (pDiN) types, each one further divided into two modules, distinguished by the central base pair. The homogeneous sector, •RR:•YY, with extreme pDiN hydropathies, is divided into module 1, •GR:•CY, composed of non hydropathy-correlated attributions, and module 2, •AR:•UY together with the Arg•CU attribution, composed of highly correlated attributions (8 points, y=1.001 x - 0.002, R^2=0.991). The mixed sector, •RY:•YR, with intermediate pDiN hydropathies, is divided into module 3, •GY:•CR, and module 4, •AY:•UR; these jointly conform to a steeper regression line (11 points, y=2.172 x – 0.443, R^2=0.849). Members of the pDiN pairs are linked by (a) solid or (b) dashed lines. Adapted from Guimarães and Moreira (2004), Farias et al. (2007), Guimarães et al. (2008a, b)

bases (the wRR:wYY sector) and those in the two last subsets of mixed types of bases (the wRY:wYR sector); each subset was further subdivided into the specific pDiN pairs, starting from the GC-richer and ending with the GC-poorer. The non-correlated set (Module 1): (1a) wCC Gly : wGG Pro [amino acids are designated by the three-letter abbreviations], (1b) wGA Ser : wCU Ser. The set with a moderate inclination of the regression line (Module 2): (2a) wUC Asp, Glu : wAG Leu, (2b) wUU Asn, Lys : wAA Phe, Leu; the yCU Arg clusters with this set. The central G:C dimers with steep inclination regression line (Module 3): (3a) wCG Arg : wGC Ala, (3b) wGU Thr : wCA Cys, Trp, X. The central A:U dimers with steep inclination regression line (Module 4): (4a) wAC Val : wUG His, Gln, (4b) wAU Ile, Met, iMet : wUA Tyr, X. Development of these studies with inclusion of a number of other amino acid and protein properties allowed the proposition of the model (Guimarães et al. 2008a, b).

Amino Acid Biosynthesis

Another line of studies relating the genetic code with physiological characters is the coevolution theory of Wong (1975, 2005), introducing the rationale that the two 'first principles' of biology—metabolism and the genetic strings—are interwoven. The observation bases were the mutual mapping of amino acid biosynthesis pathways, where some amino acids are precursors to others, forming families, and of their correspondent triplets in the code, which are also related to one another, with small changes in composition. General lines are: Ser is derived from glycerate and precursor to Gly, Cys and Trp; pyruvate is precursor to Ala, Val and Leu; Asp is derived from oxaloacetate and precursor to Asn, Lys (this also receives contribution from Glu), Thr, Met and Ile; Glu is derived from α-keto glutarate and precursor to Gln, His (this may also have an origin independent from Glu), Arg and Pro; Phe is precursor to Tyr. Variations on the theme are many, mainly dependent on differences on the biosynthesis routes in different organisms (Knight et al. 2001; Klipcan and Safro 2004) and on choices that can be made in examination of the routes. Another backing to the hypothesis is derived from the finding of some metabolic transformations of amino acids being dependent on their binding to tRNAs (DiGiulio 2005). The work of Davis (1999) follows the biosynthesis idea but with a different rationale, starting with the ammonia assimilation amino acids (Glu, Asp) and the whole central U column of anticodons.

Prebiotic Amino Acids

A different line of approach has been followed apparently even more extensively in studies of formation of the code, considering the lists of quantities of prebiotic amino acids. As happened with the biosynthesis approach, there are also many variations among authors following this trend but the main focus is centered on the group of Val, Ala, Gly and the pair of acidic amino acids Asp and Glu. Enthusiasm with this group stems from their commonality in corresponding to triplets of the anticodonic 3′ C row. A recent evaluation of this trend, also adequately reviewing other approaches, was presented by Higgs (2009). A reasonable correlation between the list of prebiotic amino acids and the list of Wong's early or precursor has been taken to be mutually supportive, which lead to an attempted consensus proposed by Trifonov (2004). The approach of Higgs (2009) takes into account a procedure of optimization of the code, relative to minimization of the consequences of

translational errors, ending up with a proposal for starting the encodings with the 3′ C-headed four columns of the anticode matrix. The approach of Trifonov (2004) adds a consideration of the thermodynamic stabilities of the codon pairs and starts the encodings with the prebiotically most abundant amino acids Ala and Gly (codons GCC : GGC, respectively), followed by Val and Asp (GUC : GAC). It is clear that the order of entry of amino acids in the code is widely different in the prebiotic-list trend of research and in the SRM, so that the latter had to be excluded from the data leading to the consensus of Trifonov (2004). Possible roles for the prebiotic amino acids in the process of formation of the SRM are discussed below.

Present and Future

The present report reexamines some metabolic routes of amino acid biosynthesis with the specific intent of showing that there is definite support for the encodings having started with the Gly and Ser pair of amino acids, in accordance with the SRM. In the presentations of the SRM, it was already considered, on the basis of the known consensual routes of amino acid derivation, that Pro and Arg would have entered the code only after their precursor Glu had been encoded, so that Module 1 would have been composed by Gly and Ser only. The proposition of the Gly and Ser metabolic relatedness through the activities of the system formed by the glycine decarboxylase complex (GDC) coupled with the serine hydroxymethyltransferase (SHMT) was only suggested for further investigation (Guimarães et al. 2008b), which is now accomplished. This search was thought necessary due to the apparent inconsistency of (a) the derivation of a simpler amino acid (Gly) from the more complex Ser, with the most common trend in biological evolution of deriving complex from simple forms, and of (b) the rationales in the SRM, such as the properties of stability and of RNA-binding, which pointed to a precedence of Gly over Ser, with the route deriving Gly from Ser, considered in the coevolution model.

A general principle is adopted that any amino acid entering the code at a successive module should have its main family precursor already encoded in a previous module. There is no need for the encodings having to follow precisely the precursor-product relationships in the linear metabolic derivations of the families of amino acids. The linear pathways are sets inside complex networks containing many intermediates and reticulated connections so that deriving time-ordered successions from single linear derivations could be misleading. No attempt is made to reinvestigate the biosynthesis pathways of other amino acids.

A separate report (in preparation) is dedicated to examination of the thermodynamics of triplet dimer formation, to rationalize the relatedness of this with the constitution of the modules and with the succession of modules in the coding system. Suggested experimental tests of the SRM would be on the postulated ability of mini-tRNA (Beuning and Musier-Forsyth 1999) dimers to facilitate the peptidyltransferase reaction, utilizing mini-tRNAs with the adequate complementary loops and sufficiently flexible aminoacylated ends (see e. g. Tamura and Schimmel 2006). Such a system could be put to evolve spontaneously—to obtain truly synthetic organisms, in spite of having to obey the experimentally defined environmental constraints. Dimerization of present day tRNAs has been extensively studied (Grosjean and Houssier 1990) but their possible ability in producing the peptidyltransferase reaction has not been tested; it is thought that full length tRNAs might be too rigid or bulky for the suggested tests.

Simple to Complex Biosynthesis Pathways

The genetic code might be considered a record of some of the steps followed in the development of metabolic complexity. A compilation of the established basic pathways is presented in Fig. 2, ordered according to an approximate succession of sizes of the carbon

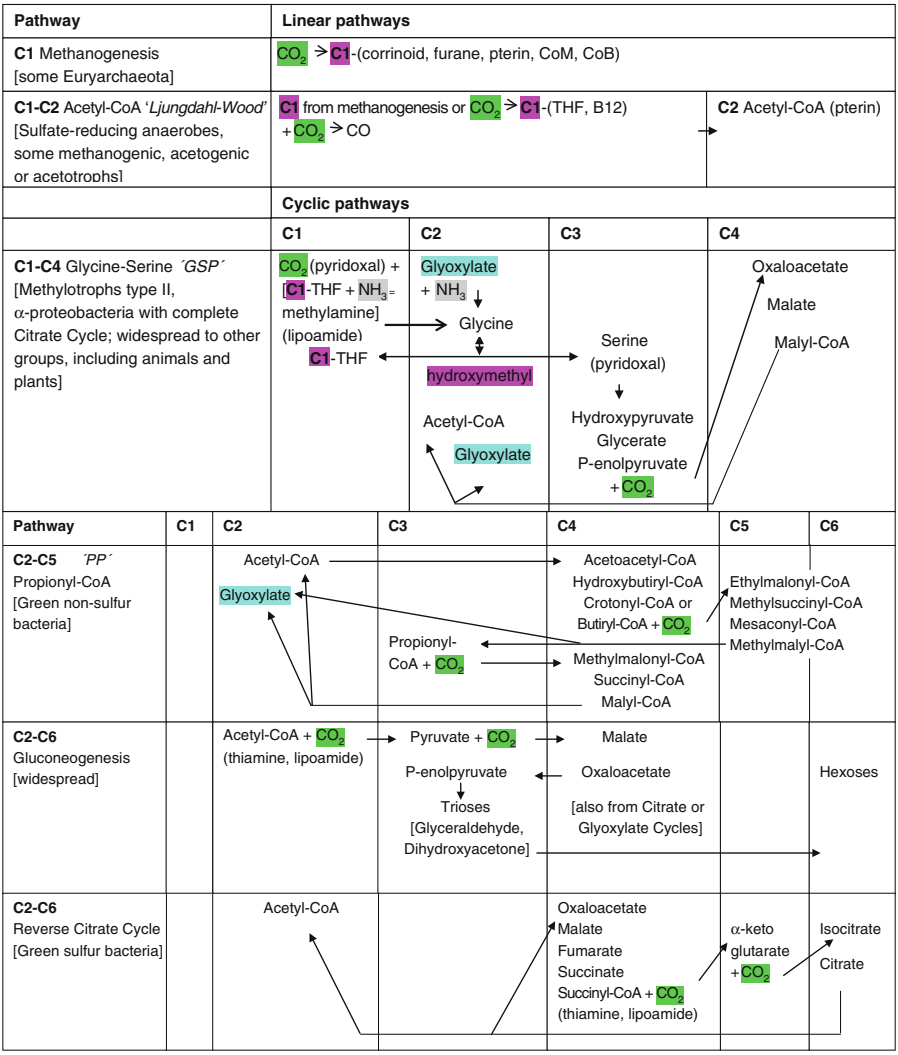

Fig. 2 Sketch of some basic biosynthesis pathways. Some key acyl-carrier cofactors are shown in parenthesis. In the methanogenesis and in the acetyl-CoA pathways, the pterin cofactor is a simple form of tetrahydrofolate (THF); the CoM (thioethylsulfate) and CoB (thioethylamine-pantothenate, a simple form of CoA) work in association, the couple possibly representing an early form of the lipoamide (dithio); the corrinoid (related to the B12) works in some methanogenesis pathways. For brevity, only some of the phosphate groups of compounds and of the variety of interconversions between sugars are shown. Three routes for synthesis of Glycine are shown (from simple molecules, by the glycine decarboxylase complex, from glyoxylate, by transamination, and from de-hydroxymethylation of Serine); there are various pathways generating glyoxylate. Highlights: green, assimilation of CO_2; pink, of C1 compounds; grey, of ammonia; turquoise, of glyoxylate

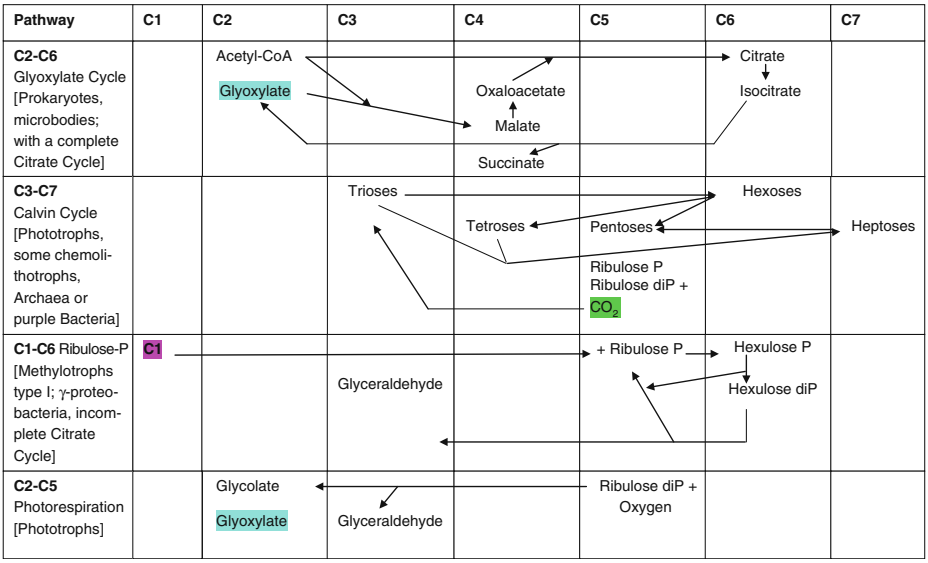

Fig. 2 (continued)

chains of the compounds involved in the routes. Most data are taken from Madigan et al. (2003). The first pathway is of methanogenesis (C1 compounds), which shares many reactions with the second, the acetyl-CoA biosynthesis (C2 compounds; Ljungdahl-Wood) pathway; these routes utilize a large variety of cofactors and some of these may represent simple forms of cofactors that participate in reactions of the more complex pathways. C1 assimilation via the tetrahydrofolate (THF) cofactor would be an interesting candidate for early routes due to its lower energetic requirements (about 20 Kj/mol at hydrolysis, near the equivalent to one unit of the proton motive force—one biochemical energy 'quantum', compared with the ~30–40 for ATP or acetyl-CoA; about 3 'quanta' are required for ATP synthesis; Heimann et al. 2010). C1 transfer belongs among the 11 kinds of basal reaction mechanisms of biochemistry (Kyte 1995b). C3 compounds are indicated to have arisen initially from condensation of C1 and C2 compounds, as seen in the Glycine-serine pathway (GSP) and its associated activities (Fig. 2). The first two pathways are linearly feeding-forward and cyclic configurations come up in the ≥ C3 realm. It is apparent that the C3 step is at the origin of the more complex pathways, e. g., allowing the formation of C6 through condensation in the gluconeogenesis pathway, from which glycolysis and the pentose-phosphate pathways are derived.

The GSP seems to occupy, in association with the propionate pathway (PP; Peyraud et al. 2009), a central role among the variety of assimilation routes. They might be considered the simplest routes for assimilation of CO_2 and C1 compounds and would represent the earliest instances of development of the circularly self-centered (a higher order of self-reference) character of the metabolic network. Some components of GSP are shared with the PP (the C2 glyoxylate and the C4 malate), and other components of these two routes (the C4, oxaloacetate from GSP and succinate from PP) are shared with the most ubiquitous citrate cycle. This may work in the reductive direction, assimilating CO_2, or in the more common oxidative direction. It is thought to have evolved by joining and enchaining smaller pathways, the C4 half outlined above with the more complex half, containing the larger C5 and C6 compounds, finally acquiring the circular configuration (Wood et al.

2004). Besides the most obvious C2 component—acetyl-CoA, another key C2 compound, glyoxylate, is highlighted in these routes. This shows up again in the anaplerotic glyoxylate cycle and, in photorespiration, it drives the rescue of the GSP into the metabolic network.

In the GSP, an important aspect is the interconvertible nature of the Gly and Ser derivations. A simple form of THF is seen in the methanogenesis routes and it is also at work in the de novo synthesis of Gly by the GDC, plus in the GDC-coupled SHMT activity, which synthesizes Gly and Ser through C1 exchanges. Involvement of the GSP in so many and varied processes, points to the overwhelming metabolic role of Gly and Ser, and of the C1 pathways, in cellular economy (Bauwe and Kolukisaoglu 2003). In fact, calculations (Reitzer 2003) on the biosynthetic requirements of the amino acids for *Escherichia coli* growth highlighted these two, together with the acidic (Asp, Glu), as responsible for 89% of the carbons donated to compose the bacterial cell mass (39% and 50% for each of the couples of donors, respectively, considering only the donated carbons of the nine donor amino acids; Fig. 3). Metabolic calculations for Ser are remarkably close to the nutritional studies of Pizer and Potochny (1964), describing that about 15% of the glucose carbons pass through Ser. It can be indicated that the predictions of the SRM are plainly supported

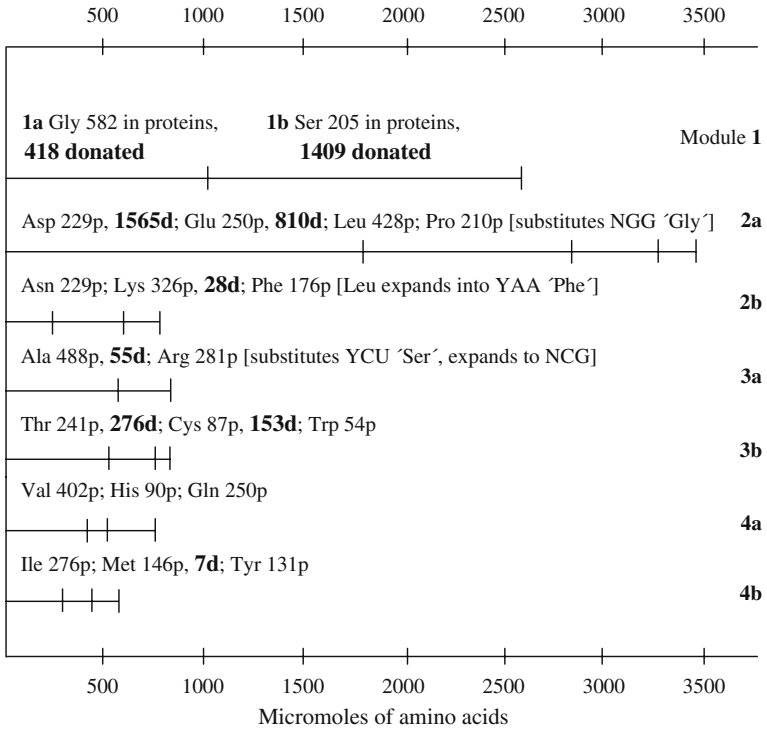

Fig. 3 Requirements on carbons of amino acids to synthesize *Escherichia coli* mass. Adapted from Reitzer (2003; micromoles of amino acids per one gram dry weight of cells). Carbons in the skeletons of amino acids are calculated: p, for the protein composition and d (highlighted; nine of the amino acids), when they are donated to the structure of other compounds, such as other amino acids or nucleotide bases. Amino acids are displayed according to the self-referential model for the formation of the genetic code (Guimarães et al. 2008a, b). Module 1 started octacodonic for both Gly (NCC : NGG anticodons) and Ser (NGA : NCU). Notes inside square brackets: the NGG (previously 'Gly') became occupied by Pro after its precursor Glu was encoded, at the Module 1-2a transition; Leu expanded from the NAG into the YAA (previously 'Phe'), at the Module 2a-2b transition; Arg entered the YCU (previously 'Ser') and expanded to NCG, at the Module 2b-3a transition

by metabolic evidence: these four amino acids enter the composition of exactly Module 1 and the first half of Module 2. The messages derived from these observations are that (a) these four amino acids are of high value both for general metabolic and protein structural (and functional) purposes and (b) some quantitative aspects of general metabolic pathways, as seen through amino acid metabolic requirements, are coincidental with the pathways leading to the temporal sequence of encodings. The carbon skeleton complexity series indicates that the encoding of Gly and Ser (Module 1) would have been settled in the C1-C3 stage of biosynthetic developments. The derivation of these amino acids from C3 intermediates of glycolysis, which stems from C6 compounds, is indicated to be a late addition, after the development of gluconeogenesis. It is noteworthy that the derivation of Ser from the glycerate of glycolysis takes a reverse configuration in the GSP (Fig. 2).

Metabolic Functions and the Protein Synthesis Sink

A rationale for the role of protein synthesis inside a developing proto-metabolic system is necessary. A scenario could be constructed considering mainly the aspect of dissipation of free energies by the proto-metabolic system through formation of some stable chemical compounds. Dominant features in such systems are of two orders: (a) the chains of transformations, with beginnings and ends, which might be considered propulsive to the ends and propelled by excesses of free energies at the beginnings; (b) the cyclical or reticulated connections, uniting the chains at different levels and forming networks. These contribute to the system with integration of the various activities and compounds, and provide it with stabilization mechanisms, when excesses or scarcities of some components could become balanced through the utilization of lateral routes of processing. In the network terminology, stability becomes robustness and, lateral processing, a kind of redundancy. The self-feeding property of the cyclic configurations could run into problems of sustainability, if they enter uncontrolled auto-catalysis, or could follow a more tamed course, only producing chemical transformations with some possible diversification, following the free energy inputs and being limited by them. These configurations would suffer from the lack of evolutionary propulsion mechanisms. The development of protein synthesis inside such systems provides an internal drive, the nucleic acids accomplishing the roles of templates and of polymerization machinery, and the protein products working as a sink of amino acids. Stability of the proteins against degradation would assure their role of a metabolic sink, which is also enforced by the irreversibility of translation, and by the diversification and the accumulative powers of proteins.

The SRM indicates a main role of Gly in composing the first peptides with the properties of intrinsic stability against degradation (Guimarães et al. 2008a, b) and of RNA binding (Guimarães and Moreira 2004), therewith building the early RNPs. These properties, among others, founded the rationale to identify, among the modules demarcated by the hydropathy correlation, the start with Module 1, whose core components are Gly and Ser. With the present study of metabolic routes, a justification is reached for the self-feeding interplay between the SRM and metabolism. Early peptides leading to formation of the RNP system should have been mainly composed by Gly and Ser, and their consumption would have driven the fixation of the GSP as a fundamental pathway. It seems adequate to postulate that the earliest route composing the GSP would have been a GDC-like activity, synthesizing Gly from elemental units (CO_2 plus NH_3 coupled to a reduced C1 unit, as methylamine), and then the Ser synthesis through condensation of Gly with another C1 unit, also derived from Gly, by a SHMT-like activity. The present-day system of coupled

GDC-SHMT activities is freely reversible, being thought of as a system for integrating the C1 to C2 and C3 metabolism, generating the two amino acids and C1 units for a large diversity of metabolic pathways.

An Evolutionary Continuum

A succession could be proposed to reconcile the two apparently contradictory lines of investigation on the origin of the coding system. In a prebiotic scenario, there would be the monomers for oligomerization of the proto-tRNAs and these would be developing dimer-directed peptide synthesis utilizing amino acids following the influence of their concentrations as reviewed by Trifonov (2004) and Higgs (2009). This could be considered, according to the SRM rationale, as a stage of mutual adjustment (evolutionary ´learning´) between peptides and proto-tRNAs, where some improvement in abundances of different kinds of molecules would have occurred, dictated by e. g., stability and association abilities. It is possible that peptides produced in this era would also be contributing with catalytic abilities but it is not known how much development or richness would have reached the necessary geochemical (proto-metabolic) processes, to provide support to the system. At this stage, there could be some kinds of proto-codes established, among which those modeled by Trifonov (2004) and Higgs (2009). A specific consideration of the SRM is that the prebiotic sources of amino acids would have been poor and rapidly exhausted, so that the proto-codes would not have been sustainable. A turning point was reached when, inside such systems, the productive self-referential loops became established, forming a sink for substrates that could find an adequate feeding source, and this had to be provided for also by the components of the loops. If the observations gathered by the SRM capture a reasonable portion of the historical reality, it can be indicated that the initial source was based on the C1 and ammonia assimilation pathways that generate Gly and interconvertibly Ser. The proposition of an evolutionary continuum may seem to reside upon weak evidence, considering only one strong consistency on the amino acid side (Gly), but it gathers strength on the mechanism side, the dimer-directed peptide synthesis and the metabolic support.

Metabolic Complexity

While the proposition that early peptides had a composition dominated by Gly and Ser could be submitted to experimental tests as parts of aggregates with known nucleic acids, it is possible to indirectly evaluate it on the basis of metabolic relevance. Calculations on the high metabolic requirements for Gly and Ser were not detailed by Reitzer (2003; Fig. 3), but most of their components can be rescued from general biochemistry textbooks. Gly and C1 units are main components in the biosynthesis of the purine skeleton and of creatine, and C1 units enter the composition of the pantothenate moiety of coenzyme A, highlighting their close relationships with the nucleic acids and energetic processes. The pterin ring of THF (pyrimidine-like; Kyte 1995b), cofactor in most C1 transfer pathways, is biosynthetically derived from guanosine, introducing one more circular puzzle to studies of metabolic origins. Purines seem to behave like end-products of the GDC; the *Escherichia coli* glycine cleavage system operon is repressed in the absence of Gly but only reaches full repression in the presence of purines in the medium (Heil et al. 2002). Since the metabolic roles of Gly seem to be much more varied than those of Ser, it would be adequate to indicate that a main

metabolic role of Ser would be of a reservoir for replenishing cellular requirements for Gly and C1 units. The threonine aldolase pathway adds to the diversity of the Gly biosynthesis routes. When the GDC is active mainly in the catabolic direction, it would contribute to an adequate balance of Gly and C1 unit pools, besides generating NADH. Gly contributes to the biosynthesis of tetrapyrroles, glutathione, sarcosine and glycine betaine, among other compounds. Gly and Ser participate, together with L- and D-Ala, in the constitution of the non-translated peptides of the cell wall peptidoglycans (Volmer et al. 2008). The enchained pathways of C1 transfers include the biosynthesis of His (interwoven with that of purines) and of thymine and Met, this spreading further into various types of methylations of proteins and nucleic acids. Ser is source to both hydroxypyruvate, via transaminations, and pyruvate, through oxidation, which feed into gluconeogenesis; in the GSP, Ser generates glycerate and phosphoenolpyruvate through hydroxypyruvate (Fig. 1); Ser also participates in the structure of some phospholipids, such as phosphatidylserine, and of sphingosines. In accordance with such wide range of metabolic implications, the Gly and Ser biosynthesis system, as well as the glyoxylate precursor, are regulated in response to various kinds of stresses, including anoxia and nutrient deprivation, such as iron deficiency (Peterson 1982). In eukaryotic cells, glyoxylate production generates H_2O_2 so it is confined to organelles (peroxysomes, glyoxysomes; components of the microbodies category); the GDC is mitochondrial while SHMT has mitochondrial and cytosolic isoforms. Activation of the GSP is a remarkable feature of photorespiration, fed from the generation of glycolate and glyoxylate after the oxygenase activity of RubisCO. The photorespiration pathway is considered a mode of activating 'housekeeping' functions (associated with the C1 assimilation routes), more important than focusing on photosynthesis (CO_2 assimilation) alone. Deficiencies in photorespiration correspond to growth retardation in plants (Bauwe and Kolukisaoglu 2003) and, in the GDC system, to some severe forms of hyperglycinemias in humans (Hamosh and Johnston 2000).

Conclusions and Perspectives

Development of a simple and experimentally testable model for the formation of the genetic code is a necessary prerequisite for instructing the endeavors of producing fully synthetic biological organisms and for probing into plausible scenarios on the spontaneous origin of living systems from geochemical proto-metabolism. The SRM attempts to fulfill the conditions above. It starts from proto-tRNAs, whose constitution is still enigmatic but are supposed to be approximately mimicked by the known mini-tRNAs. Oligomers about the size of mini-tRNAs are apparently the only components required from the RNA world hypothesis. Dimers of the proto-tRNAs would be able to accomplish the function of primitive translation machinery, before the advent of the long RNA chains, these being considered proper to the RNP realm.

The enigma of the proto-tRNA composition extends into the characters of the proto-metabolic system that is required for having produced such RNA-like oligomers. While we have to wait for advancements in this field, one research option (a) is devoted to discern, among known metabolic pathways, some that would seem adequate to support the model. If such support—looking from the model up into biology—is lacking, the model itself would remain plainly speculative. Similar problems affect partially the SRM, when attempts (b) are made to find clear continuities between its propositions and prebiotic chemistry—looking from the model back into geochemistry. The SRM is clear in proposing that the first encoded amino acids were Gly and Ser, but only Gly is an abundant component among

prebiotic organic compounds; another abundant prebiotic compound is Ala, which is encoded late in the SRM scheme. While only a partial fit was obtained in pursuing approach (b), it is now reported that approach (a) achieved full consistency.

Ser and Gly are the only amino acids belonging in the central and simpler metabolic routes, among a large variety of mainly organic acids. They would constitute a *bona fide* subsystem inside the GSP, relatively self-contained via amination and C1transfer loops. Gly may be synthesized by the anabolic GDC activity utilizing one-carbon compounds and NH_3, also indicating that this activity could have been an early mode for assimilation of NH_3 into organic molecules. The in vivo working of GDC in the biosynthetic direction is well documented in eukaryotes (Maaheimo et al. 2001; Hiraga et al. 1972; Motokawa and Kikuchi 1974) and Gly may serve as sole nitrogen source for yeast cultures (Sinclair and Dawes 1995). There are various metabolic routes generating glyoxylate (Ensign 2006), that may be aminated to form Gly. A degree of division of labor between Gly and Ser is apparent (Fig. 3): (a) Gly is more abundant than Ser in protein sequences; (b) Gly is involved with a large variety of metabolic syntheses including most remarkably the nucleotide bases, but Ser is a quantitatively more important metabolite in the network, mainly as source for the variety of C1 transfers. The SRM provides a neat indication for the early Gly-Ser association and also records the next most important metabolic development. Module 2a contains precisely Asp and Glu, and these four amino acids together comprise 56% of the total amino acid metabolic requirements for cellular growth (considering donated and protein carbons of the 20 amino acids; Fig. 3). The trend depicted by the succession C2 Gly, C3 Ser, C4 Asp and C5 Glu would reflect the rationale of enzyme evolution towards building binding sites for increasingly larger substrates. There is no clear quantitative metabolic trend along the following steps of the SRM, where mainly the localization of the punctuation system in the last modules becomes noticeable.

The proposition that the GSP is a strong candidate for a central position among the basic biosynthesis pathways is backed by various observations, besides its record in the first module of the SRM. An attempted chronology would be based on the rationale that the protein synthesis sink, as the main driving force to pull the development of the supporting metabolic pathways, is sensed directly by the Gly and Ser biosynthesis routes. The simplest component in these, considering a stepwise succession of carbon skeleton size and an original NH_3 assimilation source, is a GDC-like activity synthesizing the C2 Gly. The GDC-associated SHMT activity enters with transfer of C1, derived from the degradation of Gly by the GDC, to other Gly molecules therewith generating the C3 Ser. Transformations in the C3 realm are many and include the formation of trioses, from which gluconeogenesis and glycolysis can start. After the C6 level is reached, the citrate and the glyoxylate cycles develop. An intermediate step in composing these higher level pathways would be the fixation of the PP, introducing the glyoxylate precursor to Gly. The GSP and PP work concertedly, sharing glyoxylate and malate; this plus the other C4 components in these pathways contribute to form the C4 half of the citrate and glyoxylate cycles. Malate stands up as the hub connecting the three routes (Peyraud et al. 2009). It is consistent with most observations that, when glyoxylate sources are abundant, this becomes a preferred route for generation of Gly, and the GDC activity may work preferentially in the degradation direction (Maaheimo et al. 2001; Peyraud et al. 2009). A precursor to glyoxylate—glycolate—is known from the photorespiration pathway and is a likely prebiotic compound (Nuevo et al. 2010). The rationale above presupposes, among early routes, the concomitant formation of acetyl-CoA from C1 elements, such as found in methanogenesis and in the Ljungdahl-Wood pathway, which is consistent with the hydrogen hypothesis on metabolic origins (Martin and Russell 2003).

Protein Evolution

Starting the encodings with Gly is a common feature of the SRM and of other models (Trifonov 2004; Higgs 2009). Otherwise, Trifonov (2004) and the SRM disagree about the second encoding. Trifonov´s second example would be Ala but he could reportedly find only a Gly clock (Trifonov 1999; Trifonov, personal communication). The second encoding of the SRM is Ser, which now receives metabolic support, and Ala remains a late addition to the code. Ser is a C3 compound but really belongs in the realm of biosynthesis from C1 elements. Ala belongs in the fully C3 realm, deriving mainly from pyruvate; other Ala sources are branched-chain amino acids and Cys. Ala is a main source of pyruvate for gluconeogenesis and participates more heavily than Ser or Gly in the structure of bacterial cell wall peptidoglycans (Volmer et al. 2008). Similarly, our view is that the derivation of Ser from the glycolysis intermediate glycerate occurred later than the derivation of Ser from Gly; therefore, the glycolysis route to Ser would not be relevant to the order of fixation of the encodings.

Possible roles of Ser in the early Gly- and Ser-rich peptides and RNPs would be of modulating the strong RNA-binding and stabilization properties of Gly. It is a weaker RNA-binder (Guimarães and Moreira 2004) and it can be reasoned that excessive stabilization could be impeditive to functions that would require flexibility and plasticity of the complexes. In present day proteins, Ser demonstrates 'informational' attributes, as judged by the large variety of post-translational modifications it is subjected to (Kyte 1995a); these are less frequent with Gly and it is noteworthy that Ala does not show up in the long list of known post-translational modifications. The indicated significant roles of Gly and Ser in proteins can be contrasted with the roles of Ala, which would be mainly related to formation of α-helices and to participation more in DNA-binding than in RNA-binding motifs (Guimarães and Moreira 2004). Mutational records indicate Ala to be a mostly neutral amino acid, showing high rates of exchangeability with a large diversity of other amino acids and at different sites (Kyte 1995a; Zuckerkandl 1975).

An aspect still unresolved in the SRM refers to the precise partition of components between the third and fourth modules. The 'windmill' structure (Guimarães et al. 2008a, b) considered the possibility of the biosynthesis of His being independent from other amino acids and that it should have entered the code before its companions in the NYR quadrant of anticodons, which are all derived in amino acid families. Following this rationale, the windmill configuration emerged, with the two boxes containing the termination codes becoming encoded last. We are now considering revising the configuration of the last modules (see the section Hydropathy correlation), giving more weight to the self-contained structure of the modules of triplet dimers (in preparation).

Acknowledgments The enlightening discussions in the Self-organization Group, Logics and Epistemology Center, UNICAMP, and the continued collaboration with Carlos Henrique Costa Moreira, are gratefully acknowledged.

References

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

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

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. Braz J Med Biol Res 22:931–944

Davis BK (1999) Evolution of the genetic code. Prog Biophys Mol Biol 72:157–243

DiGiulio M (2005) The origin of the genetic code: theories and their relationships, a review. BioSystems 80:175–184
Ensign SA (2006) Revisiting the glyoxylate cycle: alternate pathways for microbial acetate assimilation. Mol Microbiol 61:274–276
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
Grosjean H, Houssier C (1990) Codon recognition: evaluation of the effects of modified bases in the anticodon loop of tRNA using the temperature jump-relaxation method. In: Gehrke CW, Kuo KCT (eds) Chromatography and modification of nucleotides. Elsevier, Amsterdam, pp A255–A295
Guimarães RC, Moreira CHC (2004) Genetic code—a self-referential and functional model. In: Pályi G, Zucchi C, Caglioti L (eds) Progress in biological chirality. 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 69–110
Hamosh A, Johnston MV (2000) Nonketotic hyperglycinemia. In: Scriver CR, Beaudet AL, Sly WS, Valle D (eds) The metabolic and molecular bases of inherited disease. McGraw-Hill, New York, pp 2065–2078
Heil G, Stauffer LT, Stauffer GV (2002) Glycine binds the transcriptional accessory protein GcvR to disrupt a GcvA/GcvR interaction and allow GcvA-mediated activation of the *Escherichia coli* gcvTHP operon. Microbiology 148:2203–2214
Heimann A, Jakobsen R, Blodau C (2010) Energetic constraints on H_2-dependent terminal electron accepting processes in anoxic environments: a review of observations and model approaches. Environ Sci Technol 44:24–33
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
Hiraga K, Kochi H, Motokawa Y, Kikuchi G (1972) Enzyme complex nature of the reversible glycine cleavage system of cock liver mitochondria. J Biochem 72:1285–1289
Klipcan L, Safro M (2004) Amino acid biogenesis, evolution of the genetic code and aminoacyl-tRNA synthetases. J Theor Biol 228:389–396
Knight RD, Freeland SJ, Landweber LF (2001) Rewiring the keyboard: evolvability of the genetic code. Nat Rev Genet 2:49–58
Kyte J (1995a) Structure in protein chemistry. Garland, New York
Kyte J (1995b) Mechanism in protein chemistry. Garland, New York
Maaheimo H, Fiaux J, Çakar ZP, Bailey JE, Sauer U, Szypersky T (2001) Central carbon metabolism of *Saccharomyces cerevisiae* explored by biosynthetic fractional ^{13}C labeling of common amino acids. Eur J Biochem 268:2464–2479
Madigan MT, Martinko JM, Parker J (2003) Brock biology of microorganisms. Pearson Education Inc.—Prentice Hall, Upper Saddle River
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. Philos Trans R Soc Lond B 358:59–85
Motokawa Y, Kikuchi G (1974) Glycine metabolism by rat liver mitochondria. Reconstruction of the reversible glycine cleavage system with partially purified protein components. Arch Biochem Biophys 164:624–633
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
Peterson RB (1982) Regulation of the glycine decarboxylation complex and of the L-serine hydroxymethyltransferase activities by glyoxylate in tobacco leaf mitochondrial preparations. Plant Physiol 70:61–66
Peyraud R, Kiefer P, Christen P, Massou S, Portais JC, Verholt JA (2009) Demonstration of the ethylmalonyl-CoA pathway by using ^{13}C-metabolomics. Proc Natl Acad Sci USA 106:4846–4851
Pizer LI, Potochny ML (1964) Nutritional and regulatory aspects of serine metabolism in *Escherichia coli*. J Bacteriol 88:611–619
Randau L, Münch R, Hohn MJ, Jahn D, Söll D (2005) *Nanoarchaeum equitans* creates functional tRNAs from separate genes for their 5′ and 3′ halves. Nature 433:537–541
Reitzer L (2003) Nitrogen assimilation and global regulation in *Escherichia coli*. Annu Rev Microbiol 57:155–176
Sinclair DA, Dawes IW (1995) Genetics of the synthesis of serine from glycine and the utilization of glycine as sole nitrogen source by *Saccharomyces cerevisiae*. Genetics 140:1213–1222

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
Volmer W, Blanot D, Pedro MA (2008) Peptidoglycan structure and architecture. FEMS Microbiol Rev 32:149–167
Wong JTF (1975) A co-evolution theory of the genetic code. Proc Natl Acad Sci USA 72:1909–1912
Wong JTF (2005) Coevolution theory of the genetic code at age thirty. BioEssays 27:416–425
Wood AP, Aurikko JP, Kelly DP (2004) A challenge for the 21st century molecular biology and biochemistry: what are the causes of obligate autotropohy and methanotrophy? FEMS Microbiol Rev 28:335–352
Zuckerkandl E (1975) The appearance of new structures and functions in proteins during evolution. J Mol Evol 7:1–57