CMLS, Cell. Mol. Life Sci. 55 (1999) 1626–1635
1420-682X/99/121626-10 $ 1.50 + 0.20/0

CMLS Cellular and Molecular Life Sciences

Review

Structure and evolution of the genetic code viewed from the perspective of the experimentally expanded amino acid repertoire in vivo

N. Budisa[a,b,*], L. Moroder[b] and R. Huber[b]

[a]Proteros Biostructures GmbH, Am Klopferspitz 19, D-82152 Martinsried (Germany), Fax + 49 89 7007 6115, e-mail: budisa@proteros.com
[b]Max-Planck-Institut für Biochemie, Am Klopferspitz 18a, D-82152 Martinsried (Germany)

Received 30 June 1999; accepted 9 July 1999

Abstract. Much effort has been devoted recently to expanding the amino acid repertoire in protein biosynthesis in vivo. From such experimental work it has emerged that some of the non-canonical amino acids are accepted by the cellular translational machinery while others are not, i.e. we have learned that some determinants must exist and that they can even be anticipated. Here, we propose a conceptual framework by which it should be possible to assess deeper levels of the structure of the genetic code, and based on this experiment to understand its evolution and establishment. First, we propose a standardised repertoire of 20 amino acids as a basic set of conserved building blocks in protein biosynthesis in living cells to be the main criteria for genetic code structure and evolutionary considerations. Second, based on such argumentation, we postulate the structure and evolution of the genetic code in the form of three general statements: (i) the nature of the genetic code is deterministic; (ii) the genetic code is conserved and universal; (iii) the genetic code is the oldest known level of complexity in the evolution of living organisms that is accessible to our direct observation and experimental manipulations. Such statements are discussed as our working hypotheses that are experimentally tested by recent findings in the field of expanded amino acid repertoire in vivo.

Key words. Amino acid repertoire; evolution; genetic code; metabolism; protein folding.

An expanded amino acid repertoire and the genetic code

New terminology

In our first attempts to interpret new experimental findings in the context of the structure and evolution of the genetic code, we found inconsistent terminology for amino acids to be a major stumbling block. In other words, we have been convinced that understanding the genetic code at a deeper level cannot be attained with the current taxonomy of amino acids. Thus, we proposed a new nomenclature which should not be difficult to integrate into the already existing biochemical terminology [1]. In brief, the well-known standard set of 20 amino acids represents *canonical amino acids*; other amino acids outside this standard set which can be introduced in a codon-dependent manner are *non-canonical amino acids*. Those amino acids whose introduction is not only codon dependent, but also dependent on context (e.g. selenocysteine or formyl-methionine) are *special canonical amino acids*. There are also experimental procedures, such as in vitro suppression, that led to context-dependent introduction of

* Corresponding author.

some amino acids with special properties (e.g. cages, sensors) that are named *special non-canonical amino acids*. Finally, numerous amino acids resulting from secondary metabolism, precursors or post-translational modifications are *special biogenic amino acids*. This distinction is not a semantic issue, but a biological one. In fact, only such terminological clarification convinced us to propose the concepts about the genetic code where the amino acid repertoire is a central criterion for dissecting its nature, evolution and establishment.

Steps in the flow of genetic information

The flow of genetic information is a complex process. It starts with self-replicating DNA that contains instructions that are converted into biological activity through transcription followed by translation, as

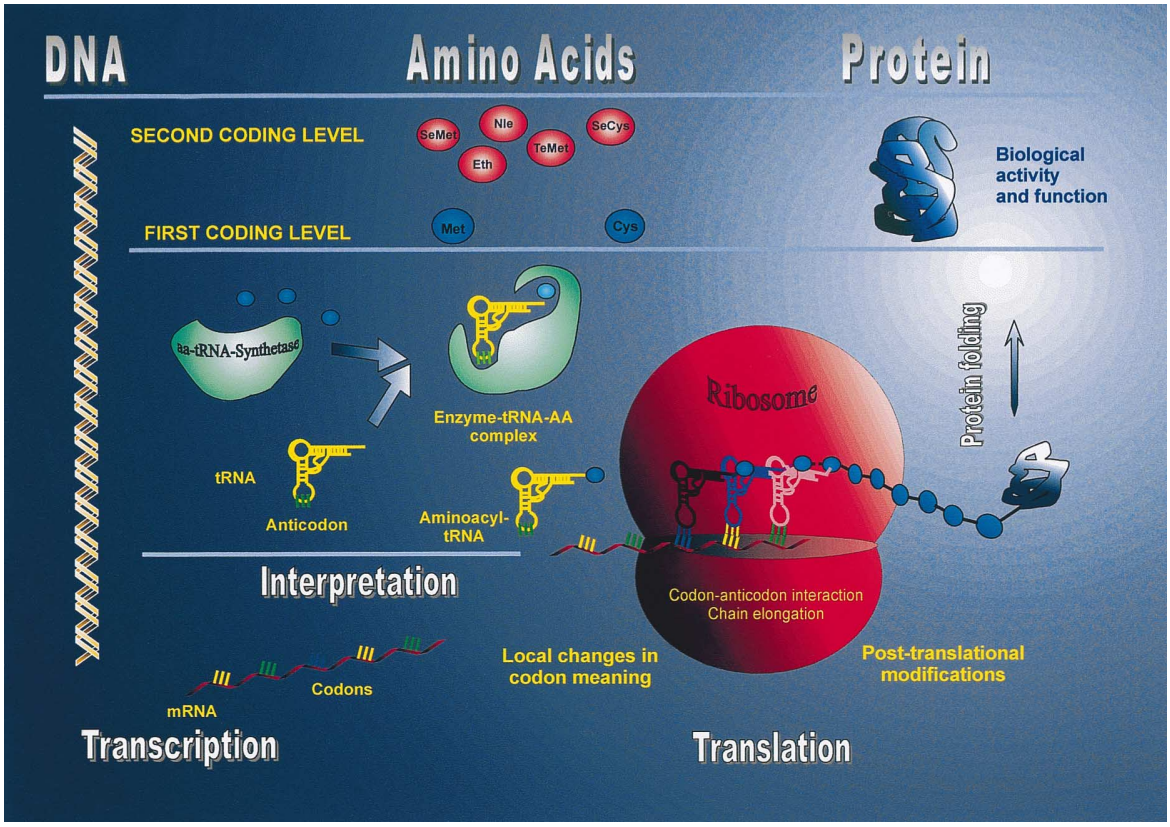

Figure 1. An emerging new structure of the universal code. The universal genetic code is structured in such a way that the genetic message written in the DNA sequence is always precisely transmitted into biological activity. There are intrinsic determinants and constraints that always lead to proper and optimal stability and folding of functional proteins [1]. The message is decoded via mRNA-tRNA codon-anticodon specific pairing; the meaning of each triplet is interpreted by specific recognition and pairing of the appropriate tRNA and amino acids by aminoacyl-tRNA synthetase enzymatic activity. This is the interpretation step in the flow of genetic information. In general, the code is organised so that the codon meanings in all species are not universal, but the amino acid repertoire is universal. This means that living organisms can exceptionally survive codon reassignments in the genetic code, but not new additions into its amino acid repertoire. The deterministic nature of the genetic code defined in such a way stems from: (i) specific rules for protein self-assembly in the cellular milieu, (ii) interrelated cellular metabolic and bioenergetic networks in which each of the 20 standard amino acids is caught. In experimental conditions, using recombinant DNA technology combined with artificially achieved strong selective pressure and avoidance of metabolic toxic effects, redefinitions of the codon meanings are possible without changing aminoacyl-tRNA synthetases or tRNAs [48]. For example, in vivo, the AUG codon is interpreted as the canonical amino acid methionine (this is the 'restricted' or 'first' coding level) while under experimental conditions, AUG allows for introduction into stable and functional proteins of the following non-canonical amino acids: selenomethionine (SeMet), telluromethionine (TeMet), norleucine (Nle) and ethionine (Eth). This represents the 'relaxed' or 'second' coding level of the universal code. Similarly, two cysteine codons, UGU and UGC, can be interpreted at the second level as the non-canonical amino acid selenocysteine (SeCys). Note that codon redefinitions will always be to similar amino acids; for example, AUG (methionine) could be reassigned to isoleucine or norleucine but never to arginine, as this reassignment would be fatal for efficient folding and consequently for biological activity.

shown in figure 1. Note that a novel feature is added to this scheme: an interpretation step. In fact, such a claim should not be surprising at least for researchers in the field of aminoacyl-tRNA synthetases since this idea has already been circulating in this scientific community [2]. Thus, it is included in our scheme as a legitimate step in the process of genetic information transfer, not only for academic interest but also because it represents a new level accessible to our experimental efforts and subsequently practical applications. Namely, at this level, it is possible to intervene experimentally and under specific conditions to change coding meanings of the codon triplets. In this scheme, the 'first' or 'restricted' part of the universal code works only for canonical amino acids. The 'second' or 'relaxed' coding level includes various non-canonical amino acids that can be translationally integrated in vivo under defined experimental conditions. With this level of conceptual resolution in mind, we can conceive an emerging new structure of the genetic code as will be discussed below.

General considerations

The reader will probably note that in this review the experimental work performed in our laboratory and those of other groups working in the same field is often cited. That does not mean, however, that we are not aware of many other equally valuable and important contributions in the field of genetic code evolution. But, it is not our intention to provide a comprehensive review in each of these subjects, since excellent, extensive and recent treatments of each topic in this field are already available [3–6]. Nevertheless, we must express some reservations. It is highly probable that in this field of many authors some of our ideas and concepts will not be surprising. For example, our statements come very close to the very early proposal made by Crick ('frozen accident' theory) [7]. Our concepts fit quite well in the frame of this theory because we fully agree with the following statements of Crick's theory: (i) the code is universal; (ii) the code does not change; (iii) all life evolved from a single organism. But we disagree with the final statement that (iv) allocation of codons at this point was entirely a matter of chance. According to our criteria (as will be discussed later), the process of codon allocations to particular amino acids cannot be entirely random. It should be noted, additionally, that Crick was concerned about codon reassignments and not about the amino acid repertoire itself. To our knowledge, there are no studies which consider the repertoire of amino acids for protein synthesis in living cells as the main criterion for considerations of the structure and evolution of the genetic code. Moreover, our approach to this matter is in full agreement with Popper's principle of explanatory power [8]. It requires an explanation of the large number of facts (in this case relevant to the genetic code) with the smallest possible number of evolutionary assumptions, as was elaborated by Wächterhäuser [9].

Statement 1: the nature of the genetic code is deterministic

First major determinant: protein self-assembly

Over the previous four billion years, proteins have evolved to fold in specific and compact three-dimensional conformations. Such structures generate 'active sites' through precise arrangements of functional groups that are able to carry out sophisticated chemical reactions [10]. The principle of protein self-assembly (folding) pioneered by Anfinsen states that the steric information for newly synthesised protein chains to fold correctly within cells resides solely in their primary structure or sequence [11]. It was argued that a general and deterministic set of rules (that remains to be understood) for protein self-assembly must exist because it is unlikely that each protein folds into correct conformation by its own specific rules [12]. On this basis, we postulate that the protein self-assembly rules are one of the basic principles behind establishment of the genetic code.

In previous work, we found that all steps of the translational process in living cells "are mutually interdependent and intrinsically coupled with the proper folding of the resulting protein" [1]. Therefore, experimental attempts to introduce non-canonical amino acids will always face stringent substrate specificity of the enzymes (i.e. aminoacyl-tRNA synthetases) in the interpretation step of the code. Those non-canonical amino acids that are in the range of this substrate specificity are further submitted to editing mechanisms. Even then, if they pass this 'checkpoint', ribosome editing is the next stumbling block for successful translational integration in vivo and even in vitro, despite its rather broad substrate specificity [13, 14]. Attempts to relax the substrate specificity of aminoacyl-tRNA synthetases have been described [15], and it has been even possible to translate phenylalanine codons as chloro-phenylalanine and bromo-phenylalanine into protein sequences. However, it was not possible to fold this protein into an active structure. In the context of the evolution of life on earth, chlorine and bromine were readily available in early prebiotic conditions [16], as were, probably, chlorinated and brominated aromatic amino acids. But they have never been used for protein building, due to the difficulty in accommodating such bulky atoms into the tightly packed protein core. In this case, the bulkiness of the amino acids represent the major limiting factor while other possible limits are discussed elsewhere [1]. In any case, the physical-chemical determinants (e.g. bulki-

ness, polarity, stereochemistry) in protein building and folding processes are very strong.

The lessons here are obvious: they enable us to learn which new amino acids are allowed as 'proteinogenic' amino acids in the translational process. There is no reason to believe that natural selection escaped these rules for building proteins with sufficient and balanced stability and functionality in living beings in the course of evolution.

Second major determinant: metabolism

It is not difficult to imagine that many other non-canonical amino acids outside the canonical 20 would actually enable protein building and folding according to the rules discussed above. For example, it was demonstrated in the 1960s [17] that experimental replacement of canonical methionine residues into staphylococcal nuclease with the non-canonical amino acid norleucine occurs without significant effects on structure and activity. In the last few decades, other groups have repeated this experiment with other proteins [18, 19]. Norleucine is abundant in carbonaceous Murchison meteorites and apparently represents a product of abiotic synthesis in greater amounts than most canonical amino acids [20]. Why then is it not present in the code? Norleucine itself is highly toxic for all cells since it inhibits cell growth [18] and its incorporation into cellular proteins would be lethal for living cells and it is therefore excluded from the genetic code.

This conclusion is fully supported by recent experiments in the field of the expanded amino acid repertoire in vivo. Based on these experiments, two levels in the structure of the genetic code have been proposed. The first or 'restricted' part of the universal code encodes only those amino acids that are optimally integrated into the metabolic chemistry of the living cells allowing their reproduction, growth and differentiation. In vivo integration into proteins of non-canonical amino acids like norleucine is possible only after efficient elimination of their toxic effects on cellular metabolism. This is the 'second' or 'relaxed' part of the genetic code. At this level it is possible to establish why many other non-canonical but 'proteinogenic' amino acids are prevented from entering the code. Namely, canonical amino acids are involved in the finely tuned metabolic and bioenergetic network of living cells, while non-canonical ones are not. Such reasoning is reconcilable with the co-evolution hypothesis [21] which proposes that the genetic code co-evolved with the prebiotic biosynthetic pathways for production of amino acids. This theory postulates the existence of only a few prebiotically abundant amino acids which were used in the evolutionary time course in inventive biosynthetic processes for production of other amino acids that are currently present in all living organisms. According to this theory, allocations of amino acids in the genetic code stem primarily from the biosynthetic relationships between them.

Evolution and establishment of the code could have been bi-phasic

Taking into account the arguments discussed above, we postulate that the evolutionary process of code expansion was determined by rules and principles that govern proper protein folding, and metabolic and bioenergetic requirements of primitive cells. The evolution of the code to its established form was probably bi-phasic: at the beginning, the physico-chemical determinants played a dominant role while towards the end, bioenergetic and metabolic constraints became increasingly stringent (fig. 2). In this way, a deeper understanding of the origin of the genetic code would derive from integrating various relevant aspects and facts responsible for its establishment, as recently discussed by DiGiulio [22]. In this context, the meaning of the statement 'deterministic nature of the genetic code' could be explained as follows. Under the specific conditions and according to principles of protein self-assembly, in the framework of defined patterns of metabolism, it might be possible to create a universal genetic code experimentally. We are not alone in such reasoning: in their study of factors responsible for the occurrence of the 20 coded protein amino acids, Weber and Miller [23] reached similar conclusions.

The arrangement of the code before conservation: chance or necessity?

It is difficult to explain the preservation of characteristic folds and stable conformations in proteins (despite their great sequence variation) over the course of evolution by accidental association between amino acids and codons. Indeed prior to the advent of an adapter system, the initial codon allocation may have occurred in an accidental manner in a primitive system where a few (two or three) amino acids and nucleotides interacted (directly or indirectly) with each other. However, for further protein building, the selection pressure for rational use of available biosynthetic and metabolic resources simply cannot be ignored.

It has been demonstrated experimentally that the relative distribution of amino acids between the surfaces and the interiors of native globular proteins is associated with a sharp bias in the genetic code [24]. Accordingly, a mutation introducing hydrophilic amino acids at interior hydrophobic locations, which is especially damaging, is not likely since the codons of hydrophobic amino acids are grouped together (the XUX group). Such arrangements may have been crucial during evolu-

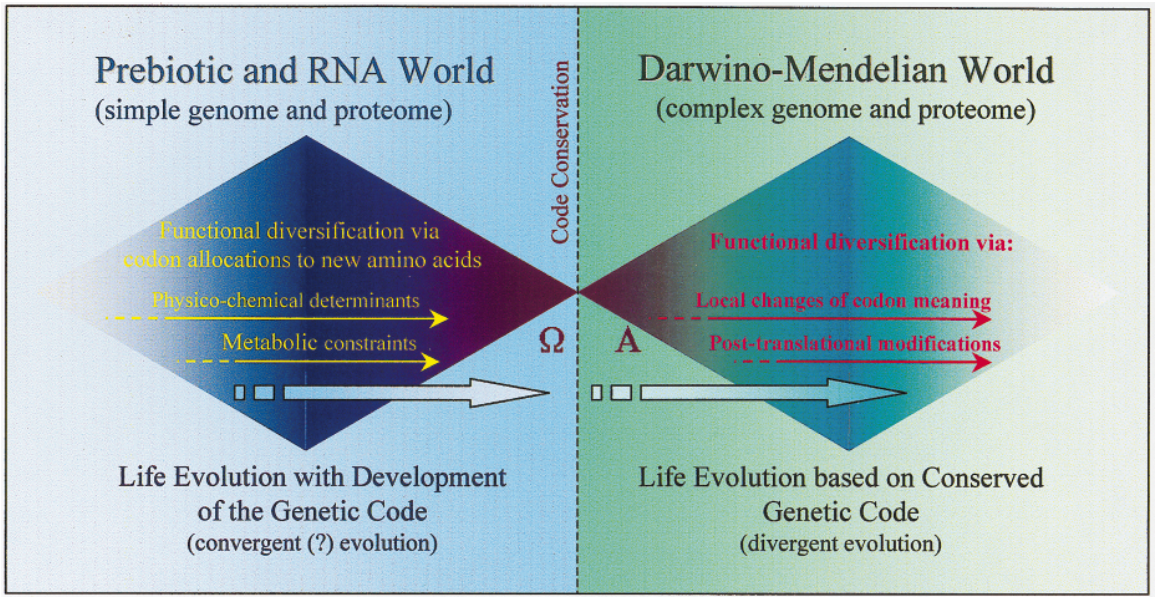

Figure 2. The evolutionary process of developing and establishing a universal code. Physico-chemical determinants and metabolic constraints during code expansion resulted in its optimisation and conservation ('omega' point) in the context of the prebiotic and RNA world. Living cells which adopted the universal code entered a new Darwino-Mendelian world ('alpha' point) with highly organised and relatively more stable genomes and proteomes. This complexity was further increased first by using local changes of the codon meaning in prokaryotes, and later by the evolution of post-translational modifications in eukaryotes, giving rise to a vast outburst in life diversity on earth.

tion in maintaining the structural stability of globular proteins. Furthermore, the existence of synonymous quotas (a number of synonymous codons allocated to particular amino acids) for different amino acids [25] may also have ensured the forced distinction of hydrophobic versus hydrophilic residues. Indeed, the general principle of globular protein organisation (polar-out, apolar-in) indicates that first additions to the code were either strictly apolar or strictly polar. This rigid mechanism for maintaining the protein structure was very important for the primitive proto-cells, due to the absence of efficient proof-reading and editing mechanisms. These particular amino acids (e.g. Arg, Ser, Leu) have highly synonymous quotas (increased redundancy) in order to achieve a low mutation frequency. In their recent study, Xia and Li [26] found that primitive amino acids differ in polarity and hydropathy and little in other properties.

Therefore, primitive cells with a relatively simple organisation started code expansion first through additions that were probably strictly apolar (in the core or more probably as the integral transmembrane part) and strictly polar (at the surface). Later additions were probably more 'promiscuous', for which methionine is an example: methionine is distributed mainly in the protein core but about 15% is also found at the surfaces and in minicores [27]. Such amino acids are characterised by smaller synonymous quotas and, correspondingly, there is a greater probability that these may be substituted without significant disturbance of the protein structure. In fact, it has been argued that the code has evolved toward minimising differences in polarity and hydropathy of its amino acids [26], confirming the pioneering concept of Sonneborn [28] that it is intrinsically structured to enable minimisation of deleterious mutation effects.

Moreover, under the conditions of such a primitive living world, even our term 'species' would be inadequate, as there simply would not be interbreeding populations which could be genetically isolated from other populations. Genetic barriers were probably not so strict due to the smaller genomes and proteomes, and in this context further codon captures were possible via a combination of codon usage, gene exchange and recombination, and random mutagenesis of both tRNA and aminoacyl-tRNA synthetases. Each codon capture was probably random; but once favourable, it was cemented in the code by the selective pressure that acted in a feedback-like manner.

'Error catastrophe' and the 'central dogma' of molecular biology

According to the proposal of Woese et al. [29], primitive cells started with a completely random, highly ambiguous set of codon assignments with very inaccurate translation. Such early systems would produce bad copies of proteins (faulty proteins) with a higher average translational error and this in turn would produce even worse proteins, so that the advantage gained by the mutational chance event for the system could be quickly lost in an 'error catastrophe'. How could stable translation be achieved in this primitive, inaccurate translational machinery? We find such reasoning unlikely for several reasons. First, such a catastrophic situation should never prevail, because selection acts in a feedback-like manner eliminating bad copies from the population. Second, the postulated ribozyme-mediated translation [7] was probably improved constantly through recombination and gene transfer among populations. Finally, the smaller genome and proteome contained simple but efficient proto-enzymes in the primitive cells, which would allow for further codon captures without deleterious effects for the cell. It has always to be kept in mind that from the very beginning, the chemistry of life was well organised [30].

The 'central dogma' of molecular biology prescribes unidirectional flow of genetic information from self-replicating DNA through intermediate mRNA to protein [31]. Indeed, such unidirectional flow without control would certainly lead to 'error catastrophe' even in the conditions before code establishment. It has often been argued that this 'central dogma' neglects possible cellular feedback mechanisms [32], and thus leads to a simplified scheme of mechanisms that control gene expression and protein synthesis [33]. At the same time, we have no experimental evidence that there are feedback mechanisms in translation. In modern cells, the fidelity and precision of this process are ensured by the numerous proof-reading and editing steps [1]. In the context of the postulated primitive prebiotic [16, 30] and RNA world [7, 34–36] (in which the code is believed to have been established), the mechanisms of selective pressure which acted in a feedback-like manner can sufficiently explain the appearance of stable translation, i.e. how 'error catastrophe' was avoided. Combined, this constitutes a strong argument for replacing the oversimplified and naive 'central dogma' concept with a more realistic scheme, as shown in figure 1.

Statement 2: the genetic code is conserved and universal

Which factors caused conservation of the code?

From an evolutionary viewpoint, entry of a novel amino acid into the code is reserved for those new residues whose translational integration results in a protein with superior cellular function and biological activity. The introduction of other amino acid building blocks, i.e. those that do not bring any substantial advantage or that are even damaging for cells, renders them inferior and susceptible to elimination from the population. In addition, each successful expansion step is probably conserved in the code and additionally 'cemented' by efficient proof-reading and even editing mechanisms accompanied with optimal integration in the cellular metabolic chemistry. Such early selective pressure also explains how the interpretation level of the code was established. When hereditary molecules acquired intrinsic properties to ensure proper physical, chemical and genetic stability, and when a further amino acid repertoire expansion was not possible without lethality for living cells, its conservation was inevitable. The appearance of the code with such properties was the turning point in evolution [37]. From this point on, the genome and proteome complexity allowed further functional diversification by other means.

Three important questions

The discussion about universality and conservation of the genetic code should address at least three important questions. First, is codon specificity reserved only for the 20 canonical amino acids [23]? Second, why are amino acids such as selenocysteine not a '21st amino acid' [38] since such residues are sometimes introduced into proteins as a result of the UGA codon. And finally, why were several hundreds of 'other' amino acids (special canonical, special biogenic, non-canonical) excluded from the coding process.

Beyond the 20-amino-acid repertoire: local changes in codon meaning

The process of code establishment included a physico-chemical phase accompanied by a bioenergetic-metabolic phase, making it increasingly difficult to extend the code vocabulary in terms of new amino acids. The process reached the number 20 and stopped. Within this rigid structure, it became advantageous to alter single individual proteins which would confer a selective advantage to the whole cell, while keeping the change specific, i.e. separate form other proteins. For example, one could improve the catalytic properties (e.g. nucleophilicty) of a protein by replacing the amino acid cysteine in the active site with selenocysteine. Indeed, selenocysteine is introduced as a response to the UGA codon in a variety of organisms, by distinct mechanisms that keep an internal UGA codon separated from other UGA stop codons [38]. In this way,

functional performance of the target protein is improved without disturbing cellular viability by local change in codon meaning. This process includes a special context characterised by a series of enzymatic changes, the presence of special elongation factors, and loops in the mRNA where tRNA serves as substrate carrier. This last feature indicates that this mechanism is probably old and universal, i.e. it is present in the range from Archaea to mammals [39]. Similarly, N-terminal methionines are separated from other identical residues in the structure by unique formylation. Other examples of context-dependent deciphering of codons are exceptional phenomena of suppression [40, 41], where the meaning of one codon as stop is changed such that it encodes a canonical amino acid (nonsense suppression), or a canonical amino acid is replaced by another (missense suppression).

Further diversification beyond the standard repertoire: post-translational modifications

The processes of local changes in codon meaning are not strictly separated from the coding process, and thus are probably as old as the basic coding itself. New evolutionary inventions that led to strict separation of the newly produced biogenic amino acids from the genome-coding process and RNA world are post-translational modifications. The appearance of these mechanisms in eukaryotes is accompanied by vast protein functional diversification. Enzyme-mediated production of novel biogenic amino acids is made possible by reactions like phosphorylation, glycosylation, acetylation or by oxidative modifications (i.e. hydroxylations) of aspartate, proline, lysine and tyrosine. In this way, hundreds of new biogenic amino acids, which are completely separated from the coding process, are produced in the cytoplasmic functional proteins. All these operations are energetically expensive but eukaryotes can afford them by dropping more and more basic synthetic activities. While all prokaryotes have a simple synthetic machinery that uses raw chemical building blocks and thus are independent of any other life form, eukaryotes are freed from the task of completing many basic biological syntheses. Thus, eukaryotes which retain the basic cytoplasmic chemistry of prokaryotes can use other life forms as a source of essential chemicals like fats, essential amino acids, co-enzymes and minerals [30].

Genetic code evolution or evolution based on a conserved code?

Together, local changes in codon meaning and post-translational modifications are good strategies to compensate for coding repertoire expansion, leading to increasing complexity of the living systems, i.e. to evolution. Evolution has thereby produced increasingly advanced cellular forms with superior protective, sensitive and cognitive mechanisms, allowing a greater survival advantage, without the need for substantial changes in the established code design. In this context, it is hard to anticipate any significant changes in the structure of the basic code with a standardised and conserved amino acid repertoire. Even if we accept arguments of 'recent evidence for evolution of the genetic code' [20] it would be extremely difficult to imagine how all these codon reassignments would spread across all life kingdoms without globally destructive effects, taking into account the interdependence of life forms on earth. Thus, arguments that the code further evolves via recruiting mechanisms of local changes in codon meaning (e.g. SeCys introduction) or post-translational modifications (e.g. phosphoserine incorporation) [42] are rather anecdotal, since the evolution of life based on a conserved code is at work (fig. 2).

'Copernican turn' in reasoning about the 'standard' genetic code

During the last few decades, numerous changes in the 'standard' genetic code have been found in both prokaryotes and eukaryotes [20]. Here the question arises as to what is the 'standard' genetic code? One where AUA reads as Ile or one where it reads as Met? Those codon reassignments found in *Escherichia coli* or in eukaryotic nuclear genes are normally assumed as 'standard code'. On this basis, many species (even in the same genus) have their 'own code'. Very instructive examples of codon reassignments are found and now well documented for several pathogenic species in the genus *Candida* where they are used for adaptation to new ecological niches [43]. Does this mean that several genetic codes exist in this species? Or better, does it mean that in every eukaryotic cell there are at least two genetic codes? Such reasoning is the consequence of the dogma that a particular codon cannot have more than one meaning. Conversely, the lesson from the multiple meaning of e.g. the UGA codon (reads sometimes as SeCys, in some organisms as Trp, and mainly as a termination signal) is that the codon meanings are not universal, they are flexible to a certain extent in the range of the interpretation level of the code. Moreover, there is now solid evidence that two distinct canonical amino acids can be assigned by a single codon and this type of codon is even named a 'polysemous' codon [44]. These ambiguities disappear when assuming an amino acid repertoire for protein synthesis in living beings as the central criterion for possible genetic code changes in taxonomically different biological categories (e.g. species, genus, families, phyla). This new reasoning is a sort

of 'Copernican turn' in our conceptual perception of the genetic code since codon reassignments are now of second importance. They are allowed to a certain extent in the context of the genetic code only if they do not break (i) specific rules for protein self-assembly in the cellular milieu, and (ii) interrelated networks of metabolic and bioenergetic relationships in living organisms. The amino acid repertoire is of prime importance. The introduction of a new amino acid into the conserved set of 20 amino acids is not allowed since (i) the organisms cannot survive changes that would insert new amino acids into the repertoire of the universal genetic code, and therefore (ii) there is no evolution 'standard' → 'alternative' amino acid repertoire in the genetic code.

Following this logic of reasoning, it is obvious that the genetic code is conserved and universal. Its universality could only be questioned seriously in the case of the existence of species that regularly, in a codon-dependent manner, build proteins with amino acids like homoserine, homocysteine, ornithine, α-aminobutyric acid, norleucine, selenocysteine, phosphoserine and norvaline. To our knowledge, there are no such examples in the living world. There are no codon captures which introduce new amino acids, but only codon reassignments in various organisms from one terminal signal or one canonical amino acid to another, e.g. STOP → Gln, Cys → Trp, Ile → Met, always in the framework of the same, standardised amino acid repertoire. In other words, there is no experimental evidence for amino acid repertoire extension in all living beings.

Species-specific and experimental codon reassignments

As discussed above, species-specific reassignments or changes in particular codon meaning are often reported. This is especially pronounced in certain rapidly evolving cellular organelles such as mitochondria [20] where, in the context of the smaller genome and proteome, such reassignments are more likely, since lethal effects are less possible. In other living organisms, these species-specific reassignments are sometimes 'destructive'. An example is the Leu → Ser change for the CUG codon in some *Candida* species that probably enables better survival of these obligatory parasitic species in their hosts, i.e. they function for adaptation to the new ecological niches [43]. Disturbing effects of such reassignments combined with codon usage are avoided in these cells since the CUG codon is not used as a codon in the cellular mRNA but only in a stress response like heat shock [45]. It is therefore plausible to expect that future investigations will reveal many other examples of such or similar phenomena. But even then, it is not difficult to anticipate that the only possible context for their exceptional occurrence will be in the framework of the evolutionarily established amino acid repertoire of the genetic code.

Possibilities for intervening experimentally at the interpretation level of the code are now well established. Introduction of numerous non-canonical amino acids into proteins can be accomplished in vivo and in vitro as well [46]. In this context, the question arises as to how such findings can be interpreted in the light of the discussed concepts? Such findings do not contradict the above concepts and hypotheses since this artificially achieved coding level ('second' code) is also within the framework of the universal code. It follows an 'amino acid replacement model' [47] which postulates a 'similar replaces similar' mode of reassignment [1]. Finally it has to be kept in mind that such reassignments are possible only under special laboratory conditions using recombinant DNA technology combined with artificially achieved strong selective pressure and avoidance of toxic effects for metabolism [48].

Statement 3: the genetic code is the oldest known level of complexity in the evolution of living organisms that is accessible to our direct observation and experimental manipulations

Oldest molecular fossil known

In the context of the evolution of life on earth, the development and establishment of the genetic code represent one of the most fascinating aspects. Although in the 1960s, Crick prophetically stated that "discussion of the actual amino acids used in the code may not be very profitable. It might be more useful to consider which amino acids are not used in the code", until recently there were no systematic attempts to study which amino acids in addition to the canonical 20 can be introduced into proteins during biosynthesis in vivo. After studying the genetic code in the context of experimental attempts to expand its amino acid repertoire, we have arrived at a major conclusion: that the genetic code is only one, but basic, level in the developing scheme of complex life structures. In other words, it represents the oldest molecular fossil known. The structure of the universal code reveals the complexity of the living world existing at the time of code conservation. This structure is not unknown to us—it is actually accessible to our experimental examination and even manipulations. In fact, experimental study of this level has revealed the rules and determinants that 'cemented' 20 canonical amino acids as the standard repertoire in the genetic code. Such a set of conserved building blocks was the basis for establishing other levels of complexity that generated immense biodiversity in the course of the evolution of life on earth. It is beyond the scope of this discussion to consider the constraints on the selection of the many

other biogenic and special canonical amino acids into the bioenergetic and metabolic chemistry of living organisms. Long after the standard set of the amino acid building blocks was established, these amino acids originated at other levels above the basic genetic code level of complexity. In such a conceptual framework, their generation represents the introduction of biological novelties that can certainly be better understood after experimental dissection of these levels.

'Alpha' and 'omega' points

It is widely accepted that conditions for the appearance of life existed on earth around four billion years ago [30]. The genetic code was established relatively quickly after that event, either through convergent or divergent evolution in the context of the postulated (and widely accepted) prebiotic [16, 23, 30] and RNA world [7, 34–36]. Alternatively, its appearance might be viewed in the context of the less widely accepted hypothesis of 'cosmic panspermia' [49] which suggests that information necessary for the beginning of the evolution of life was transmitted to the earth. In any case, it has not escaped the attention of many researchers in the field that the uniformity of biochemistry in all living organisms argues strongly that all modern organisms descend from the 'last-common ancestor' as recently discussed by Orgel [50]. Furthermore, such conceptualisation is also in agreement with Crick's early 'frozen accident' theory [7] which states that 'all life evolved from a single organism (single closely interbreeding population)'. Woese et al. [29] argued that "all the previous stages… might be considered 'convergent' by virtue of having as the sole or main 'goal' the improvement of some features of information transfer". Thus, the emergence of the 'last common ancestor' with an established and standardised amino acid repertoire is the beginning or the 'alpha' point of the Darwino-Mendelian world as outlined in figure 2. This point as border between two worlds also represents a rather sharp transition from the postulated prebiotic and RNA world to the recent Darwino-Mendelian world of modern cells. It can be alternatively seen as a tunnel through which some actors from the old RNA world succeeded in entering the new Darwino-Mendelian world. Examples of this are tRNA molecules, which play a crucial role as passive adapters in protein synthesis (they sometimes even have other functions) or rRNA molecules which, together with ribosomal proteins, are tuned and optimised in the finest way possible to enable protein translation. The characteristics of the Darwino-Mendelian world are cells that possess a basic genetic code level which can be defined as a set of instructions stored in nucleic acid chemical form, which are transmitted into specific biological activities. This is enabled by the conserved (standardised) amino acid repertoire in ribosomal protein biosynthesis and by the presence of established proofreading and editing control mechanisms from DNA replication to translation [51]. For primitive cells from the postulated prebiotic and RNA world not having such a code, this was the 'omega' point since it marked their elimination or disappearance from the face of the earth. While characteristics of the Darwino-Mendelian world are observable, those of the extinct prebiotic and RNA world might possibly be deduced only indirectly.

The best of all possible codes?

Arguing that the code for the standard 20 canonical amino acids resulted at least in part from historical accident, Crick [7] wrote: "There is no reason to believe, however that the present code is the best possible, and it could have easily reached its present form by a sequence of happy accidents. …it may be frozen at local minimum which it has reached by a rather random path". In contrast to this hypothesis, we strongly believe that our code is the best of all possible codes [1]. But we are also well aware that we might easily be wrong. Specifically, if the evolutionary establishment of the code was a random historical event or an accidental 'chain of happy events', we will probably never be able to rationally explain it by causal analysis. All our conceptual, methodological and experimental efforts to deduce the causes for the existing structure might be futile. On the other hand, if the evolution of the code was a mechanistically specific process, where statistically random but favourable mutational events were successfully recruited and amplified by selection, then we may be able to grasp the principles behind its establishment. In this way, it should be possible to challenge the enigma of the origin of the code assuming that its genesis might not have been a random historical process, but rather a mechanistic one governed by universal laws of physics and chemistry in evolution.

Acknowledgements. We are very grateful to Sarah Teter for useful corrections, critical remarks and suggestions that resulted in substantial improvements to the initial form of this manuscript. Caroline Minks is acknowledged for assistance in figure preparations.

1 Budisa N., Minks C., Alefelder S., Wenger T., Dong F., Moroder L. et al. (1999) Toward the experimental codon reassignment in vivo: protein building with an expanded amino acid repertoire. FASEB J. **13:** 141–151
2 Schimmel P. (1996) Origin of genetic code: a needle in the haystack of tRNA sequences. Proc. Natl. Acad. Sci. USA **93:** 4521–4522
3 Alberti S. (1997) The origin of the genetic code and protein synthesis. J. Mol. Evol. **45:** 352–358

4 Amirnovin R. (1997) An analysis of the metabolic theory of the origin of the genetic code. J. Mol. Evol. **44:** 473–476
5 DiGuilio M. (1997) On the origin of the genetic code. J. Theor. Biol. **187:** 573–581
6 Yarus M. (1998) Amino acids as RNA ligands: a direct-RNA-template theory for the code's origin. J. Mol. Evol. **17:** 109–117
7 Crick F. H. C. (1968) The origin of the genetic code. J. Mol. Biol. **38:** 367–369
8 Popper K. R. (1972) Objective knowlegde: an evolutionary approach, Clarendon, Oxford
9 Wächtershäuser G. (1997) The origin of life and its methodological challenge. J. Theor. Biol. **187:** 483–494
10 Gelman S. H. (1998) Foldamers: a manifesto. Acc. Chem. Res. **31:** 173–180
11 Anfinsen C. B. (1971) Principles that govern the folding of protein chains. Science **181:** 223–230
12 Miranker A. D. and Dobson C. M. (1996) Collapse and cooperativity in protein folding. Curr. Opin. Struct. Biol. **6:** 31–42
13 Cornish V. W., Benson D. R., Altenbach C. A., Hideg K., Hubbell W. L. and Schultz P. G. (1994) Site-specific incorporation of biophysical probes into proteins. Proc. Natl. Acad. Sci. USA **91:** 2910–2914
14 Mendel D., Cornish V. W. and Schultz P. G. (1995) Site-directed mutagenesis with an expanded genetic code. Annu. Rev. Biophys. Biomol. Struct. **24:** 435–462
15 Ibba M. and Hennecke H. (1995) Relaxing the substrate specificity of an aminoacyl-tRNA synthetase allows in vitro and in vivo synthesis of proteins containing unnatural amino acids. FEBS Lett. **364:** 272–275
16 Miller S. L. (1986) Current status of the prebiotic synthesis of small molecules. In: Molecular Evolution of Life, pp. 5–11, Baltscheffsky H., Jörnvall H. and Riger R. (eds), Cambridge University Press, Cambridge
17 Anfinsen C. B. and Corley L. G. (1969) An active variant of staphylococcal nuclease containing norleucine in place of methionine. J. Biol. Chem. **244:** 5149–5152
18 Bogosian G., Violand B. N., Dorward-King E. J., Workman W. E., Jung P. E. and Kane J. F. (1989) Biosynthesis and incorporation into proteins of norleucine by *Eschericha coli*. J. Biol. Chem. **264:** 531–539
19 Gilles A., Marliere P., Rose T., Sarfati R., Longin R., Meier A. et al. (1988) Conservative replacement of methionine by norleucine in *Escherichia coli* adenylate kinase. J. Biol. Chem. **263:** 8204–8209
20 Osawa S., Jukes T. H., Watanabe K. and Muto A. (1992) Recent evidence for evolution of the genetic code. Microbiol. Rev. **56:** 229–264
21 Wong J. T. F. (1975) A co-evolution theory of the genetic code. Proc. Natl. Acad. Sci. USA **72:** 1909–1912
22 DiGuilio M. (1997) The origin of the genetic code. Trends Biochem. Sci. **22:** 49
23 Weber A. L. and Miller S. L. (1981) Reasons for the occurence of the twenty coded protein amino acids. J. Mol. Evol. **17:** 273–284
24 Wolfenden R. V., Cullis P. M. and Southgate C. C. F. (1979) Water, protein folding and the genetic code. Science **206:** 575–577
25 Dufton M. J. (1997) Genetic code synonym quotas and amino acid complexity: cutting the cost of proteins. J. Theor. Biol. **187:** 165–173
26 Xia X. and Li W. H. (1998) What amino acid properties affect protein evolution? J. Mol. Evol. **47:** 557–564
27 Dayhoff M. O. (1972) Atlas of protein sequence and structure, vol. 5, National Biomedical Research Foundation, Washington, DC
28 Sonneborn T. M. (1965) Degeneracy of the genetic code: extent, nature and genetic implications. In: Evolving Genes and Proteins, pp. 337–397, Bryson V. and Vogel H. J. (eds), Academic Press, New York
29 Woese C. R., Dugre D. H., Dugre S. A., Kondo M. and Saxinger W. C. (1966) On the fundamental nature and evolution of the genetic code. Cold Spring Lab. Symp. Quant. Biol. **31:** 723–736
30 Williams R. J. P. (1997) The natural selection of the chemical elements. Cell. Mol. Life Sci. **53:** 816–829
31 Crick F. H. C. (1958) On protein synthesis. Symp. Soc. Exp. Biol. **12:** 138–163
32 Blomberg C. (1997) On the appearance of function and organisation in the origin of life. J. Theor. Biol. **187:** 541–554
33 Thieffry D. and Sarkar S. (1998) Forty years under the central dogma. Trends Biochem. Sci. **23:** 312–316
34 Benner S. A., Ellington A. D. and Tauer A. (1989) Modern metabolism as a palimpsest of the RNA world. Proc. Natl. Acad. Sci. USA **86:** 7054–7058
35 Gilbert W. (1986) The RNA world. Nature **319:** 618
36 Orgel L. E. and Crick F. H. C. (1993) Anticipating an RNA world: some past speculations on the origin of life. Where are we today? FASEB J. **7:** 238–239
37 Jimenez-Sanchez A. (1995) On the origin and evolution of the genetic code. J. Mol. Evol. **41:** 712–716
38 Böck A., Forchhammer K., Heider J., Leinfelder W., Sawers G., Veprek B. et al. (1991) Selenocysteine: the 21st amino acid. Mol. Microbiol. **5:** 515–520
39 Cedergren R. and Miramontes P. (1996) The puzzling origin of the genetic code. Trends Biochem. Sci. **21:** 199–200
40 Garen A. (1968) Sense and nonsense in the genetic code. Science **160:** 149–159
41 Enghlhardt D. L., Webster R., Wilhelm R. and Zinder N. (1965) In vitro studies on the mechanism of suppression of a nonsense mutation. Proc. Natl. Acad. Sci. USA **54:** 1791–1797
42 Wong J. T. F. (1988) Evolution of the genetic code. Microbiol. Sci. **5:** 174–181
43 Santos M. A. S., Cheesman C., Costa V., Moradas-Ferreira P. and Tuite M. F. (1999) Selective advantages created by codon ambiguity allowed for the evolution of an alternative genetic code in *Candida* spp. Mol. Microbiol. **31:** 937–947
44 Suzuki T., Ueda T. and Watanabe K. (1997) The 'polysemous' codon – a codon with multiple amino acid assignment caused by dual specificity of tRNA identity. EMBO J. **16:** 1122–1134
45 Tuite M. F. and Santos M. A. S. (1996) Codon reassignment in *Candida* species: an evolutionary conundrum. Biochimie **78:** 993–999
46 Ibba M. and Hennecke H. (1994) Towards engineering proteins by site-directed incorporation in vivo of non-natural amino acids. Biotechnology **12:** 678–682
47 Woese C. R. (1965) On the evolution of the genetic code. Proc. Natl. Acad. Sci. USA **54:** 1546–1552
48 Budisa N., Steipe B., Demange P., Eckerskorn C., Kellerman J. and Huber R. (1995) High level biosynthetic substitution of methionine in proteins by its analogues 2-aminohexanoic acid, selenomethionine, telluromethionine and ethionine in *Escherichia coli*. Eur. J. Biochem. **320:** 788–796
49 Crick F. H. C. and Orgel L. E. (1973) Directed pansperima. Icarus **19:** 341–346
50 Orgel L. E. (1998) The origin of life – a review of facts and speculations. Trends Biochem. Sci. **23:** 491–495
51 Fersht A. R. (1981) Enzymatic editing mechanisms and the genetic code. Proc. R. Soc. Lond. B **212:** 351–379