INSTITUT PASTEUR

Research in Microbiology 160 (2009) 481–486

www.elsevier.com/locate/resmic

The origin of the genetic code and of the earliest oligopeptides

Edward N. Trifonov [a,b,*]

[a] *Genome Diversity Center, Institute of Evolution, University of Haifa, Mount Carmel, Haifa 31905, Israel*
[b] *Division of Functional Genomics and Proteomics, Faculty of Science, Masaryk University, Kamenice 5, Brno CZ-62500, Czech Republic*

Received 2 March 2009; accepted 27 May 2009
Available online 11 June 2009

Abstract

Reconstruction of the earliest proteins in the ancient binary alphabet [glycine family *G*, alanine family *A*] leads to repeats of *G* alternating with repeats of *A*. In addition, omnipresent motifs can be assembled in two of the earliest genes involved in energy supply, crucial for Life, i.e. ATP/GTP binding and ATPase activity. They are an almost perfect match to the alternating *G* and *A* and are complementary to each other.

Keywords: Early molecular evolution; Origin of the genetic code; Binary amino acid alphabet; Coding in both strands

1. Introduction

Enormous databases of nucleotide and amino acid sequences of modern organisms, a collective labor of already two generations of molecular biologists, may be viewed as a depository of invaluable experimental material, which awaits large scale theoretical work to unravel the loads of unknown meanings hosted by these sequences. The theory, in general, is lagging well behind the mounting data.

One exception is the rapidly developing theory of early molecular evolution [32,33,28] that started with the assumption that the sequences, active in the earliest eras of Life, are still around, some even, perhaps, unchanged. For example, the repeating sequences $(GCU)_n$ and $(GCC)_n$ of triplet expansion diseases may have survived 3.9 billion years of Life exactly because of their exceptional ability to expand [34]. Below, I briefly review the theory and outline the latest developments.

2. Evolutionary chart of codons

The speculation that ancient sequence motifs have been conserved, combined with a consensus of now over 100 different opinions about the temporal order of appearance of the amino acids on the evolutionary scene, resulted in the first, though still largely speculative reconstruction of the evolution of the triplet code [30,31]. Indeed, the evolutionary chart of codons displays several highly relevant features. The temporal order of amino acids on which the chart is built turned out to correspond with the thermostability of the respective codon/anticodon pairs (Fig. 1). This does not necessarily mean a hot start of Life, since the stability of molecular structures is of natural importance at every stage of evolution. The consecutive codons entering the evolving chart come as complementary pairs, confirming, again, the role of stability. The amino acids of the imitation experiments of Miller [21] appear first in the chart. The newcomers (codon capture amino acids) [23] could be incorporated in the codon table only by reassignment of already assigned codons. E.g. AUG, one of the early codons for isoleucine, had been reassigned to methionine. Indeed, these newcomers are at the end of the ''chronology''. But perhaps the most striking feature is that the codon assignments in the chart are practically all the same as in the extant codon table. This would suggest that the codons, most of them, did not change their identity since the times when they were first introduced in the system. They are still around.

3. Ancient binary alphabet of protein sequences

It also turned out that every new codon in the chart is a point-mutated version of one or several codons already engaged. These

* Tel.: +972 4 828 8096; fax: +972 4 824 6554.
E-mail address: trifonov@research.haifa.ac.il

0923-2508/$ - see front matter © 2009 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.resmic.2009.05.004

→ **chronological order of amino acids appearance** →

1	2	3	4	5	6	7	8	9	10	11	12	13	14
Gly	Ala	Asp	Val	Ser	Pro	Glu	Leu	Thr	Arg	TRM	Ile	Gln	TRM	Asn	Lys
GGC	GCC	GAC	GUC	UCC	CCC	GAG	CUC	ACC	CGC	UGC	AUC	CAC	UAC	AAC	AAG

→ **descending thermostability of triplet pairs** →

Newcomers (codon capturers):

15	16	17	18	19	20	21	22
His	Phe	Cys	Met	Tyr	Trp	Sec	Pyl
CAC	UUC	UGC	AUG	UAC	UGG	UGA	UAG

Fig. 1. Simplified version of the original codon evolution chart [31]. Sec (selenocysteine) and Pyl (pyrrolysine) are recently discovered additional amino acids [5,29]. Complementary pairs of codons are underlined. TRM: terminators of translation.

are either transitions in second position (GGC for glycine to GAC for aspartic acid in the first step), or changes in degenerate third positions, simultaneously turning to changes in first positions of the respective complementary codons. This holds for the remaining 30 steps, until completion of all 64 assignments, before the codon capture stage [31]. In other words, *all* purines in the second codon positions are descendants of **G** in G**G**C (glycine), and *all* codons with pyrimidines in the middle descend from G**C**C (alanine). Respectively, amino acids are split into two families, i.e. the Gly-family (*G*): C, D, E, G, H, K, N, Q, R, S, W and the Ala-family (*A*): A, F, I, L, M, P, S, T, V. It appears that during the evolution of the codon table the two families of amino acids (codons) evolved separately, without mixing between the two lineages. The protein sequences of that time had glycine-type amino acids in place of the original G, and alanine-type amino acids in place of the original A.

The striking discovery is that this rule still holds for modern amino acid substitutions, that is, the *A*-type residues predominantly change to *A* type, and changing *G*-type residues stay within the *G*-family. This results in splitting standard amino acid substitution matrices into two independent boxes [11,32]. The conservation may be partially explained by the fact that the *A*-family is largely hydrophobic, while amino acids of *G*-family are mostly hydrophilic. The nature of the conservation appears, however, to be deeper than just a balance between hydrophilic and hydrophobic residues. Indeed, even in hypervariable regions, the loyalty to the *A*- or *G*-family is kept. For example, 12 of the 13 different substitutions of R (AGG) in the epitope TLYCVHQR of HIV-1 keep the middle purine of the replacement codons unchanged [14]. Whatever the reason, due to this conservation rule, those original glycines and alanines that have been present in the very first ancestral versions of modern sequences could well be ''still around'', in the form of *A*-type and *G*-type residues. The binary presentations of modern sequences could correspond to earlier (or even the earliest) versions of the sequences, and this insight is the basis for the reconstruction described below.

4. The sizes of the earliest mini-genes and oligopeptides

With the pair of the very first, and complementary, codons GGC and GCC, and the amino acids that correspond to them, i.e. glycine and alanine, the emerging translation system could only synthesize oligopeptides GGGGG… (G_n) and AAAAA… (A_n), encoded by the two complementary strands $(GGC)_n$ and $(GCC)_n$. Presumably, the process of expansion of the respective codons, may have very much resembled the process as it is known from contemporary triplet expansion diseases [16,22]. The oligopeptides must have been rather short, to stay within their solubility limits. Presumably, the primitive mini-proteins for some time accumulated changes, each keeping with the conservation of their binary character. The G_n and A_n oligopeptides turned to G_n and A_n (note *italics* which indicate the presence of glycine-like resp. alanine-like amino acids, besides Gly resp. Ala). The eventual (multiple) shuffling/fusion of respective mini-genes would have led to the appearance of more complex mosaic proteins consisting of $G_nA_nG_nA_n$… As *G* residues are largely hydrophilic, while *A* residues mainly hydrophobic, the alternation of runs of *G* and *A* rather than their respective longer runs should have been preferable, because long runs of *G* would not be good for folding, while long runs of *A* would increase the risk of aggregation. That brings us to the following prediction: if modern sequences still keep remnants of the hypothetical ancient alternation of the short runs of *G* and *A*, then the analysis of the preferred distances between *G* and *A* residues should reveal the unknown preferential size n of the ancient mini-proteins. Such calculations on a large ensemble of prokaryotic protein sequences have been carried out, and the preferential size of approximately 7 amino acid residues has been found [35]. Hence, the size of the respective mini-genes would have been approximately 21 bases. An independent estimate of the size of the earliest mini-genes was derived from the study of the earliest hairpins in mRNA still detectable in modern sequences and yielded an identical size of seven triplets [11,36].

5. The earliest, omnipresent sequence motifs

Another, independent route of extrapolation to the protein's past is the derivation from modern protein sequence databases of so-called omnipresent motifs [26–28]. These are believed to be literally ''still around'' from the time of the LCA. This is well supported, indeed, by the higher content of chronologically earlier amino acids in the omnipresent motifs [26]. The list of 27