Computational and Structural Biotechnology Journal xxx (2018) xxx–xxx

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Protein Sequences Recapitulate Genetic Code Evolution

Hervé Seligmann [a,b,*]

[a] Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes, UMR MEPHI, Aix-Marseille Université, IRD, Assistance Publique-Hôpitaux de Marseille, Institut Hospitalo-Universitaire Méditerranée-Infection, 19-21 boulevard Jean Moulin, 13005 Marseille, France
[b] The National Natural History Collections, The Hebrew University of Jerusalem, 9190401 Jerusalem, Israel

ARTICLE INFO

Article history:
Received 17 January 2018
Received in revised form 14 May 2018
Accepted 17 May 2018
Available online xxxx

Keywords:
Codon directional asymmetry
Genetic code structure
Gene punctuation
Secondary structure formation
Antiparallel betasheets
tRNA synthetases

ABSTRACT

Several hypotheses predict ranks of amino acid assignments to genetic code's codons. Analyses here show that average positions of amino acid species in proteins correspond to assignment ranks, in particular as predicted by Juke's neutral mutation hypothesis for codon assignments. In all tested protein groups, including co- and post-translationally folding proteins, 'recent' amino acids are on average closer to gene 5′ extremities than 'ancient' ones. Analyses of pairwise residue contact energies matrices suggest that early amino acids stereochemically selected late ones that stablilize residue interactions within protein cores, presumably producing 5′-late-to-3′-early amino acid protein sequence gradients. The gradient might reduce protein misfolding, also after mutations, extending principles of neutral mutations to protein folding. Presumably, in self-perpetuating and self-correcting systems like the genetic code, initial conditions produce similarities between evolution of the process (the genetic code) and 'ontogeny' of resulting structures (here proteins), producing apparent teleonomy between process and product.

1. Introduction

The structure of biological molecules includes imprints of ancient evolution at life's dawn. For example, comparisons between protein and RNA structures suggest affinities between viruses and hypothetical bacterial-like cellular ancestors (as described for protein structural families, [61], Nasir et al. 2017; and for RNA secondary structures, [101]). The ribosome's structure testifies to even more ancient events: ribosomal protein amino acids interact preferentially with ribosomal RNA trinucleotides that correspond to that amino acid's assigned anticodon(s) according to the standard genetic code [42]. This striking fossilization of the process that determined some codon-amino acid assignments in the ribosome's structure confirms that at least some codon-amino acid assignments result from stereochemical affinities between RNA and amino acids [118–120].

1.1. Steps in the Evolution of the Genetic Code and the Translational Apparatus

Johnson and Wang [42] suggest that several processes structured the genetic code, meaning determined codon-amino acid assignments. Indeed, structurally simple amino acids tend to associate with rRNA nucleotide triplets corresponding to their genetic code codon assignments, while complex amino acids associate with their anticodons (stereochemical complexity according to Dufton [19]). This indicates a primary phase of direct codon-amino acid contact, and secondarily evolution of mRNA, anticodon and from there the proto-tRNA [97].

Several hypotheses predict the order of inclusion of amino acids in the genetic code. These orders tend to be consensual among hypotheses, and usually consider that structurally simple amino acids were included early, and complex one's late [36,56,113,114]. Considering 40 hypotheses about the inclusion order of amino acids in the genetic code reviewed by Trifonov [114], the strength of association between amino acids and their anticodons in rRNA (data from [42], therein figure 1) increases with their order of inclusion in the genetic code. This correlation is strongest with the inclusion order predicted by the tRNA-Urgen hypothesis ([20,21], here Fig. 1).

1.2. Imprints of the Genetic Code Evolution in Modern Protein Sequences

Above observations about the ribosome's structure suggest that imprints of the genetic code's evolution might remain also in protein structures. Here I test the hypothesis that the inclusion order of amino acids in the genetic code correlates with average positions of amino acids in proteins.

This working hypothesis is derived from principles of the biogenetic law or Meckel-Serres law, formulated by Haeckel as 'ontogeny recapitulates phylogeny' [50]. As in that evo-devo hypothesis, the history of a

* Corresponding author at: Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes, UMR MEPHI, Aix-Marseille Université, IRD, Assistance Publique-Hôpitaux de Marseille, Institut Hospitalo-Universitaire Méditerranée-Infection, 19-21 boulevard Jean Moulin, 13005 Marseille, France.
E-mail address: varanuseremius@gmail.com.

https://doi.org/10.1016/j.csbj.2018.05.001

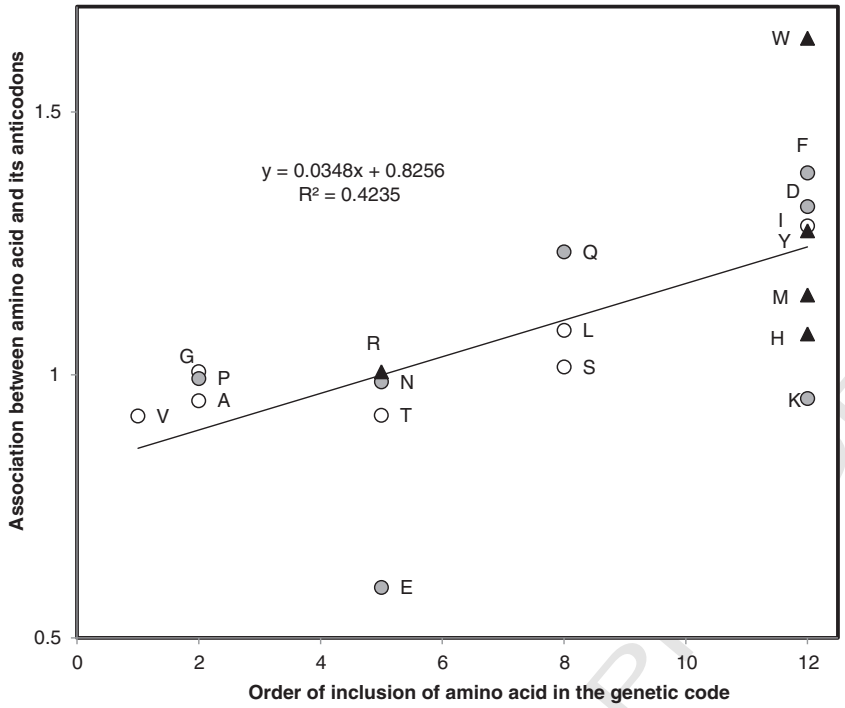

Fig. 1. Strength of association of amino acids with ribosomal RNA triplets corresponding to their anticodons in the ribosome's structure, based on contacts between proteins and rRNAs in crystallized ribosomes [42], as a function of the order of inclusion of amino acids in the genetic code according to the tRNA Urgen hypothesis which has only 12 ranks (all 'late' amino acids get rank 12, [20,21,114]). Association strengths are ratios between observed numbers of amino acid contacts with anticodon triplets and expected random contacts, after data in Fig. 1 of Johnson and Wang [42]. Amino acids are classified according to three levels of structural complexity [19]: low (hollow circles), intermediate (gray circles) and high (filled triangles). The latter group would include cysteine, for which the ribosome's structure does not include contacts between residues and rRNA.

process might be imprinted in the structures produced by that process [44]. The reason to expect this apparent teleonomy frequently observed in biological processes is that self-organizing and self-perpetuating processes such as the genetic code are by definition self-correcting [49]. Structures resulting from early historical initial conditions are frequently conserved or recovered by resulting processes and structures. Hence historical/evolutionary processes would be conserved as imprints in modern structures because self-corrections towards the least error-prone structures conserve or recover the same initial structures/constraints. Accordingly, protein structures should also reflect the evolution of the genetic code.

1.3. Evolution for Coding Versatility

The genetic code evolved to include more complex amino acids, which are also more diverse in physicochemical terms than randomly selected potential amino acids [31,40,67]. Directional evolution of genetically coded amino acids towards diversification and greater complexity corresponds to the most recently integrated amino acids in the genetic code, selenocysteine and pyrrolysine [122], complex amino acids with peculiar properties (i.e. selenocysteine includes a selenium atom (doesn't occur in other natural amino acids) where cysteine has a sulfur atom (occurs only in one other natural amino acid)).

This suggests constraints towards increasing the genetic code's versatility for diverse types of specialized proteins. The evolutionary need to develop proteins with new functions would have driven inclusion of complex and physicochemically outstanding amino acids. Presumably, RNA secondary structure-based punctuation signals initiated translation before the genetic code assigned start codons [22,70]. The presumably late assignment of methionine, a structurally complex and 'special' amino acid, to initiation codon(s) would suggest that 'late' amino acids would tend to be coded close to gene 5′ extremities, and ancient amino acids closer to their 3′ extremities.

The working hypothesis expects that the genetic code evolved to include complex amino acids to stabilize protein structures, beyond increasing the diversity of potentially coded proteins. Predictions are tested versus lack of bias in average locations of amino acid species in genes/proteins.

2. Materials and Methods

Analyses focus on eight groups of proteins, seven from the *Escherichia coli* proteome (downloaded from GenBank entry NC_002695). The two groups consist of all tRNA synthetases of *Escherichia coli* (as used previously, [92]), subdivided in tRNA synthetase class I and class II (10 amino acid species per class, 10 proteins for class I and 13 for class II (including both subunits alpha and beta for tRNA synthetases Phe and Gly)). Class II tRNA synthetases are completed by the tRNA synthetase for pyrrolysine found in some archaea [69,109].The tRNA synthetases are chosen because these conserved proteins essential to translation occur in all organisms [66,74], including some viruses (Megavirales, [1,2,71,75]), and because within each class they are related among each other, facilitating comparative analyses [30,60,65] The two tRNA synthetase classes differ in their structures: class I are usually monomeric proteins with a Rossman fold catalytic domain. Class II tRNA synthetases are usually di- or multimeric with an anti-parallel betasheet fold flanked by alpha helices.

Other protein groups from *E. coli*'s proteome are: 67 ribosomal proteins, 36 polymerases, 119 membrane-linked proteins. Using predictions on *E. coli* protein folding modes [15], a group of 63 proteins folding cotranslationally is compared with another group of 101 proteins folding post-translationally. These were chosen from a longer protein list because predicted folding mode in these proteins does not vary with specific conditions as computationally tested by Ciryam et al. [15]. Identities and sequences of the 408 analyzed *E. coli* proteins are available in the supplementary data. The *E. coli* proteome is translated from