BioSystems 129 (2015) 36–43

Contents lists available at ScienceDirect

BioSystems

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

Phylogeny of genetic codes and punctuation codes within genetic codes

Hervé Seligmann

Unité de Recherche sur les Maladies Infectieuses et Tropicales Émergentes, Faculté de Médecine, URMITE CNRS-IRD 198 UMER 6236, Université de la Méditerranée, Marseille, France

ARTICLE INFO

Article history:
Received 10 October 2014
Received in revised form 2 January 2015
Accepted 14 January 2015
Available online 16 January 2015

Keywords:
Initiation codon
Termination codon
Natural circular code
Ribosomal frameshift
Coding frame retrieval
Circular code
Off frame stops
Principal component analyses

ABSTRACT

Punctuation codons (starts, stops) delimit genes, reflect translation apparatus properties. Most codon reassignments involve punctuation. Here two complementary approaches classify natural genetic codes: (A) properties of amino acids assigned to codons (classical phylogeny), coding stops as X (A1, antitermination/suppressor tRNAs insert unknown residues), or as gaps (A2, no translation, classical stop); and (B) considering only punctuation status (start, stop and other codons coded as −1, 0 and 1 (B1); 0, −1 and 1 (B2, reflects ribosomal translational dynamics); and 1, −1, and 0 (B3, starts/stops as opposites)). All methods separate most mitochondrial codes from most nuclear codes; Gracilibacteria consistently cluster with metazoan mitochondria; mitochondria co-hosted with chloroplasts cluster with nuclear codes. Method A1 clusters the euplotid nuclear code with metazoan mitochondria; A2 separates euplotids from mitochondria. Firmicute bacteria *Mycoplasma*/*Spiroplasma* and Protozoan (and lower metazoan) mitochondria share codon-amino acid assignments. A1 clusters them with mitochondria, they cluster with the standard genetic code under A2: constraints on amino acid ambiguity versus punctuation-signaling produced the mitochondrial versus bacterial versions of this genetic code. Punctuation analysis B2 converges best with classical phylogenetic analyses, stressing the need for a unified theory of genetic code punctuation accounting for ribosomal constraints.

1. Introduction

Alphabetical languages use two codes, letters coding for sounds, and other symbols for punctuation. The punctuation code can be further subdivided into an inter-sentence punctuation code (including structure between sub-sentences) that delimits sentences (and subsentences) by indicating their boundaries, and symbols indicating limits between words, word boundary coding.

In genetic codes, all codons code for amino acids (these compare to letters in this context), besides stop codons that terminate translation (unless antiterminator (or suppressor) tRNAs (especially in mitochondria, Seligmann, 2010a) are active (Faure et al., 2011; Seligmann, 2011a, 2012a,b)). Initiation codons have both roles, they code for an amino acid and initiate translation. Stops and starts correspond to inter-sentence boundary punctuation symbols in human written languages.

Genetic codes include a further punctuation code, which matches the inter-word boundary code of written human languages, the circular code (Arquès and Michel, 1996). The mathematical properties of this group of 20 codons enable coding frame retrieval, hence indicates codon boundaries within the gene's boundaries for any sequence window longer than 12 nucleotides (Michel, 2012, therein Fig. 2). This defines the two major components of the genetic code's punctuation code theory, an intra- and an inter-gene boundary component. Pro- and eukaryotes have the same circular code, the latter emerged when chiral preference for L-amino acids evolved in proto-organismic systems (Michel and Seligmann, 2014). Overall, plastids and viruses have a very similar or identical circular code as eukaryotes and prokaryotes (Michel, 2015). Trinucleotides with sequence identical to stop codons also play a role in the within gene-between codon boundary context (Seligmann and Pollock, 2003, 2004; Seligmann, 2007; Itzkovitz and Alon, 2007; Seligmann, 2010b, 2012c; Tse et al., 2010). At this stage, the existence of a separate circular code in mitochondria (Arquès and Michel, 1997) remains uncertain (Michel and Seligmann, 2014). The circular code can be used to detect/confirm overlapping genes (Ahmed et al., 2007, 2010; Ahmed and Michel, 2011; Seligmann, 2012b,d,e; Seligmann, 2013a,b,c), including such coded by codons expanded by a fourth, silent nucleotide, called tetracodons (Seligmann, 2012e, 2013d, 2014; Seligmann and Labra, 2013).

Unlike the circular code for within-gene punctuation, the punctuation code for inter-gene boundaries varies between

E-mail address: varanuseremius@gmail.com (H. Seligmann).

http://dx.doi.org/10.1016/j.biosystems.2015.01.003
0303-2647/

taxonomic groups. Here, analyses of nineteen genetic codes from different taxonomic groups (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils,Elzanowski and Ostell, 2013, completed, for the missing *Blepharisma* nuclear genetic code (Lozupone et al., 2001, but see Eliseev et al., 2011) found at http://www.bioinformatics.org/jambw/2/3/TranslationTables.html) reconstruct phylogenetic relationships between these codes. Analyzes follow two main approaches. Approach A uses classical phylogenetic reconstruction, based on alignments between homologous codons (Table 1), using their amino acid assignments, and usual amino acid replacement matrices as a model of evolution. Approach B focuses on variation between codes in codons assigned to gene boundary roles, hence as initiation and termination codons. This information is unaccounted by approach A. Most codon reassignments involve gene boundary punctuation (Osawa and Jukes, 1989; Johnson, 2010; Johnson et al., 2011), suggesting that accounting for the initiation codon status of a codon may be useful in the context of the study of the evolution of genetic codes.

2. Results and discussion

2.1. Codon-amino acid reassignments

Table 1 presents the codon-amino acid assignments for all 64 codons in the 19 different recognized genetic codes (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils, Elzanowski and Ostell, 2013 and http://www.bioinformatics.org/jambw/2/3/TranslationTables.html). These data are considered as an alignment that is analyzed by the ClustalW2-Phylogeny software of EMBL-EBI (http://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=-clustalw2_phylogeny-I20140920-152606-0768-96439267-oy, Larkin et al., 2007) to build a phylogeny, using distance correction for multiple replacements, not excluding gaps, and using the neighbor-joining clustering method. The software uses the Gonnet transition matrix as evolutionary model. Note that this differs from the approach of mapping different genetic codes on a phylogeny derived from independent data (i.e., Telford et al., 2000; Sengupta et al., 2007; Cocquyt et al., 2010; Mateus et al., 2013; Fucikova et al., 2014; Muehlhausen and Kollmar, 2014; including genetic code coevolution for virus and host, Taylor et al., 2013). Here the phylogeny is derived from the genetic codes themselves, based on the variation in codon-amino acid assignments.

Including stop codons in this analysis is not straightforward as these do not code for any amino acid. Two alternative methods are used to solve this problem.

The first approach (A1) inserts 'X' for stops. This implies that any amino acid could be inserted for a stop. This approach seems somehow unrealistic, but some evidence indicates that the antisense sequence of mitochondrial tRNAs (Seligmann, 2010a), whose anticodon would match stop codons, might translate stops, inserting various amino acids (Seligmann, 2011a, 2012a,b; Faure et al., 2011). The predicted specificity of antisense tRNAs for their cognate amino acid is lower than for their sense counterpart (Seligmann, 2010c, 2011b), potentially justifying 'X' that suggests insertion of any amino acid. In addition, 'X' corresponds to assuming ambiguity in the assignment of the stop codon. Ambiguity during codon assignment evolution is an accepted concept (Schultz and Yarus, 1994; Sengupta and Higgs, 2005), especially for mitochondrial genetic codes (Knight et al., 2001; Sengupta et al., 2007).

Further physiological information suggests stop codon ambiguity in mitochondria. For example, AGA and AGG codons, which usually code for arginine in nuclear genetic codes, seem involved in termination signaling in vertebrate mitochondria (Temperley et al., 2010; Chrzanowska-Lightowlers et al., 2011; Lind et al., 2013). Mitochondrial import of cytosolic tRNAs (Schneider, 2011; Salinas et al., 2012) with anticodons matching these two codons, and loaded with arginine, can alleviate the termination role, creating ambiguity in their codon assignment.

The second approach (A2) inserts gaps '–' for stops. Gaps indicate lack of residue insertion, representing an actual stop. Hence A1 stresses that residues are inserted for stops, perhaps in the context of translation of alternative reading frames, or of codon ambiguity during genetic code evolution, while A2 does not enable this, and considers stops as unambiguous translational termination signals. Note that, as suggested by an anonymous reviewer, both A1 and A2 are first, preliminary steps delimitating different approaches in integrating stops into the reconstruction of genetic code phylogenies. This is discussed in an ulterior section.

2.1.1. Stops as 'X'

Fig. 1 plots the topology of the phylogeny obtained according to A1. This analysis yields three main branches. One branch consists solely of the "standard" genetic code, identical with the bacterial, archaeal and plant plastid genetic code. The second branch includes two among three ciliate nuclear codes, the mitochondrial code of *Scenedesmus obliquus* and the chlorophycean mitochondrial genetic code, two organisms co-hosting chloroplasts along mitochondria, and the mitochondrial genetic code of *Thraustochytrium* (Eukaryota: Stramenopiles). The third branch includes the bulk of (mainly metazoan) mitochondrial genetic codes, but also three nuclear genetic codes (from yeast, euplotids (ciliates), and Gracilibacteria (and candidate division SR1)). It also includes the protozoan mitochondrial genetic code, which is identical to the nuclear code used in the bacterial Firmicute group formed by *Mycoplasma/Spiroplasma*. This phylogeny has some details that are congruent with the known phylogeny of these taxa: mitochondrial codes co-hosted with chloroplasts form a monophyletic group, the vertebrata and ascidian mitochondrial genetic codes are monophyletic, and all invertebrate mitochondrial codes are monophyletic. The positions of the yeast mitochondrial genetic code and of Pterobranchia mitochondrial genetic codes are (relatively) minor incongruencies (Pterobranchia are expected closer to chordata (ascidians and vertebrates) than other metazoans), but the inclusion in branch 3 of three nuclear genetic codes is a major indication that the phylogeny is far from accurate, and probably reflects convergences.

2.1.2. Stops as gaps

Fig. 2 plots the topology of the phylogeny obtained according to A2, coding stops as gaps. This phylogeny also includes three main branches. On the first main branch, the standard and bacterial genetic codes form a monophyletic group with the mitochondrial code of *Thraustochytrium*, and the protozoan mitochondrial + *Mycoplasma/Spiroplasma* genetic code (genetic code 4). Hence the presence of the mitochondrial code of *Thraustochytrium* remains incongruent with the known tree of life. The most dramatic change in Fig. 2 as compared to Fig. 1 is the presence in this cluster of genetic code 4. Coding stops as gaps clusters this genetic code with other bacteria (A2), while coding stops as unknown/ambiguous residues (A1) clusters this code with mitochondria. Hence this suggests that the convergence between the code detected in these two taxa is due to different constraints: one relates to the translation of stops by unknown amino acids, perhaps due to an ambiguous period in the evolution of the mitochondrial code, or due to frequent translation of stops in mitochondria; the other stresses the role of stops in translation termination, clustering code 4 with bacterial and standard genetic codes. Hence these results indicate that A1 and A2 imply both realistic evolutionary scenarios of codon-amino acid reassignment, matching different cases, as previously suggested (Sengupta et al., 2007).