BioSystems 164 (2018) 16–25

Contents lists available at ScienceDirect

BioSystems

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

Review article

On universal coding events in protein biogenesis

Vladimir Kubyshkin[a], Carlos G. Acevedo-Rocha[b], Nediljko Budisa[a,*]

[a] Berlin Institute of Technology/TU Berlin, Department of Chemistry, Biocatalysis Group, Müller-Breslau-Straße 10, D-10623 Berlin, Germany
[b] Biosyntia ApS, 2100 Copenhagen, Denmark

ARTICLE INFO

Article history:
Received 10 July 2017
Received in revised form 2 October 2017
Accepted 3 October 2017
Available online 10 October 2017

Keywords:
Biocontainment
Code biology
Engineering and expansion
Genetic isolation
Horizontal gene transfer
Minimal genome
Operational RNA code
Protein folding
Ribosomal translation
tRNA
Synthetic biology
Xenobiology

ABSTRACT

The complete ribosomal protein synthesis cycle and codon-amino acids associations are universally preserved in all life taxa on Earth. This process is accompanied by a set of hierarchically organized recognition and controlling events at different complexity levels. It starts with amino acid activation by aminoacyl tRNA synthetases (aaRS) followed by matching with the acceptor units of their cognate tRNAs ("operational RNA code") and ribosomal codon-anticodon pairing of messenger RNA ("triplet code"). However, this codon-anticodon matching is possible only when protein translation machinery (translation factors, ribosome) accepts an esterified amino acid. This capacity ("charge code") correlates mainly with the amino acid nature and the identity elements in the tRNA 3D structure. A fourth potential "folding code" (also referred as "stereochemical code") between the translation dynamics, sequence composition and folding of the resulting protein can also be defined in the frame of the 'Anfinsen dogma' followed by post-translational modifications. All these coding events as well as the basic chemistry of life are deemed invariant across biological taxa due to the horizontal gene transfer (HGT) making the 'universal genetic code' the 'lingua franca' of life of earth. When cells (or organelles) are prevented from transmitting genetic information (i.e., HGT) the deviations in the above-mentioned coding events become inevitable. A better understanding of these codes, in particular the mechanisms of their conservation in the context of HGT could provide a guide for the experimental engineering[1] of the ribosomal protein biosynthesis machinery. This is highly relevant, among others, in attempts to create synthetic life forms in genetic isolation by using tailored "minimal genomes" and may explain the necessity for multiple coding evens in nature.

Contents

* Corresponding author at: Berlin Institute of Technology/TU Berlin, Department of Chemistry, Biocatalysis Group, Müller-Breslau-Straße 10, D-10623 Berlin, Germany.
E-mail address: nediljko.budisa@tu-berlin.de (N. Budisa).
[1] *"What I cannot create – I do not understand"* – Richard Feynman

https://doi.org/10.1016/j.biosystems.2017.10.004

Numerum combinationum in terminis etiam numero finitis esse infinitum; si rite omnia expendantur. In ipso immenso combinatorium numero in immensum plures esse combinationes inordinatas, quam ordinatas.[2]

R. J. Boscovich (1763) Theoria philosophie naturalis redacta ad unicam legem virium in natura existenttium. p. 254–256. Venetiis. Ex Typographia Remondiana.

1. Introduction – the coding levels in the ribosomal cycle of protein biosynthesis

Coding events are ubiquitous in living systems. For example, it is believed that prokaryotes developed the genetic code and signal processing codes, while eukaryotes further developed the histone code, splicing codes, tubulin code, apoptosis code, compartment codes and many others (Barbieri, 2015a). A code refers to the set of rules establishing a correspondence between the objects of two independent worlds (Barbieri, 2016). For example, the genetic code describes the rules between the world of nucleic acids (codons) and that of amino acids to build proteins. The first code that appeared in living systems is the genetic code, which is well documented by the existence of genes and genomes, which encode sequences of cellular proteins that can operate a genetic program. Living systems can thus create, modify and amplify information, but this capacity requires the existence of codes connecting objects from different levels of molecular self-organization.

Life can thus be defined as a process shaped by two forms of causality: (i) deterministic laws of chemistry and physics (Natural laws) and (ii) a genetic program that determines all biological activities and phenomena (Budisa, 2012). Importantly, the genetic code is compatible with the laws of physics and chemistry – but to which extent it is defined by them is rather a philosophic issue. This program generally operates such that nucleic acids encode information and proteins execute it.[3] The syntax of the genetic program is reflected in a precise sequence of bases in nucleic acids.

The correspondence between the sequences of nucleotides and amino acids is mechanistically well-established and generally described in biochemistry as “molecular recognition”. However, the development of the genetic code can be also studied by the establishment of various sub-coding events that brought different objects of the organic worlds into coordination. These events, which are discussed in more detail in the sections below, include the operational RNA code, charging code and triplet code as well as the protein folding code and the post-translational modifications (PTM) code (Fig. 1).

The fact that the informational flow between nucleic acids and proteins is unidirectional (i.e., a polynucleotide is never determined by a protein sequence) is often termed as the “central dogma of molecular biology” (Crick, 1970). This flow of information operates under the regime of invariant basic chemical organization and has not substantially changed since the last universal common ancestor (LUCA) (Woese, 1998).

Notably, the most conserved molecular machine in cellular life is the ribosome, which is the protein synthesis apparatus that translates genetic sequences into proteins, thus enabling the execution of all essential metabolic processes encoded in the genetic program. Throughout millions of years of natural selection and evolution, these processes were shaped by the most diverse environmental niches with extreme conditions of temperature, pressure, acidity, salinity, light intensity, etc (Weiss et al., 2016). Despite this diversity, a fundamental question is whether it would be possible to create experimentally an organism with a different biochemistry as well as genetic code different from life on our planet (Acevedo-Rocha and Schulze-Makuch, 2015). For example, non-natural building blocks, i.e. organic molecules containing boron, fluorine, silicon of other elements rarely used in earth biochemistry might be potentially advantageous for living cells surviving in synthetic media or artificial environments (Kubyshkin and Budisa, 2017). In fact, this is one of the goals of the research field of Xenobiology (XB), which aims to create artificial biological diversity by changing the chemical make-up of living cells in order to understand life, its origins and evolution (Budisa, 2014). Another objective of XB is to engineer the genetic code beyond the 20 α-amino acids in order to expand the functionality of proteins for applied research such as biomaterials or biocatalysis (Agostini et al., 2017).

The structure and function of the genetic code is determined by its amino acid repertoire: no organism either natural or artificial operates by using more than 20 (+2) amino acids (Wiltschi and Budisa, 2007). However, a specific challenge for XB is to engineer organisms that encode not only 20 (+2) natural amino acids, but also more synthetic ones. This goal requires a deeper understanding of the associated chemical and biochemical factors as well as the coding principles underlying the ribosomal protein biosynthesis. Thus, it will be beneficial for XB to understand the different sub-coding events that gave rise to the actual genetic code in order to define experimental efforts that could allow for design of synthetic organisms endowed with new useful biochemistries.

In this context, we are motivated by the idea of studying the genetic code at different complexity levels as well as hierarchies for controlling and recognition of events in the genetic information flow. Our aim at this point is to understand the set of boundaries, which restrict the canonical repertoire to the 20 (+2) amino acids. Furthermore, these considerations should correct what is meant by the term “genetic code”, while offering important indications of how this can be constructed. Herein, we will provide an overview of the various recognition or coding steps in course of protein translation, the universal process which leads to interpretation and execution of the genetic information.

2. Transfer RNA recognition

During the protein translation process, the polymerization reaction of α-amino acids monomers takes place at the ribosome. Beforehand, a specific selection of the amino acid side chains is carried out in a two-step reaction by a class of enzymes known as aminoacyl-tRNA synthetase (aaRS) (Hoagland et al., 1957). The aaRSs constitute a family of 20 cellular enzymes (divided in two distinct classes of 10 members each based on the presence of exclusive sets of sequence motifs) that operate at the interface between nucleic acids, proteinsand metabolites (Francklyn, 2003). In the cellular milieu, these enzymes must recognize the correct amino acid and tRNA(s) from a large cellular pool of similar molecules. The pairing of amino acids and tRNAs defines the associations of codon(s): amino acid, making aaRSs essential molecules in maintaining the

[2] Careful examination reveals the infinite number of combinations between the finite numbers of terms. In this large number of combinations there are far more without order than with it.

[3] It is generally difficult to compare fundamental features of computer programming with the flow of the genetic information as there are parts which cannot be compared directly. Programmers define a program (written by using a programming language) as a well-defined abstract model that serves as a prescription (not as a description!). The program is compiled through a systematic process that transforms it by using a hierarchy of increasingly detailed models, preserving thereby the prescribed behavior. The executive part (executable) is the lowest level specification, which has a direct physical implementation. In the context of the flow of the genetic information (DNA → RNA → Protein) the DNA (i.e. gene sequence) can be described as a program (prescribed information). The ‘compilation’ of this genetic information (which includes the use of a different RNA transcripts and associated events, see Fig. 1) results in functional proteins (i.e., proteins are the executors of the genetic program).