BioSystems 155 (2017) 10–19

Contents lists available at ScienceDirect

BioSystems

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

A binary representation of the genetic code

Louis R. Nemzer

Department of Chemistry and Physics, Halmos College of Natural Sciences and Oceanography, Nova Southeastern University, Davie, FL, USA

ARTICLE INFO

Article history:
Received 16 October 2016
Received in revised form 3 March 2017
Accepted 6 March 2017
Available online 12 March 2017

Keywords:
DNA
Genetic code
Binary
Mutations
Evolution

ABSTRACT

This article introduces a novel binary representation of the canonical genetic code based on both the structural similarities of the nucleotides, as well as the physicochemical properties of the encoded amino acids. Each of the four mRNA bases is assigned a unique 2-bit identifier, so that the 64 triplet codons are each indexed by a 6-bit label. The ordering of the bits reflects the hierarchical organization manifested by the DNA replication/repair and tRNA translation systems. In this system, transition and transversion mutations are naturally expressed as binary operations, and the severities of the different point mutations can be analyzed. Using a principal component analysis, it is shown that the physicochemical properties of amino acids related to protein folding also correlate with certain bit positions of their respective labels. Thus, the likelihood for a point mutation to be conservative, and less likely to cause a change in protein functionality, can be estimated.

1. Introduction

"The virtue of binary is that it's the simplest possible way of representing numbers. Anything else is more complicated."

Because of its central role in biological information processing, the canonical genetic code – which maps DNA codons onto corresponding amino acids – has been closely scrutinized for underlying symmetries. In this article, a novel binary representation of the code is introduced that accounts for both the chemical structures of the nucleotides themselves, as well as the physicochemical properties of the amino acids encoded. An accurate evaluation of the adaptive advantage of the code, which is robust to many point mutations and mRNA/tRNA mispairings, must consider not only the relatedness of amino acids separated by a single letter mutation, but also the probability of such a change or mispairing occurring in the first place based on the similarities of the nucleotides.

The primary addition this research makes to the existing literature is that the current work provides quantitative support for its novel binary classification system. That is, the choice of binary labels, as well as the order of the bits, have meaningful relationships with both the chemical structures of the nucleotides themselves, as well as the amino acids corresponding to codons in which they appear, as demonstrated with physicochemical data. This stands in contrast with many previous studies, which fixated on using the degree of degeneracy in the third letter as the primary or sole metric, and more crucially, treated the nucleotides as interchangeable labels for group theory analysis. This had the effect of strongly deemphasizing or obliterating entirely the physical reality of these biomolecules and their physicochemical similarities.

One of the first dichotomous divisions of the genetic code was due to theoretical physicist Yuri Rumer (Rumer, 1966; Fimmel and Strüngmann, 2016), who noticed complete third-letter degeneracy in exactly half of the codon quartets [those of the form NCN or SKN. Refer to Figs. 1 and 2 for nucleotide abbreviations]. More recent generalizations (Gumbel et al., 2015; Fimmel et al., 2013; Gonzalez et al., 2008) explored additional ways to bisect the genetic table. However, these works remained focused on classification according to the metric of third-letter degeneracy, along with hidden symmetries revealed by transformation rules involving the nucleotides. These rules show patterns under the interchange of nucleotides (Barbieri, 2008), while the current research takes account of the actual chemical properties of the nucleotides, as well as the amino acids they encode.

Modern computing (Atanasoff, 1984), which is built on the foundation of a binary system, provides a fertile analogy for the information conveyed by the quaternary encoding (Chechetkin, 2003) of DNA. That is, each of the four possible nucleotide bases of DNA represents a maximum of $\log_2(4) = 2$ bits of information. However, this comparison extends far beyond a superficial similarity; the canonical genetic code, represented by a correspondence table between codons and amino acids, has a manifestly hierarchical organization (Nemzer, 2017; Bashford et al., 1998). For example, the code distinguishes most clearly between pyrimidine (Y, Uracil or Cytosine) and purine (R, Adenine or Guanine) bases (Wilhelm

E-mail address: lnemzer@nova.edu

http://dx.doi.org/10.1016/j.biosystems.2017.03.001
0303-2647/

	Y_{o-}	R_{1-}
W_{-o}	U_{oo}	A_{1o}
S_{-1}	C_{o1}	G_{11}

Fig. 1. Nucleotide bases and their 2-bit identifiers as subscripts, along with the IUPAC letter abbreviations for duos. The four bases are assigned a binary identifier in which the first bit designates whether it is pYrimidine (0-) or puRine(1-). The second bit shows if the base is Weak (-0), forming two hydrogen bonds during Watson-Crick pairing, or Strong (-1), forming three. These pairings were chosen to prioritize the same physiological characteristics most recognized by the DNA repair and amino acid translation systems.

	U	C	A	G
	00	01	10	11
Y o-	1	1	0	0
R 1-	0	0	1	1
W -o	1	0	1	0
S -1	0	1	0	1
K o XOR	1	0	0	1
M 1 XOR	0	1	1	0

Fig. 2. Nucleotide truth table. The four nucleotides (U, C, A, G) are listed according to their respective 2-bit identifiers. Each base can join with each of the three others to make a duo based on physicochemical similarities. Complimentary duos (Y vs. R, then W vs. S, then K vs. M) are ordered according to their physiological relevance.

and Friedel, 2004). The number of heterocyclic rings differ in Y and R bases, and mutations that preserve this classification, called transitions, are more likely, but less damaging, than transversions between classifications.

Multiple factors lead to the observed excess of transition mutations relative to transversion mutations. Natural selection plays a role by disfavoring organisms with unfit mutations, which are shown here to be more likely to be transversions than transitions. However, even in pseudogene regions of DNA that are never transcribed into proteins, the excess remains (Zhang and Gerstein, 2003; Li et al., 1984). This can be attributed both to the spontaneous rate of mutation (Wakeley, 1996), which will be more frequent between chemically similar nucleotides, as well as recognition by the DNA repair machinery. It requires fewer chemical changes for mutations to arise in the first place between structurally similar nucleotides – particularly if they have the same number of rings. On the other hand, once transversion mutations, which change the number of rings from one to two (or vice versa) do occur, they are more likely to cause bulging or other distortions in the DNA double-helix, which may then be targeted by repair proteins.

The binary identifiers chosen here to represent the nucleotide bases are not arbitrary. They are selected to reflect the molecular similarities exhibited by the nucleotides themselves. And, as will be shown, these labels have significant correlations with the physicochemical properties of the amino acids to which they correspond. The system presented accords with the theory that the genetic code has been shaped by natural selection (Knight et al., 2001; Higgs and Sengupta, 2015; Koonin and Novozhilov, 2009), and that its evolution (Higgs, 2009; Wong, 1975) alongside the DNA mutation repair system (Modrich, 2006; Fukui, 2010) and tRNA translation mechanism (Saint-Leger et al., 2016) has produced a table with the adaptive benefit (Itzkovitz and Alon, 2007; Vetsigian et al., 2006) that single-nucleotide mutations (Haig and Hurst, 1991) most likely to cause a loss of protein function are also the most likely to be avoided (Berleant et al., 2009) or fixed. Frameshifted "hidden" stop codons (Seligmann, 2012) can also quickly terminate protein synthesis upon ribosomal slippage. The protein translation machinery provides further robustness to error, in that non-cognate amino acids most likely to be misloaded by a tRNA molecule tend to have codons with the largest probability of being mismatched by the anticodon in the first place (Seligmann, 2010, 2011).

Ancestral versions of the genetic code may have already exhibited clustering of related amino acids as a result of stereochemical or biosynthetic similarities (Knight et al., 1999). The inherent redundancy (Ardell and Sella, 2001) in the code provides a measure of fault-tolerance (Freeland et al., 2000), but also reduces the information (Nemzer, 2017) conveyed by each base. A fundamental irreducible representation (Sciarrino and Sorba, 2012) can be constructed as the direct sum of two special unitary groups, one corresponding the Y/R dichotomy, and the other corresponding to W/S. Throughout this work, the "/" notation represents a binary choice between the two elements. For example, "Y/R" means "a choice between pyrimidine and purine." This "crystal basis model" (Sciarrino and Sorba, 2014) was presented to explain the observed minimization of the number of unique tRNA molecules required to complete protein translation by allowing "wobble pairing" (Crick, 1966) of certain similar codons to the same tRNA molecule. Hypothesized ancestral codes in which "expanded" codons had more than three bases (Baranov et al., 2009; Seligmann, 2016) may also lend themselves to binary representations (Gonzalez et al., 2012).

Here, the standard amino acid correspondence table entries are recast as 6-bit binary messages. Due to the clustering of amino acids with similar physicochemical properties – the most critical (Lu and Freeland, 2008) for proper protein folding and function being size, hydropathy (Butler et al., 2009), and charge – individual bit positions are correlated with specific properties. The classification system introduced here is not arbitrary; it places the most "determinative" bits first, and prioritizes the same nucleotide molecular features that nature does. This should be contrasted with some methods that attempt to solve the reverse problem – encoding binary data using DNA (Bornholt et al., 2016) – that implement an arbitrary revolving code in order to minimize the occurrence of repeated bases, irrespective of the structures of the nucleotides. Following conventional codon tables, the system introduced here focuses on mRNA, so it uses uracil instead of thymine, but since these bases differ only by a single methyl group, it is likely that the same or similar physicochemical properties that are recognized by the mRNA-to-peptide translation machinery are also utilized by the DNA replication and repair mechanisms.

2. Method

A set of four elements, such as the nucleotides, can be divided into (4 choose 2)= 6 unique duos. Only two bits are needed to unambiguously identify each element, which leaves some freedom of