Physica A 454 (2016) 117–128

Contents lists available at ScienceDirect

Physica A

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

Codon information value and codon transition-probability distributions in short-term evolution

M.A. Jiménez-Montaño [a], H.F. Coronel-Brizio [a,b], A.R. Hernández-Montoya [a,c,*], A. Ramos-Fernández [d,e]

[a] Center for Research on Artificial Intelligence, University of Veracruz, Sebastián Camacho No. 5, Col. Centro, C.P. 91000, Xalapa, Ver., Mexico
[b] Faculty of Physics, University of Veracruz, Circuito Gonzalo Aguirre Beltrán s/n, Zona Universitaria, CP 91000, Xalapa, Veracruz, Mexico
[c] Ph. D. Program on Science, Technology and Society (DCTS), Centro de Investigación y de Estudios Avanzados del IPN (Cinvestav), Av. Instituto Politécnico Nacional 2508, Col. San Pedro Zacatenco, Delegación Gustavo A. Madero, Código Postal 07360 Apartado Postal: 14-740, 07000 México, D.F., Mexico
[d] Institute of Biotechnology and Applied Ecology (INBIOTECA), University of Veracruz, Av. de las Culturas Veracruzanas No.101, Col. E. Zapata, C.P. 91090, Xalapa, Veracruz, Mexico
[e] Red Biodiversidad y Sistemática, Instituto de Ecología, A. C. (INECOL), A.P. 63, Xalapa, Veracruz, Mexico

HIGHLIGHTS

- Inverse power-laws fit two-fold degenerated codons; they have high information values.
- Exponentials fit Four-fold degenerated codons; they have low information values.
- Six-fold degenerated codons are considered to be doubly assigned.
- We propose a codon-level model for acid substitutions in short-term protein evolution.
- Ordering codons according to their value, orders them according to their degeneracy.

ARTICLE INFO

Article history:
Received 21 May 2015
Received in revised form 19 December 2015
Available online 27 February 2016

Keywords:
Codon
Value of information
Fitting
Power law
Exponential
Reduced empirical matrix

ABSTRACT

To understand the way the Genetic Code and the physical–chemical properties of coded amino acids affect accepted amino acid substitutions in short-term protein evolution, taking into account only overall amino acid conservation, we consider an underlying codon-level model. This model employs codon pair-substitution frequencies from an empirical matrix in the literature, modified for single-base mutations only. Ordering the degenerated codons according to their codon information value (Volkenstein, 1979), we found that three-fold and most of four-fold degenerated codons, which have low codon values, were best fitted to rank-frequency distributions with constant failure rate (exponentials). In contrast, almost all two-fold degenerated codons, which have high codon values, were best fitted to rank-frequency distributions with variable failure rate (inverse power-laws). Six-fold degenerated codons are considered to be doubly assigned. The exceptional behavior of some codons, including non-degenerate codons, is discussed.

* Corresponding author at: Center for Research on Artificial Intelligence, University of Veracruz, Sebastián Camacho No. 5, Col. Centro, C.P. 91000, Xalapa, Ver., Mexico.
E-mail addresses: ajimenez@uv.mx (M.A. Jiménez-Montaño), alhernandez@uv.mx (A.R. Hernández-Montoya).

http://dx.doi.org/10.1016/j.physa.2016.02.043
0378-4371/© 2016 Elsevier B.V. All rights reserved.

1. Introduction

1.1. Motivation and background

In a former publication [Jiménez-Montaño M.A. and Matthew He (2009)] [1], one of the present authors underscored the importance of short-term protein evolution for the study of some problems in molecular biology and biotechnology, such as virus variability [2], somatic hyper mutation [3], *in vitro* exploration of protein sequence space [4], enzyme, and drug design by directed evolution [5,6] and human genetic mutations [7], among others. However, the general attention has been mainly devoted to distantly-related protein sequences, in studies oriented to molecular phylogenetics analysis [Li, W.H. (1997) [8]; Lió, P.N. Goldman (1998)] [9] rather than to closely-related sequences, which are dominant in the short-term. Contrary to this tendency, in the present paper we employ the empirical codon substitution model proposed by Gaston Gonnet and his group [10], adapted for single-point mutations in Ref. [1]. We will not consider inversions, duplications, indels or other mutational changes.

The ubiquity of power-law distributions in physical, biological and social sciences is so well known (Martínez-Mekler et al., 2009 [11]; Newman, 2006 [12] and references therein) that it is not necessary to add further comments here. For the case of the frequency of word use in a large written text, Zipf (1949) [13] found the power-law that bears his name (see below). In the nineties of the last century, mainly motivated by issues of mathematical modeling of abstract dynamical systems, stochastic processes and noise, several research groups contributed to the characterization of the correlation structure of DNA sequences, as described in a review of that period by Wentian Li (1997) [14]. Since most of these studies of correlation in DNA sequences were not biologically-motivated they were established on base-to-base statistical correlations, which are analogous to statistical correlations between letters in an English text (Schürmann and Grassberger, (1996)) [15]. These correlations rarely reveal correlations at the syntax level in a sentence. In human languages words are the shortest meaningful units, while in a DNA (or RNA) polymers base triplets only become the smallest meaningful units, called codons, in the context of the translation apparatus of the cell. Physically there is no such thing as a codon. There are merely nucleotide triplets because codons can only be defined with respect to a genetic code (e.g. the Universal Code, the Mammalian Mitochondrial Code, etc.). Besides, changing the reading-frame redefines codons in the same polynucleotide (i.e., the same physical object).

Pursuing the analogy between human and artificial languages with genetics, Mantegna et al. (1995) [16] studied correlations in DNA sequences using tools previously developed by linguists for the quantitative analysis of natural language and symbolic sequences. In particular, one problem of interest is to find the distribution of meaningful codons (base triplets that code for amino acids) over the coding regions of DNA (Mantegna, 1995; Som et al., 2001) [16,17]; that is, along a given sequence. The approach is based on conventional Zipf analysis (Zipf, 1949) [13] in which the frequency of occurrence of words in a given text is measured by counting the number of occurrences of each word throughout the text and dividing this value by the number of words. The frequency of occurrence f of each word is ordered from the most frequent to the least frequent value. The position of each word in this ordered list is called its rank R. Log–log plots of word frequency versus word rank are called Zipf-plots. Since we are not interested in correlations *per se* in long DNA sequences, in contrast to the main analysis in (Mantenga et al., 1995) [16], in this work we will not discuss their generalization to overlapping n-tuples in DNA sequences, which they called n-tuple Zipf analysis.

At variance with naive expectation, apparently the distribution of non-overlapping 3-tuples (i.e. codons) of protein-coding DNA segments does not obey Zipf's law but an exponential distribution. In an interesting paper, Som et al., [17] compared two-parameter power-law fits with two-parameter exponential fits and found that the exponentials provide better fits. Analogous results were obtained in (Kim, 2005) [18]. In this case one gets a straight line in a semi log frequency-rank plot, in place of the straight line found in Zipf plots [19].

Following the statistical linguistic approach, we investigate codon variation in a given evolutionary period; that is, a vertical analysis of an alignment of homologous sequences. Specifically, we consider codon transition probabilities, from a given codon to its single-nucleotide mutation neighbors, in order to separate the ones that can be fitted with a memory-less distribution (the exponential distribution for a continuous variable, or the geometrical distribution for a discrete variable), with constant failure rate, from those that follow distributions with variable failure rate (such as power laws, lognormal, negative binomial, Weibull distributions, etc.). Furthermore, we relate our fitted distributions with a mathematical measure of the information value of a codon, introduced by Volkenstein [20] thirty five years ago, which takes into account both physical–chemical factors of the coded amino acids and the structure of the genetic code (see Methods). We show that ordering the codons according to their value automatically orders them according to their degeneracy. Besides, we found that there is a strong correlation between their degeneracy and the model to fit the transition probability distribution, from a given codon to its single-nucleotide mutation neighbors.

1.2. Related problems

We first discuss, with respect to their failure rate, several probability distributions of amino acids proposed in the literature for different sites in a multiple-sequence alignment (MSA). A MSA is a historical record (Valdar (2002)) [21], with different positions in the alignment having different structural and functional roles. Thus, the amino acid in each position will be subject, both, to different physicochemical constraints and selective pressures. Therefore, the probability distribution