algorithms

MDPI

Article

Use of the Codon Table to Quantify the Evolutionary Role of Random Mutations

Mihaly Mezei

Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; Mihaly.Mezei@mssm.edu

Abstract: The various biases affecting RNA mutations during evolution is the subject of intense research, leaving the extent of the role of random mutations undefined. To remedy this lacuna, using the codon table, the number of codons representing each amino acid was correlated with the amino acid frequencies in different branches of the evolutionary tree. The correlations were seen to increase as evolution progressed. Furthermore, the number of RNA mutations that resulted in a given amino acid mutation were found to be correlated with several widely used amino acid similarity tables (used in sequence alignments). These correlations were seen to increase when the observed codon usage was factored in.

Keywords: codon table; amino acid propensity; mutation probability; amino-acid substitution matrix

Citation: Mezei, M. Use of the Codon Table to Quantify the Evolutionary Role of Random Mutations. *Algorithms* **2021**, *14*, 270. https://doi.org/10.3390/a14090270

Academic Editor: Frank Werner

Received: 18 August 2021
Accepted: 15 September 2021
Published: 17 September 2021

1. Introduction

It is a fundamental tenet of evolutionary biology that evolution is driven by mutations in the carrier of the genetic code, the DNA, and, consequently, the RNA. While a simplistic view of Darwinian evolution would assume all mutations to be equally probable, the current work in this field is dominated by characterizing various biases in the mutation rates. A recent paper reviewed the research on these biases indicated that biases are categorized as mutation bias, i.e., the difference in the rate of changes at the RNA level (e.g., C–G bias [1]), and composition bias, i.e., the difference in the fitness of the mutated structure, showing bias at the protein level, and showed that they act synergistically [2]. Another recent work [3] emphasized the non-randomness of mutations.

The aim of this paper is to divert the attention from the various biases, and instead attempts to quantify the randomness of the mutations observed during evolution. For this purpose, the results of the mutations will be examined at the protein level. In support of the protein-centric view of mutation, it should be noted that while mutations at the RNA level are interesting biochemical problems, the biological significance of the mutations manifests itself at the protein level.

It has been recognized that if evolution were driven by unbiased random mutations, then the amino acid (AA) frequencies in proteins should fully correlate with the number of codons that code each AA; a correlation was indeed found, but it was not very strong [4]. The number of codons were also found negatively correlated with the AA size and were correlated with the AAs hydrostatic pressure asymmetry index [5]. This paper will examine the codon multiplicity–AA frequency correlation according to families of organisms.

One important use of sequence alignment algorithms is for the detection of the evolutionary relations of proteins. Sequence alignments are based on so-called substitution matrices that measure the similarity of various AAs; in the evolutionary context they should measure the likelihood of one AA to be mutated into another one. The more the RNA mutation are unbiasedly random, the more these substitution matrices should be correlated with the likelihood of AA mutations resulting from unbiased random RNA mutations. To get at these correlations, the codons representing each AA will be used to derive the unbiased likelihood of AA mutations resulting from single or double nucleotide changes

followed by the calculation of the correlation between the mutation likelihoods derived from the codon table and various AA substitution matrices. The larger the correlation, the larger the role of randomness in the AA mutations.

2. Materials and Methods

The AA frequencies were obtained for fungi, protozoa, invertebrate, vertebrate (mammals), vertebrate (non-mammals), and plants from the work of Gaur [6]; Tables 1 and 2 of [6] show the AA frequencies for membrane and non-membrane proteins, respectively. The codon table used is from Wikipedia, using the AA coding established for the majority of organisms, with the understanding that there are some organisms where different coding is used. The codon usage frequencies cfreq(*i*,*j*,*k*) were obtained from the GenScript (GenScript Inc., Piscataway, NJ, USA) website (https://www.genscript.com/tools/codon-frequency-table, accessed on 16 September 2021), which provides usage frequencies for a number of organisms, including homo sapiens. The list of AA substitution matrices is described in [7] and was downloaded from the AAindex database (http://www.genome.jp/dbget-bin/www_bfind?aaindex, accessed on 16 September 2021).

The number of single ($M = 2$) or double ($M = 1$) RNA mutations, possibly weighted, that change AA a to AA b, $m(a,b)$, is obtained as follows

$$m(a,b) = \sum_{i,j,k=1}^{4} \delta(a, aa(i,j,k)) \\ * \sum_{i',j',k'=1}^{4} \delta(\delta(i,i') + \delta(j,j') + \delta(k,k'), M)\delta(b, aa(i',j',k')) \\ * W(i,j,k,i',j',k')$$

where $\delta(i,j)$ is the Kroenecker delta; $aa(i,j,k)$ is the amino acid number ($\leq$20) coded by the nucleotides i, j, and k; and $W(i,j,k,i',j',k')$ is a weight depending on the codons (i,j,k) and (i',j',k') involved.

All of the calculations described in this paper were performed by the program codon.f, available at the URL https://mezeim01.u.hpc.mssm.edu/codon (accessed on 16 September 2021).

3. Results

3.1. Correlation between Codon Multiplicities and AA Propensities

The Pearson correlations (r) between the codon multiplicities and propensities of the AAs (shown in Table 1) were calculated for protozoa, invertebrates, fungi, plants, vertebrates (non-mammals), and vertebrates (mammals); the calculations were performed separately for the membrane and non-membrane proteins. In addition, all correlations were also calculated using only the 11 AAs that were found to be proximal to each other in [4]; for this subset, the correlations were stronger. With the observation in mind that in the evolutionary tree both fungi and plants are on a separate branch of the branch formed by protozoa, invertebrates, and vertebrates, we can conclude that the correlations increased as evolution progressed.

3.2. The Number of RNA Mutations per AA Mutation

Tables 2 and 3 show for each pair of AAs, the number of single and double RNA mutations, respectively, that result in the AA mutation. First, these data reflect the observation that single mutations are likely to result in small chemical changes [8,9]. While it is not surprising that not all of the $20 \times 19/2 = 180$ AA mutations can be achieved by a single RNA mutation, it is important to notice that a significant number of AA mutations cannot be achieved even by double nucleotide changes.

Table 1. Pearson correlation *r* between amino acid frequencies and codon multiplicities.

	All 20		Top 11	
	Membrane	**Non-Membrane**	**Membrane**	**Non-Membrane**
Vertebrates (mammals)	0.87	0.81	0.97	0.98
Vertebrates (non-mammals)	0.76	0.73	0.93	0.97
Plants	0.72	0.71	0.88	0.81
Fungi	0.67	0.75	0.86	0.94
Invertebrates	0.52	0.52	0.82	0.88
Protozoa	0.50	0.23	0.78	0.70

Legend: membrane and non-membrane: correlations calculated for all membrane proteins and all non-membrane proteins, respectively; All 20 and Top 11: correlations calculated for all 20 AAs and for the top 11 AAs in the ranking of [4], respectively.

Table 2. Number of single mutations that result in each AA mutation.

Amino Acid	Amino Acid Code	Number of Single Mutations																			
GLY	**G**	12																			
ALA	**A**	4	12																		
VAL	**V**	4	4	12																	
LEU	**L**	0	0	6	18																
ILE	**I**	0	0	3	4	6															
SER	**S**	2	4	0	2	2	14														
THR	**T**	0	4	0	0	3	6	12													
ASP	**D**	2	2	2	0	0	0	0	2												
GLU	**E**	2	2	2	0	0	0	0	4	2											
ASN	**N**	0	0	0	0	2	2	2	2	0	2										
GLN	**Q**	0	0	0	2	0	0	0	0	2	0	2									
LYS	**K**	0	0	0	0	1	0	2	0	2	4	2	2								
HIS	**H**	0	0	0	2	0	0	0	2	0	2	4	0	2							
ARG	**R**	6	0	0	4	1	6	2	0	0	0	2	2	2	18						
PHE	**F**	0	0	2	6	2	2	0	0	0	0	0	0	0	0	2					
TYR	**Y**	0	0	0	0	0	2	0	2	0	2	0	0	2	0	2	2				
TRP	**W**	1	0	0	1	0	1	0	0	0	0	0	0	0	2	0	0	0			
CYS	**C**	2	0	0	0	0	4	0	0	0	0	0	0	0	2	2	2	2	2		
MET	**M**	0	0	1	2	3	0	1	0	0	0	0	1	0	1	0	0	0	0	0	
PRO	**P**	0	4	0	4	0	4	4	0	0	0	2	0	2	4	0	0	0	0	0	12
STP		1	0	0	3	0	3	0	0	2	0	2	2	0	2	0	4	2	2	0	0	4
		G	**A**	**V**	**L**	**I**	**S**	**T**	**D**	**E**	**N**	**Q**	**K**	**H**	**R**	**F**	**Y**	**W**	**C**	**M**	**P**	**STP**

Table 3. Number of double mutations that result in each AA mutation.

Amino Acid	Amino Acid Code	Number of Double Mutations																			
GLY	**G**	0																			
ALA	**A**	12	0																		
VAL	**V**	12	12	0																	
LEU	**L**	6	6	18	12																
ILE	**I**	3	3	9	14	0															
SER	**S**	10	14	6	12	7	4														
THR	**T**	4	12	4	6	9	18	0													
ASP	**D**	6	6	6	2	2	4	2	0												
GLU	**E**	6	6	6	4	1	2	2	0	0											
ASN	**N**	2	2	2	2	4	4	6	2	4	0										
GLN	**Q**	2	2	2	8	1	2	2	4	2	4	0									
LYS	**K**	2	2	2	4	5	6	6	4	2	0	2	0								
HIS	**H**	2	2	2	6	2	4	2	2	4	2	0	4	0							
ARG	**R**	18	6	6	18	8	12	10	2	4	6	8	4	6	12						
PHE	**F**	2	2	6	6	4	8	2	2	0	2	0	0	2	2	0					
TYR	**Y**	2	2	2	6	2	8	2	2	4	2	4	4	2	2	2	0				
TRP	**W**	3	1	1	2	0	5	1	0	1	0	1	1	0	4	2	2	0			
CYS	**C**	6	2	2	6	2	8	2	2	0	2	0	0	2	10	2	2	0	0		
MET	**M**	1	1	3	4	0	3	3	0	1	2	1	1	0	2	2	0	1	0	0	
PRO	**P**	4	12	4	14	3	14	12	2	2	2	2	2	6	14	2	2	1	2	1	0
STP		5	3	3	6	2	11	3	4	3	4	3	5	4	8	6	2	1	4	1	3	2
		G	**A**	**V**	**L**	**I**	**S**	**T**	**D**	**E**	**N**	**Q**	**K**	**H**	**R**	**F**	**Y**	**W**	**C**	**M**	**P**	**STP**

3.3. Correlation between the Number of RNA Mutations and AA Substitution Matrices

The larger the role of random mutations in the genetic evolution, the more the number of mutations per AA substitution should correlate with the AA similarity measures. The observation suggests the examination of the correlation between the number of RNA mutations that can change a given AA into another one and the corresponding elements of the various substitution matrices used in the AA sequence alignments.

For the calculation of these correlations, 98 substitution matrices were downloaded from the AAindex database [7]. The list of matrices involves several variants of the two widely used matrix types: PAM [10–13] and BLOSUM [14]. PAM stands for "point accepted mutation", the matrices are based on scoring all amino acid positions in the related sequences; the different versions are based on different timeframes for the mutations. BLOSUM stands for "BLOcks SUbstitution Matrix", and the matrices are based on substitutions and conserved positions in blocks; the closer the compared sequences are, the higher the number version of the matrix that is supposed to be used. It is important to note that these substitution matrices were developed from data at the AA level without considering the requisite RNA mutations. Several of the rest were designed for specific situations. Most, but not all, were symmetric.

Table 4 shows the Pearson correlations between the matrix of the number of mutations, both unweighted and weighted (with the codon usage frequency from the human genome), as shown in Tables 2 and 3, respectively, and the PAM and BLOSUM matrices; the correlations with the full set of 98 substitution matrices are provided in the Supplementary Materials. For the matrices that were not symmetric, the correlations were calculated for the lower triangular matrix. In addition, the correlation calculations were also performed using only AA mutations that could be achieved with only one or two RNA mutations.

Table 4. Correlation between PAM or BLOSUM matrices and the number of mutation paths.

	All AA Mutations Included				AA Mutations Requiring Two/Three RNA Mutations Excluded			
	Single Mutations		Double Mutations		Single Mutations		Double Mutations	
Code [1]	Corr_n [2]	Corr_w [2]	Corr_n	Corr_w	Corr_n	Corr_w	Corr_n	Corr_w
ALTS910101 [3]	0.508	0.540	0.080	−0.091	0.389	0.364	0.360	0.172
BENS940101 [4]	0.488	0.626	0.073	−0.075	0.306	0.507	0.378	0.250
BENS940102 [5]	0.469	0.577	0.051	−0.090	0.333	0.486	0.346	0.230
BENS940103 [6]	0.447	0.518	−.007	−0.114	0.363	0.462	0.270	0.221
DAYM780301 [7]	0.449	0.480	0.055	0.085	0.338	0.326	0.307	0.167
JOND920103 [8]	0.476	0.596	0.057	−0.096	0.318	0.486	0.376	0.237
JOND940101 [9]	0.404	0.591	−0.043	−0.111	0.169	0.444	0.216	0.243
DAYM780302 [10]	0.562	0.565	0.071	−0.104	0.481	0.350	0.339	0.247
HENS920101 [11]	0.487	0.491	−0.093	−0.224	0.420	0.407	0.222	0.125
HENS920102 [12]	0.506	0.518	−0.072	−0.206	0.440	0.450	0.235	0.139
HENS920103 [13]	0.521	0.532	−0.061	−0.205	0.441	0.447	0.252	0.139
HENS920104 [14]	0.483	0.498	−0.103	−0.214	0.413	0.401	0.114	0.187

[1] Code: the AAindex database identifier; [2] Corr_n, Corr_w: number of mutation paths calculated with $W = 1$ and $W = \text{cfreq}(i,j,k) \times \text{cfreq}(i',j',k')$, respectively; [3] PAM-120 [11]; [4] PAM-log-odds_6.4–8.7 [12]; [5] PAM-log-odds_22–29 [12]; [6] PAM-log-odds_74–100 [12]; [7] PAM-250-log-odds [10]; [8] PAM-250_PET91 [13]; [9] PAM-250_TM [13]; [10] PAM-40–log-odds [10]; [11] BLOSUM45 [14]; [12] BLOSUM62 [14]; [13] BLOSUM80 [14]; [14] BLOSUM50 [14].

4. Discussion

The variation in the number of codons coding for a given amino acid suggested a number of ways to quantify the role of random mutations in genetic evolution. In particular, the work of Mittal and Jayaram was refined to study the correlation between the number of codons and the propensity of the corresponding AAs to calculate it for different branches of the evolutionary tree and the correlations between the number of mutations changing each AA to other AAs with the corresponding substitution matrices used in sequence alignments. Significant correlations were indeed found, confirming that random mutations indeed form one contribution to genetic evolution, but are not enough to account for it completely. At the same time, the results confirmed the existence of significant biases, as referred to in the Introduction.

The interesting observation from the present calculations is the increase of correlation moving up in the evolutionary tree; this observation also holds for the subset of 11 AAs defined by Mittal and Jayaram. If one accepts the role of randomness of mutations, then this observation provides experimental support for the idea of Darwinian evolution (not that it lacks other, more convincing support). If one accepts Darwinian evolution, then this observation provides experimental support and a measure of the role of random mutations in the evolution.

The correlations found also raise an intriguing question. It has been shown recently that a significant contribution to the foldability of an AA sequence is its adherence to the non-uniform distribution of AAs [15]. This leads to the following question: was there, at the early stages of genetic evolution, a mechanism that biased the establishment of the codon table to use more codons for the more frequent AAs? It is not impossible, as it has already been suggested, that the establishment of the genetic code was influenced by the chemistry of the AAs [16], as it has been found to be likely that the genetic code was established at a later stage of evolution [17].

The fact that the codon table as finalized during evolution ended up with a number of AA mutations that can only be achieved with three RNA mutation can be considered an imperfection. It could have been the result of either too fast evolution (like an optimization process using simulated annealing that was not run slow enough) or of competing require-

ments (like minimizing the number of RNA mutations, which would result in a drastic AA switch).

In the study examining the correlations between the number of RNA mutations per AA mutations and the substitution matrices, the results were different for AA mutations that required single and those that required double RNA mutations. For single RNA mutations, the results showed significant correlations, both when using all matrix elements and when excluding the matrix elements that correspond to AA mutations that cannot be achieved with a single RNA mutation, although the correlations were weaker for the latter. Furthermore, it is gratifying that the correlations increased when the experimental codon usage probabilities were taken into account. The increased correlations provide additional confirmation for the role of random mutations.

For the double RNA mutations, however, the results were different. When using all matrix elements and without using frequency weighting, the correlations were small, but at least positive, but they become negative when the usage frequencies were factored in. Omitting the "forbidden" AA mutations, contrary to the single RNA mutation case, the correlations increased, but remained small. This suggests that when an alignment indicates AA mutations requiring more than one RNA mutation, it is unlikely to be the result of just a pair of mutations.

One explanation for the low or even negative correlations, using the assumption of random double mutations, could be that such AA changes are the result of insertions and/or frame shifts; this could also explain the occurrence of AA mutations requiring three RNA mutations (which is even more unlikely). The incorporation of this idea could be an alignment algorithm that is done simultaneously at the RNA and the AA level, which could verify the presence of a frame shift.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/a14090270/s1: the correlations of the mutation numbers with the full set of 98 substitution matrices.

Funding: This research received no external funding. It was supported in part through the computational resources and staff expertise provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai.

Acknowledgments: George Rose, B. Jayaram, and Aditya Mittal are thanked for their helpful suggestions.

Conflicts of Interest: The author declares no conflict of interest.

References

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921. [CrossRef] [PubMed]
2. Cano, A.V.; Payne, J.L.P. Mutation bias interacts with composition bias to influence adaptive evolution. *PLoS Comput. Biol.* **2020**, *16*, e1008296. [CrossRef] [PubMed]
3. Caporale, L.H.; Doyle, J. In darwinian evolution, feedback from natural selection leads to biased mutations. *Ann. N. Y. Acad. Sci.* **2013**, *1305*, 18–28. [CrossRef] [PubMed]
4. Mittal, A.; Jayaram, B. A possible molecular metric for biological evolvability. *J. Biosci.* **2012**, *37*, 573–577. [CrossRef] [PubMed]
5. Giulio, M.D. The origin of the genetic code: Theories and their relationships, a review. *BioSystems* **2005**, *80*, 175–184. [CrossRef] [PubMed]
6. Gaur, R.K. Amino acid frequency distribution among eukaryotic proteins. *IIOAB J.* **2014**, *5*, 6–11.
7. Kawashima, S.; Pokrowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. AAindex: Amino acid index database, progress report 2008. *Nucleic Acids Res.* **2008**, *36*, D202–D205. [CrossRef] [PubMed]
8. Giulio, M.D. The extension reached by the minimization of the polarity distances during the evolution of the genetic code. *J. Mol. Evol.* **1989**, *29*, 288–293. [CrossRef] [PubMed]
9. Wong, J.T.-F. Role of minimization of chemical distances between amino acids in the evolution of the genetic code. *Proc. Natl. Acad. Sci. USA* **1980**, *77*, 1083–1086. [CrossRef] [PubMed]
10. Dayhoff, M.O.; Schwartz, R.M.; Orcutt, B.C. A model of evolutionary change in proteins. *Atlas Protein Seq. Struct.* **1972**, *5*, 89–99.
11. Altschul, S.F. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* **1991**, *219*, 555–565. [CrossRef]

12. Benner, S.A.; Cohen, M.A.; Gonnet, G.H. Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng.* **1994**, *7*, 1323–1332. [CrossRef]
13. Jones, D.T.; Taylor, W.R.; Thornton, J.M. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **1992**, *8*, 275–282. [CrossRef]
14. Henikoff, S.; Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 10915–10919. [CrossRef]
15. Mezei, M. On predicting foldability of a protein from its sequence. *Proteins* **2020**, *88*, 355–356. [CrossRef]
16. Wong, J.T.-F. A co-evolution theory of the genetic code. *Proc. Nat. Acad. Sci. USA* **1975**, *72*, 1909–1912. [CrossRef] [PubMed]
17. Giulio, M.D. Reflections on the origin of the genetic code: A hypothesis. *J. Theor. Biol.* **1998**, *191*, 191–196. [CrossRef]