BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net **Hypothesis**

Volume 8(3)

Codon usage bias as a function of generation time and life expectancy

Rami N Mahdi[1] & Eric C Rouchka[2]*

[1]Weill Cornell Medical College, Department of Genetic Medicine, New York, NY USA 10065; [2]University of Louisville, Speed School of Engineering, Department of Computer Engineering and Computer Science, Duthie Center Room 208, Louisville, KY USA 40292; Eric C Rouchka - Email: Eric.rouchka@louisville.edu; *Corresponding author

Received September 21, 2011; Accepted January 27, 2012; Published February 03, 2012

Abstract:
It has recently been demonstrated that human natural codon usage bias is optimized towards a higher buffering capacity to mutations (measured as the tendency of single point mutations in a DNA sequence to yield the same or similar amino acids) compared to random sequences. In this work, we investigate this phenomenon further by analyzing the natural DNA of four different species (human, mouse, zebrafish and fruit fly) to determine whether such a tolerance to mutations is correlated with the life span and age of sexual maturation for the corresponding organisms. We also propose a new measure to quantify the buffering capacity of a DNA sequence to mutations that takes into account the observed mutation rates within every genome and the effect of the corresponding mutation.

Our results suggest there is a propensity for tolerance to mutations that is positively correlated with the life expectancy of the considered organisms. Moreover, random sequences that are constrained to produce the same protein as the naturally occurring sequences are found to be more buffered than completely random sequences while being less buffered than the natural sequences. These results suggest that optimization toward protective mechanisms tolerant to mutations is correlated with both life expectancy and age to sexual maturity at both the levels of codon usage bias and the bias of the natural sequence of codons itself.

Keywords: Buffering capacity, Codon bias, Sequence evolution

Background:

In gene coding regions, DNA is put together in triplets to form 64 distinctive codons. 61 of those codons synthesize for the 20 amino acids, each of which can be synthesized by as few as one or as many as six synonymous codons. An intuitive assumption is that codons and amino acids are evenly used throughout different genomes. However, it has been discovered that amino acid incorporation as well as codon usage are biased in all organisms from bacteria to mammals **[1-3].** It is anticipated then that selective forces are acting to maintain the current balance between mutation and selection, resulting in optimal coding regions. Some of the proposed factors include degree and timing of gene expression, codon-anticodon interaction, transcription and translation rate and accuracy, codon context, and global and local G+C content.

In humans, bias has been reported on nucleotide abundance in the three codon positions, dinucleotide usage, and di-codon usage within coding regions **[4]**. This bias leads to a high usage of acidic amino acids as well as reflects the avoidance of stop codons. Bias in codon usage has been attributed to selection towards a more efficient translation model. This suggests codon usage in highly expressed genes is biased toward optimal codons corresponding to more abundant tRNAs. Such models take into consideration both the elongation rate and the translation accuracy **[5-6]**. In more recent studies **[7-8]**, it has been shown that the codon usage bias is found to relate to gene expression with highly expressed genes favoring codons corresponding to the more abundant tRNAs (thus increasing translation efficiency) and lower expressed genes favoring less abundant tRNAs. In addition, highly expressed genes in humans are found to be shorter in general and contain less intronic content **[9].** This study also reports a bias in the amino acid usage where highly expressed genes avoid complex amino acids, where complexity is based on the weight and shape of the amino acid. Other selection factors proposed include the effects of DNA polymerase and repair mechanism, methylation, CpG islands **[10-11],** tissue or organelle specificity **[12],** increasing mRNA stability, transcription rate **[13-14]** and evolutionary age **[15]**. None of the previously mentioned factors manage to provide a complete explanation for codon usage bias. It is likely that all of these forces work together competitively with selection, organelle and organism specificity to produce the current bias found in every gene and every genome.

In a previous paper **[16]**, we demonstrated that in humans, bias in codon usage makes the coding DNA significantly more tolerant to point mutations than random sequences. If only substitution events are considered in a codon, the degeneracy of the genetic code allows for a single nucleotide conversion to result in a codon representing the same amino acid, a different amino acid (missense mutation), or a stop codon (nonsense mutation). Analysis based on single nucleotide mutation rates and similarity among amino acids showed that point mutations in the natural coding sequence of humans are significantly more likely to yield the same or similar amino acids than would be the case in random DNA sequences. Moreover, the coding sequence based on the natural transcription reading frame was shown to be more buffered to mutations than the other two possible shifted reading frames. Here, we say a codon is buffered to or tolerates point mutations when it tends to translate to the same or similar amino acid as the original one more often than would be expected by random when exposed to point mutations.

Methodology:

Buffering capacity

Buffering capacity refers to the tendency of a coding sequence to allow for point mutation events that result in a sequence of the same or similar amino acids, as opposed to missense or nonsense mutations. In order to quantify this property, we have developed a mathematical model that approximates the buffering capacity for an individual sequence by taking into account the codon composition, individual nucleotide mutation rates, and a distance matrix between amino acids that reflects the resulting damage caused by individual amino acid substitutions**(See supplementary material).**

Selection toward error minimization or increased buffering is not a new concept. For example, the natural coding or the natural mapping between the 20 amino acids and the 61 coding codons is believed to have come to its current state by evolution and selection and it has been shown to provide high tolerance to mutations or translation errors **[17-19]**. In **[19],** the authors have shown that only two in a billion randomly generated mappings would provide better error minimization than the natural genetic code.

In this paper, we test for a possible correlation between the expected life span of an organism and the buffering capacity in its coding DNA. This hypothesis is motivated by the fact that an organism that lives long will require genes buffered against mutations to protect against harmful changes to protein sequences while a species with a short life span does not need such a property since it will not be subjected to nearly as many point mutations throughout its life. Furthermore, organisms that have a long life span usually reach sexual maturity at a later age at which time their DNA gets passed to the next generation. Therefore, it is expected that the longer the life span or age to sexual maturity of an organism, the higher the buffering.

To calculate the buffering capacity of a given coding DNA sequence, we used a similar measure to the one used in [16]. The computed buffering capacity takes into consideration both the organism-specific rates of mutation and the consequences of the corresponding mutations. For mutation rates, we have used the neighbor-dependent substitution rates reported in a recent study based on observed neutral mutations within inserted retrotransposable elements **[20].**

To estimate the consequence of a mutation, an amino acid similarity matrix is used to measure the distance between the original amino acid and the amino acid resulting from the mutation **[19]**. This similarity matrix is based on computations of the change in the structure and folding free energy of a protein when a single amino acid is mutated to another one at all positions in a set of 141 different proteins. On the other hand, a nonsense mutation is thought to be the worst type of mutation; nonetheless there is no natural way to weigh a nonsense mutation compared to other types of missense mutations. As an approximation, in this paper we consider it to be three times as detrimental as the worst missense mutations.

Dataset

The species selected include *Drosophila melanogaster* (fruit fly), *Danio rerio* (zebrafish), *Mus musculus* (mouse), and *Homo sapiens* (humans). These were selected based upon varying rates of average life expectancy and time to sexual maturity as well as the availability of sequence data and species specific mutation rates. The rates for life expectancy and time to sexual maturity were derived from the Human Aging Genome Resources **[21].** These rates are provided in **Table 1 (See supplementary material).**

For the purpose of this study, human gene sequences from GenBank build 35.1 were downloaded from the human Exon-Intron Database (EID) **[22]** from the website http://www.utoledo.edu/med/depts/bioinfo/database.html. 16,800 human genes in this dataset were considered. These sequences were further filtered to exclude genes whose nucleotide sequence is not a multiple of three, or who do not begin with the start codon ATG. In addition, gene sequences that do not end in one of the three stop codons (TAA, TAG, or TGA) were removed. For zebrafish, mouse, and fruit fly, mRNA sequences were downloaded from the RefSeq database. The total number of genes used for each organism is provided in **Table 2 (See supplementary material).**

CDS categorization

The data downloaded represents known mRNA sequences from the coding sequences (CDS) of genic regions. These datasets were categorized and labeled as “natural”. For each known CDS, a randomly generated mRNA sequence was constructed that yields the exact same protein. These correspond to a group labeled as “constrained random” since they are randomly generated sequences constrained by the property that they must yield an identical amino acid sequence upon translation. A third set of sequences labeled “random” was constructed consisting of mRNA sequences of the same length as the natural group.

Analysis of buffering capacity

For each of the natural, constrained random and random sequences, the buffering capacity was calculated as described in the buffering capacity section. Mutation rates for fruit fly, zebrafish, and human were downloaded from an existing dataset **[20].** Mutation rates for the mouse were calculated using the publicly available web server using the methods described in **[20]** with the mouse B1_Mur1 repetitive element from RepBase as the ancestral sequence and a set of 4,010 annotated B1_Mur1 occurrences in the mouse genome as the descendant sequences. A distribution of the buffering capacities for each of these three groups was examined. A distribution close to zero would indicate a set of sequences that are highly buffered to point mutations at the mRNA level that do not significantly affect the corresponding translated protein sequence. The main hypothesis is that if a species has a built-in tolerance to mutations, the distribution of buffering capacity scores between the naturally occurring mRNA sequence and the random sequence sets should be significantly different.

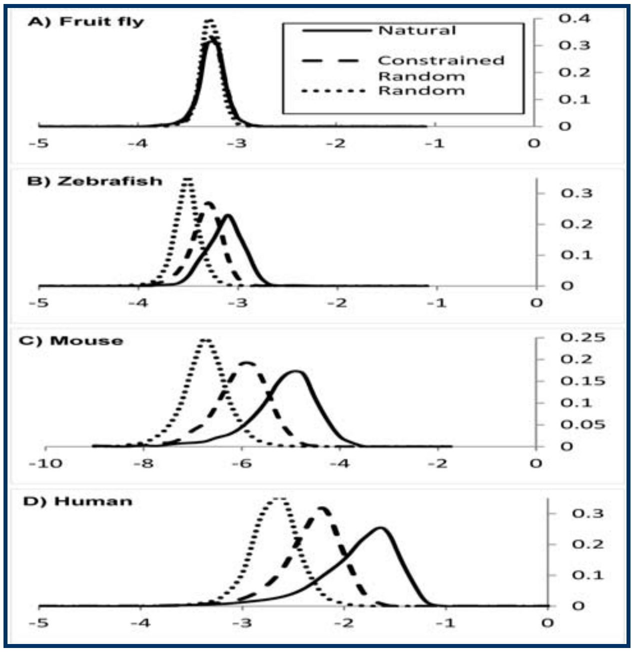

Figure 1: Buffering capacity distributions. Shown are the distribution of buffering capacities (larger x-value = more buffering) for the naturally occurring mRNA sequences (solid line), constrained random sequences (dashed line), and random sequences (dotted line). The species represented are fruit fly (panel A), zebrafish (panel B), mouse (panel C), and human (panel D).

Statistical analysis

For further analysis of the results two different statistical analyses tools are used. A t-test is used to measure the significance of the increased buffering in the natural coding sequences as compared to the random sequences. Second, for each of the four species, we use Cohen’s d measurement to quantify the effect size of codon usage bias in increasing the tolerance to mutations in the natural coding sequences compared to the random sequences and to quantify how much more the constrained random sequences are buffered than the completely random sequences. Note that a direct comparison between the three organisms’ coding sequences would be invalid because the mutation rates are time independent and were computed within every genome separately. The mutation rates are relative values within the same genome only. However, the comparison between the natural coding and the random sequences under the same mutation rates should tell us how much bias is there in the natural DNA structure toward buffering within every genome separately. Therefore, the effect size within every genome should be a valid indicator for comparison between the different genomes.

Discussion:

Our results show that both human and zebrafish DNA are significantly more buffered than random sequences. Buffering is more apparent in human while being absent in fruit fly. Furthermore, random sequences that are constrained to produce the same protein as the natural ones are found to be more buffered than completely random sequences while being less buffered than the natural sequences themselves in both human and zebrafish. This property is also more apparent in human than zebrafish while being absent in fruit fly. These results suggest evidence that bias in the DNA structure toward tolerance to mutations is correlated with life span and is realized on two levels: protein selection and codon usage bias.

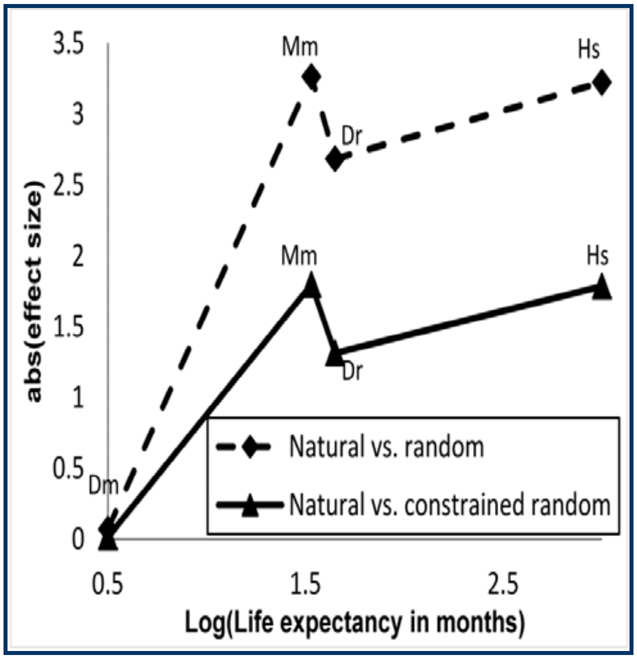

Figure 2: Distribution effect as a function of life expectancy. Dm: *Drosophila melanogaster* (fruit fly), Mm: *Mus musculus* (mouse), Dr: *Danio rerio* (zebrafish), Hs: *Homo sapiens* (human).

Distributions of buffering capacities for fruit fly, zebrafish, mouse, and human are given in **Figure 1**. Note that the x-axis for each of these cannot be compared directly due to the different mutation rates in each of these organisms. As **Figure 1** clearly indicates, a difference in the distribution of buffering capacity for the random, constrained random and natural mRNA sequences cannot be observed in the fruit fly while the separation is evident in zebrafish and more pronounced in the mouse and human.

Analysis of the effect size **Table 3 (see supplementary material)** shows that the separation between the naturally occurring mRNAs and random sequences is greatest in the mouse and in the human. When the effect size is compared to the life expectancy **(Figure 2)** and time to sexual maturity **(Figure 3)**, a clear trend results with an exception occurring with the mouse buffering distribution. This may be due to a difference in the method for calculating mutational rates within the mouse.

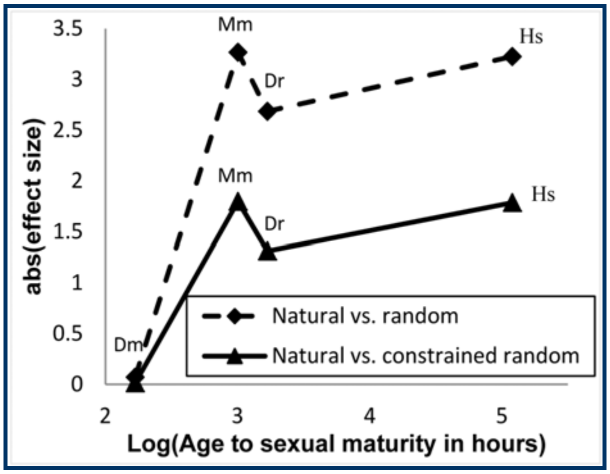

Figure 3: Distribution effect as a function of age to sexual maturity. Dm: *Drosophila melanogaster* (fruit fly), Mm: *Mus musculus* (mouse), Dr: *Danio rerio* (zebrafish), Hs: *Homo sapiens* (human).

In order to provide a more accurate measure of the relationship between life expectancy and codon buffering, a more thorough analysis is needed for organisms with a wider range of life expectancies and age to sexual maturity. Nonetheless, the presented results demonstrates an overall significant correlation between buffering capacity to mutations as a result of codon usage bias and both the life expectancy and the time to sexual maturity of the corresponding organisms. To our knowledge, this is the first time such a hypothesis has been tested. Our analysis also shows that the optimization toward higher tolerance to point mutations is accomplished at both levels the codon usage and the sequence of codons themselves. These results add a new additive explanation to current theorems about codon usage bias in the natural transcripts of these organisms. Finally, as more organisms get sequenced and as more accurate and comprehensive studies of mutation rates become available, we hope to continue this work on a larger scale, with additional phylogenetic information included as well.

Acknowledgements:
Support provided by NIH NCRR grants P20RR16481 and P20RR16481S1. The contents of this manuscript are solely the responsibility of the authors and do not represent the official views of NCRR or NIH.

References:

[1] Eyre-Walker AC, *J Mol Evol.* 1991 **33**: 442 [PMID: 1960741].
[2] Ikemura T & Wada K, *Nucleic Acids Res.* 1991 **19**: 4333 [PMID: 1886761].
[3] Grantham R *et al. Nucleic Acids Res.* 1980 **8**: r49 [PMID: 6986610].
[4] Karlin S & Mrazek J, *J Mol Biol.* 1996 **262**: 459 [PMID: 8893856].
[5] Akashi H. *Genetics.* 1994 **136**: 927 [PMID: 8005445].
[6] Bulmer M. *Nucleic Acids Res.* 1990 **18**: 2869 [PMID: 2190183].
[7] Lavner Y & Kotlar D, *Gene* 2005 **345**: 127[PMID: 15716084].
[8] Urrutia AO & Hurst LD, *Genetics* 2001 **159**: 1191 [PMID: 11729162].
[9] Urrutia AO & Hurst LD, *Genome Res.* 2003 **13**: 2260 [PMID: 12975314].
[10] Hanai R & Wada A, *J Mol Evol.* 1988 **27**: 321 [PMID: 3146642].
[11] Tazi J & Bird A, *Cell* 1990 **60**: 909 [PMID: 2317863].
[12] Sharp PM & Matassi G, *Curr Opin Genet Dev.* 1994 **4**: 851[PMID: 7888755].
[13] Andersson SG & Kurland CG, *Microbiol Rev.* 1990 **54**: 198 [PMID: 2194095].
[14] Chamary JV & Hurst LD, *Genome Biol.* 2005 **6**: R75 [PMID: 16168082].
[15] Smith TF *et al. Mol Biol Evol.* 1985 **2**: 390 [PMID: 3870868].
[16] Mahdi RN & Rouchka EC, *Proc IEEE Int Symp Signal Proc Inf Tech.* 2008 **2008**: 29 [PMID: 20622995].
[17] Haig D & Hurst LD, *J Mol Evol.* 1991 **33**: 412 [PMID: 1960738].
[18] Freeland SJ & Hurst LD, *J Mol Evol.* 1998 **47**: 238 [PMID: 9732450].
[19] Gilis D *et al..Genome Biol.* 2001 **2**: RESEARCH0049 [PMID: 11737948].
[20] Arndt PF *et al.J Comput Biol.* 2003 **10**: 313 [PMID: 12935330].
[21] de Magalhaes JP *et al.. Aging Cell* 2009 **8**: 65 [PMID: 18986374].
[22] Saxonov S *et al. Nucleic Acids Res.* 2000 **28**:185 [PMID: 10592221].

Edited by P Kangueane

Citation: Mahdi & Rouchka, Bioinformation 8(3): 158-162 (2012**)**

Supplementry material:

Methodology:

Buffering capacity

Buffering capacity refers to the tendency of a coding sequence to allow for point mutation events that result in a sequence of the same or similar amino acids, as opposed to missense or nonsense mutations. In order to quantify this property, we have developed a mathematical model that approximates the buffering capacity for an individual sequence by taking into account the codon composition, individual nucleotide mutation rates, and a distance matrix between amino acids that reflects the resulting damage caused by individual amino acid substitutions.

For a given DNA or mRNA coding sequence $S = s_1 s_2 s_3 \ldots\ldots s_n,\ \ n = 3 \times m$, where each s_i is a nucleic acid residue and m is the number of codons, the buffering capacity is estimated by:

$$B(S) = \frac{1}{m}\sum_{i=1}^{n}\sum_{j\neq i}^{4} P\left(s_j \mid L_{s_i}, s_i, R_{s_i}\right) \times D\left(C_{s_i}, C_{s_j}\right) \qquad (1)$$

In (1), $P\left(s_j \mid L_{s_i}, s_i, R_{s_i}\right)$ is the probability that the nucleotide s_i will mutate to s_j given that L_{s_i} and R_{s_i} are the left and the right nucleotides of s_i respectively in the given sequence *S*. C_{s_i} is the codon in which the nucleotide s_i occurs while C_{s_j} is the same codon where s_i is replaced by s_j. $D\left(C_{s_i}, C_{s_j}\right)$ is the distance between the two codons C_{s_i} and C_{s_j}

Here, we define the distance between two codons as a measure of the difference between the encoded amino acids and to approximate this distance, we use the amino acid similarity matrix proposed in [19] as follows:

$$D\left(C_{s_i}, C_{s_j}\right) = max(Sim) - Sim\left(A\left(C_{s_i}\right), A\left(C_{s_j}\right)\right) \qquad (2)$$

In (2), $A\left(C_{s_i}\right)$ and $A\left(C_{s_j}\right)$ are the amino acids encoded by the codons C_{s_i} and C_{s_j} respectively. $max(Sim)$ is the maximum similarity value in the same similarity matrix. If the new codon C_{s_j} is a stop codon (nonsense mutation), the distance will be considered three times the highest distance given by (2). Note that other values for the nonsense mutation distance produce similar results (not shown).

As an approximation of the neighbor dependent mutation probability $P\left(s_j \mid L_{s_i}, s_i, R_{s_i}\right)$, we assume the probability is dependent on a function of time period t and proportional to the relative neighbor dependent mutation rate $R\left(s_j \mid L_{s_i}, s_i, R_{s_i}\right)$. Since we are comparing natural coding against random DNA, replacing the mutation probabilities by the relative missing period mutation rates should have no effect on the significance test and the effect size measurement since both statistical measures are scale independent

Table 1: Life expectancy and sexual maturity rates

Species	Life expectancy (years)	Sexual maturity (days)
fruit fly	0.3	7
zebrafish	3.5	70
mouse	3	42
humans	66	5000

Table 2: Number of mRNA sequences used

Species	mRNAs
fruit fly	2,479
zebrafish	1,294
mouse	9,476
human	16,016

Table 3: Buffering capacity distribution effect size

Species	Natural vs. Constrained random	Natural vs. random
fruit fly	-0.0141	-0.0707
zebrafish	1.3126	-2.6846
mouse	1.7976	-3.2664
human	1.7849	3.2245