Published in final edited form as:

Proc IEEE Int Symp Signal Proc Inf Tech. 2008 December 16; 2008: 29–34. doi:10.1109/ISSPIT.2008.4775640.

Evidence of bias towards buffered codons in human transcripts

Rami N. Mahdi and

University of Louisville, Department of Computer Engineering and Computer Science,
ramimahdi@yahoo.com

Eric C. Rouchka

University of Louisville, Department of Computer Engineering and Computer Science,
eric.rouchka@louisville.edu

Abstract

Codon usage bias is well established in different species from bacteria to mammals. A number of models have been proposed to show this bias as a balance between mutation and selection. Most of these models emphasize controlling the speed of protein translation from the mRNA and increasing the accuracy where this bias is dependent on the abundance and properties of the available tRNA. In this work, codon usage bias in general is considered from a different angle based on a new hypothesis where selection is expected to act in a direction to favor codons that are more buffered, or protected, from mutation than those more sensitive to mutation. It is anticipated that the more buffered the original coding sequence, the higher the survival chance for the whole organism since the resulting protein sequence remains unchanged. Two different complementary measures are developed to compute the average buffering capacity in a given sequence. We show that the buffering capacity of coding sequences in humans is in general higher than that of randomly generated sequences and that of shifted reading frames. Highly expressed genes are shown to have an even higher buffering capacity than non-housekeeping genes. This result is to be expected due to the necessity of housekeeping genes.

1. Introduction

Deoxyribonucleic Acid (DNA) is composed of four building blocks (A, C, G, T). 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 which are the bases for all proteins. Every amino acid can be synthesized by as few as one or as many as six synonymous codons. Synonymous mutations occur when a change in the codon sequence results in a second codon coding for the same amino acid. Synonymous mutations are thought to be neutral and unaffected by natural selection [1;2]. An intuitive assumption was that codons and amino acids were evenly used through the 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 [3-5]. It is expected 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.

Bias has been reported on nucleotide abundance in the three codon positions, dinucleotide usage, and di-codon usage within coding regions [6]. The authors reported that G is the most frequent nucleotide at codon position 1 (32%) while A is the most frequent at position 2. GA is found to be unusually high at position (1,2) reflecting the high usage of acidic amino acids. TA is the most infrequent dinucleotide beginning at position 1, reflecting the avoidance of stop codons. In addition, CG and TA are strongly avoided at position (3,4)

(inter-codon junction). The di-codon GTC.GAA was found to be highly underrepresented while the di-codon GCG.GCG was frequent.

Bias in codon usage has been attributed to the selection towards a more efficient translation model. This suggests codon usage in highly expressed genes is biased toward optimal codons corresponding to the more abundant tRNAs. Such models take in consideration both the elongation rate and the translation accuracy [7;8]. In more recent studies [9;10], it has been shown that the codon usage bias has a U-shaped relationship with the expression level. It turns out a greater bias is found in highly and lowly expressed genes. Highly expressed genes favor the codon corresponding to the more abundant tRNAs resulting in more efficient translation while lowly expressed genes favor the less abundant tRNAs. Another demonstration of the selection-translation model was shown where highly expressed genes in humans are found to be shorter in general and contain less intronic content [11]. 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 [12;13], tissue or organelle specificity [14], increasing mRNA stability, transcription rate [15;16] and evolutionary age [17].

None of the previously mentioned factors manage to provide a complete explanation for codon 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. We propose a different force that could play a role in selection toward the current bias.

Point mutations are events occurring at a single nucleotide as a result of an insertion, deletion, or substitution. 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). Since mutational events are going to occur, it is only logical that natural selection will favor substitution mutations in codons resulting in the same or similar amino acids. A consequence of this preference would be an observation of codons that can accept point mutations without drastically changing the resulting amino acid. We refer to the property of codons to sustain these mutations as its buffering capacity.

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 [18-20]. In [20], the authors has shown that only two in a billion randomly generated mappings would provide better error minimization than the traditional genetic code.

A previous attempt [21] to attribute the bias in the usage of codons to minimizing translation errors has failed and unexpectedly has shown that the natural bias of codons usage increases the sensitivity to errors or mutations. The authors had used generalized mutation rates (transition/transversion) and applied them to estimate the sensitivity to errors in different genes in the three families (Archaea, Bacteria and Eukaryotes). In contrast, through this paper we use specific mutation rates published recently for human and apply them to human coding regions.

We have developed two complementary measures to calculate the buffering capacity of a given DNA sequence. Both methods use the nucleotide substitution rates observed in humans to evaluate the possibility and the consequences of the mutations in the given

sequence. The first measure considers the probability of nonsense mutations occurring within a codon while the second measures the probability of missense mutations, taking into consideration favorable versus non-favorable amino acid substitutions. Our results indicate a buffering capacity greater in the reading frame of known gene sequences than randomly generated sequences. In addition, this capacity is greater than the buffering capacity for the same sequences shifted by one or two nucleotides. A separate observation shows a difference in buffering capacity between housekeeping and non-housekeeping genes in the human genome.

2. Methods

Two measures combining the mutational probability and the consequence of the mutations were constructed in order to derive an estimate of the buffering capacity and mutational tolerance in a given coding region. The first buffering measure estimates the average probability of codons in a given sequence mutating into stop codons, thus producing nonsense mutations. The buffering capacity, B_1 , of a given codon C_i is computed using:

$$B_1(C_i) = \sum_{j=1}^9 \text{Cost}(C_i, C_i^j) \cdot P_{ij} \quad (1)$$

where

$$\text{Cost}(C_i, C_i^j) = \begin{cases} -1 & \text{if } C_i^j \text{ is a stop codon} \\ 0 & \text{otherwise} \end{cases}$$

where C_i^j are the codons that can be derived from C_i by a single nucleotide mutation j with a probability of P_{ij} . Nucleotide substitution rates recently published [22] were used for the values of P_{ij} . These rates were gathered by looking at pseudogene sequences originating as copies from ribosomal protein coding genes. Since the pseudogene mutations are not within coding sequences, they are thought to be unaffected by selective pressure.

Each nucleotide in a codon has the potential to mutate to one of the three other nucleotides, discounting back-mutations. Since there are three positions in a codon, there are a total of nine different codons that can result from single point mutations. The buffering capacity for a given sequence, Seq , is then calculated as the average codon score using:

$$B_1(Seq) = \frac{1}{n} \sum_{i=1}^n B_1(C_i) \quad (2)$$

Where n is the number of non-overlapping codons in the given sequence, Seq . Therefore, all possible single point mutations are considered for every codon in the sequence. The occurrence of a stop codon mutation is given a penalty of -1 weighted by the probability of the corresponding nucleotide substitutions occurring. As a result of equations (1) and (2), a high score corresponds to a higher buffering capacity against nonsense mutations for a given sequence.

The second buffering measure estimates the average cost of a codon mutating to another codon by taking into account the resulting changes in the amino acid. To make an estimate of the amino acid change, a similarity matrix between amino acids derived in [20] is

employed. 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 at all positions in a set of 141 different proteins. The buffering of a given codon C_i is computed using:

$$B_2(C_i) = \sum_{j=1}^9 \text{Sim}(C_i, C_i^j) \cdot P_{ij} \quad (3)$$

And the buffering capacity of a sequence becomes:

$$B_2(\text{Seq}) = \frac{1}{n} \sum_{i=1}^n B_2(C_i) \quad (4)$$

In equation (3), Sim is the similarity score between the amino acid coded by the codon C_i and the amino acid coded by the codon C_i^j as given by the employed similarity matrix described above. P_{ij} is the probability of a single nucleotide mutation j mutating C_i into C_i^j . Afterward, the buffering of a given sequence is computed as the average score of its non-overlapping codons.

These two measures are considered independently since the cost of nonsense to missense mutations is not inherently known. In addition, the cost of nonsense mutations is likely to have an additional complexity as a result of the length positioning (i.e. one near the end of the protein will still produce an almost complete product, while one at the beginning will result in a severely truncated protein). This further complexity is not considered in the given equations.

The similarity matrix employed in (3) and described in more detail in [20] gives similarity values ranging from -5 as most dissimilar to +7 as most similar which is the similarity of the amino acid to itself. As a result, a mutation causing a change to the same amino acid or similar one will increase the score in (3) while mutations to dissimilar amino acids will decrease it. Therefore, a high score in (3) represents high buffering capacity while a low score indicates sensitivity to missense mutations.

3. Coding Sequence Data

For the purpose of this study, human gene sequences from GenBank build 35.1 [23] were downloaded from the human Exon-Intron Database (EID) [24] from the website <http://hsc.utoledo.edu/bioinfo/eid/>. 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.

A total of 16,016 genes remained for consideration. Of these, 439 were classified as housekeeping genes according to [25]. Housekeeping genes are those genes involved in routine processes that are constitutively expressed. This list of 439 genes is compared to the remaining set, since housekeeping genes are involved in essential processes and are continuously transcribed.

4. Results

Random sequences were generated corresponding to each of the 16,016 gene sequences studied. These random sequences were given the same length as the gene, and were further constrained to translate into the exact same amino acid sequence. Buffering capacity scores were then calculated for each of these sequences using equations (2) and (4).

Gene coding regions show a slight propensity to be buffered against nonsense mutations when compared to the corresponding random sequence (Figure 1). A sequence-by-sequence comparison shows that 77% of the time, the coding sequence has a higher buffering capacity than its equivalent randomly generated sequence. Furthermore, a statistical paired t-test at ($P < 0.05$) has shown the difference between the two populations to be significant.

A comparison of the buffering capacity of the 439 housekeeping genes to the rest of the genes reveals a propensity for housekeeping genes to have higher buffering capacity in general (Fig. 2). The difference between the two distributions is found to be significant by two samples statistical t-test at ($P < 0.05$).

When considering the buffering capacity against missense mutations in all genes, the distribution is shown to have a greater variance for gene coding regions (Fig. 3). Gene coding regions show a greater buffering capacity than their random equivalents 65% of the time. If only housekeeping genes are considered, this percentage increases to 85%. This shows that not only are housekeeping genes more buffered against missense mutations than their random equivalents, they are also more buffered than non-housekeeping genes (Figure 4). Furthermore, figure 3 shows the buffering of random sequences that were generated with all even codons and amino acid usage. It reveals that random sequences with the same amino acid usage as the natural sequence has a higher buffering than random sequences with evenly distributed amino acid usage indicates that the bias in amino acid usage by itself increases the buffering capacity. A statistical paired t-test with ($P < 0.05$) has shown that the differences between the three distributions are significant.

In the second set of experiments, only the actual reading frame in the house keeping genes was considered in calculating the buffering capacity. This capacity is then compared against two alternative reading frames shifted by one and two nucleotides. The results are very striking, showing that the actual reading frame is far more buffered than the shifted frames contained within them (Fig. 7 and 8). We find that 84.7% of the time the actual reading frame has a higher buffering capacity than both of the two alternative frames for nonsense mutations (Fig. 7). 84.5% of the time, the actual reading frame has a higher capacity in housekeeping genes for missense mutations (Fig. 8). Both of these results illustrate a buffering capacity far beyond that of their random equivalents.

5. Discussion

For further analysis, statistical tests were performed to measure the significance of the difference between the means. Paired t-tests were performed to compare the buffering of natural sequences against random sequences and the actual reading frame against the two shifted frames in non-sense and missense mutations buffering separately. The difference was significant in all cases with ($P\text{-Value} < 0.05$). A paired t-test was also performed to compare buffering capacity in random sequences with the same amino acid usage bias as the natural sequences against all random sequences. Surprisingly the difference in the buffering to missense mutation was found to be significant while the difference in the non-sense mutation buffering was not significant with ($P\text{-Value} < 0.05$).

A two sample t-test was performed to compare the housekeeping genes against the rest of the genes in both buffering cases and the difference was also found to be significant with (P-Value<0.05).

The buffering capacity in coding sequences is not always higher than that of randomly generated sequences. This could be due to some other factors playing a role in the selection process. However, the fingerprint of the proposed selection scenario is obvious, especially when comparing the actual reading frame to the two shifted reading frames.

The increased buffering capacity and mutational tolerance found in coding regions provides strong evidence that selection favors codons with a higher buffering. This in turn plays a role in codon usage bias. Furthermore, the higher buffering found in highly expressed, housekeeping genes is consistent with our hypothesis. This is to be expected, since the functionality of those genes is more critical in the day-to-day functioning of the organism.

Acknowledgments

The authors would like to thank the members of the University of Louisville Bioinformatics Laboratory as well as the members of the UofL Bioinformatics Journal Club for important insights and feedback. We would also like to thank the anonymous reviewers for their time and efforts.

This project was made possible by NIH – NCRR grant P20RR16481 and NIH – NIEHS grant P30ES014443. Its contents are solely the responsibility of the authors and do not represent the official views of NCRR, NIEHS, or NIH.

References

1. King JL, Jukes TH. Non-Darwinian evolution 4. *Science* May;1969 164(881):788–798. [PubMed: 5767777]
2. Kimura M. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genet Res* June;1968 11(3):247–269. [PubMed: 5713805]
3. Eyre-Walker AC. An analysis of codon usage in mammals: selection or mutation bias? *J Mol Evol* Nov;1991 33(5):442–449. [PubMed: 1960741]
4. Ikemura T, Wada K. Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers; relation between nucleotide sequence data and cytogenetic data 3. *Nucleic Acids Res* Aug;1991 19(16):4333–4339. [PubMed: 1886761]
5. Grantham R, Gautier C, Gouy M, Mercier R, Pave A. Codon catalog usage and the genome hypothesis 16. *Nucleic Acids Res* Jan;1980 8(1):r49–r62. [PubMed: 6986610]
6. Karlin S, Mrazek J. What drives codon choices in human genes? *J Mol Biol* Oct;1996 262(4):459–472. [PubMed: 8893856]
7. Akashi H. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* Mar;1994 136(3):927–935. [PubMed: 8005445]
8. Bulmer M. The effect of context on synonymous codon usage in genes with low codon usage bias. *Nucleic Acids Res* May;1990 18(10):2869–2873. [PubMed: 2190183]
9. Lavner Y, Kotlar D. Codon bias as a factor in regulating expression via translation rate in the human genome 2. *Gene* Jan;2005 345(1):127–138. [PubMed: 15716084]
10. Urrutia AO, Hurst LD. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection 7. *Genetics* Nov;2001 159(3):1191–1199. [PubMed: 11729162]
11. Urrutia AO, Hurst LD. The signature of selection mediated by expression on human genes 3. *Genome Res* Oct;2003 13(10):2260–2264. [PubMed: 12975314]
12. Hanai R, Wada A. The effects of guanine and cytosine variation on dinucleotide frequency and amino acid composition in the human genome. *J Mol Evol* 1988;27(4):321–325. [PubMed: 3146642]

13. Tazi J, Bird A. Alternative chromatin structure at CpG islands 1. *Cell Mar*;1990 60(6):909–920. [PubMed: 2317863]
14. Sharp PM, Matassi G. Codon usage and genome evolution 3. *Curr Opin Genet Dev* Dec;1994 4(6): 851–860. [PubMed: 7888755]
15. Andersson SG, Kurland CG. Codon preferences in free-living microorganisms. *Microbiol Rev* June;1990 54(2):198–210. [PubMed: 2194095]
16. Chamary JV, Hurst LD. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol* 2005;6(9):R75. [PubMed: 16168082]
17. Smith TF, Ralph WW, Goodman M, Czelusniak J. Codon usage in the vertebrate hemoglobins and its implications 2. *Mol Biol Evol* Sept;1985 2(5):390–398. [PubMed: 3870868]
18. Haig D, Hurst LD. A quantitative measure of error minimization in the genetic code (vol 33, pg 412, 1991). *Journal of Molecular Evolution* Nov;1999 49(5):708. [PubMed: 10552053]
19. Freeland SJ, Hurst LD. The genetic code is one in a million. *Journal of Molecular Evolution* Sept; 1998 47(3):238–248. [PubMed: 9732450]
20. Gilis, Dimitri; Massar, Serge; Cerf, Nicolas J.; Rومان, Marianne. Optimality of the genetic code with respect to protein stability and aminoacid frequencies. *Genome Biology* 2001;2(0049)
21. Marquez, Roberto; Smit, Sandra; Knight, Rob. Do universal codon-usage patterns minimize the effects of mutation and translation error? *Genome Biology* 2005;6(91)
22. Zhang Z, Gerstein M. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res* Sept;2003 31(18):5338–5348. [PubMed: 12954770]
23. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank 1. *Nucleic Acids Res.* Dec;2007
24. Saxonov S, Daizadeh I, Fedorov A, Gilbert W. EID: the Exon-Intron Database-an exhaustive database of protein-coding introncontaining genes 7. *Nucleic Acids Res* Jan;2000 28(1):185–190. [PubMed: 10592221]
25. Eisenberg E, Levanon EY. Human housekeeping genes are compact. *Trends Genet* July;2003 19(7):362–365. [PubMed: 12850439]

Figure 1. Buffering of all genes from becoming a stop codon. X-axis: buffering capacity (equations 1 and 2); y-axis: total number of observations.

Figure 2. Buffering of housekeeping and non-housekeeping genes from becoming a stop codon. X-axis: buffering capacity (equations 1 and 2); y-axis: total number of observations normalized by the number of genes in the same category.

Fig. 3.
Buffering of gene coding regions from missense mutations. X-axis: buffering capacity (equations 3 and 4); y-axis: total number of observations.

Fig. 4. Buffering of housekeeping gene coding regions and all gene coding regions from missense mutations. X-axis: buffering capacity (equations 3 and 4); y-axis: total number of observations normalized by the number of genes.

Fig. 5. Buffering of housekeeping gene coding regions in all three reading frames from nonsense mutations. X-axis: buffering capacity (equations 1 and 2); y-axis: total number of observations.

Fig. 6. Buffering of housekeeping gene coding regions in all three reading frames from missense mutations. X-axis: buffering capacity (equations 3 and 4); y-axis: total number of observations.