EC MICROBIOLOGY
Research Article

Prevalence and Length of Open Reading Frames Vary Across Randomly Generated Sequences of Different Nucleotide Compositions

Chuan Yang Neo[1] and Maurice HT Ling[1,2]*

[1]*School of Applied Sciences, Temasek Polytechnic, Singapore*

[2]*HOHY PTE LTD, Singapore*

***Corresponding Author:** Maurice HT Ling, School of Applied Sciences, Temasek Polytechnic and HOHY PTE LTD, Singapore.

Received: May 27, 2020; **Published:** June 18, 2020

Abstract

The emergence of open reading frames is an important step in the origination of *de novo* genes. However, the conditions leading to the origination of *de novo* genes is not well-understood. This study aims to determine the effect of nucleotide composition on the length and occurrence of ORFs by examining various ORF parameters using randomly generated sequences from 85 different nucleotide compositions. Our results suggest that various ORF parameters are significant across different nucleotide compositions (p-value < 1E-120). The average length, standard error of the average length, average maximum length and standard error of the average maximum length of ORFs can be moderately predictable ($0.43 < r^2 < 0.59$) by nucleotide compositions. These results suggest that the prevalence and length of ORFs may be function of the underlying nucleotide composition.

Keywords: *Open Reading Frames (ORF); Nucleotide Compositions*

Introduction

Open Reading Frames (ORF) are series of nucleotides with the potential for transcription [1] and translation into peptides. ORFs starts with a start codon and vary in length, from as little as dozens [2,3] to several hundreds of codons, before encountering a stop codon. The shortest ORF that has been found to date consists of 33 nucleotides [4]. Hence it serves as a minimum benchmark when determining whether a sequence is capable of encoding for functional proteins [5].

However, the conditions leading to the origination of *de novo* genes is not well-understood [6], which are novel genes that arises from ancestrally non-coding DNA sequences. Compared with older genes, *de novo* genes are shorter [7], consist of fewer exons, coupled with low expression levels [8] and tissue- and condition-specificity [9].

This suggests that the length of the ORFs could be regulated by certain factors. Thong-Ek., *et al.* [10] suggested that an average of 196 ORFs was found in every 10 kilobases of random nucleotides. In another study [5], it was found that there was an average of 184 ORFs per 10 kilobases of random nucleotide sequences. Although the occurrences of ORFs are not substantially different, Thong-Ek., *et al.* [10] and Kwek., *et al.* [5] used different nucleotide compositions to generate random sequences. This suggests that the occurrence of random ORFs may be an effect of nucleotide composition as it is plausible that nucleotide composition may affect the prevalence of start and stop codons.

Aim of the Study

This study aims to determine the effect of nucleotide composition on the length and occurrence of ORFs. By generating sequences for 85 different nucleotide compositions with ten thousand kilobases per sequence, our results suggest that various ORF parameters are significant across different nucleotide compositions. The average length, standard error of the average length, average maximum length and standard error of the average maximum length of ORFs can be moderately predictable by nucleotide compositions.

Methods

Random nucleotide sequence generation and ORF identification

A total of 85 sets of nucleotide composition, each consisting of 10 replicates, were generated using RANDOMSEQ [11]. 10% to 70% of the four nucleotides, with increment of 10%, were used to generate the set of 85 possible nucleotide compositions. A set of potential ORFs from the generated random sequences were identified using SeqProperties [12] where each were between 33 nucleotides, which corresponds to the shortest ORF known [4] and 105000 nucleotides. Three alternative start codons (TTG, CTG and ATG) for *Escherichia coli* [13] and the three stop codons in standard genetic code (TAA, TAG and TGA) were used.

Statistical analysis of ORFs

The average numbers, average lengths and average maximum lengths of ORFs identified across various nucleotide compositions were statistically analyzed using Microsoft Excel®. Regression tests were also performed to determine the relationships between the composition of nucleotides and the various properties of ORFs.

Results and Discussion

Frequencies of ORFs are significant across nucleotide compositions

Our results show that the average number of ORFs vary with the change in nucleotide compositions (Figure 1) and the differences in average ORF counts are significant (F = 449.8, p-value < 1E-260). Of the 85 nucleotide compositions, the lowest average count of ORFs is 253.0 for composition 10/10/70/10 (10% adenine, 10% thymine, 70% guanine, and 10% cytosine), while the largest average count is 792.4 ORFs for composition 10/50/30/10 (10% adenine, 50% thymine, 30% guanine, and 10% cytosine). In particular, the average number of ORFs increases as the concentration of guanine increases, with a peak was seen at 30 - 40% concentration, followed by a decreasing trend. In a study [14], it was found that the total number of random ORFs significantly decreases with increasing guanine and cytosine (GC) content. This was based on the relationship that a higher GC content leads to a lower probability of stop codons present in the genome. As a result of lesser stop codons, the length of ORFs present would be longer and lower in numbers [15]. This suggests the possibility that the frequencies of ORFs may be affected by nucleotide compositions [16], particularly the concentrations of guanine and cytosine. However, the average ORF counts cannot be predicted by the composition of nucleotides (Table 1, r^2 = 0.0035, p-value = 0.983). This suggests that nucleotide composition may not directly affect the frequencies of ORFs; instead, nucleotide composition may indirectly affect the frequencies of ORFs via its effect on codon usage [17,18].

Regression Equation	r^2	F	p-value
Average ORF count = 527.5903 + 0.147302 (%T) - 0.30107 (%G) - 0.39996 (%C)	0.0035	0.097	0.983
ORF Length = 321.002 - 3.35309 (%A) - 3.30093 (%T) + 0.071138 (%C)	0.0900	14000	< 1E-260
Average ORF Length = -32.0608 + 0.043555 (%T) + 3.814364 (%G) + 3.903749 (%C)	0.5799	37.268	1.4E-17
Standard Error of Average ORF Length = -13.0857 + 0.027793 (%T) + 0.536654 (%G) + 0.505394 (%C)	0.4305	20.411	1.3E-11
Average Maximum ORF Length = -451.355 - 0.2481(%T) + 26.63726 (%G) + 26.8575(%C)	0.5925	39.262	3.7E-18
Standard Error of Average Maximum ORF Length = -128.251 - 0.39923 (%T) + 6.814077 (%G) + 6.396495 (%C)	0.5409	31.813	7.4E-16

***Table 1:** Regression analysis. ANOVA is performed to test the null hypothesis of no correlation (r^2 = 0).*

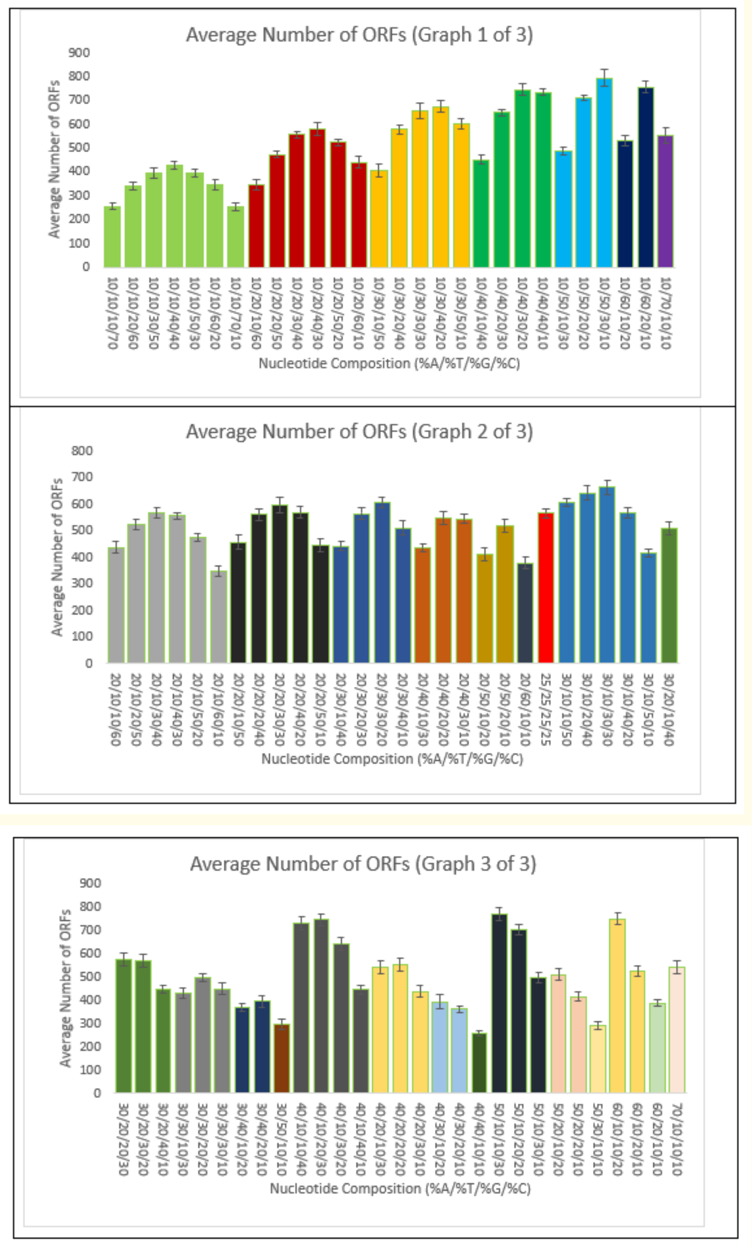

Figure 1: *Average frequencies of ORFs across nucleotide compositions. Error bars denotes standard error of ORF counts.*

Length of ORFs are significant across nucleotide compositions

Our results show that the average length of ORFs (Figure 2; F = 206.3, p-value < 1E-260) and the average maximum length of ORFs (Figure 3; F = 58.6, p-value = 1.2E-120) vary significantly with the change in nucleotide compositions. There appeared to be a decreasing trend in the average length and average maximum length of ORFs as overall adenine and thymine (AT) content increased. This could be due to a decrease in GC content, which leads to a lower probability in the number of stop codons present, resulting in an increase in the length of ORF [15]. A similar result was shown by McCoy., *et al.* [14] where 75% of the ORFs were shorter than 15 codons in a genome of 78.6% AT content while another genome with 24.1% AT content had 75% of its ORFs shorter than 195 codons.

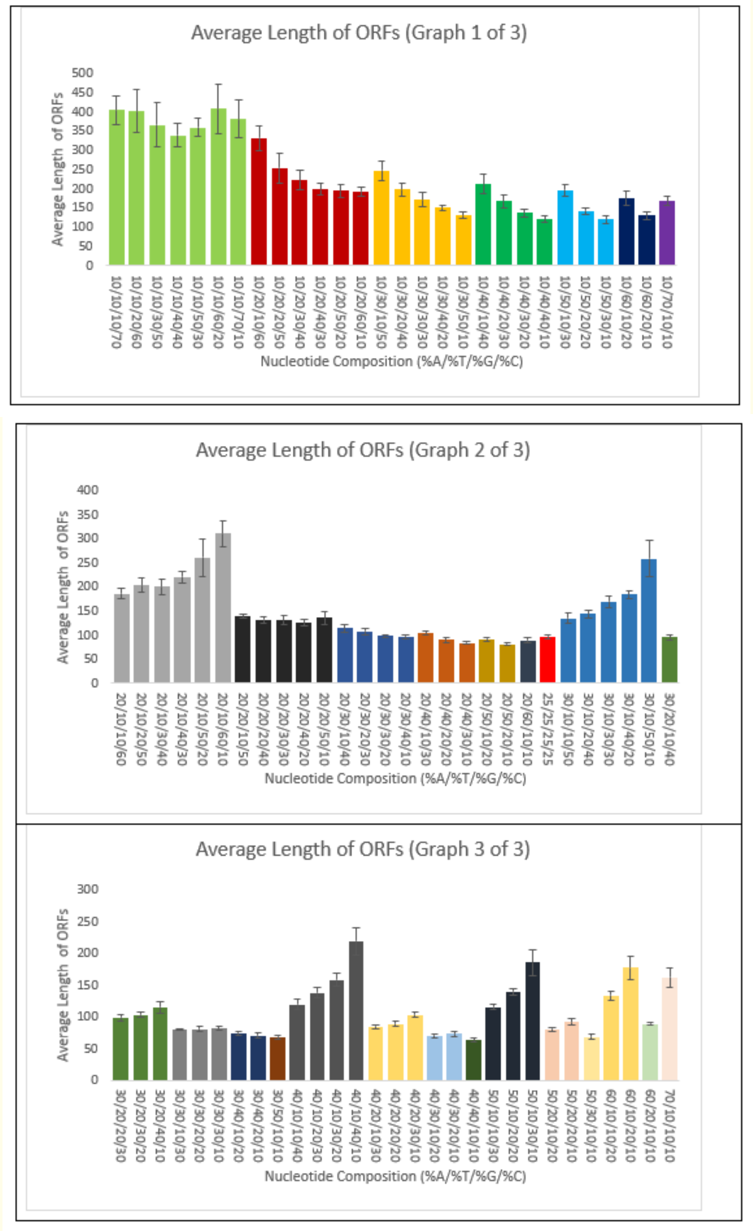

Figure 2: *Average lengths of ORFs across nucleotide compositions. Error bars denotes standard error of ORF lengths.*

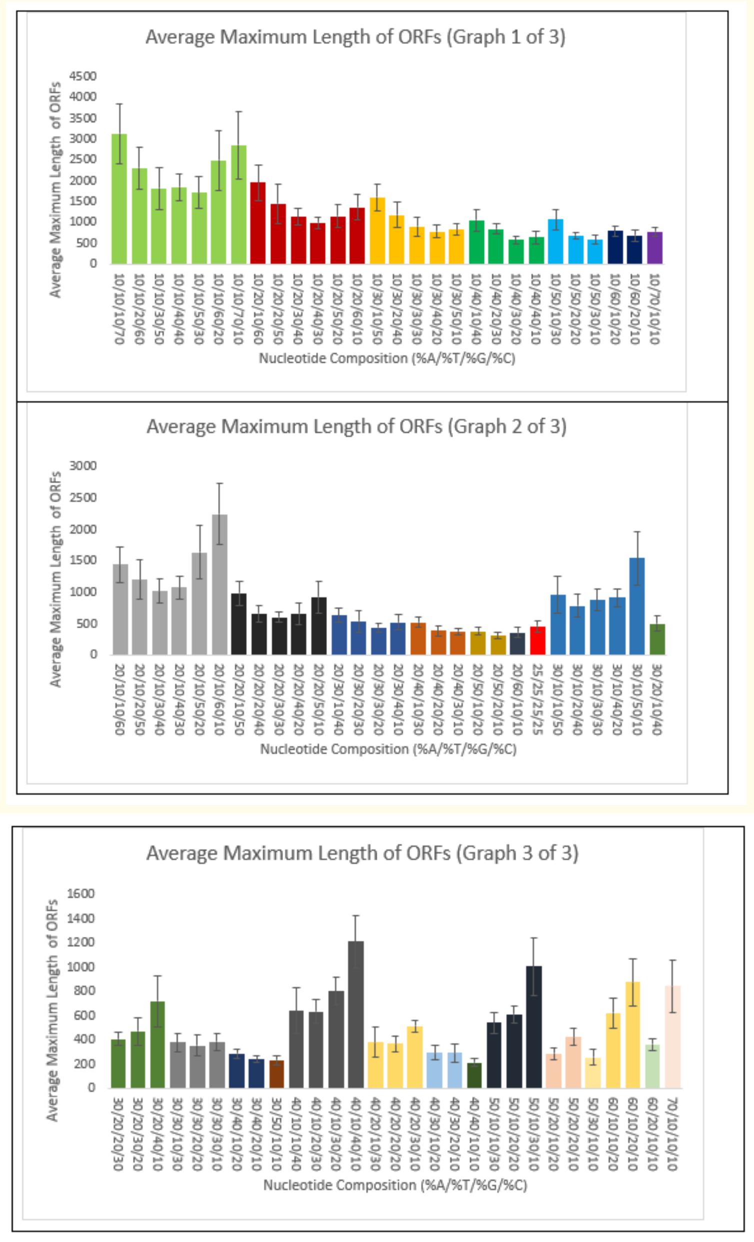

Figure 3: *Average maximum lengths of ORFs across nucleotide compositions. Error bars denotes standard error of maximum ORF lengths.*

Our results also show the ratio of adenine and thymine may interact with the ratio of guanine to cytosine to have an impact on the average ORF length (Figure 2) as adenine to thymine ratio of more than 1 increases average ORF lengths with increasing guanine to cytosine ratio. On the other hand, the average ORF lengths decreases with increasing guanine to cytosine ratio when the adenine to thymine ratio of less than 1. This trend is not seen in average maximum ORF length (Figure 3). This suggests the need to account for each nucleotide rather than GC versus AT content.

Regression analyses show that 4 parameters are moderately predictable ($0.43 < r^2 < 0.58$, $F > 20.4$, p-value < 1E-11) from the nucleotide compositions (Table 1); namely, the average and standard error of the average length of ORFs, as well as the average and standard error of the average maximum length of ORFs. This is consistent with studies demonstrating that nucleotide compositions may affect the prevalence of specific codons [17,18]. Although the correlation between individual ORF length is significant from the nucleotide compositions (F = 14000, p-value < 1E-260), its predictive value is low ($r^2 = 0.09$). These results suggest that nucleotide composition is predictive of average ORF lengths rather than individual lengths. Moderate predictability of variance in average ORF lengths also suggests that there can be large variance between individual ORF lengths; thus, making the prediction of individual ORF lengths difficult.

Conclusion

This study confirms previous study that ORFs can emerge from random sequences [5,10]. Our results demonstrate that the prevalence and average lengths of ORFs vary significantly (p-value < 1E-120) across different nucleotide compositions and the average lengths of ORFs is moderately predictable ($0.43 < r^2 < 0.59$) by nucleotide compositions. Therefore, these results suggest that the prevalence and length of ORFs may be function of the underlying nucleotide composition.

Conflict of Interest

The authors declare no conflict of interest.

Bibliography

1. Sieber P., *et al*. “The Definition of Open Reading Frame Revisited”. *Trends in Genetics* 34.3 (2018): 167-170.
2. Cheng H., *et al*. “Small Open Reading Frames: Current Prediction Techniques and Future Prospect”. *Current Protein and Peptide Science* 12.6 (2011): 503-507.
3. Ladoukakis E., *et al*. “Hundreds of Putatively Functional Small Open Reading Frames in Drosophila”. *Genome Biology* 12.11 (2011): R118.
4. Pueyo JI and Couso JP. “The 11-Aminoacid Long Tarsal-Less Peptides Trigger a Cell Signal in Drosophila Leg Development”. *Developmental Biology* 324.2 (2008):192-201.
5. Kwek BZ., *et al*. “Random Sequences may Have Putative Beta-Lactamase Properties”. *Acta Scientific Medical Sciences* 3.7 (2019): 113-117.
6. Van Oss SB and Carvunis A-R. “De Novo Gene Birth”. *PLOS Genetics* 15.5 (2019): e1008160.
7. Wolf YI., *et al*. “The Universal Distribution of Evolutionary Rates of Genes and Distinct Characteristics of Eukaryotic Genes of Different Apparent Ages”. *Proceedings of the National Academy of Sciences of the United States of America* 106.18 (2009): 7273.
8. Palmieri N., *et al*. “The Life Cycle of Drosophila Orphan Genes”. *eLife* 3 (2014): e01311.
9. Zhang J-Y and Zhou Q. “On the Regulatory Evolution of New Genes Throughout Their Life History”. *Molecular Biology and Evolution* 36.1 (2019): 15-27.
10. Thong-Ek C., *et al*. “Potential De Novo Origins of Archaebacterial Glycerol-1-Phosphate Dehydrogenase (G1PDH)”. *Acta Scientific Microbiology* 2.6 (2019): 106-110.
11. Ling MH. “RANDOMSEQ: Python Command–Line Random Sequence Generator”. *Journal of Proteomics and Bioinformatics* 7.4 (2018): 206-208.

12. Ling MHT. "SeqProperties: A Python Command-Line Tool for Basic Sequence Analysis". *Acta Scientific Microbiology* 3.6 (2020): 103-106.

13. Blattner FR., *et al*. "The Complete Genome Sequence of Escherichia Coli K-12". *Science* 277.5331 (1997): 1453-1462.

14. McCoy MW., *et al*. "The Random Nature of Genome Architecture: Predicting Open Reading Frame Distributions". *Plos One* 4.7 (2009): e6456.

15. Mir K., *et al*. "Predicting Statistical Properties of Open Reading Frames in Bacterial Genomes". *Plos One* 7.9 (2012): e45103.

16. Zahdeh F and Carmel L. "Nucleotide Composition Affects Codon Usage Toward The 3'-End". *Plos One* 14.12 (2019): e0225633.

17. Novembre JA. "Accounting for Background Nucleotide Composition When Measuring Codon Usage Bias". *Molecular Biology and Evolution* 19.8 (2002): 1390-1394.

18. Zhang J., *et al*. "Analysis of Codon Usage and Nucleotide Composition Bias in Polioviruses". *Virology Journal* 8.1 (2011): 146.

Volume 16 Issue 7 July 2020