ACTA SCIENTIFIC MICROBIOLOGY (ISSN: 2581-3226)

Volume 2 Issue 11 November 2019

Research Article

De novo Origination of *Bacillus subtilis* 168 Promoters from Random Sequences

Keerthana D Ardhanari-Shanmugam[1,2], Khadija Shahrukh[1,2], Vicnesh B[1,2], Jun Hong Woo[1,2], Chakrit Thong-Ek[1,2], Sharlene Usman[1,2], Brenda ZN Kwek[1,2], Jing Wen Chua[1,2] and Maurice HT Ling[1,2,3]*

[1]*Department of Applied Sciences, Northumbria University, United Kingdom*

[2]*School of Life Sciences, Management Development Institute of Singapore, Singapore*

[3]*HOHY PTE LTD, Singapore*

***Corresponding Author:** Maurice HT Ling, Department of Applied Sciences, Northumbria University, United Kingdom and School of Life Sciences, Management Development Institute of Singapore and HOHY PTE LTD, Singapore.

Received: September 16, 2019; **Published:** October 09, 2019

DOI: 10.31080/ASMI.2019.02.0390

Abstract

How the first promoters may have originated is of evolutionary curiosity. Several studies have shown that new promoters arise by copying over an existing promoter sequence. Although *de novo* origination of promoters has also been suggested, there has been limited evidence. Hence, we investigate the possibility of *de novo* origination of promoters in this study using the model organism *Bacillus subtilis* 168. 10,000 random sequences were generated and alignment to known promoter sequences from *B. subtilis* 168 were used to assess their probability of being putative promoters. Results showed that 380 out of 10,000 random sequences have ≥97% probability. *In silico* evolution was performed to test the possibility of promoter selection using selective pressure and our simulation results suggest that the functionality of a random sequence may increase overtime. Therefore, *de novo* origination of promoters from random sequences is possible.

Keywords: *Bacillus subtilis; In silico; De novo*

Introduction

Promoters are responsible for directing RNA polymerase towards the transcription of genes; hence, are key factors of transcriptional regulation. A typical prokaryotic promoter consists of a Pribnow box and -35 region [1], which lies between 9-18 base pairs (bp) and 35 bp upstream from the start point of transcription. The Pribnow box has a sequence similar to 5′-TATAAT-3′ while the -35 region has a sequence similar to 5′-TTGACA-3′. These consensus sequences are not strictly conserved in prokaryotes; however, the promoters of *Bacillus subtilis* are known to have similar sequences [2]. Initiation factors known as sigma (σ) factors determine the specific promoter sequences RNAP bind to in prokaryotes [1]. Different σ factors are produced in response to specific stimuli; such as, stress, stationary and growth phases, starvation and morphological differentiation [3]; therefore, driving the transcription of necessary genes. Some genes may have more than one promoter, which allows binding of multiple σ factors under different conditions [4].

New promoters may originate either through the mobilisation of existing promoters to upstream of the gene to be expressed or *de novo* from random sequences for the activation of new or silent genes [5,6]. Promoters originating de novo are referred to as *de novo* promoters [5]. Horwitz and Loeb [7] first demonstrated using *Escherichia coli* that chemically synthesised, random sequences may mimic or even promote transcription much more strongly than wild type promoters. This corroborates the recent findings by Yona., *et al.* [5] that randomly generated sequences can produce gene expressions comparable to that of wild type promoters with-

out mutations or with minimal mutations. Hence, it is of evolutionary curiosity as to how the first *B. subtilis* 168 promoter may have originated, which may then undergo mutation and selection into a set of promoters of varying strengths.

In this study, we investigate the possibility of *de novo* origination of promoters from random sequences using the model Gram-positive organism, *B. subtilis* 168. Our results show 3.8% of random sequences have ≥97% probability of functioning as putative promoters and *In silico* evolution of a random sequence with low probability showed an increasing promoter probability over time.

Methods

Baseline promoter sequence data set

In this study, we used the data set of *B. subtilis* promoters by Coelho., *et al* [1]. The data set of 769 promoters, from now referred to as baseline sequences, was used for comparison against the random sequences generated to determine putative promoters.

Random sequence data set

10,000 random sequences were generated to investigate their probability of being putative promoters. The lengths of these random sequences were set between 38 and 93 nucleotides, with nucleotide compositions of 3535 adenine (A), 3218 thymine (T), 1763 guanine (G) and 1485 cytosine (C) per 10,000 bases; in accordance to the length and sequence compositions of *B. subtilis* promoters [1]. Start codons and stop codons were excluded. The random sequence data set of variable sequence lengths was generated using RANDOMSEQ [8].

Pairwise alignment of baseline sequences

The Smith–Waterman algorithm [9], known as 'local' alignment, and the Needleman–Wunsch algorithm [10], known as 'global' alignment, were used for pairwise alignment of the baseline sequences. The distribution of the alignment scores was used as a measure of putative promoter sequences. Pairwise alignment was performed using SEQPROPERTIES in Bactome (https://github.com/mauriceling/bactome).

Pairwise alignment of random sequences against baseline sequences

Each random sequence generated was pairwise aligned with every baseline sequence to determine the probability of putative promoter sequences. Minimum and average pairwise alignment scores were generated for each of the random sequences. The probability of a random sequence being a putative promoter sequence was determined by the proportion of alignment scores of baseline sequences below the minimum and average alignment scores of the random sequences. Minimum alignment score was used for high stringency while average alignment score was used for low stringency. Pairwise alignment of random sequences against baseline sequences was performed using SEQPROPERTIES.

In silico evolution using digital organism simulation environment (DOSE) [11]

In silico evolution was performed to investigate the potential of random sequences to evolve over several generations to achieve characteristics similar to baseline sequences using the method described by Kwek., *et al* [12]. Briefly, a random promoter sequence with a minimum alignment score identical to that of baseline promoter sequences was randomly selected as the ancestral genome. The ancestral genome is used for cloning the initial population of 100 digital organisms and deployed in one eco-cell. A 10% background point mutation rate was used [13,14]. Fitness of each organism was calculated as the average pairwise alignment score of the organism's genome against a data set of 250 randomly selected baseline sequences. Organisms in the lowest 10th percentile by organism fitness were removed from the population after each generation. Alternatively, if less than 50% of the population remained after the removal, 10 random organisms were randomly removed. The remaining organisms were randomly selected for replication to maintain a population of 100 organisms at each generation. A total of 30 simulation repeats were performed and 500 generations were simulated each time.

Results and Discussion

Characterisation of *B. subtilis* 168 Promoters

The minimum and maximum nucleotide lengths for the baseline sequences were determined to be between 38 bp and 93 bp, with an average of 53 bp and standard deviation of 9.56 bp. The average distribution of nucleotides in the baseline sequences was calculated to be 35.35% A, 32.18% T, 17.63% G and 14.85% C.

The pairwise alignment results (Figure 1) of both 'local' and 'global' alignments were identical (p-value = 1.0). This could have been probably due to the short nucleotide lengths of the promoter sequences. Both algorithms yielded a total of 221,445 alignments. The minimum and maximum alignment scores were determined to be 17 and 60 respectively. The average alignment score was 32 with a standard deviation of 4.60.

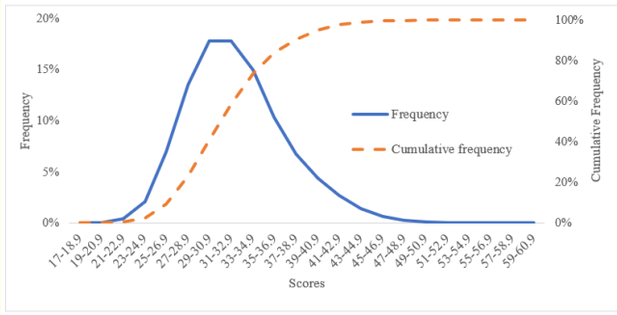

Figure 1: Distribution of pairwise alignment scores of *B. subtilis* 168 baseline sequences.

3.8% Random sequences with more than 97% probability as putative *B. subtilis* 168 promoters

Our results show that approximately all the random sequences (99.7%) have a minimum alignment score that is equal or greater than that of the baseline sequences (Table 1). The range of pairwise alignment scores among baseline sequences represents the sequence diversity of *B. subtilis* 168 promoters. Using the same argument from previous studies [12,17]; if a random sequence is not likely to be a putative promoter, then its minimum alignment score against the baseline sequences should be lower than the minimum alignment score of the baseline sequences. Our results show that almost all (99.9%) the random sequences have 9.5% probability of being putative *B. subtilis* 168 promoters (Table 1). Our results also show that at high stringency (using minimum alignment score), 21 random sequences (0.2%) have 73.5% probability while at low stringency (using average alignment score), 380 random sequences (3.8%) have 97.6% probability (Table 1) of being putative *B. subtilis* 168 promoters.

Threshold score (Baseline)	Minimum fitness score	Average fitness score	Probability of B. subtilis 168 Function
>16.9	99.7%	100.0%	0.0%
>18.9	98.0%	100.0%	0.0%
>20.9	90.5%	100.0%	0.4%
>22.9	77.1%	100.0%	2.5%
>24.9	62.1%	99.9%	9.5%
>26.9	45.1%	96.8%	23.0%
>28.9	25.6%	87.4%	40.8%
>30.9	7.3%	76.3%	58.6%
>32.9	0.2%	64.2%	73.5%
>34.9	0.0%	51.0%	83.9%
>36.9	0.0%	35.9%	90.6%
>38.9	0.0%	19.5%	94.9%
>40.9	0.0%	3.8%	97.6%
>42.9	0.0%	0.0%	99.0%

Table 1: Probability of random sequences functioning as *B. subtilis* 168 promoters.

These results are consistent with Horwitz and Loeb [7] demonstrating that a small portion of chemically synthesised sequences can serve as active promoters. Similarly, Yona., *et al*. [5] had demonstrated that 10% of the randomly generated sequences were able to produce a gene expression comparable to wild-type promoters prior to evolution. Therefore, our findings suggest that *de novo* origination of *B. subtilis* 168 promoters is possible.

We also screened the random sequences with the highest probability for similarity to the consensus sequences. We discovered similarity to the promoter motifs in all 21 sequences with 38% (n = 8) having only a difference of 1 nucleotide in both or either of the consensus sequences. This supports the findings of Yona., *et al*. [5] who discovered that random sequences which were active without mutation and after mutation showed high similarity to the consensus sequences. In another study [15], it was observed that the strength of *Lactococcus lactis* promoters was dependent on both the similarity to promoter consensus sequences and the lengths of the spacer sequences. This suggests that random sequences with similarity to consensus sequences show greater functionality; thus, may be preferentially selected to form strong promoters.

Putative *B. subtilis* 168 promoters can evolve under selective pressure

A random sequence, labelled as Test_9826, which has the minimum alignment score of 17 (average and maximum score of 26 and 37 respectively) to the baseline sequences was selected for *In silico* evolution. Its sequence is "AAAACAAACAATTTACATTTTTTACGTTTATTTTCCATCTCC". Our simulation results (Figure 2) show a sharp increase followed by plateauing in the fitness level of Test_9826. The fitness of the organism increased from a score of 26 to 27.4 (highest average fitness score) and to 28.7 (highest maximum fitness score). The average and maximum fitness scores were used to interpret the fitness of the promoter sequence at high and low stringencies respectively. Based on these results, the probability of Test_9826 being a putative promoter increased from 9.5% to 23% (Table 1). Despite the increase in fitness, both fitness scores do not surpass the average alignment score of the baseline sequences. This could have been due to a low mutation rate or fewer number of simulated generations. Despite Test_9826 not evolved to achieve the average fitness of wild-type promoters, the increase in its fitness level is evident of how random sequences may evolve into highly functional promoters. Our results are consistent with previous findings that evolution increases the functionality of promoters. Yona., *et al*. [5] had demonstrated that 60% of the random sequences evolved expression similar to wild-type promoters with a single mutation.

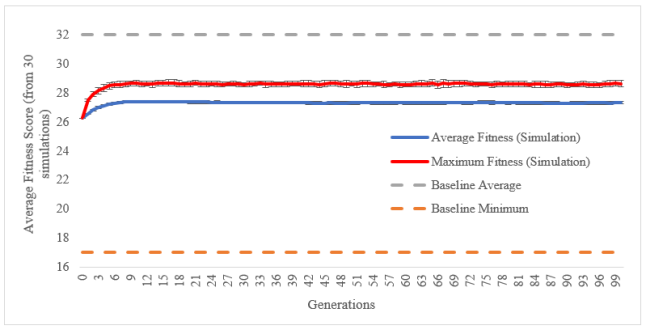

Figure 2: Simulation results for random sequence Test_9628. Error bars represent standard errors.

Conclusion

Carvunis., *et al*. [16] had described that in order for a proto-gene originating *de novo* to be expressed, it must be first transcribed. Therefore, having a functional promoter is an essential step for the expression of genes originating *de novo* [12,17]. The same applies for silent genes acquired through mechanisms like horizontal gene transfer (HGT) or already existing without a compatible promoter [6]. Our results demonstrate the possibility for *de novo* origination of promoters from random sequences; thus, fulfilling the requirement for the transcription of proto-genes and existing genes without a compatible promoter. Since gene expression is also dependent on other events such as recognition of promoters by σ factors, binding of transcription factors and RNA polymerases, future studies could assess gene expression and strengths of random sequences with the highest probability of being putative *B. subtilis* 168 promoters.

Conflict of Interest

The authors declare no conflict of interest.

Bibliography

1. Coelho RV., *et al*. "Bacillus subtilis promoter sequences data set for promoter prediction in Gram-positive bacteria". *Data in Brief* 19 (2018): 264-270.

2. Amaya E., *et al*. "Analysis of promoter recognition in vivo directed by sigma (F) of Bacillus subtilis by using random-sequence oligonucleotides". *Journal of Bacteriology* 183 (2001): 3623-3630.

3. Souza BM., *et al*. "σ (ECF) factors of gram-positive bacteria: a focus on Bacillus subtilis and the CMNR group". *Virulence* 5 (2014): 587-600.

4. Dostálová H., *et al*. "Overlap of promoter recognition specificity of stress response sigma factors SigD and SigH in Corynebacterium glutamicum ATCC 13032". *Frontiers in Microbiology* 9 (2019): 3287.

5. Yona AH., *et al*. "Random sequences rapidly evolve into de novo promoters". *Nature Communications* 9 (2018): 1530.

6. Matus-Garcia M., *et al*. "Promoter propagation in prokaryotes". *Nucleic Acids Research* 40 (2012): 10032-10040.

7. Horwitz MS and Loeb LA. "Promoters selected from random DNA sequences". *Proceedings of the National Academy of Sciences of the United States of America* 83 (1986): 7405-7409.

8. Ling MH. "RANDOMSEQ: Python command–line random sequence generator". *MOJ Proteomics and Bioinformatics* 7 (2018): 206-268.

9. Smith TF and Waterman MS. "Identification of common molecular subsequences". *Journal of Molecular Biology* 147 (1981): 195-197.

10. Needleman SB and Wunsch CD. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". *Journal of Molecular Biology* 48 (1970): 443-453.

11. Castillo CF and Ling MH. "Digital Organism Simulation Environment (DOSE): a library for ecologically-based in silico experimental evolution". *Advances in Computer Science: an International Journal* 3 (2014): 44-50.

12. Kwek BZ., *et al*. "Random sequences may have putative beta-lactamase properties". *Acta Scientific Medical Sciences* 3 (2019): 113-117.

13. Rattray AJ and Strathern JN. "Error-prone DNA polymerases: when making a mistake is the only way to get ahead". *Annual Review of Genetics* 37 (2003): 31-66.

14. Lee DF., *et al*. "Mapping DNA polymerase errors by single-molecule sequencing". *Nucleic Acids Research* 44 (2016): e118.

15. Jensen PR and Hammer K. "The sequence of spacers between the consensus sequences modulates the strength of prokaryotic promoters". *Applied and Environmental Microbiology* 64 (1998): 82-87.

16. Carvunis A-R., *et al*. "Proto-genes and de novo gene birth". *Nature* 487 (2012): 370-374.

17. Thong-Ek C., *et al*. "Potential de novo origins of archaebacterial glycerol-1-phosphate dehydrogenase (G1PDH)". *Acta Scientific Microbiology* 2 (2019): 106-110.

Volume 2 Issue 11 November 2019