Short Communication

Open Access

Pseudomonas balearica DSM 6083^T promoters can potentially originate from random sequences

Abstract

Recent studies and researches have proposed that many genes are plausibly emerged from previously non-coding genomic regions. However, how a promoter can emerge and function properly from *de novo* genes remain debatable as this has not been show in large numbers of organisms. Therefore, this study aims to explore the possibility of *de novo* evolution of a promoter from random sequences by using *Pseudomonas balearica* DSM 6083^T as the model organism. Our result shows that 39.3% of the generated random sequences have 68.6% probability to be a functional promoter. Evolution simulation was carried out to observe the effect of evolution in the putative *P. balearica* promoter over generations. The simulation result proves that selection enhances the functionality of the generated random sequences overtime. Therefore, it is plausible that *P. balearica* promoter could emerge from random sequences, which is consistent with findings from previous studies.

Volume 8 Issue 2 - 2019

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

[#]Equal contributors

Correspondence: Maurice HT Ling, School of Life Sciences, Management Development Institute of Singapore, 501 Stirling Road, Singapore 148951, Republic of Singapore, Singapore, Email mauricelin@acm.org

Received: November 04, 2019 | **Published:** December 31, 2019

Introduction

Genetically, a promoter is a region of DNA which responsible as a key factor in initiating the transcription of genes. In prokaryotes, promoters contain specific response elements as a secure initial binding site for sigma factors and RNA polymerase.[1] A typical prokaryotic promoter consists of two short sequence elements called Pribnow Box and -35 region which locates approximately 10 and 35 nucleotides upstream from the transcription start site (+1 region).[2] New promoters may originate through the mobilization of existing promoters to upstream of the gene or *de novo* from random sequences for the activation of new or silent genes.[3,4] Changes in the promoter sequences are crucial in evolution, indicated by the relatively stable number of genes in many lineages. As an example, most vertebrates have roughly the same number of protein-coding genes, which are often highly conserved in sequence and thus, most of evolutionary change might be originated through changes in the gene expression.[5,6]

Recently, evidence of *de novo* genes has been suggested.[7] Kaessmann et al.,[8] reported that new protein-coding and RNA genes can evolve from non-functional genomic sequences, various types of gene fusion, and even RNA intermediates. This is supported by Andersson and Schlötterer which emphasized that a pre-existing gene can be emerged by modification of the existing gene by divergence or evolved *de novo* from noncoding DNA.[9,10] At the peptide level, Ling[11] suggests that around 27% of the randomly generated amino acid chains may contain putative protein domains. Whilst at the genetic level, Begun and Levine[12] identified *de novo* genes in *Drosophila yakuba* and *Drosophilia erecta*. Subsequently, *de novo* genes had been identified in multiple eukaryotic genomes; such as yeast,[13,14] plants,[15] mammals, primates,[16] and even in the human.[17] Recently, studies by Thong-Ek et al.,[18] and Kwek et al.,[19] suggested the possibility of *de novo* emergence of putative archaebacterial genes and beta-lactamases respectively. A study by Zhang et al.,[20] also reported that an average emergence of 51.5 *de novo* genes per million years in Oryza by studying the genomes of 13 closely related Oryza species and more importantly, 56.6% of the *de novo* genes identified are translated. However, how a promoter can emerge *de novo* remain debatable.[3]

The concept of random sequences lies on the sequence that contains adenosine, cytosine, guanosine, and thymine in equal probabilities composed of no information and represented the non-active sequence space without prejudices. *De novo* evolution of promoters is prevalently using purely random sequences with genomes preferably contain about 50% GC content; such as, the *Escherichia coli* genome with 50.8% GC content. Such genomes are advantageous because random sequences serve as a null model in the functionality test without proposing any perplexing factors due to diverging from the natural GC content of the studied genome.[3]

Pseudomonas balearica is a marine microorganism with methylmercury decomposition capabilities[21] and had been proposed for use in wastewater treatment and bioremediation.[22] Therefore, the promoter of *P. balearica* DSM 6083^T was considered as a worthy subject to investigate further since it contains 64.1 to 64.4% GC content.[23] The study of *de novo* evolution of *P. balearica* DSM 6083^T promoter was conducted through the evaluation of pairwise alignment score of known *P. balearica* promoter against randomly generated sequences. Our results suggest that substantial percentage of randomly generated sequences may exhibit functional properties of *P. balearica* promoter.

Methods

Baseline promoter sequence data sets: The genome sequence of *P. balearica* DSM 6083^T was extracted from DDBJ/ENA/GenBank under the accession number CP007511, which consist of 4,383,480 base pairs.[24] A set of potential promoters, hereafter known as baseline sequences were retrieved from Berkeley Drosophila Genome Project version 2.2 with prokaryote as the type of organism and minimum promoter score of 0.8 for both forward and reverse strands.[25] A set of 10,000 random sequences with an average of 50 nucleotides in length; which has 2300 Adenine, 2550 Thymine, 2710 Guanine, and 2440 Cytosine per 10,000 bases; without start and stop codons within the sequences; were generated using RANDOMSEQ.[26]

Determining putative *P. balearica* promoter from random sequences: Pairwise sequence alignment was done to match the regions in sequences for identifying probable similarities. Two series of pairwise alignments were done by using Smith-Waterman algorithm[27] or also known as local alignment, and Needleman-Wunsch algorithm[28] or also known as global alignment, via SEQPROPERTIES in Bactome (https://github.com/mauriceling/bactome). The distribution of the alignment scores was used as a measure of putative promoter sequences. In the first series, each baseline sequences were pairwise aligned to each of the other baseline sequences and the distribution of scores were used as measure for putative *P. balearica* promoter sequences. In the second series, each of the 10,000 random sequences was pairwise aligned to each baseline sequence. A minimum and average alignment score were generated for each random sequence. Based on the bootstrap statistics,[29] the probability of each random sequences being a putative promoter sequence was determined by the proportion of baseline alignment scores below the minimum and average score of the random sequence for stringent and relaxed criteria respectively.

***In silico* evolution of putative *P. balearica* promoter:** Putative promoters were simulated to go through evolution over 500 generations to investigate whether these sequences could enhance more characteristics of the original promoters. The mutation was carried out using DOSE[30,31] along with previously described methods.[32,33] Initially, a single population of 100 digital organisms was created, deployed in the same ecological cell and simulated for 500 generations. A random sequence with minimum alignment scores just above that of the baseline sequences was used as genome for the ancestral organism, which would be cloned into the initial population of 100 organisms. The overall point mutation rate used was 10%.[34,35] Organism fitness was calculated as average pairwise alignment of its genome to a random selection of 250 baseline sequences (original *P. balearica* promoters). The lowest decile of the organisms by fitness were removed. However, in event where more than 50% of the population were removed, a random selection of 10 organisms were removed instead. A random selection of remaining organisms after removal was replicated to top up the population to 100 organisms for the next generation. The simulation was repeated 30 times.

Results and discussion

Characterization of *P. balearica* promoter: Promoter predictor of complete *P. balearica* genome sequences yielded 4,079 potential promoters within the desired threshold; forming the set of baseline sequences. Excluding start and stop codons, the minimum and maximum nucleotide length for baseline sequences lie between 17 and 50 respectively. The distribution of nucleotide composition of the 4,079 baseline sequences were analyzed and showed to be 23.00% adenine, 24.40% cytosine, 27.10% guanine, and 25.50% thymine. The baseline sequences were pairwise aligned locally and globally and yielded a total of 8,317,081 alignments (Figure 1). The mean alignment score is 30.511 with a standard deviation of 2.015 and a median score of 31. The minimum and maximum scores respectively are 17 and 50 as shown in Figure 2. Since both local[27] and global[28] alignment results are identical, the rest of the subsequent analysis are obtained based on the local alignment only. An analysis of the predicted promoter sequences using sequence logo (weblogo.berkeley.edu) suggests that thymine at position 35 to be critical (Figure 3).

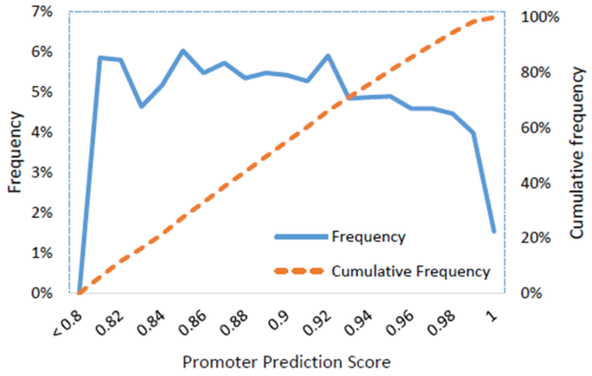

Figure 1 Distribution of pairwise baseline sequence alignment scores. This graph suggests that the probability of predicted promoters is likely to be uniformly distributed.

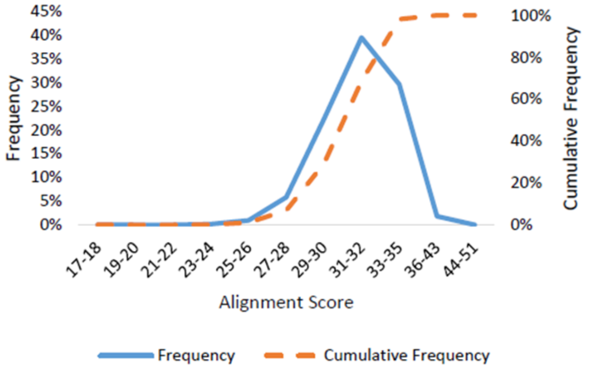

Figure 2 *P. balearica* promoter local pairwise alignment score. Taking pairwise alignment scores, this graph indicates that a minimum score of 17 is indicative of a potential promoter sequence.

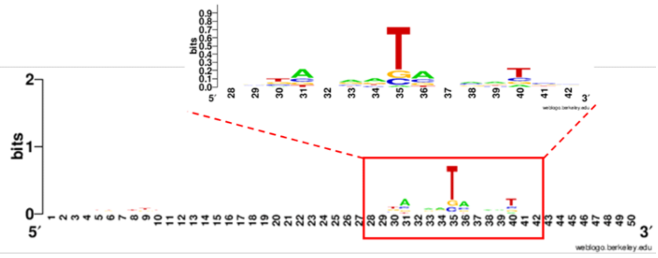

Figure 3 Sequence logo of *P. balearica* promoter suggest thymine at position 35 as the most important.

39% of Random sequences have more than 68% probability to function as a *P. balearica* promoter: Our results that all 10,000 randomly generated sequences have minimum pairwise alignment score equal or higher to the minimum local pairwise alignment score of 17 from baseline sequences. As the range of pairwise alignment scores among baseline sequences represents the sequence diversity of *P. balearica* promoter; hence, the arguments used in previous studies,[18,19,36] if a random sequence is not likely a putative promoter, then its minimum pairwise alignment score with original promoter (baseline sequences) should be lower than the minimum pairwise alignment score among known promoters. This is supported by several studies suggesting correlation between sequence and functional similarities.[37–39] Our results show that all 10,000 randomly generated sequences have at least 6.8% probability of being putative *P. balearica* promoter (Table 1), based on the probability of average pairwise alignment scores. At high stringency level (minimum alignment score), 21.9% of the randomly generated sequences have 1% probability of possessing promoter functionality while at high leniency level (maximum alignment score), 100% of the sequences have at least 98.2% probability of being a putative functional promoter. A possibility can be a left-skew of baseline alignment scores due to short sequences in the baseline set. Despite so and using the same argument, 39.3% of the 10,000 randomly generated sequences have 68.6% probability of being putative *P. balearica* promoter as their pairwise alignment scores are at or above 68th percentile of the baseline pairwise alignment scores (Table 1).

Table 1 Prediction of random sequences probability of functionality. Random sequences minimum, average, and maximum pair wise alignment scores are projected against baseline sequences alignment scores

Threshold score	Minimum score	Average score	Maximum score	Probability of *P. balearica* promoter function
> 17.9	100.00%	100.00%	100.00%	0.000%
> 19.9	99.78%	100.00%	100.00%	0.001%
> 21.9	96.42%	100.00%	100.00%	0.012%
> 23.9	74.31%	100.00%	100.00%	0.131%
> 25.9	21.93%	100.00%	100.00%	1.068%
> 27.9	0.25%	100.00%	100.00%	6.863%
> 29.9	0.00%	99.13%	100.00%	29.166%
> 31.9	0.00%	39.27%	100.00%	68.603%
> 34.9	0.00%	0.00%	100.00%	98.195%
> 42.9	0.00%	0.00%	0.00%	99.998%
> 50.9	0.00%	0.00%	0.00%	100.000%

This is supported by the first study of *de novo* promoters from Horwitz and Loeb back in 1986.[40] Using random sequences generated from *E. coli* promoters (where 19 base pairs have been replaced at the -35-promoter region), they demonstrated that random sequences may mimic or even promote transcription much stronger than the wild-type promoters.[40] This technique is further applied in a study by Yona et al.,[3] where it is reported that around 60% of random sequences have wild-type promoter efficiency with only 1 base mutation out of 103 bases. This is supported by a recent study by Keerthana et al.,[36] suggesting that 380 out of 100,000 random sequences generated from *Bacillus subtilis* promoters have ≥97% probability to behave as functional promoters.

Putative *P. balearica* promoter can evolve under a silico mutation: Next, a putative *P. balearica* promoter is simulated under selective pressure by selecting a random sequence with the lowest average pairwise alignment score with the baseline sequences using silico evolution. The selected sequence, Test_8147, has the sequence length of 56 nucleobases with lowest average score of 18. The simulation results show a possibility that Test_8147 can evolve into a functional promoter as its average maximum score cross the baseline average score of 30.5 at the 1st simulated generation (Figure 4). This is consistent with the previous studies[32,33] which shows a rapid increase in fitness under selective pressure. At the stricter criteria of grand mean (mean of means), the fitness of Test_8147 increases rapidly and surpasses the baseline average score of 30.5 at the 2nd simulated generation. This suggests that Test_8147 reaches the average functionality of a putative *P. balearica* promoter. Based on the grand mean plateau between 33.3 and 33.4, which has more than 68% probability of being putative *P. balearica* promoter (Table 1), from 9th to 500th generation. This is similar to the simulation results from Keerthana et al.,[36] suggesting that the functionality of a random sequence may even increase over time under appropriate selective pressure. This suggests that a putative *P. balearica* promoter may rapidly evolve into a functional promoter under selective pressure.

A study by Carvunis et al.,[41] coined the term "proto-gene" to describe a gene that born from non-genic sequence by random processes without selection, however, it must fulfil 3 criteria; namely, the DNA sequence must be transcribed and translated, and the protein product must be beneficial to the organism. A proto gene is the first stage of a *de novo* gene origin with a beneficial and selectable phenotype where selection pressure can act on.[42] As Yona et al.,[3] suggest that substantial proportion of random sequences can evolve into functional promoters, it is plausible to consider that the requirement for transcription has substantial chance of being randomly fulfilled. As our results illustrate the feasibility of *de novo* origination of putative *P. balearica* promoter from random sequences; hence, the requirement for transcription is fulfilled. More importantly, it is plausible to consider *de novo* origination of *P. balearica* promoters, which is consistent with findings from Yona et al.[3]

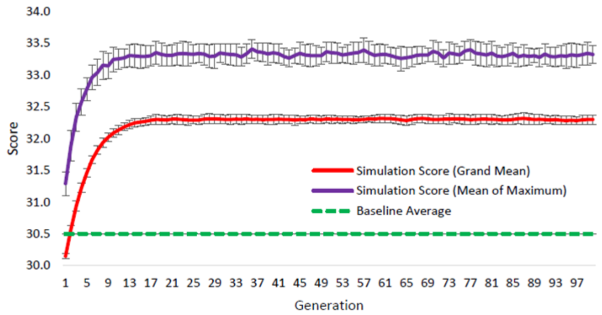

Figure 4 Simulation result of random sequence Test_8147. Error bars represent standard errors.This graph indicates that selective pressure is likely to improve a putative promoter sequence.

Acknowledgements

None.

Conflicts of interest

The authors declare no conflict of interest.

Funding

None.

References

1. Barnard A, Wolfe A, Busby S. Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes. *Curr Opin Microbiol*. 2004;7(2):102–108.
2. Haugen S, Ross W, Gourse R. Advances in bacterial promoter recognition and its control by factors that do not bind DNA. *Nature Reviews Microbiology*. 2008;6(7):507–519.
3. Yona A, Alm E, Gore J. Random sequences rapidly evolve into de novo promoters. *Nature Communications*. 2018;9(1):1–10.
4. Matus-Garcia M, Nijveen H, Van Passel M. Promoter propagation in prokaryotes. *Nucleic Acids Research*. 2012;40(20):10032–10040.
5. Gagniuc P, Ionescu-Tirgoviste C. Gene promoters show chromosome-specificity and reveal chromosome territories in humans. *BMC Genomics*. 2013;14(1):278.
6. Gagniuc P, Ionescu-Tirgoviste C. Eukaryotic genomes may exhibit up to 10 generic classes of gene promoters. *BMC Genomics*. 2012;13(1):512.
7. McLysaght A, Guerzoni D. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2015;370(1678):20140332.
8. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. *Genome Res*. 2010;20(10):1313–1326.
9. Andersson D, Jerlström-Hultqvist J, Näsvall J. Evolution of New Functions De Novo and from Preexisting Genes. *Cold Spring Harb Perspect Biol*. 2015;7(6):17996.
10. Schlötterer C. Genes from scratch–the evolutionary fate of de novo genes. *Trends Genet*. 2015;31(4):215–219.
11. Ling M. De novo putative protein domains from random peptides. *Acta Scientific Microbiology*. 2019;2(4):109–112.
12. Begun D, Lindfors H, Thompson M, et al. Recently evolved genes identified from Drosophila yakuba and D. erecta accessory gland expressed sequence tags. *Genetics*. 2005;172(3):1675–1681.
13. Cai J, Zhao R, Jiang H, et al. De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. *Genetics*. 2008;179(1):487–496.
14. Wu B, Knudson A. Tracing the de novo origin of protein-coding genes in yeast. *mBio*. 2018;9(4):e01024.
15. Donoghue M, Keshavaiah C, Swamidatta S, et al. Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. *BMC Evolutionary Biology*. 2011;11(1):1–23.
16. Xie C, Zhang Y, Chen J, et al. Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. *PLoS Genetics*. 2012;8(9):e1002942.
17. McLysaght A, Guerzoni D. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. Philos Trans R Soc Lond B Biol Sci. 2015;370(1678):20140332.
18. Thong-Ek C, Usman S, Woo J, et al. Potential de novo origins of archaebacterial glycerol-1-phosphate dehydrogenase (G1PDH). *Acta Scientific Microbiology*. 2019;2(6):106–110.
19. Kwek B, Thong-Ek C, Usman S, et al. Random sequences may have putative beta-lactamase properties. *Acta Scientific Microbiology*. 2019;3(7):113–117.
20. Zhang L, Ren Y, Yang T, et al. Rapid evolution of protein diversity by de novo origination in Oryza. *Nature Ecology & Evolution*. 2019;3(4):679–690.
21. Lee S, Chung J, Won H, et al. Removal of methylmercury and tributyltin (TBT) using marine microorganisms. Bulletin of Environmental Contamination and Toxicology. 2012;88(2):239–244.
22. Ahmadi M, Jorfi S, Kujlu R, et al. A novel salt-tolerant bacterial consortium for biodegradation of saline and recalcitrant petrochemical wastewater. *Journal of Environmental Management*. 2017;191:198–208.
23. Rossello R, Garcia-Valdes E, Lalucat J, et al. Genotypic and phenotypic diversity of Pseudomonas stutzeri. *Systematic and Applied Microbiology*. 1991;14(2):150–157.
24. Bennasar-Figueras A, Salvà-Serra F, Jaén-Luchoro D, et al. Complete genome sequence of Pseudomonas balearica DSM 6083T. *Genome Announcements*. 2016;4(2):217.
25. Reese M. Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome. *Comput Chem*. 2001;26(1):51–56.
26. Ling M. RANDOMSEQ: Python command–line random sequence generator.. *MOJ Proteomics Bioinform*. 2018;7(4):206–208.
27. Smith T, Waterman M. Identification of common molecular subsequences. *Journal of Molecular Biology*. 1981;147(1):195–197.
28. Needleman S, Wunsch C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*. 1970;48(3):443–453.
29. Boos D. Introduction to the Bootstrap World. *Statistical Science*.2003;18(2):168–174.

Citation: Usman S, Chua JW, Ardhanari-Shanmugam KD, et al. *Pseudomonas balearica* DSM 6083^T promoters can potentially originate from random sequences. *MOJ Proteomics Bioinform*. 2019;8(2):66–70. DOI: 10.15406/mojpb.2019.08.00268

30. Castillo C, Ling M. Digital Organism Simulation Environment (DOSE): a library for ecologically-based in silico experimental evolution. *Advances in Computer Science: an International Journal*. 2014;3(1):44–50.

31. Lim J, Aw Z, Goh D, et al. A genetic algorithm framework grounded in biology. *The Python Papers Source Codes*. 2010;2(6):1–15.

32. Castillo C, Ling M. Resistant traits in digital organisms do not revert preselection status despite extended deselection: implications to microbial antibiotics resistance. *BioMed Research International. 2014;2014:648389.*; 648389.

33. Castillo C, Chay Z, Ling M. Resistance maintained in digital organisms despite guanine/cytosine-based fitness cost and extended de-selection: implications to microbial antibiotics resistance. *MOJ Proteomics Bioinform*. 2015;2(2):39.

34. Rattray J, Strathern J. Error-prone DNA polymerases: when making a mistake is the only way to get ahead. *Annual Review of Genetics*. 2003;37:31–66.

35. Lee D, Lu J, Chang S, et al. Mapping DNA polymerase errors by single-molecule sequencing. *Nucleic Acids Research*. 2016;44(13):e118.

36. Ardhanari-Shanmugam K, Shahrukh K BV, Woo J, et al. De novo origination of Bacillus subtilis 168 promoters from random sequences. *Acta Scientific Microbiology*. 2019;2(11):7–10.

37. Louie B, Higdon R, Kolker E. A Statistical Model of Protein Sequence Similarity and Function Similarity Reveals Overly-Specific Function Predictions. *PLoS ONE*. 2009;4(10):e7546.

38. Higdon R, Louie B, Kolker E. Modeling sequence and function similarity between proteins for protein functional annotation. *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing-HPDC*. 2010;499–502.

39. Joshi T, Xu D. Quantitative assessment of relationship between sequence similarity and function similarity. *BMC Genomics*. 2007;8(1):222.

40. Horwitz M, Loeb L. Promoters selected from random DNA sequences. *Proc Natl Acad Sci U S A*. 1986;83(19):7405–7409.

41. Carvunis A, Rolland T, Wapinski I, et al. Proto-genes and de novo gene birth. *Nature*. 2012;487(7407):370–374.

42. Weisman M, Eddy S. Gene evolution: getting something from nothing. *Current Biology*. 2017;27(13):R661–R663.