ACTA SCIENTIFIC MICROBIOLOGY (ISSN: 2581-3226)

Volume 2 Issue 6 June 2019

Research Article

Potential *De Novo* Origins of Archaebacterial Glycerol-1-Phosphate Dehydrogenase (G1PDH)

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

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

Received: May 17, 2019; **Published:** May 23, 2019

Abstract

Eubacterial glycerol-1-phosphate dehydrogenase (G1PDH) may originate from archaebacteria by horizontal gene transfer; however, the origins of archaebacterial G1PDH remains unanswered. While recent studies show possible *de novo* origination of protein encoding genes and functional promoters, the mechanism of *de novo* origins of functional genes remains debatable. In this study, we examine the probability of *de novo* emergence of putative G1PDH from random sequences. Our results show that high number of open reading frames in random sequences and 71.8% of randomly generated sequences have 9.88% probability of being putative G1PDH. Hence, *de novo* origination archaebacterial G1PDH from random sequences is plausible.

Keywords: *De Novo*; G1PDH; Archaebacteria

Introduction

Glycerol-1-phosphate dehydrogenase (G1PDH), which catalyses nicotinamide adenine dinucleotide hydrogen (NADH) or nicotinamide adenine dinucleotide phosphate hydrogen (NADPH) into sn-glycerol-1-phosphate (G1P); is a crucial enzyme for *de novo* synthesis of archaea phospholipid [1,2]. G1P lipid backbone enables archaebacteria to possess well-defined lipid membrane [3]; providing its host with crucial survival advantages, such as thermal tolerance [4]. Archaeal G1PDH was first sequenced from *Methanothermobacter thermautotrophicus* and named egsA [5], consisting of 1,041 base pairs with a peptide mass of 36,963Da. Although both archaeal G1PDH and its eubacterial/eukaryal counterpart, G3PDH, catalyse the reduction of DHAP in the presence of NAD(P)H, they share no homology and belong to separate families [6]. This is supported by Daiyasu., *et al.* [7] and Koga and Morii [1]. G1PDH transfers the pro-R hydrogen of NADH instead of pro-S hydrogen in the case of G3PDH, which leads to the generation of stereospecificity of G1P backbone [8]. G1P stereospecificity was considered unique for archaeal domain, and the emergence of G1PDH has been linked to the separation between archaebacteria and eubacteria [9-11]. However, Guldan, *et al.* [12] reported a G1PDH homolog from *Bacillus subtilis*, AraM, having 31% sequence identity with *Archaeoglobus fulgidus* egsA and known to form a homodimer with G1PDH activity. Yokobori., *et al.* [13] proposed that eubacterial G1DPH may have originated from archaebacteria via horizontal gene transfer. Yet, the question on the origins of archaebacterial G1PDH remains.

Until recently, the emergence of new genes was thought to be solely driven by duplication, recombination and horizontal gene transfer of existing genetic materials [14] as the possibility of novel gene formation from non-functioning genetic sequences was considered unlikely [15]. However, there are evidences of protein-encoding genes originating from noncoding sequence, termed as *de novo* origin of genes, and this process might have been active throughout evolution [16-19]. Horwitz and Loeb [20] first showed that functional promoters can originate random DNA sequences, which is supported by Yona., *et al.* [21] reporting 10% of randomly

generated DNA sequences can serve as functional promoters in *Escherichia coli*. At the peptide level, Ling [22] suggests that random amino acid chains may contain putative protein domains. At the gene level, *de novo* genes were first identified in Drosophila *melanogaster* [23] and subsequently in multiple organisms; such as, yeast [14,16,24], primates [25] and plants [26]. Wu and Knudson [14] suggest that *de novo* genes may have originated due to mutation or DNA shuffling of non-coding DNA sequences. Most studies have highlighted the discoveries of *de novo* genes. However, how *de novo* genes can arise remains debatable [17] and can *de novo* genes be functional from the time of their emergence [28]. More specifically, what is the possibility of *de novo* origins of archaebacterial G1PDH?

In this study, we examine the possibility of *de novo* origins of archaebacterial G1PDH through the evaluation of pairwise alignment scores of known archaebacterial G1PDH against randomly generated sequences. Our result suggests that a substantial portion of randomly generated sequences may possess some properties of archaebacterial G1PDH.

Methods

Sequence Data Sets. A set of archaebacterial G1PDH sequences, hereafter known as baseline sequences, were retrieved from KEGG Genes using BLASTN search with *Methanothermobacter thermautotrophicus* (KEGG Gene ID mth:MTH_610 and UniProt ID P72010) as query sequence with default parameters and an E-value threshold of 1e-9. A set of 10,000 random sequences between 1011 and 1086 nucleotides, which is the range of sizes of baseline sequences without start and stop codons, were generated using RANDOMSEQ [29] with 2754 adenine, 2136 thymine, 2775 guanine, and 2336 cytosine per 10,000 bases; without start and stop codons within the sequence.

Determining open reading frames from random sequences. An open reading frame (ORF) can be defined as a sequence with length divisible by three and begins with a translation start and ends at a stop codon [30]. A set of 10 sequences of 10 kilobases with uniform nucleotide distribution was generated using RANDOMSEQ [29] and ORFs of at least 33 nucleotides, corresponding to one of the shortest gene known [31], were identified.

Determining putative G1PDH from random sequences. Two series of pairwise sequence alignments were performed using Bactome (https://github.com/mauriceling/bactome). Both Smith-Waterman algorithm [32], also known as local alignment, and Needleman-Wunsch algorithm [33], also known as global alignment, were used for each series. In the first series, each baseline sequence was pairwise aligned to each of the other baseline sequences and the distribution of scores were used as measure for putative G1PDH sequences. In the second series, each of the 10,000 random sequence was pairwise aligned to each baseline sequence. A minimum and average alignment score were generated for each random sequence. The probability of each random sequence being a putative G1PDH sequence was determined by the proportion of baseline alignment scores below the minimum and average alignment score of the random sequence for stringent and relaxed criteria respectively.

Results and Discussion

Characterization of Archaebacterial G1PDH. BLASTN of *Methanothermobacter thermautotrophicus* G1PDH sequence yielded 161 hits. Of which, 103 entries are within the E-value threshold and specific to archaebacteria, known to partake in glycerophospholipid metabolism pathway, and are egsA genes; forming the set of baseline sequences. Excluding start and stop codons, the minimum and maximum nucleotide length for baseline sequences are 1011 and 1086 respectively. The distribution of nucleobases of the 103 baseline sequences were analysed and shown to be 27.54% adenine, 23.36% cytosine, 27.75% guanine, and 21.36% thymine. The baseline sequences were pairwise aligned, and yield a total of 5,253 alignments (Figure 1). The mean alignment score is 740.11 (standard deviation of 52.445 and a median score of 729), with 609 and 1,065 as the minimum and maximum scores respectively.

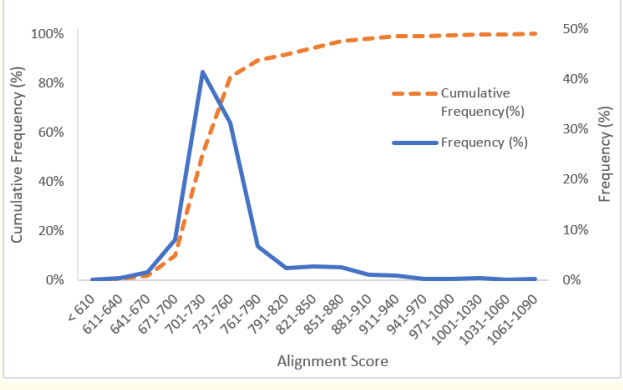

Figure 1: G1P dehydrogenase local pairwise alignment score.

196 ORFs per 10 Kilobases Found. We identified an average of 196 ORFs per random sequence of 10 kilobases (Figure 2). The

shortest ORF found consists of 36 nucleotides while the longest ORF consists of 498 nucleotides. Our result suggests that ORFs, potential protein-coding region of a gene [34], can exist randomly. Cardoso-Moreira and Long [35] had presented a model of *de novo* origins of ORFs through mutations. However, our results suggest another possible route for *de novo* ORF – emerging from random sequences without the need for mutations, which does not contradict the model proposed by Cardoso-Moreira and Long [35].

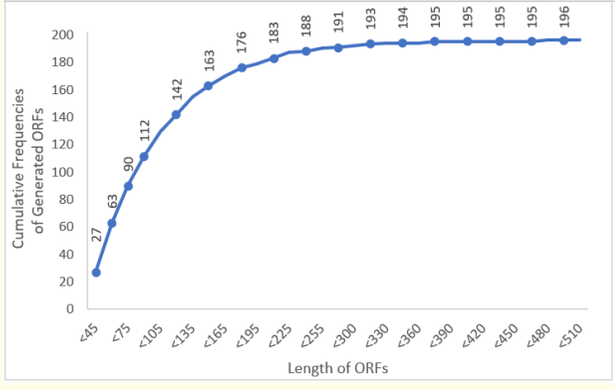

Figure 2: Total number of ORFs identified per 10 Kilobases of randomly generated sequences.

Substantial random sequences with more than 9.88% probability as putative G1PDH. Our results show that the minimum pairwise alignment score of 99.98% (n = 9,998) of the randomly generated sequences are equal or higher than the minimum local pairwise alignment score of 609 from baseline sequences. As the range of pairwise alignment scores among baseline sequences represents the sequence diversity of archaebacterial G1PDH; therefore, if a random sequence is not likely a putative archaebacterial G1PDH, then its minimum pairwise alignment score with known archaebacterial G1PDH (baseline sequences) should be lower than the minimum pairwise alignment score among known archaebacterial G1PDH. Moreover, of the 10,000 randomly generated sequences, 7,180 (71.80%) sequences have 9.88% probability of being putative G1PDH (Table 1). This is based on the probability that the average pairwise scores of 71.80% of the 10,000 randomly generated sequences are above 9.88% of the 5,253 baseline pairwise scores. At high stringency level (minimum alignment score), 37.44% of the randomly generated sequences have 1.87% probability of possess G1PDH functionality while at high leniency level (maximum alignment score), 51.24% of the sequences have more than 50% probability of being a putative functional archaebacterial G1PDH.

Score Range	Minimum Score	Average Score	Maximum Score	Probability of G1PDH function
< 610	99.98%	100.00%	100.00%	0.02%
611-640	99.95%	100.00%	100.00%	0.38%
641-670	37.44%	100.00%	100.00%	1.87%
671-700	0.00%	71.80%	100.00%	9.88%
701-730	0.00%	0.00%	51.24%	51.36%
731-760	0.00%	0.00%	0.01%	82.58%
761-790	0.00%	0.00%	0.00%	89.21%
791-820	0.00%	0.00%	0.00%	91.55%

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

However, none of the 9,998 random sequences yield average pairwise alignment score higher than the average local pairwise alignment score of 740.11 from among baseline sequences (Figure 3). As the average pairwise alignment score among baseline sequences can be taken to present the average archaebacterial G1PDH gene, our results suggest that all 9,998 random sequences may be putative archaebacterial G1PDH. This is supported by our results showing that the average pairwise alignment scores of all random sequences are more than 83% of the average pairwise alignment scores among known archaebacterial G1PDH (Figure 3). This suggests that substantial proportion of random sequences may have some properties of archaebacterial G1PDH.

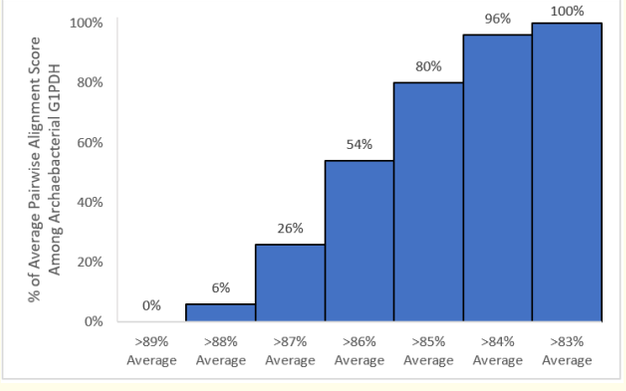

Figure 3: Percentage of 10,000 random sequences with average pairwise alignment scores above a certain percentage of average pairwise alignment score among archaebacterial G1PDH.

This is consistent with Yona., *et al*. [21] reporting about 60% of random sequence can function comparably to wild-type promoters with only 1 base mutation out of 103 bases, and is consistent with Ling [22] suggesting nearly 27% of random amino acid chains may contain putative protein domains. Although Ling [22] reports on amino acid chains, Neme., *et al*. [36] report that 25% of randomly generated 150 nucleotide sequences demonstrating beneficial effect on *E. coli* growth rate when expressed as RNA or peptide. Zhang., *et al*. [37] suggest an average emergence of 51.5 *de novo* genes per million years in Oryza by studying the genomes of 13 closely related Oryza species; importantly, 56.6% of the *de novo* genes identified are translated.

Carvunis., *et al*. [38] coined the term "proto-gene" to defined a gene born from non-genic sequence by random processes without selection, and 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 selectible phenotype that selection pressure can act on [39]. With substantial proportion of random sequences able to function as promoters [21], it is plausible to consider that the requirement for transcription has substantial chance of being randomly fulfilled. Our results demonstrate the feasibility of *de novo* origination of ORFs and putative archaebacterial G1PDH from random sequences; hence, the requirement for translation is fulfilled. Given that G1PDH is essential for membrane synthesis [1-4], the requirement for beneficial function is also fulfilled. Therefore, it is plausible to consider that archaebacterial G1PDH may originate *de novo* from random sequences.

Conflict of Interest

The authors declare no conflict of interest.

Bibliography

1. Koga Y and Morii H. "Biosynthesis of ether-type polar lipids in archaea and evolutionary considerations". *Microbiology and Molecular Biology Reviews* 71.1 (2007): 97-120.

2. Lai D., *et al*. "Reconstruction of the archaeal isoprenoid ether lipid biosynthesis pathway in Escherichia coli through digeranylgeranylglyceryl phosphate". *Metabolic Engineering* 11.3 (2009): 184-191.

3. Edidin M. "Lipids on the frontier: a century of cell-membrane bilayers". *Nature Reviews Molecular Cell Biology* 4.5 (2003) :414-418.

4. Koga Y. "Thermal adaptation of the archaeal and bacterial lipid membranes". *Archaea* 2012 (2012): 789652.

5. Koga Y., *et al*. "Did archaeal and bacterial cells arise independently from noncellular Precursors? A hypothesis stating that the advent of membrane phospholipid with enantiomeric glycerophosphate backbones caused the separation of the two lines of descent". *Journal of Molecular Evolution* 46.1 (1998): 54-63.

6. Jain S., *et al*. "Biosynthesis of archaeal membrane ether lipids". *Frontiers in Microbiology* 5 (2014): 641.

7. Daiyasu H., *et al*. "Analysis of membrane stereochemistry with homology modeling of sn-glycerol-1-phosphate dehydrogenase". *Protein Engineering* 15.12 (2002): 987-995.

8. Koga Y., *et al*. "Transfer of pro-R hydrogen from NADH to dihydroxyacetonephosphate by sn-glycerol-1-phosphate dehydrogenase from the archaeon Methanothermobacter thermautotrophicus". *Bioscience, Biotechnology, and Biochemistry* 67.7 (2003): 1605-1608.

9. Peretó J., *et al*. "Ancestral lipid biosynthesis and early membrane evolution". *Trends in Biochemical Sciences* 29.9 (2004): 469-477.

10. Payandeh J and Pai EF. "Enzyme-driven speciation: crystallizing Archaea via lipid capture". *Journal of Molecular Evolution* 64.3 (2007): 364-374.

11. Glansdorff N., *et al*. "The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner". *Biology Direct* 3 (2008): 29.

12. Guldan H., *et al*. "Identification and characterization of a bacterial glycerol-1-phosphate dehydrogenase: Ni(2+)-dependent AraM from Bacillus subtilis". *Biochemistry* 47.28 (2008): 7376-7384.

13. Yokobori S-I., *et al*. "Birth of Archaeal Cells: Molecular Phylogenetic Analyses of G1P Dehydrogenase, G3P Dehydrogenases, and Glycerol Kinase Suggest Derived Features of Archaeal Membranes Having G1P Polar Lipids". *Archaea* 2016 (2016): 1802675.

14. Wu B and Knudson A. "Tracing the De Novo Origin of Protein-Coding Genes in Yeast". *mBio* 9.4 (2018): e01024-18.

15. Schlötterer C. "Genes from scratch - the evolutionary fate of de novo genes". *Trends in Genetics* 31 (2015) :215-219.

16. Cai J., *et al*. "De novo origination of a new protein-coding gene in Saccharomyces cerevisiae". *Genetics* 179.1 (2008): 487-496.

17. Knowles DG and McLysaght A. "Recent de novo origin of human protein-coding genes". *Genome Research* 19.10 (2009): 1752-1759.

18. Wu DD and Zhang YP. "Evolution and function of de novo originated genes". *Molecular Phylogenetics and Evolution* 67.2 (2013): 541-545.

19. Neme R and Tautz D. "Evolution: dynamics of de novo gene emergence". *Current Biology* 24.6 (2014): R238-240.

20. Horwitz MS and Loeb LA. "Promoters selected from random DNA sequences". *Proceedings of the National Academy of Sciences of the United States of America* 83.19 (1986): 7405-7409.

21. Yona AH., *et al*. "Random sequences rapidly evolve into de novo promoters". *Nature Communications* 9.1 (2018): 1530.

22. Ling MH. "De novo putative protein domains from random peptides". *Acta Scientific Microbiology* 2.4 (2019): 109-112.

23. Levine MT., *et al*. "Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression". *Proceedings of the National Academy of Sciences of the United States of America* 103.26 (2006): 9935-9939.

24. Li D., *et al*. "A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand". *Cell Research* 20.4 (2010): 408-420.

25. Xie C., *et al*. "Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs". *PLoS Genetics* 8.9 (2012): e1002942.

26. Donoghue MT., *et al*. "Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana". *BMC Evolutionary Biology* 11 (2011): 47.

27. McLysaght A and Hurst LD. "Open questions in the study of de novo genes: what, how and why". *Nature Reviews Genetics* 17.9 (2016): 567-578.

28. Schmitz JF and Bornberg-Bauer E. "Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding DNA". *F1000Research* 6 (2017): 57-57.

29. Ling MH. "RANDOMSEQ: Python command–line random sequence generator". *MOJ Proteomics and Bioinformatics* 7.4 (2018): 206-208.

30. Sieber P., *et al*. "The Definition of Open Reading Frame Revisited". *Trends in Genetics* 34.3 (2018): 167-170.

31. 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.

32. Smith TF and Waterman MS. "Identification of common molecular subsequences". *Journal of Molecular Biology* 147.1 (1981): 195-197.

33. 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* 48.3 (1970): 443-453.

34. Woodcroft BJ., *et al*. "OrfM: a fast open reading frame predictor for metagenomic data". *Bioinformatics* 32.17 (2016): 2702-2703.

35. Cardoso-Moreira M and Long M. "The origin and evolution of new genes". *Methods in Molecular Biology* 856 (2012): 161-186.

36. Neme R., *et al*. "Random sequences are an abundant source of bioactive RNAs or peptides". *Nature Ecology and Evolution* 1.6 (2017): 0217.

37. Zhang L., *et al*. "Rapid evolution of protein diversity by de novo origination in Oryza". *Nature Ecology and Evolution* 3.4 (2019): 679-690.

38. Carvunis AR., *et al*. "Proto-genes and de novo gene birth". *Nature* 487.7407 (2012): 370-374.

39. Weisman CM and Eddy SR. " Gene evolution: getting something from nothing". *Current Biology* 27.13 (2017): R661-R663.

Volume 2 Issue 6 June 2019

© All rights are reserved by Chakrit Thong-Ek, *et al.*