ACTA SCIENTIFIC MICROBIOLOGY (ISSN: 2581-3226)

Volume 2 Issue 4 April 2019

Short Communication

De Novo Putative Protein Domains from Random Peptides

Maurice HT Ling[1,2]*

[1]*Colossus Technologies LLP, Republic of Singapore*

[2]*HOHY PTE LTD, Republic of Singapore*

***Corresponding Author:** Maurice HT Ling, Colossus Technologies LLP and HOHY PTE LTD, Republic of Singapore.

Received: March 12, 2019; **Published:** March 26, 2019

Abstract

How prebiotic chemistry in the primordial world becomes biochemistry, is a major question in evolutionary biology. Studies have found that biological activities from random DNA sequences are not rare and abiotically-catalyzed polymerization of 13 amino acid chains can occur. However, it is not clear whether random chains 13 amino acid or longer are biologically functional. In this study, random peptide sequences were generated and mapped to ProSite motifs and NCBI Conserved Domains Database. Results suggest that a large fraction of randomly generated 13 amino acid chains may contain putative protein domains while longer random peptide chains may contain functional protein domains. Large diversity of protein domains is observed. Hence, it is plausible for putative functions to originate from abiotically-catalyzed 13 amino acid chains. As both self-replicating RNA molecules and prion proteins have been found, it is plausible that both RNA and peptides may co-exist and synergize in the primordial world.

Keywords: De Novo; Protein; Peptides

Introduction

One of the big questions in evolutionary biology is how prebiotic chemistry becomes biochemistry [1]. Under this umbrella will be questions; such as, how the first gene and its constituting components like promoters and ribosome binding sites originates, or how the first functional peptide originates. Horwitz and Loeb [2] first found that random sequences may have varying gene expression promoting abilities in *Escherichia coli*. Recently, Yona., *et al*. [3] found about 10% of the randomly generated DNA sequences can function as promoters in *E. coli* and about 60% of the randomly generated DNA sequences only require one mutation to function as promoters. These suggests that functions from random DNA sequences may not be rare. At the gene level, Andersson., *et al*. [4] suggest that coding sequences may emerge from non-coding sequences, which is supported by Schlötterer [5]. A recent study on yeast by Wu and Knudson [6] suggests that new genes may emerge as a result of mutations or DNA shuffling.

From the peptide side of the story, Miller-Urey experiment [7] demonstrated that amino acids can originate abiotically [8]. The next step is to consider the possibility of amino acids naturally condensing into peptides. Greenwald., *et al*. [9] demonstrated that short chains of peptides consisting of 13 amino acids can originate from individual amino acids by volcanic gas carbonyl sulphide-catalyzed polymerization, suggesting that such possibility can occur. However, it is not clear whether randomly polymerized short peptides can have putative functions.

In this study, the hypothesis that random short peptides may exhibit putative functions is examined by scanning for putative protein domains at high sensitivity on randomly generated peptide sequences. Results suggest that short random peptides of 13 amino acids [9] may have putative protein functions and random peptides of average prokaryotic protein length [10] may have functional protein domains.

Method

Two sets of ten thousand random peptide sequences were generated as 5 sets of 2000 random peptides using RANDOMSEQ [11] with the following command, python randomseq.py FLS --length=<number of amino acids> --n=2000 --selection= A,825;R,553;N,406;D,545;C,137;Q,393;E,675;G,707;H,227;I,596; L,966;K,584;M,242;F,386;P,470;S,656;T,534;W,108;Y,292;V,687 --fasta=True. In the first set, each peptide is 13 amino acids each [9]. In the second set, each peptide is of 267 amino acids each, which is the average prokaryotic peptide length [10]. Amino acid composition is based on average amino acid composition per 10 thousand amino acids

as computed from Swiss-Prot [https://web.expasy.org/protscale/pscale/A.A.Swiss-Prot.html].

Each randomly generated peptide was scanned using ScanProsite [12] and NCBI-CDD [13,14]. ScanProsite [12] was executed with Release 2019_02 of 13-Feb-2019 [contains 1825 documentation entries, 1310 patterns, 1236 profiles and 1260 ProRule, totalling to 2548 unique motifs], including motifs with a high probability of occurrence and running the scan at high sensitivity. NCBI-CDD [13,14] was executed against CDD version 3.16, the superset of 30569 PSSMs [position-specific scoring matrices] including NCBI-curated domains and data imported from Pfam [15], SMART [16], NCBI Clusters of Orthologous Groups [17], NCBI Protein Clusters, and TIGRFAMs [18], with expect value [E-value] threshold of 0.01, using composition-corrected scoring, not filtered for low-complexity regions, and including retired sequences.

Results

Each peptide, 13 amino acids in length or 267 amino acids in length, were scanned for protein domains using ScanProsite [12] and NCBI-CDD [13,14] and found large numbers of peptides [Figure 1] mapped to large numbers of putative protein domains [Figures 2].

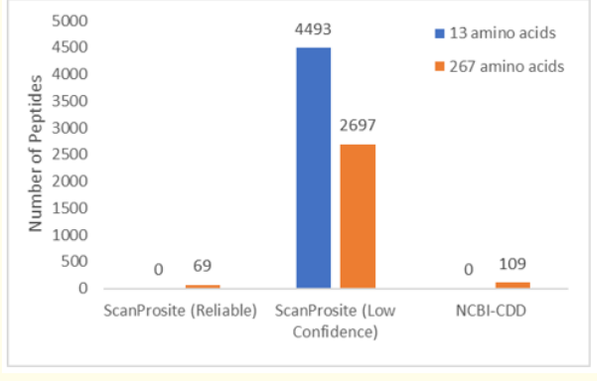

Figure 1: Number of Randomly Generated Peptides Mapped.

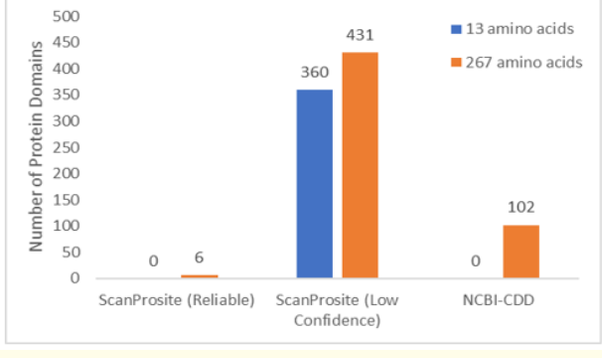

Figure 2: Number of Unique Putative Protein Domains Mapped.

Short 13 Amino Acid Peptides may have Putative Protein Domains. 4493 [44.93%] of the 10000 randomly generated 13-amino acid peptides were mapped at low confidence [confidence level = -1] to 360 [14.13%] unique ProSite motifs. None of the peptides were reliably mapped to ProSite motifs or PSSM at E-value threshold of 0.01.

Random Average Length Peptides Have Potentially Functional Properties. Of the 10000 randomly generated peptides, 69 (0.69%) were reliably mapped (confidence level = 0) to 6 (0.24%) unique ProSite motifs. From the corresponding ProSite documentation entries, these 6 are (a) bipartite nuclear localization signal (PS50079) involving in protein uptake by the nucleus, (b) Ig-like domain (PS50835) involving in binding various ligands, (c) FERM domain (PS50057) involving in localizing proteins to plasma membrane, (d) cyclic nucleotide-binding domain (PS50042) involving in binding cyclic nucleotides such as cAMP or cGMP, (e) phosphatidylinositol-specific phospholipase C profiles (PS50007) involving in eukaryotic signal transduction, and (f) gamma-carboxyglutamic acid-rich domain (PS50998) involving in blood coagulation. A further 2697 (26.97%) randomly generated peptides mapped at low confidence (confidence level = -1) to 431 (16.92%) unique ProSite motifs.

109 (1.09%) of the 10000 randomly generated peptides were mapped to 102 (0.33%) unique PSSMs by NCBI-CDD. The most significant 5 hits by E-values are (a) succinylglutamate desuccinylase/aspartoacylase family (Accession number cl27097, e-value = 1.61 x 10^{-5}) involving in arginine catabolism by the arginine succinyltransferase pathway, (b) btuF superfamily (Accession number cl28215, e-value = 5.63 x 10^{-5}) which is part of ATP-binding cassette transporters, (c) Type 1 periplasmic binding fold superfamily (Accession number cl10011, e-value = 6.37 x 10^{-5}) involving in various ligand binding including peptides, amino acids, and sugar molecules, (d) FRG domain (Accession number cl07460, e-value = 9.12 x 10^{-5}) which is functionally uncharacterized at this moment, and (e) CMAS superfamily (Accession number cl28102, e-value = 9.21 x 10^{-5}) involving in fatty acid metabolism.

Discussion and Conclusion

De novo origination of functional sequences was considered extremely unlikely [5]. Indeed, the mathematical probability of randomly generating a WW domain, one of the smallest protein domain [19], is rare. Given that the consensus sequence of WW domain (Prosite accession PS01159) is W-x(9,11)-[VFY]-[FYW]-x(6,7)-[GSTNE]-[GSTQCR]-[FYW]-{R}-{SA}-P, the random probability of generation is about 5.41 x 10^{-7}, which can be considered as rare. However, this probability does not take into account of putative or proto-domains, which may be partially functional. Results from this study suggests that substantial numbers of randomly

generated peptide sequences of average peptide length [10] may contain a diversity of known and putative protein domains. This is supported by several experimental studies [2,3] had suggested that random sequences may not be inert and functional components from random sequences may not be rare events, as Neme., *et al*. [20] have shown that large proportion of random RNA or peptide sequences are bioactive.

There is an on-going debate as to whether RNA or peptides originate first in the primordial world [21,22]. Greenwald., *et al.* [9] demonstrated that peptides of 13 amino acids in length may be synthesized abiotically. In this study, substantial proportion of the 10 thousand random peptides of 13 amino acids may contain a variety of putative protein domains. This suggests that abiotically polymerized short peptides may have putative functions, which may be partially functional. Once a partially functional sequence is present, selective pressure may act to evolve into fully functional sequences [3] or even evolve beyond environmental necessity [23]. The only required property is self-replication. Maury [22] had presented a model for self-replicating peptide. Both self-replicating RNA molecules [24] and prion proteins [25,26] have been found. This suggests that RNA and peptides may co-exist and synergize in the primordial world.

Supplementary Materials

A zip file containing the FASTA files of the randomly generated peptide sequences, as well as results from ScanProsite and NCBI-CDD can be found at http://tinyurl.com/DomainsFromRandom.

Conflict of Interest

The author declares no conflict of interest.

Bibliography

1. Big Questions in Evolution. *Cell* 166.3 (2016): 528-529.

2. 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.

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

4. Andersson DI., *et al*. "Evolution of new functions de novo and from preexisting genes". *Cold Spring Harbor Perspectives in Biology* 7 (2015): a017996.

5. Schlötterer C. "Genes from scratch--the evolutionary fate of de novo genes". *Trends in genetics : TIG* 31.4 (2015): 215-219.

6. Wu B and Knudson A. "Tracing the De Novo Origin of Protein-Coding Genes in Yeast". mBio. 9.4 (2018): e01024-1018.

7. Miller SL. "A production of amino acids under possible primitive earth conditions". *Science* 117.3046 (1953): 528-529.

8. Parker ET., *et al*. "Conducting miller-urey experiments". *Journal of Visualized Experiments* 83 (2014): e51039.

9. Greenwald J., *et al*. "Amyloid Aggregates Arise from Amino Acid Condensations under Prebiotic Conditions". *Angewandte Chemie International Edition* 55.38 (2016): 11609-11613.

10. Brocchieri L., *et al*. "Protein length in eukaryotic and prokaryotic proteomes". *Nucleic Acids Research* 33.10 (2005): 3390-3400.

11. Ling MH. "RANDOMSEQ: Python Command–line Random Sequence Generator". *MOJ Proteomics and Bioinformatics* 7.4 (2018): 206-208.

12. de Castro E., *et al*. "ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins". *Nucleic Acids Research* 34 (2006): W362-365.

13. Marchler-Bauer A and Bryant SH. "CD-Search: protein domain annotations on the fly". *Nucleic Acids Research* 32 (2004): W327-331.

14. Marchler-Bauer A., *et al*. "CDD: a Conserved Domain Database for the functional annotation of proteins". *Nucleic Acids Research* 39 (2011): D225-229.

15. El-Gebali S., *et al*. "The Pfam protein families database in 2019". *Nucleic Acids Research* 47 (2019): D427-32.

16. Letunic I and Bork P. "20 years of the SMART protein domain annotation resource". *Nucleic Acids Research* 46 (2018): D493-496.

17. Tatusov RL., *et al*. "The COG database: an updated version includes eukaryotes". *BMC Bioinformatics* 4 (2003): 41.

18. Haft DH., *et al*. "TIGRFAMs and Genome Properties in 2013". *Nucleic Acids Research* 41 (2013): D387-395.

19. Martinez-Rodriguez S., *et al*. "Crystal structure of the first WW domain of human YAP2 isoform". *Journal of Structural Biology* 191.3 (2015): 381-387.

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

21. Bernhardt HS. "The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others)". *Biology Direct* 7 (2012): 23-23.

22. Maury CPJ. "Amyloid and the origin of life: self-replicating cat-

alytic amyloids as prebiotic informational and protometabolic entities". *Cellular and molecular life sciences : CMLS* 75 (2018): 1499-1507.

23. Wistrand-Yuen E., *et al*. "Evolution of high-level resistance during low-level antibiotic exposure". *Nature Communications* 9.1 (2018): 1599.

24. Robertson MP and Joyce GF. "Highly efficient self-replicating RNA enzymes". *Chemical Biology* 21 (2014): 238-245.

25. Klimova N., *et al*. "The diversity and relationship of prion protein self-replicating states". *Virus Research* 207 (2015): 113-119.

26. Wang F., *et al*. "Self-propagating, protease-resistant, recombinant prion protein conformers with or without in vivo pathogenicity". *PLOS Pathogens* 13.7 (2017): e1006491.

Volume 2 Issue 4 April 2019