Vol. 11(1), pp. 1-8, January-June, 2021
DOI: 10.5897/JBSA2020.0114
Article Number: 88594DF66287
ISSN 2141-2464

http://www.academicjournals.org/JBSA

ACADEMIC JOURNALS
expand your knowledge

Journal of Bioinformatics and Sequence Analysis

Full Length Research Paper

Genomic profiles of *Pseudomonas aeruginosa* gene clusters based on profile HMMs, reveal novel therapeutic targets for clinical intervention

Michael Ambutsi[1]* and Patrick Okoth[2]

[1]Department of Biological Sciences, School of Natural Sciences [SONAS], Masinde Muliro University of Science and Technology, P. O. Box 190-50100 Kakamega, Kenya.
[2]Department of Biological Sciences, School of Natural Sciences [SONAS], Masinde Muliro University of Science and P. O. Box 190-50100 Kakamega, Technology, Kenya.

Received 5 December 2020; Accepted 8 February, 2021

Pseudomonas aeruginosa **has developed antibiotic resistance, a major health concern worldwide, through different mechanisms including the formation of biofilms. Thus far, typing has been primarily assay based. However, the many sequences available from the US National Center for Biotechnology Information (NCBI) and the International Consortium of Pseudomonas Database (IPCD) offer alternative ways of characterizing the biofilm formation genes which would reveal novel therapeutic targets for intervention. The current study employed profile hidden Markov models (*p*HMMs) in the characterization of biofilm formation genes and identification of possible variations based on the different ecological niches occupied by strains of *P. aeruginosa*. Statistical analyses were performed in R (v. 3.1.3) to determine the significance of these variations. Validated pHMMs identified a total of 197 hits for the 13 different ecological niches, with the human metagenomes recording 144 hits (73%) while the non-human metagenomes recorded 53 hits (27%). Human metagenomes had a significantly higher density of hits, with the abscess metagenomes indicating the highest density of hits. The overall result indicated a significant variation in density of hits between the different sites within the human metagenomes. This study successfully highlighted the significant value of already sequenced metagenomes in the identification of potential targets for novel therapeutic compounds. The profile hidden Markov model successfully identified 197 unique biofilm gene clusters emphasizing its importance in analyzing different sequenced pathogenic strains. The study recommends that experimental assays could be carried out to shed further light on the different biofilm formation gene clusters.**

Key words: Profile hidden Markov model, metagenomics, biofilm formation, antibiotic resistance.

INTRODUCTION

Conventional investigation of *Pseudomonas aeruginosa* biofilm formation genes has primarily been based on biological assays such as the evolution assays. These assay-based analyses are time consuming and labor

*Corresponding author. E-mail: ambutsim@gmail.com. Tel: +254706350662.

intensive, and may give negative results if changes occur due to mutations in the biofilm formation genes (Gong et al*.,* 2012). In the last few years *P. aeruginosa* sequences have become increasingly available, so that a database search and pair-wise comparisons are alternative ways of characterizing them.

A previous study has indicated that pair-wise sequence identification of long alignments that are greater than 40% are ideal for providing unambiguous results (Zheng et al*.,* 2019). A different study indicated that a database search may not identify the relationship between the query and target sequences if the sequence identities of related proteins are less than 30% (Kirsip and Abroi, 2019). The comparison performed for both these methods requires that a known reference strain is chosen for the analyses. To select an appropriate representative strain, one requires extensive biological knowledge of the protein family under study. The choices made at this point will inevitably affect the outcome of the downstream analyses.

Rather than using all available biofilm formation gene sequences, the study chose sequence representatives for downstream analyses. The sequences were multiply aligned and their consensus used to build the probabilistic profile hidden Markov model (pHMM). Model construction was performed using the software package HMMER (Eddy, 2011). Profile HMMs have previously been used to molecularly detect viruses within metagenomic data, search for Cry proteins expressed by Bacillus spp. genomes and characterize the subtypes of HA gene (Gong et al*.,* 2012; Skewes et al*.,* 2014; Castillo-Esparza et al*.,* 2019).

A profile HMM is an ideal model to quantitatively assess how an individual sequence belongs to a profile given that it is both probabilistically and statistically intrinsic (Eddy, 2004, 2011)**.** The Pfam database which comprises annotated protein families has relied heavily on profile HMMs in the characterization of such families. With this database, researchers can search for, characterize and classify members of these protein families (Sonnhammer et al*.*, 1997). The current study explored the use of profile HMMs in revealing the variation and conservation patterns of biofilm formation genes in different strains of *P. aeruginosa*. All non-redundant *P. aeruginosa* sequences were downloaded from the NCBI and IPCD databases for investigation. Custom python scripts were used to retrieve the biofilm formation gene sequences from representative strains of the opportunistic pathogen. We constructed individual models for the 13 biofilm formation genes that were retrieved by the custom python scripts. This made it possible to construct gene-specific profiles and evaluate how different strain sequences fitted into the specific profiles to provide an overview of how the genes are distributed in the pathogen strains occupying different ecological niches. These analyses provided insights into the relationship between the biofilm formation genes and pathogens occupying specific ecological niches.

MATERIALS AND METHODS

Identifying protein family of interest

The construction of profile hidden Markov models was based on multiple sequence alignments of DNA or proteins sequences from the same functional family. The pHMM was used to represent the patterns, motifs along with other statistical properties of the alignments. Before the actual construction was performed the protein family of interest under investigation was selected. The criteria identified by Henikoff et al*.* (1997) were used to identify the protein family of interests. In this case the protein family of interest represented a set of genes performing similar functions in different strains of *P. aeruginosa*, the pathogen under study.

Selecting sequences representative of this family

With a protein family of interest identified, representative sequences from these sequences were selected for model construction. Purposive sampling of sequences initially retrieved from the custom python scripts was performed in this case to ensure that the selected sequences provide the most informative findings in the downstream analyses. A 40% sequence similarity threshold was chosen as suggested by a previous study which indicated that long alignments greater than 40% are ideal for providing unambiguous results (Zheng et al*.,* 2019). Sequence files with similarity identities less than 40% were excluded from the downstream analyses.

Building multiple sequence alignment

To create a pHMM of each of the classes of biofilm formation genes, a global multiple sequence alignment was generated using MUSCLE (v.3.8.31) (Edgar, 2004). The MUSCLE algorithm used for these analyses is available in the UGENE platform (Okonechnikov et al*.,* 2012). The study used the default values for the gap open penalty (54.00), gap extension penalty (8.00), and terminate gap penalty (4.00). These gap penalties were used to control the positions of the conserved regions within the alignment. The consensus sequence from these global alignments informed the construction of the specific profile hidden Markov model.

Building of profile HMM

From the multiple sequence alignments, the HMMER3 toolkit available on the UGENE software was used to construct profile Hidden Markov Models (pHMMs) for the different classes of biofilm formation genes for analyses of genomes of *P. aeruginosa* strains (4). For each of the twelve biofilm formation genes, the study created a profile HMM. The *hmmbuild* algorithm within the HMMER3 tool (v 3.1b1) in the UGENE software was used to create a suitable profile HMM from the MSA aligned-FASTA file (http://hmmer.janelia.org). The algorithm generated a hmm file containing a consensus sequence for the biofilm formation genes in *P. aeruginosa*. The HMMER3 platform was accelerated by the Multiple Segment Viterbi (MSV) algorithm that is implemented in the software package (Eddy, 2011).

The study chose different threshold values for estimation of the conserved regions to facilitate different simulations of the database search. These simulations were performed to give credence to the sensitivity of the models which previous studies have shown is

Table 1. Controls used in the validation of the profile HMM.

Organism	Accession number
Pseudomonas aeruginosa PA01	NC_002516
Bat Adenovirus 2	NC_015932
Gyrovirus 4	NC_018401
Duck circovirus	NC_007220
Domestic cat hepadnavirus	NC_040719

more important than the model's specificity. The accurate models can efficiently detect protein family members even when they fail to identify ineffective residues.

Validation of profile HMM (select positive and negative controls)

Using the *hmmer search* algorithm on UGENE, sequences used to construct the model were searched for as positive controls. Sequences of unrelated microorganisms (Table 1) were also searched for as negative controls. The presence of signals for the positive control search and lack of signals for the negative control search demonstrated the efficiency in the prediction by the constructed models. The controls were also screened using a traditional BLASTn with the algC phosphomannomutase as the driver sequence (GenBank accession number NC_002516.2) to compare the sensitivity and specificity of each approach. The visualization of the constructed pHMM was performed using the HMM visual editor (HMMVE_1.2).

Target sequence translation

The study created a python script to translate the nucleotide sequences of *P. aeruginosa* into protein sequences to facilitate the comparison with the constructed protein profile HMMs. The constructed python script used the standard genetic code to translate nucleotide sequences in six frames.

Screening the selected database

The study used the HMMER3 *hmmsearch* tool with default parameters to search for the biofilm formation profile HMM against sequences of *P. aeruginosa* drawn from different ecological niches as listed in Table 2. A significance cutoff of $E \leq 1 \times 10^{-5}$ was chosen for the search by the profile HMMs. A traditional BlastP was performed on the pathogen's sequence as an additional comparison of performance.

Statistical analyses

Statistical analyses were performed in R (v. 3.1.3) (R Core Team, 2015).

RESULTS AND DISCUSSION

Homologous genes with highly similar functions are often classified as gene families. For this study, genes responsible for biofilm formation in different strains of *P. aeruginosa* were identified and selected as the protein family of interest. Using the criteria identified by Henikoff et al. (1997), the study classified these set of genes into a family of related sequences. These genes were then used to inform the construction and validation of the profile hidden Markov models. 12 sequence files were created by custom python scripts and selected for the downstream analyses as they contained sequences from different strains of the ubiquitous pathogen. Each of these sequences represented a single biofilm formation gene containing 43 records of *P. aeruginosa* sequences. Gong et al. previously reported that the choice of representative sequences of a protein family of interest inevitably affects the outcomes of downstream analyses performed on these set of sequences (Gong et al., 2012). An extensive biological knowledge of the protein family under study is necessary for one to make an informed decision. The study relied on the previous analyses to make a judgment of the sequence homology. Each of the biofilm formation gene sequences indicated a sequence similarity of 40% and above as recommended in previous studies (Zheng et al., 2019). Homologous sequence files were preferred in this case as they would contain patterns and motifs which could be identified by the pHMMs and used to analyze different strains of *P. aeruginosa*. The amino acid sequences were also preferred given that they provided adequate information that could be modelled in a pHMM. The study successfully created 12 pHMMs from the sequences retrieved using the python scripts. A representative section of one pHMM is shown in Figure 1.

Table 2. Statistics of Pseudomonas aeruginosa sequences analyzed and their respective ecological niches.

Ecological niche	Analyzed sequences
Abscess	2
Blood	15
Bronchial	6
Cell culture	4
Clinical	10
Dental	1
Environment	8
Eye	2
Lung	1
Sputum	26
Trachea aspirates	5
Urine	7
Wound	9
Total	**96**

Validation of the profile hidden Markov model

The ability of the developed profile HMM to detect biofilm formation genes was analyzed using the positive and negative controls listed in Table 3. The reference *P.*

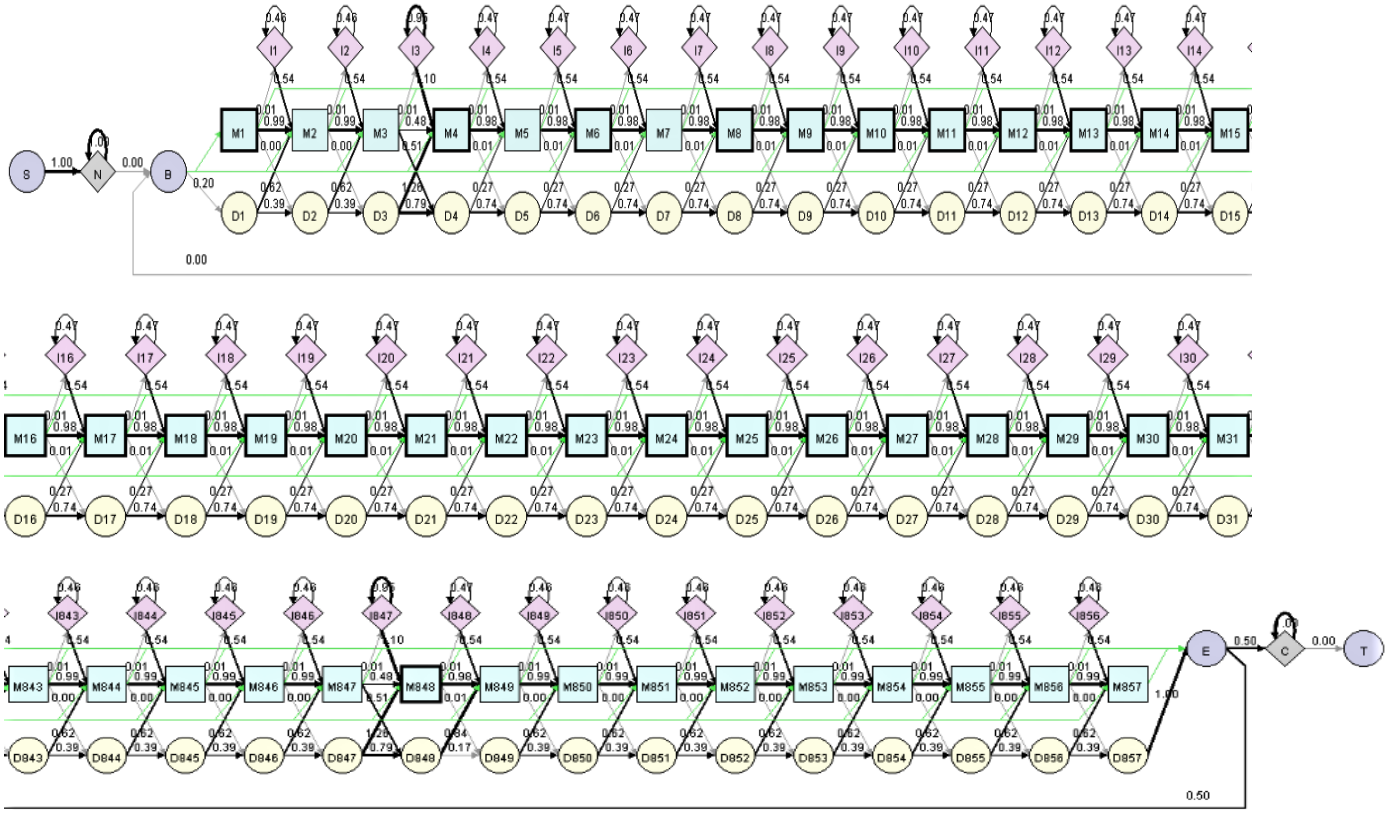

Figure 1. A representative profile HMM indicating the architecture of the constructed models. The different shapes indicate states while the arrows indicate state transitions. The 'S' represents the start position of the model. The 'N' represents the null model that the HMMER algorithm constructed first before creating the rest of the representative model. The 'M' (squares) represents the match states which indicate the frequencies of the most probable amino acid in those different locations. The 'D' (circles) represents the delete states while the 'I' (diamonds) represents the insert states.

Table 3. Comparison of the pHMM hits of the 12 biofilm formation genes across the 13 ecological niches.

Niche	algD	algU	pslJ	arnB	gshB	htpG	psl	pslE	pslG	rsaL	gshA	fliC
Abscess	2	0	2	0	0	0	0	2	2	2	1	1
Blood	10	5	2	1	0	9	4	3	4	3	3	0
Bronchial	4	0	0	0	0	0	1	0	0	2	3	0
cell culture	2	1	1	1	0	1	2	1	1	3	0	1
Clinical	7	1	0	1	0	1	3	0	0	4	2	0
Dental	1	0	0	0	0	0	0	0	0	0	0	0
environment	7	2	1	0	1	2	1	1	1	1	2	1
Eye	1	1	0	0	1	0	1	0	0	0	0	0
Lung	0	0	0	0	0	0	0	0	0	0	1	0
Sputum	5	3	1	1	0	0	6	1	1	9	8	1
Trachea	1	0	0	0	0	0	2	0	0	1	4	0
Urine	1	0	0	0	0	0	6	0	0	4	1	0
Wound	3	0	1	0	0	0	3	1	1	2	5	0

aeruginosa PA01 was chosen as the positive control. All the constructed pHMMs correctly identified different biofilm formation genes in the positive control. The negative controls were chosen because they are of different species and do not exhibit biofilm formation as one of their survival mechanisms. When searched against the four negative controls, the 12 biofilm formation pHMMs returned no false positives.

Table 4. Distribution of the density of pHMM hits (hits/MB) across the different ecological niches.

Niche	algDM	algUM	PslJM	arnBM	gshBM	htpGM	pslM	pslEM	pslGM	rsaLM	gshAM	fliCM
abscess	1.52E-04	0	1.52E-04	0	0	0	0	1.52E-04	1.52E-04	1.52E-04	7.60E-05	7.60E-05
Blood	9.69E-05	4.85E-05	1.94E-05	9.69E-06	0	8.72E-05	3.88E-05	2.91E-05	3.88E-05	2.91E-05	2.91E-05	0
Bronchial	9.77E-05	0	0	0	0	0	2.44E-05	0	0	4.88E-05	7.33E-05	0
cell culture	7.95E-05	3.98E-05	3.98E-05	3.98E-05	0	3.98E-05	7.95E-05	3.98E-05	3.98E-05	1.19E-04	0.00E+00	3.98E-05
clinical	1.06E-04	1.52E-05	0.00E+00	1.52E-05	0	1.52E-05	4.56E-05	0.00E+00	0.00E+00	6.09E-05	3.04E-05	0
Dental	1.44E-04	0.00E+00	0.00E+00	0.00E+00	0	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0
Environment	1.32E-04	3.79E-05	1.89E-05	0.00E+00	1.89E-05	3.79E-05	1.89E-05	1.89E-05	1.89E-05	1.89E-05	3.79E-05	1.89E-05
Eye	7.29E-05	7.29E-05	0.00E+00	0.00E+00	7.29E-05	0.00E+00	7.29E-05	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0
lung	0	0.00E+00	0.00E+00	0.00E+00	0	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	1.45E-05	0
sputum	2.85E-05	1.71E-05	5.69E-06	5.69E-06	0	0.00E+00	3.42E-05	5.69E-06	5.69E-06	5.12E-05	4.56E-05	5.69E-06
trachea	2.88E-05	0.00E+00	0.00E+00	0.00E+00	0	0.00E+00	5.75E-05	0.00E+00	0.00E+00	2.88E-05	1.15E-05	0
urine	2.13E-05	0.00E+00	0.00E+00	0.00E+00	0	0.00E+00	0.00E+00	0.00E+00	0.00E+00	8.51E-05	2.13E-05	0
wound	4.85E-05	0.00E+00	1.62E-05	0.00E+00	0	0.00E+00	4.85E-05	1.62E-05	1.62E-05	3.24E-05	8.09E-05	0

Screening the selected database

A search performed by the developed profile HMM against the *P. aeruginosa* sequences identified a total of 197 hits for the 13 different ecological niches as indicated in Table 3. Of the 197 hits, 144 hits (73%) belonged to the human samples while 53 hits (27%) belonged to the nonhuman samples. 38% of the human sample hits were recorded from ecological niches that were respiratory in nature. 62% of the hits were associated with non-respiratory niches within the human host. Overall, the blood ecological niche recorded the highest number of hits (44) while the lung and dental niches had the fewest number of hits (1 each).

The study sought to put these results into context and identified the density of the hits which represented the hit per Megabases as indicated in Table 4. Figure 2 indicates the distribution of the density of hits per ecological niche. In this case, the abscess ecological niche had the highest density of hits while the dental niche had the lowest density of hits. Figure 3 compares the density of hits between the human and non-human samples (environmental samples). Strains of the ubiquitous pathogen from human samples had a significantly higher density of hits compared to the strains from non-human samples. The only different observation was indicated in the density of hits for the *arnB* gene where the nonhuman samples had a significantly higher density of hits compared to their human sample counterparts. Previous studies have indicated that biofilm formation is promoted in *P. aeruginosa* strains with a loss in function mutation in the *arnB* gene (Segev-Zarko et al., 2018). This could inform the higher density of genes among environmental samples as the human-isolated strains are less likely to habour and express the *arnB* gene as they look to form biofilms. The overall results also indicated a significant variation in density of hits between the different sites within the human metagenomes. This pattern was also reflected in four of the respiratory subsites, namely bronchial, lung, sputum and trachea. The lung metagenomes had the lowest biofilm formation gene density and exhibited significantly lower densities than all the other ecological niches.

With regards to the biofilm formation genes, the *algD* gene had the highest number of hits and highest density of hits compared to the models of the other biofilm formation genes as indicated in both Figures 2 and 3. This result was consistent with previous reports which indicated higher prevalence of the *algD* gene among *P. aeruginosa* isolates from different infections within its host (Valadbeigi et al*.,* 2017). This gene has also been associated with prevalence of multi-drug resistance to different classes of antibiotics (Ghadaksaz et al*.,* 2015). The gene's influence on persisting infections could be the reason behind its high prevalence in the sequences analyzed during the present study. Wilcoxon rank test indicated significant results for the *htpG* gene model which had a p-value of 0.01736 with a rank test value of 3.

The study used profile HMM analyses to in-corporate information concerning the conservation

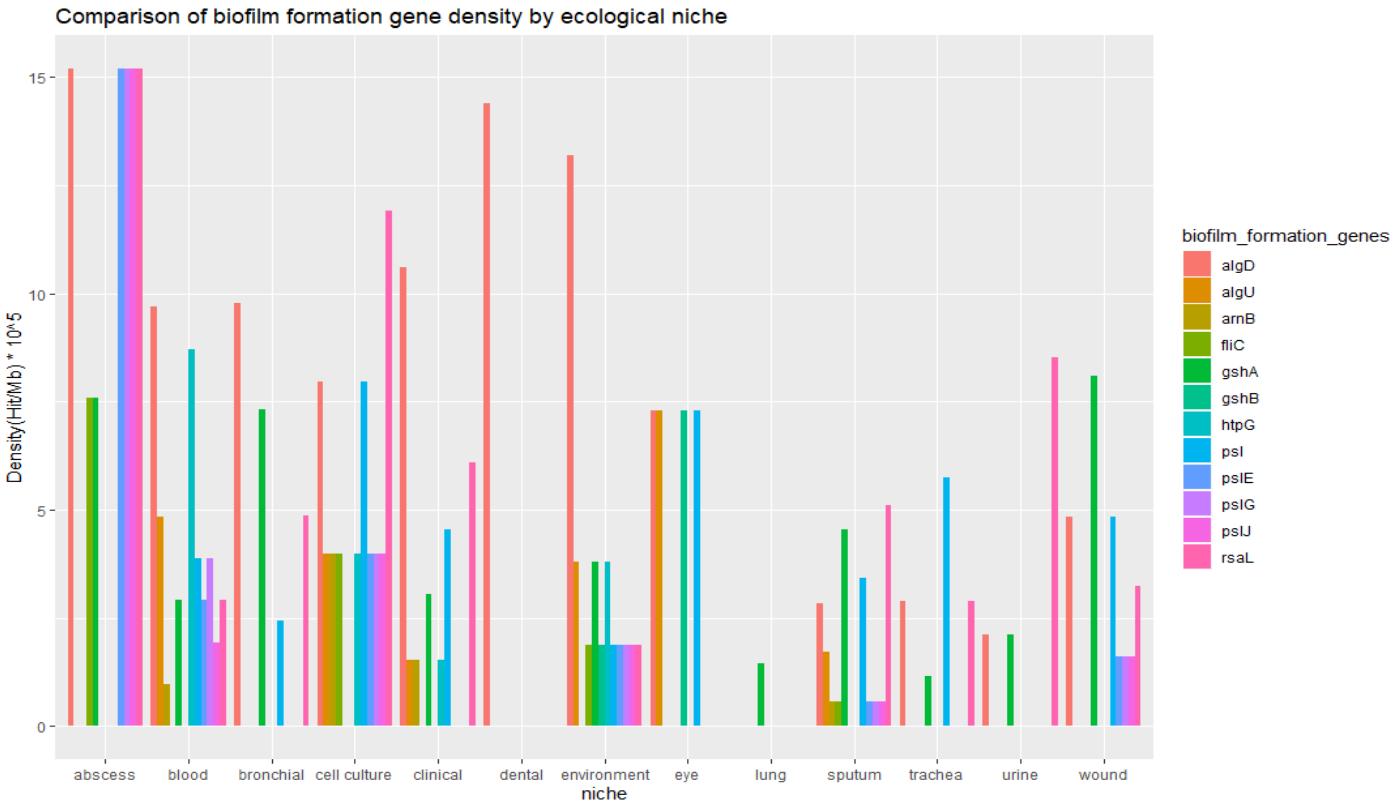

Figure 2. Comparison of biofilm formation genes by ecological niche reported by the profile Hidden Markov Models. The abscess ecological niche had a significantly high density of hits compared to the other ecological niches.

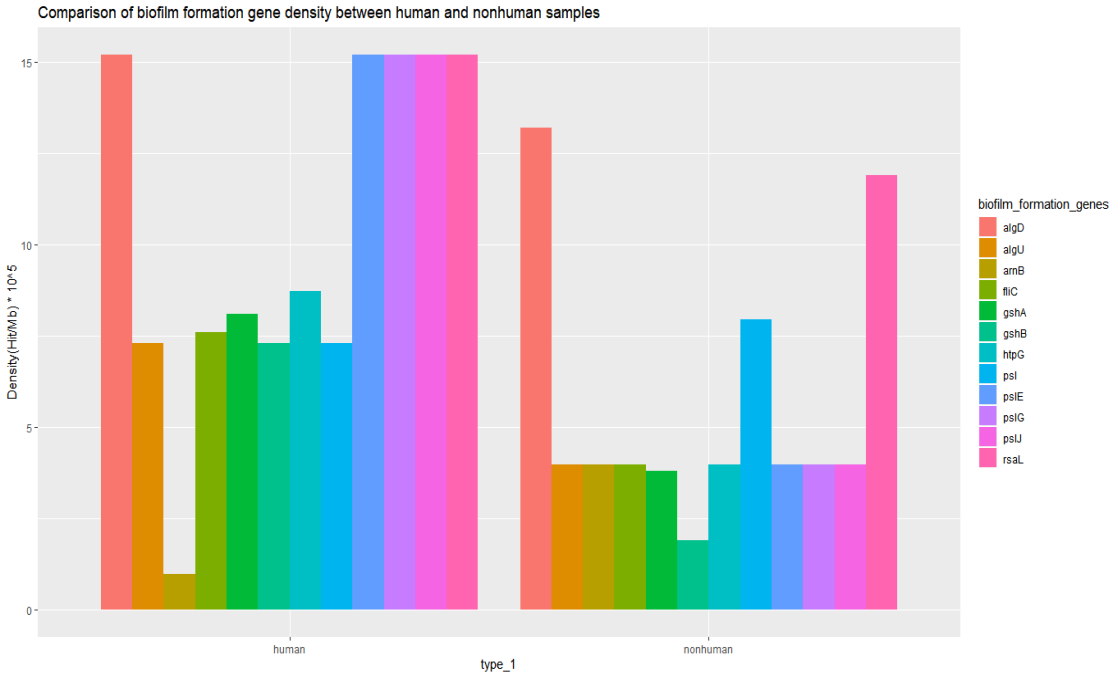

Figure 3. Comparison of biofilm formation genes between human and non-human strains reported by the profile Hidden Markov Models. The human strains had a significantly high density of hits compared to the non-human strains.

of different residues. Analyses from the constructed profiles of biofilm formation genes would be used to detect homologies and structural similarities between the sequence families of *P. aeruginosa*. From the multiple sequence alignments, the profile method built position-specific models to represent the conserved regions in the alignments. This approach has been previously applied in identifying putative applications of hidden Markov models in the analyses of biological sequence (Yoon, 2009). The state sequence, which in this case was a biologically meaningful alignment, was probabilistically inferred from the observed symbol sequence rather than simply being determined from the observed symbol sequence.

The parameters of the gene-specific models were set from the pre-aligned (pre-labeled) sequences. In this case the study assumed that the state paths were already known given that the multiple sequence alignments had been optimized. Previous studies have classified this approach as an optimal method of constructing and parameterizing profile HMMs (Bernardes et al*.,* 2007). The model converted both the state transitions and observed counts of symbol emissions into the transition and emission probabilities, respectively. These probabilities were based on the initially set transition and emission probabilities standards.

The study used the Forward algorithms to score and optimize the gene-specific pHMMs. Eddy (2011) incorporated such Forward algorithms in the development of the HMMER3 package, an accelerated pHMM construction pipeline. Alignments from the previous steps were used as input for building the profile HMMs. Building HMMs from multiple alignments was preferred in this case as the training algorithms (local optimizers) are suitable for less complex HMMs. With a less complex parameter space, there was little chance that the spurious local optima would trap the training algorithm. Given that the study was constructing profile hidden Markov models, the probability parameters were converted to additive log-odds scores. These log-odds scores were later used to score a query sequence once it is aligned against the constructed model. The profile HMM was preferred given that it is a well formulated probability model for representing similarity patterns within sequence families. These models also provide a precise method to search sequence databases using aligned sequences (Ahola et al*.,* 2003). For better accuracy of the database search using the pHMM, the study employed the efficient emission probability (EEP) estimation method to construct the gene-specific pHMMs. This estimation method ensured that the overfitting problem was overcome as signal was separated from the noise in conserved positions of the alignments, and reduced the parameter space as a result. The EEP method was preferred to the maximum likelihood estimation method for its better accuracy as indicated by a study conducted by Eddy (2011). The confidence intervals of representative emission probabilities were calculated to determine the effectiveness of the EEP estimation method. Shorter confidence intervals indicated that the model had an improved prediction power.

On screening the retrieved sequences of *P. aeruginosa* from different ecological niches, the study identified a significant number of hits for most of the ecological niches apart from the lung and dental ecological niches. These results give more credence to the importance of biofilm formation for the survival of the ubiquitous in different environments. Human niches also indicated a significantly higher number and density of hits compared to the non-human niches. Our study concluded that *P. aeruginosa* is more likely to form biofilms that increase its chances of survival once it colonizes the human host. This finding is consistent with previous studies which have correctly indicated that biofilm formation significantly contributes to the antibiotic resistance ability of this pathogen resulting in chronic illnesses for susceptible patients (Olsen, 2015).

The *algD* gene, previously described as a component of the alignate operon, demonstrated the highest number of hits compared to the other biofilm formation genes. Alignate biosynthesis, modification and export is important to chronic *P. aeruginosa* as these processes contribute significantly to antibiotic resistance and opsonization, resulting in highly potent pathogens (Okkotsu et al*.,* 2014). The significantly higher number of hits indicates an insistent need by the pathogen to express the *algD* gene. Antibiofilm therapies could be introduced to target the *algD* gene and impair alignate biosynthesis. Such therapeutic agents could limit the pathogens ability to persist when it causes infections. A clear understanding of the expression and mutation habits of this gene could prove worthwhile in the bid of developing novel treatment options against pathogenic strains of *P. aeruginosa.*

Conclusion

Across the human and non-human metagemones examined, this study successfully identified 197 unique biofilm formation gene clusters using the profile hidden Markov model, further highlighting the tremendous value of already sequenced metagenomes in identifying potential targets for novel therapeutic compounds. Given that the study was performed entirely *in silico,* experimental assays can be carried out on the different biofilm formation gene clusters to identify and characterize gene clusters that can be targeted to modulate chronic infections arising from pathogenic strains of *P. aeruginosa*.

CONFLICT OF INTERESTS

The authors have not declared any conflict of interests.

ACKNOWLEDGEMENTS

The authors appreciate Dr. Okoth Patrick for his critical comments and suggestions for improving the manuscript. Special thanks to the Department of Biological Sciences, SONAS, MMUST.

REFERENCES

Ahola V, Aittokallio T, Uusipaikka E, Vihinen M (2003). Efficient estimation of emission probabilities in profile hidden Markov models. Bioinformatics 19:2359-2368.

Bernardes JS, Davila AM, Costa VS, Zaverucha G (2007). Improving model construction of profile HMMs for remote homology detection through structural alignment. BMC Bioinformatics 8:435.

Castillo-Esparza JF, Hernandez-Gonzalez I, Ibarra JE (2019). Search for Cry proteins expressed by *Bacillus* spp. genomes, using hidden Markov model profiles. 3 Biotech 9(1):13.

Eddy SR (2004). What is a hidden Markov model? Nature Biotechnology 22:1315e6.

Eddy SR (2011). Accelerated Profile HMM searches. PLoS Computational Biology 7(10):e1002195.

Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792-1797.

Ghadaksaz A, Fooladi AAI, Hooseini HM, Amin M (2015). The prevalence of some Pseudomonas virulence genes related to biofilm formation and alignate production among clinical isolates. Journal of Applied Biomedicine 13(1):61-68.

Gong YN, Chen GW, Shih SR (2012). Characterization of subtypes of the influenza A hemagglutinin (HA) gene using profile hidden Markov models. Journal of Microbiology, Immunology and Infection 45:404-410.

Henikoff S, Greene EA, Pietrokovski S, Bork P, Attwood TK, Hood L (1997). Gene families: The taxonomy of protein paralogs and chimeras. Science 278(5338):609-614.

Kirsip H, Abroi A (2019). Protein structure-guided Hidden Markov Models (HMMs) as a powerful method in the detection of ancestral endogenous viral elements. Viruses 11(4):320.

Okkotsu Y, Little AS, Schurr MJ (2014). The *Pseudomonas aeruginosa* AlgZR two-component system coordinates multiple phenotypes. Frontiers in Cellular and Infection Microbiology 4:82.

Okonechnikov K, Golosova O, Fursov M (2012). Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics 28:1166-1167.

Olsen I (2015). Biofilm-specific antibiotic tolerance and resistance. European Journal of Clinical Microbiology 34:877-886.

R Core Team (2015). R: a language and environment for statistical computing. Vienna: The R Foundation for Statistical Computing. Available at https://www.r-project.org/.

Segev-Zarko LA, Kapach G. Josten M, Klug YA, Sahl SG, Shai Y (2018). Deficient lipid A remodeling by the arnB gene promotes biofilm formation in antimicrobial peptide susceptible *P. aeruginosa*. Biochemistry 57(13):2024-2034.

Skewes-Cox P, Sharpon TJ, Pollard KS, DeRisi JL (2014). Profile hidden Markov models for the detection of viruses within metagenomic sequence data. PLoS One 9(8):e105067.

Sonnhammer EL, Eddy SR, Durbin R (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28:405e20.

Valadbeigi H, Sadeghifard N, Salehi MB (2017). The prevalence of pilA and algD virulence genes in *P. aeruginosa* urinary tract and tracheal isolate. Infectious Disorders Drug Targets 17(2):86-89.

Yoon BJ (2009). Hidden Markov models and their applications in biological sequence analysis. Current Genomics 10:402-415.

Zheng W, Wuyun Q, Li Y, Mortuza SM, Zhang C, Pearce R, Ruan J, Zhang Y (2019). Detecting distant-homology protein structures by aligning deep neural-network based contact maps. PLoS Computational Biology 15(10):e1007411.