JOURNAL OF
EMERGING INVESTIGATORS

Article

Cutibacterium acnes sequence space topology implicates *recA* and *guaA* as potential virulence factors

Liliya Bohdan[1], Julia Platje[2]

[1] IB World School 0971, Wroclaw, Lower Silesia, Poland

[2] Biology Department, IB World School 0971, Wroclaw, Lower Silesia, Poland

SUMMARY

***Cutibacterium acnes* is a bacterium believed to play an important role in the pathogenesis of common skin diseases such as acne vulgaris. Currently, acne is known to be associated with strains from the type IA1 and IC clades of *C. acnes*, while those from the type IA2, IB, II, and III phylogroups are associated with skin health. This is the first study to explore the sequence space of individual gene products of different *C. acnes* phylogroups. Our analysis compared the sequence space topology of virulence factors to proteins with unknown functions and housekeeping proteins. We hypothesized that sequence space features of virulence factors are different from housekeeping protein features, which potentially provides an avenue to deduce unknown proteins' functions. This proposition should be confirmed based on further experimental outcomes. A notable similarity in the sequence spaces' topological features of previously known as housekeeping proteins encoded by *recA* and *guaA* genes to 'putative virulence' genes *camp2* and *tly* was observed. Our research suggests further investigation of *recA* and *guaA*'s potential virulence properties to better understand acne pathogenesis and develop more targeted acne treatments.**

INTRODUCTION

Housekeeping genes are typically defined by participation in cellular maintenance; they are essential to the functioning of a cell per se and are evolutionarily conserved (1). Virulence factors are molecules produced by pathogens that contribute to their ability to cause disease in a host organism. These factors can include toxins, enzymes, and adhesion molecules that allow the pathogen to colonize, evade the host's immune response, or inflict damage to host tissues (2). In the context of opportunistic species, studying virulence factors is particularly important because it can reveal how these species transition from harmless skin inhabitants to pathogens (2).

Cutibacterium acnes is a part of human skin microbiome and is often regarded as an opportunistic bacterium (3). It is well known that it can stimulate a wide range of inflammatory responses from host cells (3). Numerous studies mention cases when *C. acnes* inflammation was observed in connection to acne vulgaris, postoperative device-related infections, cerebrovascular, breast, spine, and cardiovascular device implants (3-6). There is also increasing evidence that the bacterium may act as an endogenous pathogen in degenerative disc disease (7).

These bacteria were historically studied as a single group, overlooking its intraspecies diversity. It is only recently that we started to tease apart how the pathogenic potential differs between phylogroups and specific strains that are associated with acne and skin health (8). *C. acnes* is most often divided into six major phylotypes: IA1, IA2, IB, IC, II, III. Other classification methods distinguish ribotypes, clonal complexes, or sequence types (9). The aforementioned classifications generally attempt to distinguish between disease and health-associated strains. Among those, sequence type approach offers the highest resolution, clonal complexes approach somewhat builds upon phylotypes, and ribotypes approach employs a fundamentally different methodology. Phylotypes are based both on disease association and phylogeny. A situation with a variety of coexisting approaches highlights the need to choose a classification method depending on the research objectives. This study required a clear distinction of the strains often noticed in association with disease for labelling sequences as 'favorable' and 'unfavorable'. Acne is known to be associated with strains from the type IA1 and IC clades of *C. acnes*, while those from the type IA2, IB, II, and III phylogroups are more frequently associated with skin health (9). In contrast, resolution was not the highest priority. As a result, the phylotype approach was chosen.

Sequence space refers to a conceptual landscape where each point represents a different sequence (such as DNA, RNA, or protein sequences), with proximity indicating similarity between sequences (10). Investigating the topology of sequence spaces enables researchers to uncover evolutionary relationships, functional convergence, and diversification among genes, including those encoding virulence factors. This approach is especially relevant since millions of genes, for example in the human gut microbiome, have been identified, yet their functions remain largely unknown. Using computational methods, researchers can categorize genes based on their sequence space features, generating hypotheses about their functions. By doing so, we can significantly advance our understanding of microbial systems and their impact on health (10).

Multidimensional Scaling (MDS) refers to a class of computational techniques used to uncover the hidden structure of data in such fields as molecular evolution, population genetics, and protein design (11). It offers the scientific community a convenient way of presenting only significant data features. That makes MDS revolutionary in two ways. Firstly, data can be re-used to discover or predict protein properties or shed light on molecular evolution. Secondly, sequence space exploration, in particular, sets a visual presentation of data as the primary goal, transforming overwhelmingly lengthy calculations results into interpretable

https://doi.org/10.59720/24-015

information. Other methods might still require a scientist to analyze the data first and then conduct additional processing steps to present it, making MDS a more efficient technique for data analysis. For this reason, MDS may be called more intuitive. MDS has been successfully utilized in a study of the local fitness landscape of the green fluorescent protein, for instance (12). With relatively fewer sequence data available for *C. acnes,* sequence space exploration was preferred over the fitness landscape.

To address our goal of understanding the virulence factor's sequence space features, we analyzed amino acid sequences of *C. acnes* proteins with the distinction between acne and health-associated lineages. Ten proteins were separated into two categories: virulence factors, and potential virulence factors with presumably housekeeping genes. The choice was motivated by the intension to utilize the molecules with the greatest number of available sequences and preferably known functions. We expected the sequence spaces of housekeeping genes to have different features from the sequence spaces of virulence factors. For instance, we expected to see distinct patterns in clusterization of acne versus health associated strains, be it reflected in the shape of the clusters, their relative positions or dispersion.

We selected Christie-Atkins-Munch-Peterson factor 2 (CAMP2) and putative hemolysin/FtsJ-like methyltransferase gene (*tly*) as virulence factors, though little is known about their exact functions due to the lack of experimental studies (13). Non-ribosomal recombinase A gene (*recA*) and guanosine monophosphate synthase gene (*guaA*) were selected into the housekeeping group (10). Other proteins we explored included the CAMP family, guanylate kinase (*gmk*), and superoxide dismutase (*sodA*) (14).

Touching upon functions of aforementioned molecules, the *gmk* gene encodes an enzyme involved in the synthesis and salvage of guanine nucleotides. It catalyzes a reaction that converts guanosine monophosphate to guanosine diphosphate, which is essential for energy conversion in a cell (15). The *sodA* product is responsible for inorganic ion transport and metabolism (16). Such ions could be either xenobiotic or co-factors in bacteria's enzymes. To summarize, while the role of most of the regarded in this work as housekeeping *C. acnes* proteins in acne pathogenesis remains unclear, they are not thought to be largely responsible for its virulent properties (15).

We selected five proteins from CAMP family for this study due to their potential for advancing our understanding of the subject matter. CAMP proteins can form pores in host membranes, leading to tissue damage (17). This general result likely applies to some of the CAMP proteins in *C. acnes*, although it does not necessarily indicate that all CAMP proteins in *C. acnes* are harmful to human host. For instance, it has been shown that CAMP2 of *C. acnes* attenuates the co-hemolytic reaction, whereas CAMP4 does not (17). CAMP1 and CAMP2 are the predominant CAMP factors produced by *C. acnes* overall. Interestingly, CAMP1 is strongly expressed by the types predominant in healthy skin, IB and II, whereas CAMP2 is mostly expressed by IA (2). Hence, the role of each protein within the CAMP family might be different and should be explored individually.

We hypothesized that sequence space features of virulence factors were different from housekeeping protein features. To test this, we obtained gene sequences from a Multilocus Sequence Typing Scheme (MLST) database, translated, aligned them, and visualized their sequence spaces. The purpose of this work was to suggest potential virulence factors by utilizing novel computational methods. This improved our understanding of the differences between *C. acnes* subtypes and demonstrated differences in the pathogenicity of strains based on virulence factors. We sought to identify the sequence space features of housekeeping and virulent proteins and compare those.

RESULTS

The primary goal of this study was to compare health-associated and harmful strains of *Cutibacterium acnes*. This required a rough division of phylogroups into 'acne-related,' 'unfavorable' (IA1 and IC) and 'health-related,' 'favorable' (IA2, IB, II, and III) to simplify visual analysis. The choice of this variant of classification was guided by the differences in pathogenic factors expression, antibiotic resistance, and the infection association (13, 18-25).

We aligned the sequences of each kind of gene in FASTA format and used multidimensional scaling to generate a plot representing the scaled Hamming distance between each point. The distance between each point represents sequence dissimilarity. The coordinates of the points that populate the sequence space (i.e., the image) are based on algorithmic projection. Their relative magnitudes were arbitrary because the nature of the sequence space concept is abstract. Moreover, a new execution of the algorithm produces a slightly different distance matrix. Therefore, the distances between the points are slightly different each time while still maintaining the general shape. Hence, axes did not have meaningful units of measurement and did not require numerical labeling.

The quality of projections is confirmed by analyzing stress (26). Stress is metrics specifically used for evaluation of projections reproducibility and is calculated within MDS Python package (see **Appendix**). If stress is less than 0.1, the picture will be nearly the same every time the code executes. With 0.0517 for CAMP2, 0.0604 for *tly*, 0.0966 for *recA*, 0.0502 for *guaA*, 0.0838 for *gmk*, 0.109 for *sodA*, 0.0855 for CAMP1, 0.0844 for CAMP3, 0.0989 for CAMP4, and 0.0832 for CAMP5, stress indicates excellent reliability of the visualizations (**Table 1**).

IA1, IA2, IB, and II phylotypes were prevalent in datasets for all gene sequences (**Figures 1-4**). IC and III were harder to draw conclusions about due to the lack of secondary data in the samples of the listed molecules.

First, we examined the virulence factors (**Figure 1**). IA1 phylotype was concentrated next to IA2, whereas II phylotype lay significantly farther from them. IB was located between IA2 and II. Generally, tight clusters of points that are well-separated from other clusters may indicate sub-populations in the data. While this offers an intuitive explanation of why representatives of one phylotype are closer to each other, the grouping of different subpopulations of *C. acnes* relative to each other is more insightful. There is a smaller sphere consisting of unfavorable IA1 and IC and an unspecified shape of favorable II and IB. This result indicates that the favorable sequence variants exhibit greater variation, whereas the less favorable ones are more tightly concentrated. In other words, the results suggest a higher sequence specificity and similarity within lineages associated with disease, regardless of their

https://doi.org/10.59720/24-015

Virulence Factors		Potential Virulence Factors		Housekeeping Genes	
Name	Stress value	Name	Stress value	Name	Stress value
CAMP2	0.0517	*recA*	0.0966	*gmk*	0.0838
tly	0.0604	*guaA*	0.0502	*sodA*	0.109
				CAMP1	0.0855
				CAMP3	0.0844
				CAMP4	0.0989
				CAMP5	0.0832

Table 1: Comparative stress values and categorization of *C. acnes* gene products. The table contains the names of the *C. acnes* molecules with the stress values of the corresponding sequence spaces visualizations and specifies the molecule's category in the experiment. CAMP2 - Christie-Atkins-Munch-Peterson factor 2, *tly* - putative hemolysin/FtsJ-like methyltransferase gene, *recA* - non-ribosomal recombinase A gene, *guaA* - guanosine monophosphate synthase gene, *gmk* - guanylate kinase, *sodA* - superoxide dismutase, CAMP1 - Christie-Atkins-Munch-Peterson factor 1, CAMP3 - Christie-Atkins-Munch-Peterson factor 3, CAMP4 - Christie-Atkins-Munch-Peterson factor 4, CAMP5 - Christie-Atkins-Munch-Peterson factor 5. The data for CAMP1, CAMP2 and CAMP4 is from Mayslich (2021) and Yu (2016), and the rest is from McDowell (2012).

phylogenetic relationships (IB is clustered away unfavorable from IA1 and IC). Thus, the less favorable phylotypes may be considered as a single group for more targeted research efforts. Contrary to our expectations, the two housekeeping genes *recA* and *guaA* displayed analogous patterns to the virulence factors (**Figure 2**).

Figures 3 and **4** present a control sample of molecules regarded as housekeeping. They show a mixed pattern, significantly different from one of the virulence factors. Here, almost no groups consist of lineages of one phylotype. Interestingly, the CAMP family has slightly more distinct clusters than products of *gmk* and *sodA* genes, although the colors of the dots within the clusters are approximately equally distributed.

At first glance, IA1, IA2, and IB phylotypes form tight clusters within the bigger grouping. However, the stress of the solution would be minimally affected by rearranging points in a tight cluster. Consequently, the arrangement of points within these larger groupings warrants more careful consideration than the relative positioning of the larger groupings. Nonetheless, the virulence factors and housekeeping proteins sequence spaces are distinct.

DISCUSSION

We explored molecules with known virulence relations and then compared the patterns to molecules with unknown properties. Established virulence factor genes CAMP2 and *tly* showed separate groupings for each phylotype within the corresponding sequence space. The observed similarity of two under-researched and presumed housekeeping genes *recA* and *guaA* to virulence factor genes CAMP2 and *tly* suggests that they might have virulent properties. In contrast, *gmk, sodA, camp1, camp3, camp4,* and *camp5* are likely to indeed be housekeeping genes and play no significant role in acne pathogenesis. In the case of *camp1*, this conclusion is supported by its increased expression in healthy-skin-associated strains (17).

It is worth keeping in mind that every sequence space is somewhat incomplete due to the limited number of sequences available for analysis. Therefore, if sequence space does not seem to have a distinguished pattern, more data points to fill it in with might clarify details for I phylotype, for example.

One point to keep in mind that the lineages were historically classified using different methods that offered varying levels of resolution, reliability, or speed (13, 27-30). This study prioritized the strength of the lineages' association with acne. Therefore, the division according to phylotypes was chosen as the most suitable, but it might not be optimal for other experiments with *C. acnes.* Such division on acne versus health-associated phylotypes is, undoubtedly, a simplification of microbial relationships with human host. Nonetheless, it was necessary and beneficial in this study.

It is clear that IA1 and IC phylotypes have more involvement in acne pathogenesis than other strains, but what exactly explains this phenomenon remains poorly understood. Our

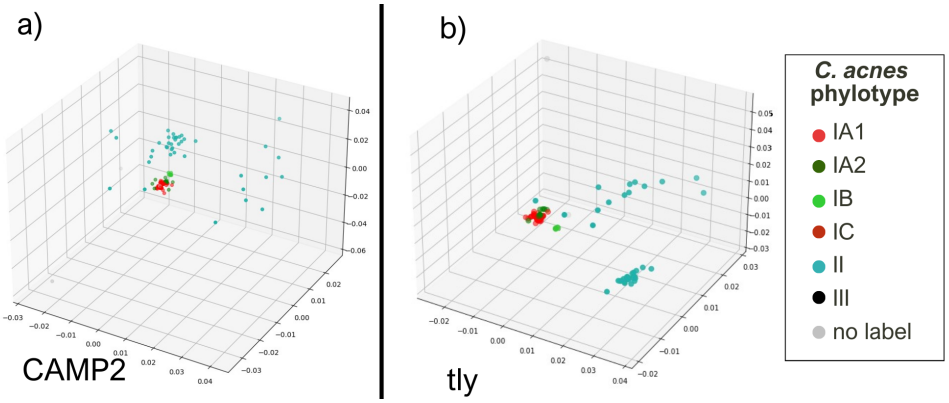

Figure 1: Disparity in the distribution of acne and healthy-skin-associated *C. acnes* strains in virulence factors a) CAMP2 and b) *tly* sequence spaces. Each point represented a protein from one bacterial cell. Different colors were used to visualize the phylotype the cell belongs to. The further the two points were located, the more different amino acids these two sequences had. Each 3D sequence space contained one type of protein but from different bacteria. The MDS plots illustrate that acne-associated strains cluster away from the health-associated ones within the two separate sequence spaces of virulence factors a) CAMP2 and b) *tly* (n=1). Here, IA1 and IC phylotypes are considered acne-associated, whereas IA2, IB, II, and III are regarded as health-associated. Stress was a) 0.0517 for CAMP2 and b) 0.0604 for *tly*.

https://doi.org/10.59720/24-015

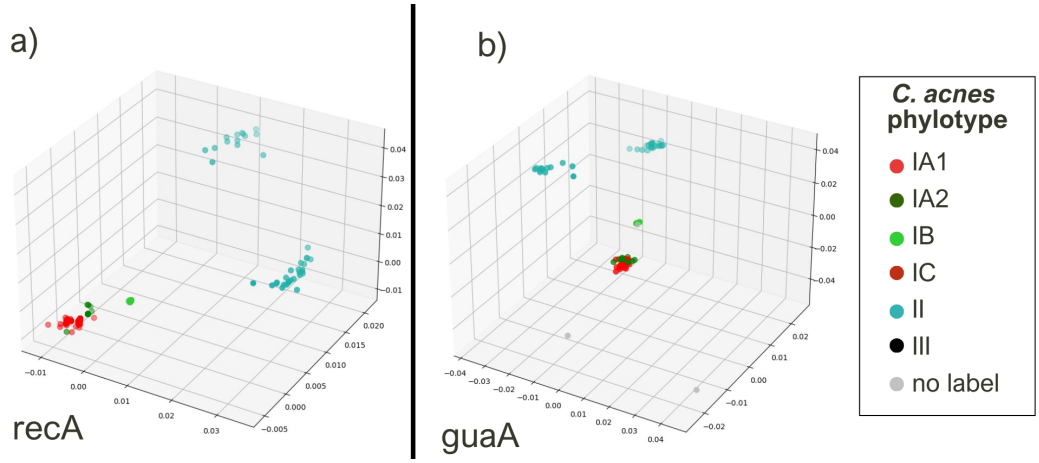

Figure 2: The sequence spaces of a) *recA* and b) *guaA* share a similar strain distribution to that of virulence factors. Each point represented a protein from one bacterial cell. Different colors were used to visualize the phylotype the cell belongs to. The further the two points were located, the more different amino acids these two sequences had. Each 3D sequence space contained one type of protein but from different bacteria. The MDS plots illustrate how the sequence spaces of a) *recA* and b) *guaA*—typically known as housekeeping genes—were filled in with gene products corresponding to different types of *C. acnes* (n=1). Here, IA1 and IC phylotypes are considered acne-associated, whereas IA2, IB, II, and III are regarded as health-associated. Stress was a) 0.0966 for *recA* and b) 0.0502 for *guaA*.

analysis revealed a notable similarity in the topological features of the sequence spaces of housekeeping proteins encoded by genes *recA* and *guaA* to virulence factors CAMP2 and *tly* gene products, highlighting the necessity of further experimental research of *recA* and *guaA* properties. If they are indeed virulence factors, they may be a valuable data point to consider when determining certain lineages association with health or disease. This might lead to further refining approaches to *C. acnes* intraspecies classification.

Additionally, it would be promising to separately examine tight clusters of IA1 and IC phylotypes within the sequence spaces of CAMP2, *tly, guaA,* and *recA*. This step would require obtaining more sequence data. A better knowledge of the evolution trajectories of listed molecules may serve as a foundation for a more holistic view of what molecular mechanisms drive the pathogenicity of certain strains. Consequently, the role of *C. acnes* in acne and other conditions may be better understood. By researching pathways these molecules are involved, we could develop ways to precisely modulate skin microbiome. Thus, therapies targeting specific strains or molecular pathways in these strains may be further developed.

MATERIALS AND METHODS

Translated sequences of each kind of gene were downloaded in FASTA format from PubMLST database isolate collection via the Sequence Export function (14). The majority of isolates were from normal human skin, followed by those sourced from acneic skin, in Europe in 2019. Isolate name, clonal complex, and phylotype were included in identifier.

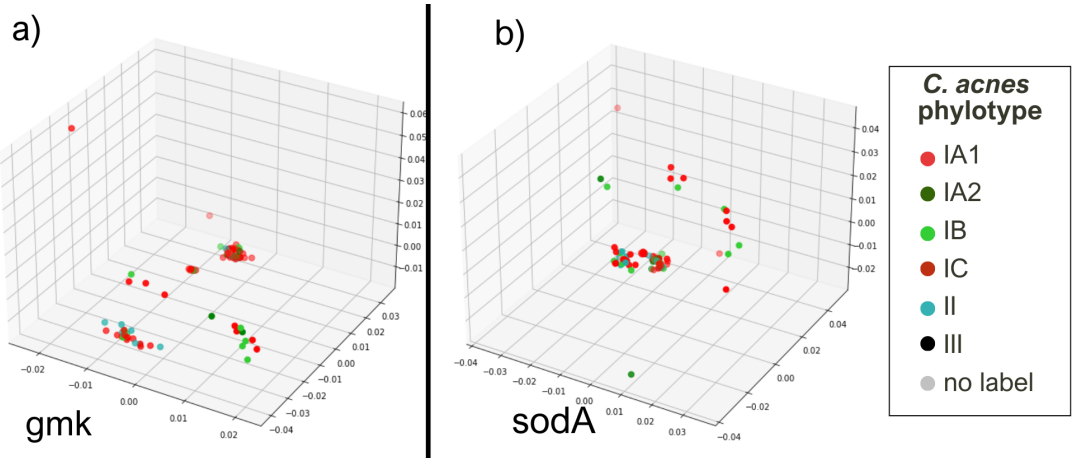

Figure 3: Distribution of housekeeping genes a) gmk and b) sodA sequence spaces in acne or healthy-skin-associated *C. acnes* strains. Each point represented a protein from one bacterial cell. Different colors were used to visualize the phylotype the cell belongs to. The further the two points were located, the more different amino acids these two sequences had. Each 3D sequence space contained one type of protein but from different bacteria. The MDS plots illustrate how the sequence spaces of a) *gmk* and b) *sodA*—proteins with presumably neutral roles in acne pathogenesis—were filled in with products corresponding to different types of *C. acnes* (n=1). Here, IA1 and IC phylotypes are considered acne-associated, whereas IA2, IB, II, and III are regarded as health-associated. Stress was a) 0.0838 for gmk and b) 0.109 for *sodA*.

https://doi.org/10.59720/24-015

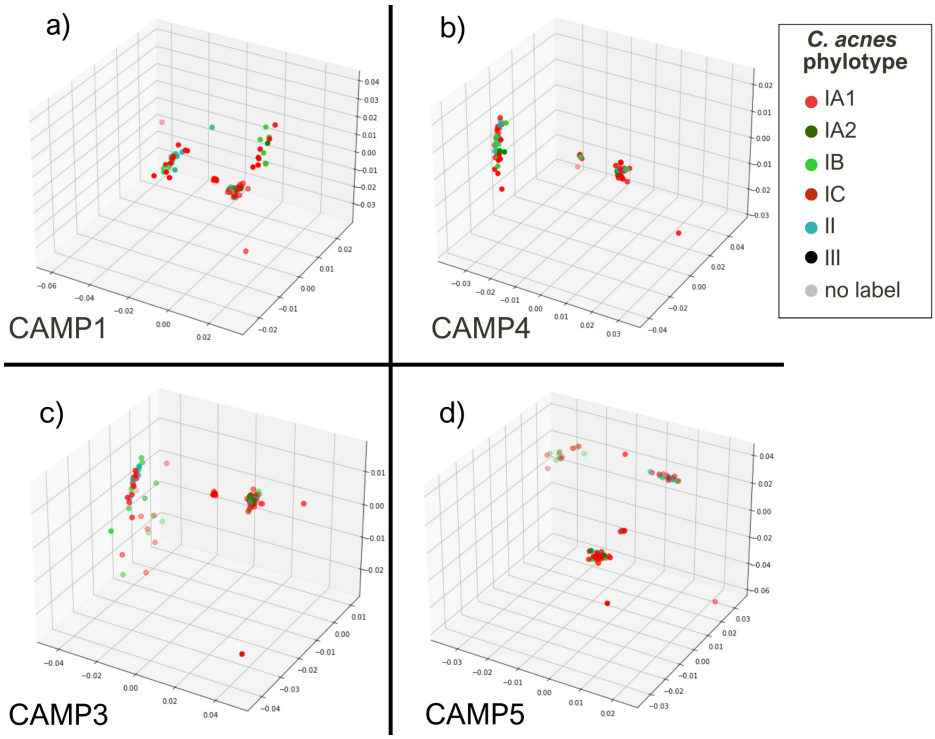

Figure 4: Distribution of housekeeping genes a) CAMP1, b) CAMP3, c) CAMP4, and d) CAMP5 sequence spaces in acne or healthy-skin-associated *C. acnes* strains. Each point represented a protein from one bacterial cell. Different colors were used to visualize the phylotype the cell belongs to. The further the two points were located, the more different amino acids these two sequences had. Each 3D sequence space contained one type of protein but from different bacteria. The MDS plots illustrate how the sequence spaces of a) CAMP1, b) CAMP3, c) CAMP4, and d) CAMP5—proteins with presumably neutral roles in acne pathogenesis—were filled in with gene products corresponding to different types of *C. acnes* (n=1). Here, IA1 and IC phylotypes are considered acne-associated, whereas IA2, IB, II, and III are regarded as health-associated. Stress was a) 0.0855 for CAMP1, b) 0.0844 for CAMP3, c) 0.0989 for CAMP4, and d) 0.0832 for CAMP5.

The initial samples contained all publicly available sequences for the listed molecules of *C. acnes*. Not more than 8% of the original sample size had to be excluded because of the flagged problems of incompleteness. The final samples include sequences from 246 bacteria for *aroE, lepA, atpD*, and *tly*, and 269 for CAMP2.

As a next step, multiple alignment was performed for each group of proteins using the Clustal Omega multiple sequence alignment tool. Obtained files were separately inputted in the Python Kaggle notebook equipped with two key modules: MDS algorithm and the visualization module (see **Appendix**). The first module calculates the distances between each of the sequences using the Hamming distance method. The MDS was implemented using Scikit Learn library. The visualization module incorporated a color-coding system, using green shades to represent 'favorable' *C. acnes* strains and red shades to show 'unfavorable' ones. Unfavorable strains belonged to IA1 and IC phylotypes; favorable were of IA2, IB, II, and III phylotypes.

3D interactive projections obtained as output were used to draw conclusions, but cannot be attached to this paper. Therefore, the 2D versions that can still demonstrate the key findings were used for illustration purposes.

ACKNOWLEDGMENTS

We would like to thank School of Molecular and Theoretical Biology (SMTB) for the opportunity to use the instruments for this scientific investigation and making this work possible. We would also like to acknowledge the 'The Biomolecular Universe: Intersection' project for providing the opportunity to optimize the Python code developed there for this study.

Received: December 22, 2023
Accepted: July 2, 2024
Published: May 01, 2025

REFERENCES

1. Joshi, C. J., et al. "What Are Housekeeping Genes?" *PLoS Computational Biology*, vol. 18, no. 7, 13 July 2022, https://doi.org/10.1371/journal.pcbi.1010295.
2. Mayslich, Constance, et al. "Cutibacterium Acnes as an Opportunistic Pathogen: An Update of Its Virulence-Associated Factors." *Microorganisms*, vol. 9, no. 2, 2 Feb. 2021, p. 303, https://doi.org/10.3390/microorganisms9020303.
3. Portillo, María Eugenia, et al. "Propionibacterium Acnes: An Underestimated Pathogen in Implant-Associated Infections." *BioMed Research International*, vol. 2013, 2013, pp. 1–10, https://doi.org/10.1155/2013/804391.
4. Conen, Anna, et al. "Characteristics and Treatment Outcome of Cerebrospinal Fluid Shunt-Associated Infections in Adults: A Retrospective Analysis over an 11-Year Period." *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, vol. 47, no. 1, 1 July 2008, pp. 73–82, https://doi.org/10.1086/588298.
5. Rieger, U. M., et al. "Sonication of Removed Breast

https://doi.org/10.59720/24-015

Implants for Improved Detection of Subclinical Infection." *Aesthetic Plastic Surgery*, vol. 33, no. 3, 26 Mar. 2009, pp. 404–408, https://doi.org/10.1007/s00266-009-9333-0.
6. Wagner, Leonie, et al. "Detection of Bacteria Colonizing Titanium Spinal Implants in Children." *Surgical Infections*, vol. 19, no. 1, Jan. 2018, pp. 71–77, https://doi.org/10.1089/sur.2017.185.
7. Rohacek, Martin, et al. "Bacterial Colonization and Infection of Electrophysiological Cardiac Devices Detected with Sonication and Swab Culture." *Circulation*, vol. 121, no. 15, 20 Apr. 2010, pp. 1691–1697, https://doi.org/10.1161/circulationaha.109.906461.
8. Lin, Yazhou, et al. "*Propionibacterium Acnes* Induces Intervertebral Disc Degeneration by Promoting Nucleus Pulposus Cell Apoptosis via the TLR2/JNK/Mitochondrial-Mediated Pathway." *Emerging Microbes & Infections*, vol. 7, no. 1, 10 Jan. 2018, pp. 1–8, https://doi.org/10.1038/s41426-017-0002-0.
9. McLaughlin, Joseph, et al. "Propionibacterium Acnes and Acne Vulgaris: New Insights from the Integration of Population Genetic, Multi-Omic, Biochemical and Host-Microbe Studies." *Microorganisms*, vol. 7, no. 5, 13 May 2019, p. 128, https://doi.org/10.3390/microorganisms7050128.
10. Sarkisyan, Karen S., et al. "Local Fitness Landscape of the Green Fluorescent Protein." *Nature*, vol. 533, no. 7603, 11 May 2016, pp. 397–401, https://doi.org/10.1038/nature17995.
11. Kruskal, Joseph B, and Myron Wish. *Multidimensional Scaling*. SAGE Publications, 1 Jan. 1978, https://doi.org/10.4135/9781412985130.
12. McDowell, Andrew, et al. "Over a Decade of RecA and Tly Gene Sequence Typing of the Skin Bacterium Propionibacterium Acnes: What Have We Learnt?" *Microorganisms*, vol. 6, no. 1, 21 Dec. 2017, p. 1, https://doi.org/10.3390/microorganisms6010001.
13. Vanni, Chiara, et al. "Unifying the Known and Unknown Microbial Coding Sequence Space." *ELife*, vol. 11, 31 Mar. 2022, p. e67667, https://doi.org/10.7554/eLife.67667.
14. McDowell, Andrew, et al. "An Expanded Multilocus Sequence Typing Scheme for Propionibacterium Acnes: Investigation of "Pathogenic", "Commensal" and Antibiotic Resistant Strains." *PLoS ONE*, vol. 7, no. 7, 30 July 2012, https://doi.org/10.1371/journal.pone.0041480.
15. "Cutibacterium Acnes." *PubMLST*, pubmlst.org/organisms/cutibacterium-acnes. Accessed 30 May 2023.
16. SIB Swiss Institute of Bioinformatics. "ENZYME - 2.7.4.8 Guanylate Kinase." *Enzyme.expasy.org*, enzyme.expasy.org/EC/2.7.4.8. Accessed 28 Sept. 2023.
17. Liu, J., et al. "Draft Genome Sequences of Propionibacterium Acnes Type Strain ATCC6919 and Antibiotic-Resistant Strain HL411PA1." Genome Announcements, vol. 2, no. 4, 14 Aug. 2014, https://doi.org/10.1128/genomea.00740-14.
18. Yu, Yang, et al. "Different Propionibacterium Acnes Phylotypes Induce Distinct Immune Responses and Express Unique Surface and Secreted Proteomes." *Journal of Investigative Dermatology*, vol. 136, no. 11, Nov. 2016, pp. 2221–2228, https://doi.org/10.1016/j.jid.2016.06.615.
19. McDowell, A., et al. "A New Phylogenetic Group of Propionibacterium Acnes." *Journal of Medical Microbiology*, vol. 57, no. 2, 1 Feb. 2008, pp. 218–224, https://doi.org/10.1099/jmm.0.47489-0.
20. Lomholt, Hans B., and Mogens Kilian. "Population Genetic Analysis of Propionibacterium Acnes Identifies a Subpopulation and Epidemic Clones Associated with Acne." *PLoS ONE*, vol. 5, no. 8, 19 Aug. 2010, p. e12277, https://doi.org/10.1371/journal.pone.0012277.
21. Eisen, JonathanA. "The RecA Protein as a Model Molecule for Molecular Systematic Studies of Bacteria: Comparison of Trees of RecAs and 16S RRNAs from the Same Species." *Journal of Molecular Evolution*, vol. 41, no. 6, Dec. 1995, https://doi.org/10.1007/bf00173192.
22. Tomida, S., et al. "Pan-Genome and Comparative Genome Analyses of Propionibacterium Acnes Reveal Its Genomic Diversity in the Healthy and Diseased Human Skin Microbiome." *MBio*, vol. 4, no. 3, 30 Apr. 2013, https://doi.org/10.1128/mbio.00003-13.
23. Nagy, István, et al. "Propionibacterium Acnes and Lipopolysaccharide Induce the Expression of Antimicrobial Peptides and Proinflammatory Cytokines/Chemokines in Human Sebocytes." *Microbes and Infection*, vol. 8, no. 8, July 2006, pp. 2195–2205, https://doi.org/10.1016/j.micinf.2006.04.001.
24. Stirling, Alistair, et al. "Association between Sciatica and Propionibacterium Acnes." *The Lancet*, vol. 357, no. 9273, 23 June 2001, pp. 2024–2025, https://doi.org/10.1016/S0140-6736(00)05109-6.
25. Olsson, Jan, et al. "Chronic Prostatic Infection and Inflammation by Propionibacterium Acnes in a Rat Prostate Infection Model." *PLoS ONE*, vol. 7, no. 12, 11 Dec. 2012, p. e51434, https://doi.org/10.1371/journal.pone.0051434.
26. McDowell, Andrew, et al. "The Opportunistic Pathogen Propionibacterium Acnes: Insights into Typing, Human Disease, Clonal Diversification and CAMP Factor Evolution." *PLoS ONE*, vol. 8, no. 9, 13 Sept. 2013, p. e70897, https://doi.org/10.1371/journal.pone.0070897.
27. Kruskal, Joseph B, and Myron Wish. *Multidimensional Scaling*. SAGE Publications, 1 Jan. 1978, https://doi.org/10.4135/9781412985130.
28. McDowell, Andrew, et al. "A Novel Multilocus Sequence Typing Scheme for the Opportunistic Pathogen Propionibacterium Acnes and Characterization of Type I Cell Surface-Associated Antigens." *Microbiology*, vol. 157, no. 7, 1 July 2011, pp. 1990–2003, https://doi.org/10.1099/mic.0.049676-0.
29. Fitz-Gibbon, Sorel, et al. "Propionibacterium Acnes Strain Populations in the Human Skin Microbiome Associated with Acne." *The Journal of Investigative Dermatology*, vol. 133, no. 9, 1 Sept. 2013, pp. 2152–2160, https://doi.org/10.1038/jid.2013.21.
30. Scholz, Christian F. P., et al. "A Novel High-Resolution Single Locus Sequence Typing Scheme for Mixed Populations of Propionibacterium Acnes in Vivo." PLoS ONE, vol. 9, no. 8, 11 Aug. 2014, p. e104199, https://doi.org/10.1371/journal.pone.0104199.

https://doi.org/10.59720/24-015

APPENDIX

Python code for one of the molecules, *tly*. For the rest of the molecules, the stages are analogical.

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

```
/kaggle/input/recaall250/tly250.txt
```

In [2]:

```
#import libraries/modules for work with data
import math
import plotly.graph_objects as go
import numpy as np
import matplotlib.pyplot as plt
from sklearn.manifold import MDS
```

In [4]:

```
#define the function for calculating the Hamming distance
def hamming(a, b):
    return sum([1 for i in range(len(a)) if a[i] != b[i]])
```

In [5]:

```
# define the function for calculating dimensions on the Hamming distance matrix
def dimension(hds_matrix, iter_split):
    x=np.linspace(0, 1, iter_split)
    x=x[1:]
    y=[]
    for k in x:
        count=0
        for i in range(len(hds_matrix)):
            for j in range(i, len(hds_matrix)):
                if hds_matrix[i][j]<k:
                    count+=1
        y.append(float(count))
    y=np.array(y)
    for i in range(len(x)):
        x[i]=math.log(x[i])
        y[i]=math.log(y[i])
    plt.plot(x, y)
    plt.show()
    dim=np.polyfit(x,y,1)[0]
    return(dim)
```

https://doi.org/10.59720/24-015

In [6]:

```
# read the file with alignments and extract the aligned sequences from it
file=open('/kaggle/input/recaall250/tly250.txt')
bulk=file.read()
file.close()
lines=bulk.split('\n>')
row_sequences=[] #includes aminoacids only
row_full=[] #includes labels too. has the same number of lines
for i in range(len(lines)):
    row_sequences.append(''.join(lines[i].split('\n')[1:]))
    row_full.append(''.join(lines[i].split('\n')))
```

In [10]:

```
#indexes of the types of bacteria I want to paint in different colors
IA1_indexes = []
for i in range(len(full)):
    if 'type_IA1' in full[i]:
        IA1_indexes.append(i)
```

In [12]:

```
IA2_indexes = []
for i in range(len(full)):
    if 'type_IA2' in full[i]:
        IA2_indexes.append(i)
```

In [13]:

```
II_indexes = []
for i in range(len(full)):
    if 'type_II' in full[i]:
        II_indexes.append(i)
```

In [14]:

```
III_indexes = []
for i in range(len(full)):
    if 'type_III' in full[i]:
        III_indexes.append(i)
```

In [15]:

```
IC_indexes = []
for i in range(len(full)):
    if 'type_IC' in full[i]:
        IC_indexes.append(i)
```

https://doi.org/10.59720/24-015

In [16]:

```
IB_indexes = []
for i in range(len(full)):
    if 'type_IB' in full[i]:
        IB_indexes.append(i)
```

In [22]:

```
counter=0
matrix=[[0 for i in range(len(sequences))] for j in range(len(sequences))]
```

In [23]:

```
# calculate the Hamming distance matrix
for i in range(len(sequences)):
    for j in range(i,len(sequences)):
        gaps=0
        counter+=1
        if counter%10000==0:
            print(counter)
        for k in range(len(sequences[i])):
            #pass
#print(i, j, k, sequences[i][k], sequences[j][k])
            if sequences[i][k]=='*' and sequences[j][k]=='*':
                gaps+=1
        matrix[i][j]=hamming(sequences[i], sequences[j])/(len(sequences[j])-gaps)
        # and gaps 1000
```

In [24]:

```
for i in range(len(matrix)):
    for j in range(i, len(matrix)):
        matrix[j][i]=matrix[i][j]
```

In [28]:

```
# calculate the average pairwise distance
sum([sum(i) for i in matrix])/(len(matrix)**2)
```

Out[28]:

```
0.017961745642724302
```

https://doi.org/10.59720/24-015

In [29]:

```
# calculate the dimensionality
dataset_dim=dimension(matrix, 1000)
print(dataset_dim)
outpit:
0.08507777916870213
```

In [30]:

```
# project points onto a space of the specified dimensionalit
mds = MDS(dissimilarity='precomputed', n_components=3, verbose=2)
matrix_fit = mds.fit_transform(matrix)
x=[i[0] for i in matrix_fit]
y=[i[1] for i in matrix_fit]
z=[i[2] for i in matrix_fit]
```

```
it: breaking at iteration 35 with stress 0.08707487471347763
```

In [31]:

```
#add colors
import itertools
A = []
for i in range(len(full)):
    if 'type_IA1' in full[i]:
         A = list(itertools.chain(A, ['firebrick']))
    if 'type_IA2' in full[i]:
         A = list(itertools.chain(A, ['forestgreen']))
    if 'type_IB' in full[i]:
        A = list(itertools.chain(A, ['forestgreen']))
    if 'type_IC' in full[i]:
        A = list(itertools.chain(A, ['firebrick']))
    if 'type_II' in full[i]:
        A = list(itertools.chain(A, ['forestgreen']))
    if 'type_III' in full[i]:
        A = list(itertools.chain(A, ['lime']))
```

In [33]:

```
import plotly.graph_objects as go
#this was added to add colors
fig = go.Figure(data=[go.Scatter3d(x=x,y=y,z=z,mode='markers',marker=dict(color=A))])
fig.show()
```