Exploration of Ultraconserved Elements in Bacteria Using Different Parameters

By: Chiemeka Nwakama, Jonathan Haak, Curtis Kokuloku

1. **Abstract/Summary paragraph**

In the world of genomics, there is often a focus on the differences between RNA sequences, especially when it comes to catching anomalies that could be causing illness in a person or, more broadly, the difference between different species of animals. Conversely, there is also a focus on similarity in the sense of how similar other organisms are to each other. Of course, in many aspects, humans are very different from a chicken, and in the same sense, both humans and chickens are very different than bacteria. Even within the same species of organisms, such as bacteria, there are many differences in how they function, infect their host, and look in shape. However, despite these differences, all living things originate from the same single-celled organisms, and as a result, even through millions of years of evolution, it can be seen that there are some things coded in very different bacteria that are 100% percent identical. These are known as ultra-conserved elements (UCEs); they represent artifacts of phylogenetic lineage from species to species. While ultraconserved elements (UCEs) first started off as having a strictly 100% match of 200 or more subsequence contiguous bases shared between two or more different genomes (Ryu, 2012), researchers unsatisfied with such strict limitations began exploring and tweaking different parameters such as the conservation threshold being lower than 100% at values such as 90% - 95% and seeing if they could still find useful information via these lesser conserved UCEs. Researchers also started adjusting how many bases quantified as UCEs, even going as low as the length of 100 or more contiguous bases instead of the original 200 (Wikipedia, 2023).

Though ultraconserved elements have been studied before, rarely has it been explored adjusting the parameters and threshold and attempting to see if any useful information can come from smaller ultraconserved elements even if they have less information and uniqueness. Our approach revealed ultraconserved k-mers in most of the 20 genomes (Genomes 0 to 19). Notably, no conserved elements were found in Genomes 4, 5, 11, 16, 17, 18, and 19. Genome 9 exhibited a conserved element with Genome 8 (E. coli strains), while Genome 3 (pneumonia strain) displayed conserved elements only with Genome 9, implying a hierarchical relation, possibly indicating a more recent taxonomy for Genomes 8 or 9. This exploration and these findings open up avenues for mapping protein functions in greater detail within these elements in the future, establishing taxonomical connections among different bacterial species and potentially extending into other organisms. Our hope is for researchers in the future to use what we learned to develop better methods that will yield more useful and impactful results.

Genomes:

---

sequence 0: NZ_NYJR01000001.1 Yersinia pestis strain 42126 42126_contig001
sequence 1: NZ_CAAKBY010000001.1 Clostridioides difficile strain LSHTM-81
sequence 2: NZ_JABBWS010000001.1 Micrococcus luteus strain MFP06 1
sequence 3: NZ_FWOJ01000045.1 Klebsiella pneumoniae strain VRCO0172
sequence 4: NZ_KK233310.1 Staphylococcus aureus VET0337R adZDQ-supercont1.1
sequence 5: NZ_CP037933.1 Flavobacterium nackdongense strain GS13 chromosome
sequence 6: NZ_CP075908.1 Clostridium perfringens strain CPM 77b chromosome
sequence 7: NZ_RJKM01000001.1 Saccharothrix texasensis strain DSM 44231 Ga0197498_11
sequence 8: NZ_NMIA01000100.1 Escherichia coli strain MOD1-EC6407
MOD1-EC6407_100_length_10555_cov_22.5471
sequence 9: NZ_CCQA01000001.1 Escherichia coli strain FHI47
sequence 10: NZ_BDOI01000001.1 Mycobacterium avium subsp. hominissuis strain Tone-12S
sequence 11: NZ_CP034662.1 Moraxella catarrhalis strain 46P58B1 chromosome
sequence 12: NZ_CYWW02000026.1 Listeria monocytogenes strain LM09-00558
sequence 13: NZ_CP007749.1 Campylobacter jejuni subsp. jejuni M129 chromosome
sequence 14: NZ_JAAIXC010000001.1 Burkholderia pseudomallei strain 47W_S67
sequence 15: NZ_BNHX01000001.1 Lactobacillus delbrueckii strain ME-790 sequence001
sequence 16: NZ_RQVD01000001.1 Veillonella sp. CHU740 contig_0001
sequence 17: NZ_JAPWCX010000001.1 Enterococcus lactis strain M426h 1
sequence 18: NZ_QBBF01000001.1 Helicobacter pylori strain CHL37 CHL37_Contig_1
sequence 19: NZ_JAHOAT010000095.1 [Clostridium] innocuum strain MSK.5.23 NODE_104_length_1019_cov_539.729

Genome Numbering Reference Table

2. **Summary of previous findings**

Describe previous or related work in this area and state any limitations. For example, you could mention and cite several prominent papers that have done the type of analysis you plan to do, discuss briefly what they did, and what is novel about your work.

Ryu et al. conducted a comprehensive study on ultraconserved elements across evolutionarily distant species, ranging from sponges to humans. They identified numerous UCEs, including those that are longer than 200 bp, indicating their presence in primitive organisms. The study suggests that UCEs cluster by sequence similarity in lineage-specific patterns, indicating independent emergence in common ancestors. The flanking genes of UCE clusters had distinct functions, often related to developmental regulation. While relevant to our bacterial UCE exploration, the lineage-specificity observed implies UCE richness may vary across bacterial clades. The study's methodology, involving whole genomes, repeat masking, and clustering, could inspire analytical approaches for bacterial UCEs. In summary, this study establishes the pervasiveness of UCEs in distant organisms, supporting exploration in bacteria, but underscores phylogenetic specificity in their emergence (Ryu, 2012).

Licastro et al. investigated the functions of ultraconserved elements, specifically their transcription, during mouse development. Using custom microarrays and next-generation sequencing, they explored the co-occurrence of enhancer and transcript functions within non-exonic UCEs. The study revealed extensive UCE transcription, particularly from non-exonic

regions, with 20% of UCEs showing both enhancer and transcript roles. This multifaceted functionality challenges previous assumptions and suggests diverse modes of action beyond enhancer functions. The findings provide valuable insights for bacterial UCE research, emphasizing the need to consider various roles beyond enhancers. Overall, the evidence of pervasive enhancer/transcript dual functionality contributes to debates on the functional and evolutionary significance of UCEs in mammalian genomes (Licastro et al., 2010)

Winker et al. explored the utility of ultraconserved elements in studying the population genomics of two bird species, McKay's bunting and snow bunting, thought to have diverged during the last glacial maximum. The study demonstrated that UCEs provided sufficient resolution to inform speciation history despite little absolute differentiation between populations. The findings indicate the importance of UCEs in analyzing recent speciation events, offering insights into demographic models and population genomics. The study's success at shallow evolutionary scales suggests the potential for similar analyses in bacterial populations. The methodological adjustments applied in this study may be valuable for uncovering UCE conservation reflecting shared fundamentals in bacterial populations despite vast phylogenetic distances (Winker et al., 2018).

Faircloth et al. explored ultraconserved elements as a source of genetic markers across amniote species separated by hundreds of millions of years of evolution. They identified over 5,000 UCEs by screening whole genome alignments of birds, lizards, and mammals. These UCEs were generally short (average 92.5 bp) but distributed throughout the genomes. The researchers designed probes targeting 2,386 UCE loci and used these to capture sequence data from nine bird species representing deep divergences. After enrichment and sequencing, they assembled upwards of 1,500 contigs per species corresponding to UCE-anchored loci. Phylogenetic analysis of over 850 loci unambiguously recovered established relationships even among these distantly related species (Faircloth et al., 2012).

Our approach differs from previous studies as they rarely explored adjusting parameters and thresholds for smaller UCEs, even if they possess less information and uniqueness. There is an argument that despite the UCEs being smaller, there can still be useful information to come out of them.

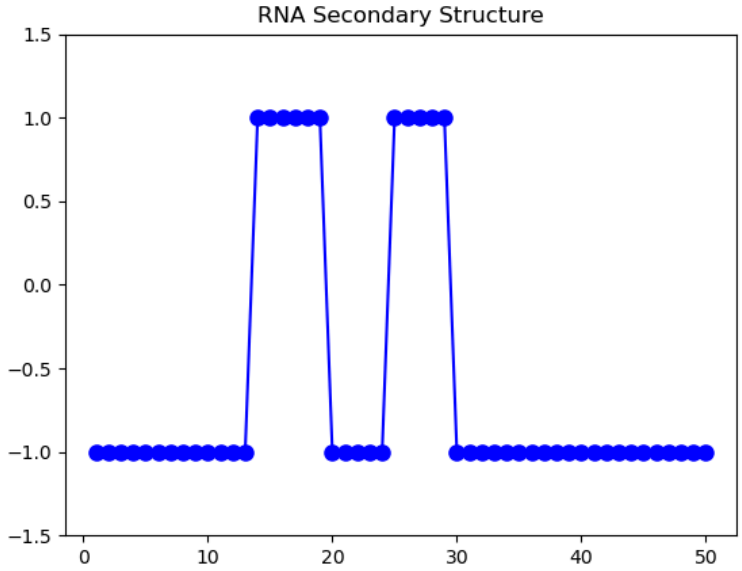
*Limitations of the study were mainly due to performance issues, which were likely Python-related. In future research, faster, more efficient algorithms written in languages such as C should be employed.

3. **Results**

Our results were limited by difficulties in computation, allowing us to only analyze a very small portion of each of the bacterial genomes that we chose, but despite this limitation what we discovered was quite interesting.

Almost all of the conserved sequences that we found mapped to hypothetical proteins, a hypothetical protein is a protein that has been predicted to exist, but has not yet been experimentally confirmed. Proteins are the key building blocks for cellular function, with many different functions related to the continued life cycle of the cell.

Many of the conserved sequences mapped to proteins with relatively simple folding structures, a protein having a simple folding structure may imply that the job that it is more fundamental, and this job is necessary for many different organisms. In contrast to a more complex folding structure which may indicate a more specialized role for the protein within the organism.

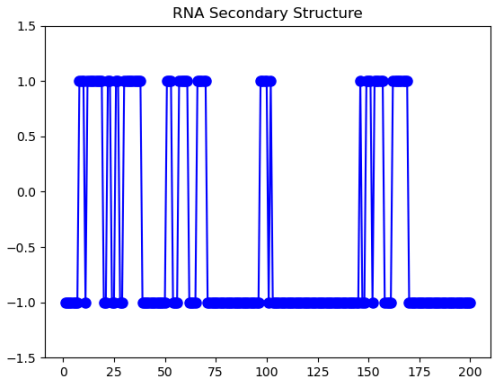


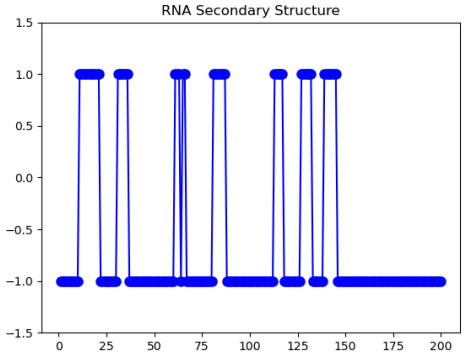
Conserved element from Genome 0

As shown in this graph, the conserved element is only 50kmers long, and the secondary structure of the protein that is coded for by the sequence is very simple.

For sequences 8 and 9, two different strains of E. Coli, we saw many more conserved sequences, which intuitively makes sense given these bacterial species have likely diverged fairly recently. Related to the other point of simple folding structures, we saw far more complicated folding structures in the sequences conserved between these two species. This seems reasonable because since the two species diverged more recently, not only are more general proteins conserved

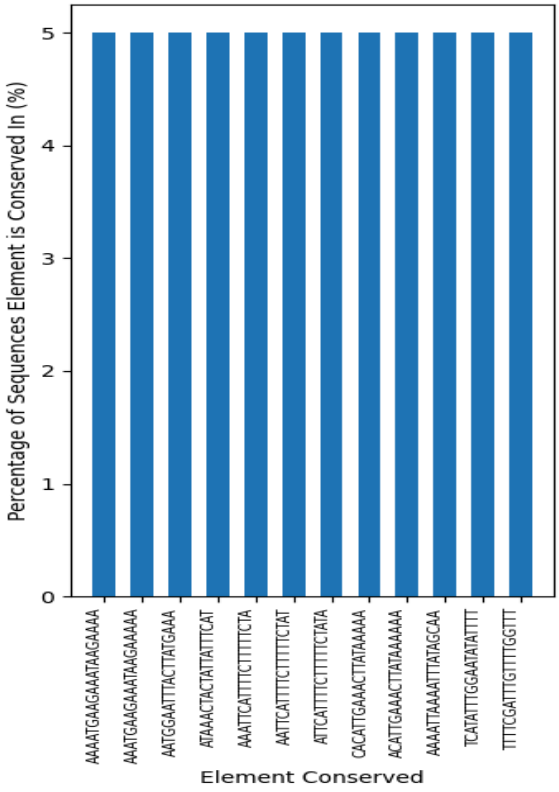
between them, but also more complex proteins as well.





Conserved element from Genome 8 **Conserved element from Genome 9**

As shown above, these two conserved elements from Genomes 8 and 9 are both 200kmers long, and contain a fairly intricate secondary structure.



Our analysis uncovered ultraconserved k-mers across the majority of the 20 genomes (Genomes 0 to 19). What was interesting was that no conserved elements were identified in Genomes 4, 5, 11, 16, 17, 18, and 19. Genome 9 demonstrated a conserved element connection with Genome 8 (E. coli strains), whereas Genome 3 (pneumonia strain) exclusively exhibited conserved elements with Genome 9. This suggests a hierarchical relationship, hinting at a potentially more

recent taxonomy for Genomes 8 or 9. Relationships like this will be useful to look at in future studies as it can give us an idea of essential functions that were conserved through potentially millions of years as well as how one organism's genome relates to another.

Results:

- Genome 0 (50mers) --- 3
- Genome 1 (20mers) ---9,5,12,13,15,3
- Genome 2 (20mers) --- 7, 14, 3, 10
- Genome 3 (200mers) -- 0, 9
- Genome 4 (20mers) --- 6, 13, 1
- Genome 5 (20mers) ---- 6, 9, 13, 1, 17
- Genome 6 (50mers) --- 2
- Genome 7 (20mers) --- 10, 14, 2
- Genome 8 (200mers) --- 9
- Genome 9 (200mers) --- 8

Resulting Reference Genomes Kmer Length and Sequences with Given Reference Genomes' Conserved Elements

4. **Conclusion**

In conclusion, our goal was to evaluate interesting patterns in ultra-conserved sequences between different bacterial species. This analysis was largely motivated by the desire to find sequences that coded to more fundamental and crucial functions of life, and we believed that the best place to look to find these fundamental functions was among different bacteria.

Due to computational limitations, our research was fairly exploratory and proof of concept, as we analyzed only a small portion of the data available for us in the genomes that we chose. We will still be able to see that more similar bacteria have longer and more complicated conserved sequences between them, and more distantly related bacteria have simpler conserved sequences.

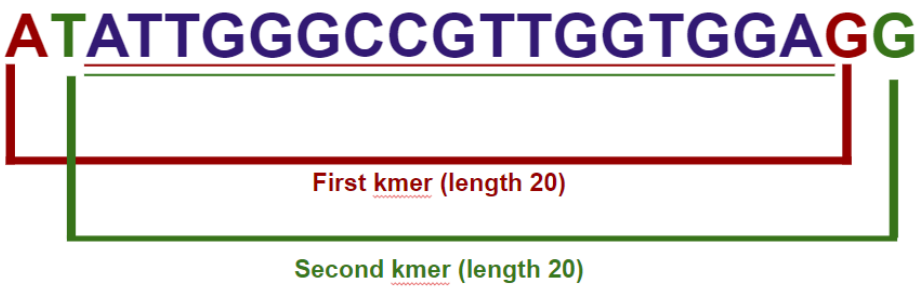
Overcoming these limitations will be essential to further gather more useful data to better understand the connection between not just different bacteria but also different organisms as a whole.

We believe that future research in this area holds a massive amount of potential, as with a large amount of computing along with the analysis of perhaps hundreds or thousands of different bacteria - or perhaps even simple eukaryotic species - analysis could be done to find highly conserved sequences across the genomes of many simpler organisms.

Existing techniques that predict protein function and folding could be leveraged on these ultraconserved sequences, and analysis in this area could help us understand what these sequences code for that is so essential and what this means regarding fundamental aspects of biological life.

5. **Methods**

In conducting the study, the initial step involved randomly selecting 20 genomes at random from the large NCBI database. Genomes were read in and then stored in a format fit for the developed algorithm. The selection of the first 10 genomes (Genomes 0 - 9) as reference genomes followed. An algorithm was produced to identify 100% conserved elements, taking into account parameters such as the reference sequence, conserved threshold for the number of sequences, number of bases per conserved element, and the number of kmers to explore. Now utilizing the algorithm to find conserved elements, parameters were tuned to find the maximal frequent set of conserved elements. In other words, tuned the kmer length parameter to the kmer length for element length until that k number >= 10 does not have any conserved elements.



How Kmers are Generated from Iteration to Iteration

This ensures that only the most significant conserved elements are kept. The search was limited to 10,000 bases of kmers due to performance issues, which was a key limitation of the study. After running the algorithm, A graphical representation of conserved elements within these genomes was then generated, and a detailed list of these elements was compiled in an FNA (FASTA Nucleic Acid) file. This file included the maximum conserved element base size with support unless it is already at 200 bp.

Subsequently, a CSV spreadsheet was created to document the position these elements appeared

in, in both the genome and the corresponding sequences containing ultra-conserved elements.

sequence 19: NZ_JAHOAT010000095.1 [Clostridium] innocuum strain MSK.5.23 NODE_104_length_1019_cov_539.729			
Ultra Conserved Elements			
ATATTGGGCCGTTGGTGGAGATATAAGTGGATCACTTTTCATCCGTCGTT			
		Position in sequence	Position in Reference Sequence
	sequence 3	35017	0
TATTGGGCCGTTGGTGGAGATATAAGTGGATCACTTTTCATCCGTCGTTG			
		Position in sequence	Position in Reference Sequence
	sequence 3	35018	1
ATTGGGCCGTTGGTGGAGATATAAGTGGATCACTTTTCATCCGTCGTTGA			
		Position in sequence	Position in Reference Sequence
	sequence 3	35019	2
TTGGGCCGTTGGTGGAGATATAAGTGGATCACTTTTCATCCGTCGTTGAC			
		Position in sequence	Position in Reference Sequence
	sequence 3	35020	3
TGGGCCGTTGGTGGAGATATAAGTGGATCACTTTTCATCCGTCGTTGACA			
		Position in sequence	Position in Reference Sequence
	sequence 3	35021	4

CSV Spreadsheet of a Reference Genomes Ultraconserved Elements

BLASTN and BLASTX algorithms were implemented during analysis on the FNA file of conserved elements for each reference genome; however, there were issues with a lack of alignment success with some of them and inability to analyze all the protein, taxonomy, and alignment findings.

Sequences producing significant alignments Download Select columns Show 10

select all *10 sequences selected* GenPept Graphics

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
hypothetical protein [Escherichia coli]	Escherichia coli	43.9	43.9	96%	7e-04	100.00%	118	MBB8949313.1
class I ribonucleotide reductase maintenance protein YfaE [Escherichia coli]	Escherichia coli	43.1	43.1	96%	0.001	100.00%	112	WP_250696342.1
hypothetical protein [Escherichia coli]	Escherichia coli	40.4	40.4	96%	0.007	100.00%	88	WP_213045624.1
hypothetical protein BFI50_10845 [Yersinia pestis subsp. microtus bv. Xilingolensis]	Yersinia pestis subsp. microtus bv. Xili...	40.4	40.4	96%	0.008	93.75%	92	OVY76039.1
hypothetical protein [Escherichia coli]	Escherichia coli	40.0	40.0	96%	0.010	93.75%	88	MCV7833527.1
hypothetical protein YPPY11_3693 [Yersinia pestis PY-11]	Yersinia pestis PY-11	39.3	39.3	96%	0.010	100.00%	50	EIR30146.1
hypothetical protein [Escherichia coli]	Escherichia coli	40.4	40.4	96%	0.010	93.75%	97	WP_261577903.1
hypothetical protein YPPY34_3577 [Yersinia pestis PY-34]	Yersinia pestis PY-34	39.3	39.3	96%	0.011	100.00%	47	EIR73473.1
hypothetical protein YPPY09_3609 [Yersinia pestis PY-09]	Yersinia pestis PY-09	39.3	39.3	96%	0.011	100.00%	49	EIR17169.1
hypothetical protein YPPY08_3600 [Yersinia pestis PY-08]	Yersinia pestis PY-08	39.3	39.3	96%	0.011	100.00%	48	EIR15359.1

BLAST Results for a Reference Genome

During the analysis, RNA secondary graphs were truncated or reduced if a significant number of conserved elements were identified. As for performance, to optimize program speed and efficiency, both multiprocessing and threading techniques were implemented. This algorithm was

carefully developed to ensure precision and reproducibility of results.
Considering the possibility of time constraints, a potential improvement for future analyses involves exploring the use of a lower-level, faster programming language like C instead of Python. Additionally, a limitation was set to analyze only the first 10,000 kmers in each search to manage computational resources effectively.

6. **Acknowledgements**

Chiemeka: did all the coding. Came up with the idea. Did much of the analysis.
Jonathan: Analyzed generated data, wrote results and conclusion and discovered insights within.
Curtis: Analyzed previous work and summarized sources, added pieces to write up.

References:

(1) Brant C. Faircloth, John E. McCormack, Nicholas G. Crawford, Michael G. Harvey, Robb T. Brumfield, Travis C. Glenn, Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales, Systematic Biology, Volume 61, Issue 5, October 2012, Pages 717–726, https://doi.org/10.1093/sysbio/sys004

(2) Licastro, D., Gennarino, V.A., Petrera, F. et al. Promiscuity of enhancer, coding and non-coding transcription functions in ultraconserved elements. BMC Genomics 11, 151 (2010). https://doi.org/10.1186/1471-2164-11-151

(3) Ryu, T., Seridi, L. & Ravasi, T. The evolution of ultraconserved elements with different phylogenetic origins. BMC Evol Biol 12, 236 (2012). https://doi.org/10.1186/1471-2148-12-236

(4) U.S. National Library of Medicine. (n.d.). Nucleotide blast: Search nucleotide databases using a nucleotide query. National Center for Biotechnology Information. https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&BLAST_SPEC=GeoBlast&PAGE_TYPE=BlastSearch

(5) Winker, K., Glenn, T. C., & Faircloth, B. C. (2018). Ultraconserved elements (UCEs) illuminate the population genomics of a recent, high-latitude avian speciation event. PeerJ, 6, e5735. https://doi.org/10.7717/peerj.5735

(6) Wikipedia. (2023, November 28). *Ultra-conserved element*. https://en.wikipedia.org/wiki/Ultra-conserved_element