Phylogeny of a Mosquitocidal *Bacillus thuringiensis* var *israelensis*

Ganga GC[1,2], Megha Raj Banjara[1], Ishan Gautam[1,3], Prakash Ghimire[1], Komal Raj Rijal*[1]

[1]Central Department of Microbiology, Tribhuvan University, Kirtipur, Kathmandu, Nepal,[2]Trichandra Multiple Campus, Ghanta Ghar, Kathmandu, Nepal,[3] Natural History Museum, Tribhuvan University, Swayambhu, Kathmandu, Nepal.

***Correspondence:** Komal Raj Rijal, Central Department of Microbiology, Tribhuvan University, Kirtipur, Kathmandu, Nepal. Email: komal.rijal@cdmi.tu.edu.np

ABSTRACT

Objectives: The study aims to compare the 16srRNA sequence of *Bacillus thuringiensis* var *israelensis* available in the National Center for Biotechnology Information, database, the sequence length, the relatedness, identity, and ancestral property by constructing a phylogenetic tree and by basic local alignment tool.

Methods: The 16srRNA sequence of all the *Bacillus thuringiensis* var *israelensis* was retrieved from the database. A basic local alignment tool was used to study the description of the sequences, the percent identity, graphical summary, and alignment. Mega software was used to draw the phylogenetic tree.

Results: The total number of mosquitocidal 16srRNA gene sequences was 19. One sequence VCRC-B17 strain was the complete genomic sequence deposited other 18 hits were partial sequences. The Expected value, percent identity, maximum, and total score of 18 hits shows a significant match with the query sequence VCRC-B17. The graphic summary result displays that the 18 hits align with the query sequence as they are coded with the highest score coding color bar, red with a ≥ 200 alignment score. The phylogenetic tree did not support the monophyletic status indicating there is a diversity in the strain.

Conclusion: The 19 sequences deposited in the database are identified as a mosquito larvicidal *Bacillus thuringiensis* var *israelensis* based on the alignment score or match even though they do not possess an equal base pair length with the query sequence.

Keywords: 16srRNA, mosquitocidal, alignment, Mega software, phylogenetic

INTRODUCTION

Bacillus thuringiensis var *israelensis* and *Bacillus sphaericus* are important bacteria used to control mosquito vectors as they produce a specific protein toxic to mosquito larvae (World Health Organization, 1985, 2012). Different countries have isolated these bacteria identified phenotypically and genotypically, especially by 16srRNA. The 16srRNA is universally present in all prokaryotes and has multiple sub-regions namely hypervariable region V1-V9 which can be used for distinct identification of various prokaryotes. Along with the hypervariable region, there are the conserved regions C1-C9 across all prokaryotes (Bukin et al. 2019; Johnson et al. 2019; Patwardhan et al. 2014; Petrosino et al. 2009). The 16srRNA gene sequencing is used as a tool to identify bacteria at the species level and assist with differentiating between closely related bacterial species.

Date of Submission: November 6, 2023
Date of Acceptance: December 15, 2023
Published Online: December, 2023
DOI: https://doi.org/10.3126/tujm.v10i1.60653

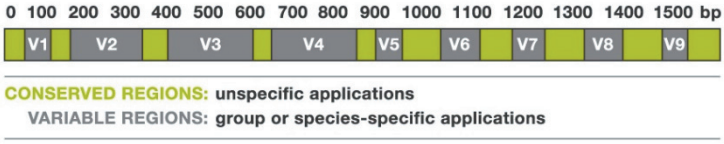

Figure 1. Structure of the prokaryotic 16srRNA gene (Source: http://1928diagnostic.com/taxonomic-classification)

It was observed that the 16srRNA sequence is mostly used for comparing it with related Genera, gene sequences present in the database for species identification or to distinguish from other related species. In GenBank, the sequences from the different countries are deposited and are compared for identification. The divergence and their evolutionary relatedness are among the sequences is not known. The size of the 16srRNA length is 1500 base pair (bp) Figure 1. but the deposited 16srRNA sequences in the database are not in this length, the short-length sequences also align with the comparative sequences or query sequence and are identified as a particular organism.

The research aims to identify why the genomic sequences 16srRNA short fragments also are identified as a particular strain or species or variety, their relatedness, identity and to know the deposited sequence of different countries have a common ancestral.

MATERIALS AND METHODS

For dry laboratory analysis, mosquitocidal, *Bacillus thuringiensis* var *israelensis* (Bti) 16srRNA gene sequences was searched in the National Center for Biotechnology Information (NCBI) GenBank by using the keywords *Bacillus thuringiensis* var *israelensis* 16s, Nucleotides (National Center for Biotechnology Information, 2006). The number of Bti sequences, the strain type, place of origin, sequence length, etc., deposited in the NCBI database was noted. The complete 16srRNA gene sequence deposited in the GenBank was used as a query sequence to perform Nucleotide Basic Local Alignment tool (nBlast) to obtain identical or similar gene sequences (Leung, 2011; Agostino, 2012). By using the FASTA sequences of Bti available in the NCBI database the sequenced were aligned by using Mega software. Then the phylogenetic tree was drawn by using the aligned sequences to note the evolutionary process (Agostino, 2012; Hall, 2013).

RESULTS

The NCBI nucleotide database shows 19 Bti's16s sequences are deposited in the GenBank from different countries. A complete 16srRNA sequence of 1563bp in length, VCRC B-17 strain of Bti deposited from India was taken as a query sequence. The other 18 sequences were partial sequences and variable in length Table 1. The E value of 17 sequences was 0.0 and the other two sequences of strain B6064, T14001, and B6064, T14001 from Russia had E values 4e-138 and 6e-57 respectively but the percent identity was 100%. Statistically generated Maximum and Total scores of each Bti strains are equal. Among the 18 hits maximum score is lower 220 in the case of the strain B6064, T14001 followed by B6064, T14001, (490), AAU2 1554, BGSCAQ2 1668, UEP234 1858.

Table 1. Bti 16srRNA sequences deposited in the NCBI database

S.N.	Sequence length bp	Strain code	Place of origin	Query coverage	E-value	Percent identity
1	1563 (query sequence)	Bti-VCRC B-17	India	100%	0.0	100.00%
2	1339	Bti	Japan	85%	0.0	99.63%
3	1405	Bti-30a	Uzbekistan	89%	0.0	100.00%
4	1006	Bti-UEP234	Turkey	64%	0.0	100.00%
5	1015	Bti-AAU2	Ethopia	61%	0.0	95.96%
6	1460	Bti-PK	Pakistan	92%	0.0	99.65%
7	1444	Bti-X-11	China	89%	0.0	94.32%
8	1444	Bti-1SPC-12	India	92%	0.0	100.00%
9	1475	Bti-AM65-52	France	94%	0.0	100.00%
10	1556	Bti-T0139	Brazil	99%	0.0	99.94%
11	903	Bti-BGSC4Q2	Uruguay	85%	0.0	100.00
12	1430	Bti-NCIM2513	India	91%	0.0	99.72%

S.N.	Sequence length bp	Strain code	Place of origin	Query coverage	E-value	Percent identity
13	1336	Bti-BGSC4Q1	UK	85%	0.0	100.00%
14	1385	Bti-1385	India	88%	0.0	99.93%
15	1473	Bti-SB67	Srilanka	93%	0.0	99.66%
16	1520	B6064	Russia	97%	0.0	99.93%
17	1523	Bti	USA	96%	0.0	99.47%
18	469	Bti-B6064, T14001	Russia	7%	6e-57	100.00%
19	265	Bti-B6064, T14001	Russia	16%	4e-138	100.00%

Percent identity 100% was shown by 8 sequences (hits) other 8 sequences showed 99% identity and the other 2 strains showed AAU2 (95.96%), and X-11 (94.32%). Query coverage means the % of subject sequence or hits that matched with the fraction of query sequence i.e., VCRC B-17 shows 100%, Other sequences 8 sequences shows 91-99% 5 sequences shows 85-89%, 61% by AAU2 strain, 64% UEP234 strain, 57% BGSC4Q2 strain, 7% B6064, T14001 strain and 16% B6064, T14001 strain.

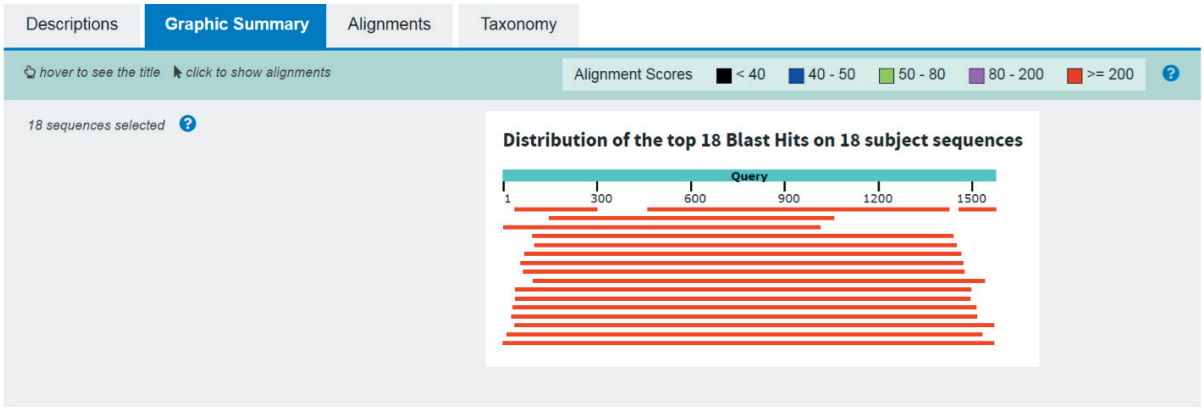

Figure 2. The graphic summary of the NCBI nBlast results. The query sequence VCRC B-17grey horizontal bar at the top is labeled as Query with the scale showing the bp length from 1 to 1500. Hits or subject sequences are presented by horizontal red bars below the scale of query sequence. The hit bars show variation in length.

The colorful graphical summary results of nBlast of Bti sequences in Figure 2. is an overview of the database sequence aligned to the query sequence. The alignment is coded ranging from black to red as indicated in the color label at the top. The hits are colored according to the obtained alignment scores. The query sequence VCRC-B17 strain appears as a grey horizontal bar at the top. The scale below shows the bp present in the query sequence from 1 to 1500 bp. Below the scale, the horizontal red bars in rows represent the hits of the 18 Bti sequences deposited in the GenBank of NCBI. The hits with the highest score are found. The color coding within the graphic is generated by the statistics of each hit. As indicated by the key at the top of the graphic hits with the highest score are red. In this blast search, all the hits are red bars, indicating the highest score for alignment. The first hit shows gaps indicated by the broken bar, 2nd hit has no full coverage along the query sequence and in 3rd 900bp are present and from the 4th hit, certain bp are missing at the left and right side of the query bar. All the hits contain the bp at the center of the query bar.

It is important to realize that these short bars show the length of the similarity between the query and the hit, and do not necessarily represent the entire length of the hit. i.e., only certain nucleotides are common with the query sequence. This is represented by the bars. The red bars appear in stacks one below the other indicating that each has some bp is common within the same place in the query sequence.

The phylogenetic tree did not support the monophyletic status a single clade is not seen Figure 3. The tree can be divided into three clades, that means there is the diversity in the strain isolated from different countries. It also indicates from a country different diverse strain can be isolated in case of Indian isolates; they are present in three clades.

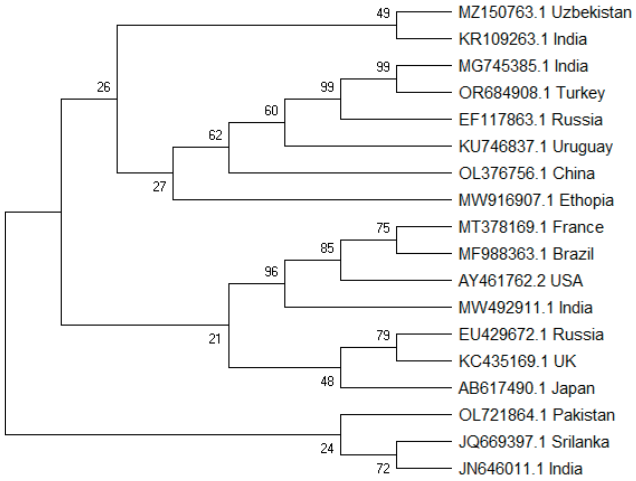

Figure 3. Evolutionary relationships of taxa

Even though the retrieved sequences have the same property i.e., they all are mosquito larvicidal organisms. But are diverse from one other. So, the sequence provides only distinguish it from other *Bacillus thuringiensis* varieties. Based on the place of origin different clades are formed. Or due to the sequence length, the clade may be formed.

DISCUSSION

Generally, the sequence length of the 16srRNA gene shown by different study and literature was approximately1500 -1550bp (Clarridge, 2004; Johnson et al. 2019). But the complete sequence of Bti -VCRC B-17 strain was found to be greater i.e., 1563bp, other 18 sequences deposited are partial sequences. The reason is due to the types of primer used for a variable region for amplification of 16srRNA gene and the sequencing platform that produce short reads or sequences (Johnson et al. 2019). Thus only 19 16srRNA sequence are retrieved from the database. The reason may be the sequencing cost or unavailability of sequencer or deposited in other databases which are not explored in this study so, few sequences available in NCBI database.

The Expected value or E value 0.0 of the sequences states that they are biologically or evolutionary significant alignment. The lower the E-value, or the closer it is to "0", the higher is the "significance" of the match so, the 17 sequences are homologous to the query sequences from the blast results but in the case of the two sequences, the sequence length or the bp is very less so, the E value so generated may be due to the alignment of the sequences in the major region of the query sequence so, the E value or p value generated statistically is significant.

Query coverage means the entire sequence or the fraction of the query sequence, matches with the parts of the subject sequences or hits present in the database. In this study as the query sequence is the complete 16srRNA sequence of bp length 1563 was 100% query coverage whereas the subject sequences or hits present in the database are partial sequences so, the query coverage displayed in the blast result page varies the query coverage of 8 strain falls in the range 91%-99% bp length (1430-1556), 5 strain 85-81% bp length (1336-1405), strain UEP234 64% (1006 bp), AAU2 strain 61% bp length (1015) and 16% and 7% B6064bp length 265, 469 respectively this is due to the length of the sequence deposited in the GenBank. Query coverage 16% and 7% of the two isolates are also identified as Bti. The two sequences deposited in NCBI database has a base pair length 469 and 265 even though the deposited sequence are low base pair showed 100% identity to Bti because 16srRNA gene are mainly sequenced for identification. Initial 500bp sequence provide an adequate differentiation for identification of most clinical bacterial isolates or other bacteria (Clarridge, 2004; Hong and Farrance, 2015). Not only the clinical bacterial isolates, in case of Bti isolates also the initial sequence help in differentiation among the *Bacillus* species. This may be the reason for 100% identity of the lower bp sequenced deposited in the database.

Maximum and total score of each Bti strains is equal. Each of the sequences has been assigned a score based on the extent of the match. The total score is obtained

by adding the maximum score from any region of the query sequence that matches any region on the subject sequences or hits.

According to proposed guidelines for bacterial classification, strains with less than 97% similarity or identity in 16srRNA gene sequence represent different bacterial species (Johnson et al. 2019). According to this statement two sequenced deposited in the Gen bank of NCBI strain. Only the following sequence have a bp length of 1500. So, to detect whether they are new strain or are the same strain for that the complete sequence of 16srRNA is required. To align so the identity % provided by the blast search doesn't indicate they have the same sequences as the reference sequence.

Graphic output gives a quick overview of the query sequence and the resulting hit sequences. All the hits are red in color indicating that all the hits have the highest score for alignment even though the sequence length is not equal to the sequence length of the query sequence. But all the hits had the main common nucleotides as the query sequence so the hits are represented by the red color bar, with the highest score for alignment. So, all the partial sequences deposited in the GenBank are the larvicidal Bti isolated from the different countries. Generally, various literature state that the 16srRNA sequence length is approximately 1500 but the Blast results indicate the 265 and 469 sequence bp also identify it as a Bti strain. The evolutionary history was inferred using the Neighbor-Joining method. The retrieved 16srRNA gene sequences were imported into MEGA for sequence alignment and phylogenetic analysis was performed. Statistical confidence on the generated tree topology was assessed with 1000 bootstrap replication. The bootstrap consensus tree inferred from 1000 replicates is taken to represent the evolutionary history of the taxa analyzed. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. The evolutionary distances were computed using the Maximum Composite Likelihood method and are in the units of the number of base substitutions per site. The analysis involved 19 nucleotide sequences. All ambiguous positions were removed for each sequence pair. There were a total of 2405 positions in the final dataset. Evolutionary analyses were conducted in MEGA X.

The reliability of a phylogenetic tree is measured by the bootstrap value if the bootstrap value is high for most branches the tree is considered to be reliable. Only few branches had a higher value and most of the branches had a lower value. Therefore, the reliability is questionable. The scale bar 0.20 was calculated means 0.20 substitution per nucleotide position. Wet laboratory analysis of the 16srRNA Bti strains that are deposited only in the NCBI GenBank. The study aims to know the types of strain, deposited in the GenBank by different countries that are different from each other or diverge from each other.

CONCLUSIONS

The 19 sequences deposited in the database are identified as a mosquito larvicidal *Bacillus thuringiensis* var *israelensis* based on the alignment score or match even though they do not possess an equal base pair length with the query sequence. The strains deposited from different countries are diverse from each other. Diversity is seen among the isolates obtained from the same country.

ACKNOWLEDGMENTS

The author would like to acknowledge the University Grants Commission for providing financial support (S&T-12) and the Central Department of Microbiology for the opportunity to attend Bioinformatics training.

CONFLICT OF INTEREST

There is no conflict of all authors in this study

REFERENCES

Agostino M (2012). Introduction to the Blast suite and BlastN. *Practical Bioinformatics* 47-71.

Bukin YS, Galachyants YP, Morozov IV, Bukin SV, Zakharenko AS and Zemskaya TI (2019). The Effect of 16srRNA Region Choice on Bacterial Community Metabarcoding Results. *Scientific Data* **6**: 1-14.

Clarridge JE (2004). Impact of 16srRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. *Clinical Microbiology Review* **17**(4): 840-862.

Hall BG (2013). Building Phylogenetic Tree from Molecular Data with MEGA. *Molecular Biology and Evolution* **30**(5): 1229-1235.

Hong S and Farrance CE (2015). Is it Essential to Sequence the Entire 16srRNA Gene for Bacterial

Identification?. *Review* 1-7. http://media.europeanpharmaceuticalreview.com/wp-content/uploads/Is-it-essential-to-sequence-the-entire-16S-rRNA-gene-for-bacterial-identification-SH-and-CF-MJ.pdf

Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, Leopold SR, Hanson BM, Agresta HO, Gerstein M, Sordergren E and Weinstock GM (2019). Evaluation of 16srRNA Gene Sequencing for species and Strain-level Microbiome Analysis. *Nature Communications* **10**(1): 1-11.

Leung W (2011). Lab week 8- An in Depth Introduction to NCBI Blast. *Washington University* **1**: 99-116.

National Center for Biotechnology Information (2006). The New Blast Result Page. *In NCBI handout series New Blast* Last updated on May 28, 2019.

Patwardhan A, Ray S and Roy A (2014). Molecular Markers in Phylogenetic Studies- A Review. *Journal of Phylogenetics & Evolutionsry Biology* **2**(02): 1-9.

Petrosino JF, Highlander S, Luna RA, Gibbs RA and Versalovic J (2009). Metagenomic Pyrosequencing and Microbial Identification. *Clinical Chemistry* **55**(5): 856-866.

World Health Organization (1985). Informal Consultation on the Development of *Bacillus sphaericus* as a Microbial Larvicide. *World Health Organization*. https://iris.who.int/handle/10665/60326?locale=ar

World Health Organization (2012). WHO Specification and Evaluation for Public Health Pesticides: *Bacillus thuringiensis* subspecies *israelensis* strain AM65-52. *World Health organization*. http://www.who.int/whopes/quality/en/