Annotation of the *Bacillus thuringiensis str. Al Hakam* Genome from Locus Tags BALH_0274 to BALH_0277

Andrew Alexis*, Tessa Begley*, Thomas Dorsey*, Oscar Wilcox, Jonathan Gambill, and Peter Hentschke
The Harley School, Rochester, New York and The Western New York Genetics in Research and Health Care Partnership

SEPA SCIENCE EDUCATION PARTNERSHIP AWARD
Supported by the National Institutes of Health

University at Buffalo

Abstract

A group of three genes from the bacterium *Bacillus thuringiensis str. Al Hakam* (BALH_0274 – BALH_0277, excluding BALH_0276) were annotated using the collaborative genome annotation website GENI-ACT. The Genbank proposed gene product name for each gene was assessed in terms of the general genomic information, amino acid sequence-based similarity data, structure-based evidence from the amino acid sequence, cellular localization data, potential alternative open reading frames, enzymatic function, presence or absence of gene duplication and degradation, and the possibility of horizontal gene transfer. The Genbank proposed gene product name did not differ from the proposed gene annotation for each of the genes in the group and as such, the genes appear to be correctly annotated by the computer database.

Introduction

Bacillus thuringiensis, also known as Bt, is a spore-forming bacterium that produces crystal proteins called δ-endotoxins (Madigan et al., 2005). It is found throughout the world in soil and it naturally occurs in the gut of caterpillars and various types of moths and butterflies (Kumar et al., 1996). Bt was first discovered in 1901 by a Japanese biologist, then rediscovered by Ernst Berliner in 1911 when he isolated it as the cause of a disease called *Schlaffsucht* in flour moth caterpillar.

Bt is an important bacteria today because of its natural pesticidal ability. Quickly after it was isolated in the early 20th century, industrial farmers began using it as a main pesticide for their crops. Bt was the best choice for a pesticide, for the crystal proteins which it forms are highly toxic to the metabolisms of pests, but have no observed effect on animals or human anatomy according to a study done by the EPA in 2001. Recently, with the introduction of genetic engineering, certain strains of corn, now called Bt corn, have been genetically modified to produce the Cry proteins that are toxic to pests on their own, without the need of added pesticides. This has opened the topic of Bt to public controversy between people who support or protest genetically modified crops.

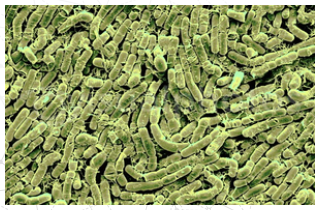

Figure 1 – This is an SEM of a colony of *Bacillus thuringiensis*.

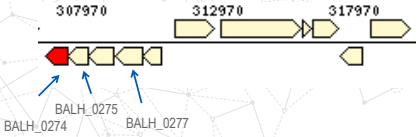

Figure 2 - These are the locus tags and relative position of the genes under investigation in this research.

Methods

Modules of GENI-ACT (http://www.geni-act.org/) were used to complete *Bacillus thuringiensis* genome annotation. The modules are described below:

Modules	Activities	Questions Investigated
Module 1- Basic Information Module	DNA Coordinates and Sequence, Protein Sequence	What is the sequence of my gene and protein? Where is it located in the genome?
Module 2- Sequence-Based Similarity Data	Blast, CDD, T-Coffee, WebLogo	Is my sequence similar to other sequences in Genbank?
Module 3- Cellular Localization Data	Gram Stain, TMHMM, SignalP, PSORT, Phobius	Is my protein in the cytoplasm, secreted or embedded in the membrane?
Module 4- Alternative Open Reading Frame	IMG Sequence Viewer For Alternate ORF Search	Has the amino acid sequence of my protein been called correctly by the computer?
Module 5- Structure-Based Evidence	TIGRfam, Pfam, PDB	Are there functional domains in my protein?
Module 6- Enzymatic Function	KEGG, MetaCyc, E.C. Number,	In what process does my protein take part?
Module 7- Gene Duplication/ Gene Degradation	Paralog, Pseudogene	Are there other forms of my gene in the bacterium? Is my gene functional?
Module 8- Evidence for Horizontal Gene Transfer	Phylogenetic Tree,	Has my gene co-evolved with other genes in the genome?

Results

BALH_0274:
The conclusion has been reached that the gene product of locus tag BALH_0274, undecaprenyl-diphosphatase, was called correctly by the computer. In the sequence-based similarity module, the top gene product name for orthologs of *Bacillus thuringiensis Al Hakam* was undecaprenyl-diphosphatase and had a high alignment score. As the genetic code is almost universal, this helps to verify that the gene was, in fact, called correctly. In creating a WebLogo, the BALH_0274 sequence had a high conservation with other similar proteins in the database. There were very few areas in the WebLogo where there was no conservation. No alternative reading frame was suggested for the gene locus, as there was a Shine-Dalgarno sequence between 5 and 15 amino acids upstream of the start codon of BALH_0274.

The protein was projected to be located in the cell membrane because of its seven transmembrane helix hits (See Figure 2.) This locus tag has been verified as an enzyme through the KEGG program and runs using the peptidoglycan biosynthesis pathway. Undecaprenyl-diphosphatase uses this pathway. Lastly, BALH_0274 is not likely to be a paralog, as there were no sequences like BALH_0274 in *Bacillus thuringiensis str. Al Hakam*. Given what is above, there is no evidence that the gene was called incorrectly by the computer, and it fits the characteristics of undecaprenyl-diphosphatase.

BALH_0275:
This gene had the coordinates 308385..309137 as predicted by GENI-ACT. The sequence is 250 amino acids long. GENI-ACT proposed that the product of this gene was a bacitracin ABC transporter and permease component. This proposal was supported by the BLAST results, the first two hits of which were bacitracin ABC transporter permease [*Streptococcus pneumoniae*] and bacitracin ABC transporter permease [*Bacillus sp.* MN5]. The CDD, T-coffee, and WebLogo results also supported this. TMHMM predicted six transmembrane helices. Each had a high probability rating, suggesting that these predictions are accurate. SignalP predicted no signal peptides. PSORTb suggested that the protein created by this gene could be found in the cytoplasm. Given these results and the results from Phobius, it is most likely that the protein is found partially in the cytoplasm, with six transmembrane helices. Also, further investigation suggests that the gene coordinates were correct as originally called.

The function pathway of this protein has been determined to be an ABC-2 transporter, though no E.C. number has been reported. No paralogs were found for this gene, and the way the raw translations of the sequence and the Genbank results matched perfectly suggest that this is not a pseudogene. Also, the phylogenetic tree, similarity of closely related bacteria in the same gene neighborhood, and the results from the chromosome viewer GC heat map all suggest that no horizontal gene transfer has occurred.

BALH_0277:
The proposed product for BALH_0277 by GENI- ACT was a sensor histidine kinase. This gene product was supported by the top BLAST hits for the amino acid sequence, the transmembrane topography of the amino acid sequence, and the cellular location of the amino acid sequence. Both PSORT-b and Phobius suggested that BALH_0277 is cytoplasmic. BALH_0277's phylogenetic tree shows that some of the *Bacillus* bacteria were more closely linked to non-*Bacillus* bacteria than to other *Bacillus*. In conclusion, the proposed product from GENI-ACT for BALH_0277 was confirmed by these findings.

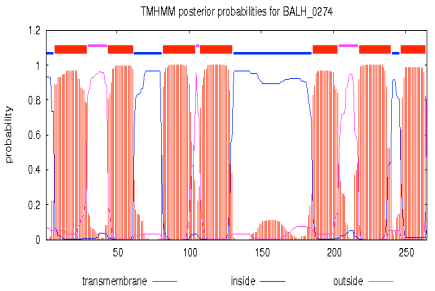

Figure 3 – TMHMM predicts how many transmembrane helices a certain protein has. If transmembrane helices are present, it is likely that the protein is located in the cell membrane. For BALH_0274, this graph shows that this protein has 7 significant predicted transmembrane helices. Therefore, it is most likely located in the cell membrane.

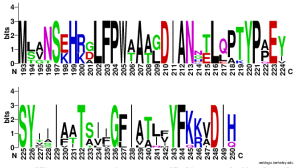

Figure 4 - This Weblogo graph from Locus Tag BALH_0275 shows the C terminal end of the protein. The first line is reasonably well-conserved, but the second one is less so. The lower line also shows an abundance of polar amino acids (in green).

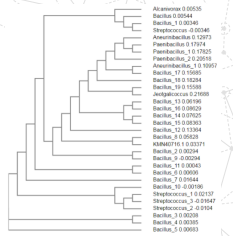

Figure 5 – The phylogenetic tree of BALH_0277 shows that this protein in some *Bacillus* bacteria is more closely linked to non-*Bacillus* bacteria than to other *Bacillus* bacteria.

Conclusion

The GENI-ACT proposed gene product did not differ significantly from the proposed gene annotation for each of the genes in the group and as such, the genes appear to be correctly annotated by the computer database.

Gene Locus	Geni-Act Gene Products	Proposed Annotation
BALH_0274	Undecaprenyl-diphosphatase	Undecaprenyl-diphosphatase
BALH_0275	Bacitracin ABC Transporter Permease	Bacitracin ABC Transporter Permease
BALH_0277	Sensor Histidine Kinase	Sensor Histidine Kinase

References

Madigan, Michael T.; Martinko, John M., eds. (2005). *Brock Biology of Microorganisms* (11th ed.). Prentice Hall. ISBN 978-0-13-144329-7.

Kumar, PA; Malik, VS; Sharma, RP (1996). "Insecticidal proteins of Bacillus thuringiensis". *Advances in Applied Microbiology*. **42**: 1–43.

Scanning Electron Micrograph of a Colony of Bacillus Thuringiensis Prior to Sporulation.

Acknowledgments

Supported by an NIH SCIENCE Education Partnership (SEPA) Award – R25ODO10536

A special thanks to Dr. Stephen Koury for his help annotating these genes

www.buffalo.edu