Annotation of the *Clostridium botulinum* Genome Strain 657/Type Ba4 at Locus Tags CLJ_0012 & CLJ_0017

Alina N. Beadle*, Amanda R. Roth* & Margaret S. Jolles

Silver Creek Central School District & The Western New York Genetics in Research and Health Care Partnership

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

National Institutes of Health
Turning Discovery Into Health

University at Buffalo

Abstract

Two students from Silver Creek Middle School participated in the Western New York Genetics in Research and Health Care Partnership Gene Annotation Research Study. Annotation is the process of assigning function or biological significance to a gene. Using the Geni-Act website, students were able to annotate their gene based off of Basic Information, Sequence Based Similarity Data, Structure Based Evidence, Cellular Localization Data & Alternative Open Reading Frame. The students were able to obtain data including protein structure, function and location.

Introduction

Clostridium botulinum is a gram positive, anaerobic, rod shaped bacterium associated with improperly produced canned goods. It produces a toxin that is lethal. This is most commonly obtained when a canned food item is consumed that was not stored at proper temperature, was not prepared properly at time of canning or when a can has begun to swell and the contents are consumed. *Clostridium botulinum* was first discovered and isolated by Emile van Ermengem in 1896, and it was later deemed to survive by forming spores, remaining in a dormant (sleep like) state until environmental conditions are perfect for growth.

Clostridium botulinum consists of seven subtypes. Each subtype produces a different botulinum toxin; with the exception of subtypes three and four, all are human pathogens. Types one and two, commonly found in soil are the primary cause of botulism outbreaks in the United States. Type five commonly found in fish is also a contributor to the cases of botulism in the United States.

Our team analyzed *Clostridium botulinum* Strain 657/ Type Ba4. Our Locus tags were CLJ_0012 and CLJ_0017, both being of similar length. Our goal was to analyze and find the functions of our two genes.

Figure 1:
Clostridium botulinum
Gram Positive
Anaerobic Bacterium

http://parasites.ftz.czu.cz/food/_data/141.jpg

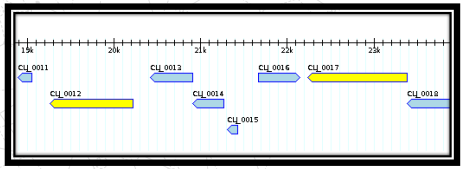

Figure 2: The locus tags (in yellow) and relative position of the genes under investigation in this research

Methods

Modules of the GENI-ACT (http://www.geni-act.org/) were used to complete *Clostridium botulinum* 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?

Results

CLJ_0012: *CLJ_0012* has a sequence of 318 amino acids with the proposed gene product of *LycA*. *LycA* is a protein with lysozyme activity in at least one strain of *Clostridium botulinum*, *Kyoto/Type A2*. *LycA* is in the glycoside transferase family of proteins. These proteins break bonds between one or more carbohydrates.

The first BLAST hit from curated Swiss Protein Data Base resulted in *Lysozyme M1*. The second BLAST hit produced *Probable N-acetylmuramoyl-L-alanine amidase*. Both proteins are in the glycoside hydrolase family. Both Non Redundant database hits resulted in *Glycoside Hydrolase family* from 2 different species of bacteria. The first COG3757 shows the result *Lysozyme M1* and COG3409 indicated a *Peptidoglycan binding domain of peptidoglycan hydrolases*. Peptidoglycan forms the cell wall in Gram Positive bacteria. Cellular localization results describe the protein as 'extracellular' with no predicted helices (TMHMM), no signal peptide (Signal P) and non-cytoplasmic (Phobius). This 'extracellular' location does make sense if it is associated with the peptidoglycan layer. T-Coffee and WebLogo results show a protein that is not well conserved at the N-terminus with 172 amino acids absent (this appears to be due to the use of a hypothetical protein WP_051082168.1 in the T-Coffee sequences), moderate conservation from amino acid 173 – 428 and poor conservation/absence from 429 – 544. TIGRFAM showed no significant hits but Pfam yielded two, a *glycoside hydrolase family 25* and *peptidoglycan binding domains*. PDB first hit results in a *glycosyl hyrdrolase* from *Bacillus anthracis*.

Possibly the most important finding from this annotation is the hypothesis that the computer called the start codon for *CLJ_0012* in error. The computer set the start codon at 19267 with the Shine-Delgarno sequence beginning at 19260. This leaves only one base pair between the end of the Shine-Delgarno sequence and the beginning of the start codon. Looking only 6 bases downstream from the called start codon is another start codon, at 19273. This codon appears to be in a much better position with 7 base pairs between the Shine-Delgarno sequence and the hypothesized start codon.

The information obtained through Geni-Act implies enzymatic function for *CLJ_0012*. Unfortunately, this research did not confirm a specific pathway Module 6 in Geni-Act.

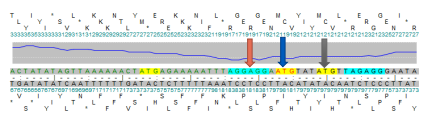

Figure 3: Location of the called start codon (blue) and the hypothesized start codon (gray) with corresponding Shine-Delgarno (red) Sequence.

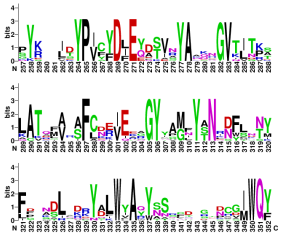

Figure 4: WebLogo showing areas of mixed conservation of sequences

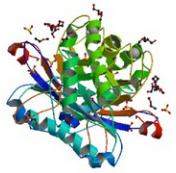

Figure 5: A glycosyl hydrolase from B. anthracis

CLJ_0017: This protein has a sequence length of 384 amino acids with a proposed gene product of *methionine adenosyltranferase*. *Methionine adenosyltransferases* (MAT) are the family of enzymes that synthesize the main biological methyl donor, *S-adenosylmethionine*. All 4 BLAST hits, 2 each from Non-Redundant and curated Swiss Protein Data Base, resulted in a match to *adenosyltransferase*. Both alignment results and e-values made these hits highly significant. Only one COG hit, #0192 was shown, with the result *S-adenosylmethionine synthetase*. 16 sequences were used in T-Coffee to compare with *CLJ_0017*. The WebLogo generated shows a very high degree of conservation across the entire amino acid sequence with the exception of the first and last few amino acids.

The location of this protein is predicted to be cytoplasmic. The evidence for this is no predicted helices by TMHMM, no signal peptide present by Signal P and a high cytoplasmic score of 9.97 from PSort B. Tigrfam & Pfam are both used to reveal structure based evidence in the annotation process. Both databases also produced results that support the previous evidence, *S-adenosylmethionine synthetase*, also known as *methionine adenosyltransferase*. Use of PDB also supports all previous evidence with the *Crystal structure of a thermostable methionine adenosyltransferase*. Work in the Alternative Open Reading Frame Module supports the fact that the computer has called the correct start and stop positions in the DNA.

The information obtained through Geni-Act implies enzymatic function for *CLJ_0017* as this is a *transferase* molecule. Unfortunately, this research did not identify a specific pathway after using the Geni-Act, enzymatic function module.

Figure 7: WebLogo from *CLJ_0017* illustrating high conservation of amino acids

Figure 6: 1st hit from Swiss Protein Data Base illustrating significant alignment between *CLJ_0017* and Subject

Figure 8: Crystal Stucture of a thermostable methionine adenosyl-transferase; *CLJ_0017* is an adenosyltransferase protein

Conclusion

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

Gene Locus	Geni-Act Gene Product	Proposed Annotation
CLJ_0012	LycA of the glycosyl transferase family	LycA of the glycosyl transferase family
CLJ_0017	Methionine adenosyltransferase	Methionine adenosyltransferase

References

- http://www.sciencedirect.com/science/article/pii/S0923250814002046
- http://www.ebi.ac.uk/interpro/protein/C1FUI1
- http://www.cazy.org/Glycoside-Hydrolases.html
- http://www.uniprot.org/uniprot/P25310
- http://www.els.net/WileyCDA/ElsArticle/refId-a0000702.html
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2643306/

Acknowledgments

- Supported by an NIH Science Education Partnership (SEPA) Award - R25OD10536-1 A1
- Stephen Koury Ph.D. – Research Associate Professor – Dept. of Biotechnical & Clinical Laboratory Sciences
- Jennifer R. Cooke, *MPS, CSA*, Program Coordinator WNY Rural Area Health Education Center (R-AHEC)