Annotation of the *Clostridium Botulinum* Genome at Locus Tags CBF_3556 and CBF_3361

Sharan Prasad, Grace Yao, and Sangeeta Gokhale
Williamsville East High School (participating in the BEAM – Buffalo Engineering Awareness for Minorities Program) and The WesternNew York Genetics in Research Partnership

Abstract

Gene annotation is a systematic procedure in which relevant information about a gene is researched and verified through research done manually using various databases. Using evidence from research, we can create changes to the existing genome. We were able to research a group of 2 consecutive genes from the microorganism Clostridium Botulinum (CBF_3556 and CBF_3361) using the collaborative genome annotation website GENI- ACT. We were able to assess our gene products through various databases using our genes' 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, as well as the possibility of horizontal gene transfer. Some of the databases we used were PFAM, TIGRFAM, BLAST, and MetaCyc, we were able to access previously researched data about our gene product and verify it with our findings. In general, most of the computer's data matched that of the manually researched data, and the expected annotations were seen.

Introduction

Clostridium botulinum is a bacterium with the ability to create toxins which may cause partial facial paralysis, degenerated vision, or even be lethal. It is rod-shaped, anaerobic, gram-positive, spore-forming, and motile. The organism is most prevalent in spores, but is not always pathogenic. Only four out of the seven neurotoxins produced are disease-causing. The toxins are only made when the spores grow out of active bacteria and can be fatal in any of its forms, but only four out of seven of the toxins Infant botulism is an intestinal infection, but the afflicted botulism can also be of an infection in a wound or foodborne (Peck, MW, 2009).

Studies have been made to research the effect of high pressure processing and heat on the spores. The test was conducted in both laboratory and pilot scale, and the spores reached total destruction after at 118°C and 700 MPa up to 10 minutes. Increasing the temperature to 121°C had only slight effects on the pilot-scale system, with only a small reduction from the previous experience. No experiments been done on the pressurizing fluid or the packaging system.

The proposed annotation after research resulted in no significant differences from the Genbank proposed genes. The sequences in question coded for proteases and no pseudogenes were found.

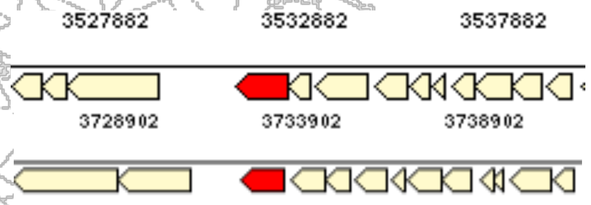


Figure 1. The position of the CBF_3361 and CBF_3556 sequences being researched, respectively

Methods

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?
Module 9- RNA	RFAM	Does my gene encode a functional RNA?

Results

GENI-ACT (http://www.geni-act.org/) modules were used to complete *Clostridium Botulinum* genome annotation . Descriptions of each module is detailed below

CCBF_3361:
The initial proposed product of this gene by GENI-ACT was an ATP-dependent Clp Protease. This gene product proposal was supported by the top BLAST hits for the amino acid sequence, the presence of well-curated protein functional domains within the amino acid sequence, the transmembrane topography of the amino acid sequence, and the cellular location of the amino acid sequence. As such, the proposed annotation is an ATP- dependent Clp Protease.

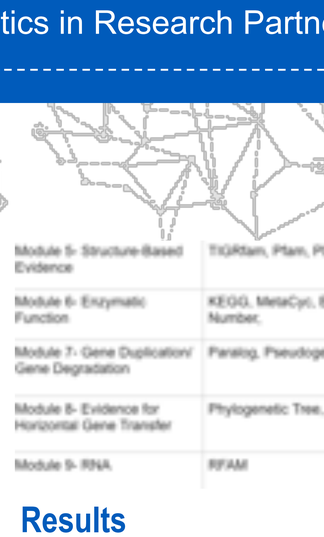


Figure 3. Structure of ATP-dependent Clpx Protease from PDB.

Figure 2. WebLogo of CCBF_3361. The beginning of the Weblogo signifies the 5' end of the sequence and the end signifies the 3' end of the sequence. Relative stack height indicates percent conservation. Relative sizes of the letters indicate the frequency of the amino acid in the alignment. CBF_3361 shows a high percent conservation throughout the sequence alignment.

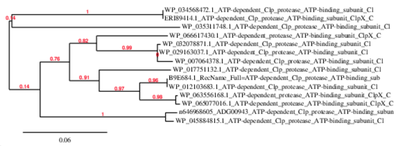
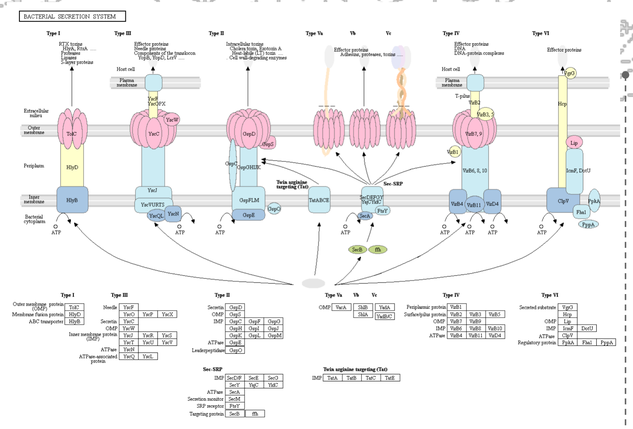


Figure 3 The phylogenetic tree shows many of the closely related gene products

Figure 4. The KEGG Pathway with CCBF_3361 shows the proposed gene product is part of a bacterial secretion system.



CCBF_3556:
The top BLAST hit and the COG name proposed a product of a carboxy-terminal processing protease. The second hit resulted in a carboxy-terminal processing peptidase. The TIGRFAM and EC name also support the c-terminal processing peptidase. An alternative EC name is a Photosystem II D1 protein processing peptidase or tail-specific protease. This data supports the final protein annotation, a carboxy-terminal processing protease.

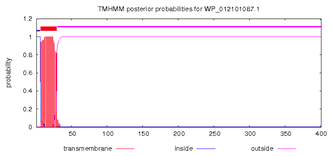
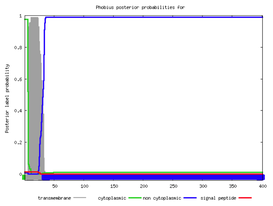


Figure 4. The THMM results show that the enzyme is a transmembrane protein with one helix

Figure 5. Phobius results support the data from THMM with the same results



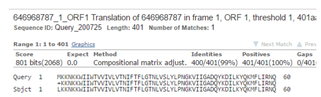


Figure 6. The comparison of the raw translation to the CBF_3356 gene shows that both are exactly the same. This suggests that there is no frameshift or mutation, and is therefore not a pseudogene.

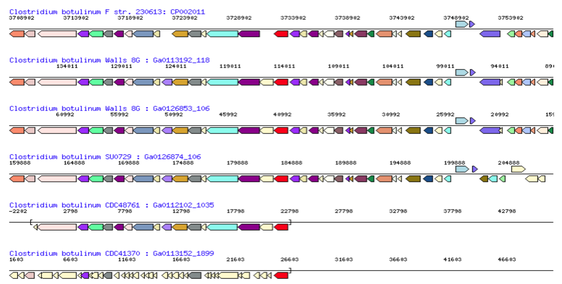


Figure 6. Gene orthology neighborhoods show very similar DNA

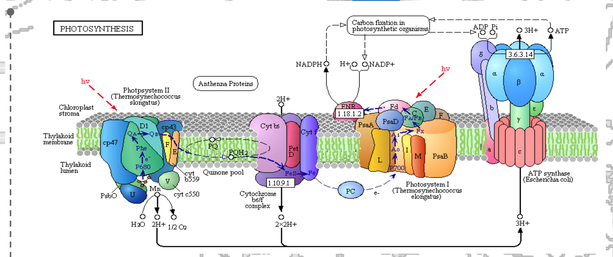


Figure 7. The *Clostridium Botulinum* bacterium is a photosynthetic, purple bacteria. It is able to metabolize its own energy using energy from the sun. In bacteria, the process occurs in a sequence of transmembrane proteins located in the cytoplasmic membrane. By analyzing the P-SORT B scores, the cytoplasmic score is 0.00, the cell wall score is 0.17, and the cytoplasmic membrane score is 9.68. The numbers show that the protein is located in the cytoplasmic membrane, where photosynthesis is occurring. MetaCyc search results for the protein support that it is in photosystem II.

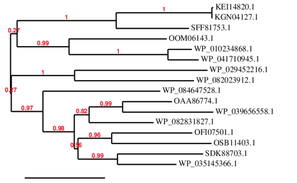


Figure 8. Phylogenic tree shows very similar evolution with most similarity over 0.80

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. It was found that the proposed gene product of CBF_3361 is an ATP-dependent Clp protease ATP-binding subunit ClpX that is involved in degradation of bacterial proteins as part of a secretion system. The proposed gene product of CBF_3556 is a carboxy-terminal processing protease and is found in Photosystem II.

Locus Tag	Geni- Act Gene Products	Proposed Annotations
CBF_3361	ATP-Dependent Clp Protease ATP-Binding Subunit ClpX	ATP-Dependent Clp Protease ATP-Binding Subunit ClpX
CBF_3556	Carboxy-Terminal Processing Protease	Carboxy-Terminal Processing Protease

References

Peck, MW (2009). "Biologycoat and genomic analysis of Clostridium botulinum". *Advances in microbial physiology*. Advances in Microbial Physiology. **55**: 183–265, 320.
J Food Prot. 2013 Mar;76(3):448-55.

Acknowledgments

Supported by an NSF Innovative Technology Experiences for Students and Teachers (ITEST) Award – 1311902
Prof: Dr. Steve Koury, Dr. Rama Dey-Rao
www.buffalo.edu