Annotation of the *Clostridium botulinum* Genome from DNA Coordinates 3444348 to 3445148 (or Locus Tag CBF_3259)

Stephen Zheng*, Raven Brock**, and Sangeeta Gokhale
*Williamsville East High School, **Buffalo Public Schools - Buffalo Engineering Awareness for Minorities at UB

BEAM Buffalo-Area Engineering Awareness for Minorities

University at Buffalo

Abstract

One gene from the microorganism *Clostridium Botulinum* (CBF_3259) was annotated using the genome annotation website GENI-ACT (http://www.geni-act.org/). The 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, the possibility of horizontal gene transfer, and the production of an RNA product. The Genbank proposed gene product name were found to be the same as the proposed gene annotation for the gene in the group and as as a result, the gene appears to be annotated accurately in the nr database.

Introduction

Clostridium Botulinum is an aerobic, non-encapsulated, motile, and pore-forming gram positive that can produce botulinum, a neurotoxin. The interest in a gene of this microorganism came from the question of foodborne illnesses relating to cafeteria food in schools.

Foodborne botulism is thought to have been present since the inception of mankind, but its symptoms or occurrences of botulism were not recorded much in the past, before the 1800's. The discovery of *Clostridium Botulinum* originates from a botulism outbreak in the smoked ham during a funeral dinner by Emile Pierre van Ermengem. Named for the association to spread of infection through sausages,, *Clostridium Botulinum* began to be researched to find a medicinal and therapeutic method to apply the neurotoxin, botulinum," produced by *Clostridium Botulinum*, leading to Alan B. Scott and Edward J. Schantz taking over and pioneering the modern application of this neurotoxin.

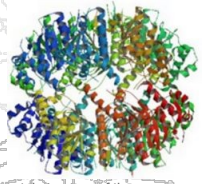

Figure 1. The crystal structure of the gene researched.

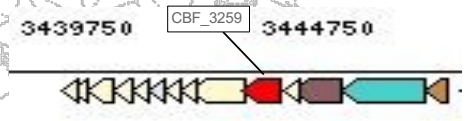

Figure 2. The locus tag and relative position of the gene researched (in red from 3444348 to 3445148).

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?
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

CBF_3259:

The gene product at this locus found by GENI-ACT was found to be a putative ATP-dependent Clp Protease. This gene product proposal was supported by the top BLAST hit for the amino acid sequence, Clp protease, the top COG hits, ATP-dependent protease ClpP and ClpP, the top TIGRFAM hits, both ClpP, and top Pfam hit, Clp Protease (a second hit was not found), all support the proposed annotation of CBF_3259 as putative ATP-dependent Clp Protease. Additional support of the annotation was provided by the WebLogo, SignalP, TMHMM, the phylogenetic tree from phylogeny.fr with TreeDyn tool , and the ortholog neigh- borhood region from IMG/ER of the gene researched.

Clp protease [Clostridium botulinum]
Sequence ID: WP_045883294.1 Length: 244 Number of Matches: 1

Range 1: 1 to 244 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
480 bits(1236)	5e-171	Compositional matrix adjust.	233/244(95%)	238/244(97%)	0/244(0%)

```
Query  23   MEIKNQTEEKAELYFYGDICSSIWGCWEEEDKCPQDVSNFLKEIINNKELDIYINSGGGS  82
            MEIKNQTEEKAELYFYGDICSSI GCWEEEDKCPQDVSNFLKEIINNK+LDIYINSGGGS
Sbjct  1    MEIKNQTEEKAELYFYGDICSSIRGCWEEEDKCPQDVSNFLKEIINNKDLDIYINSGGGS  60

Query  83   VFAGLAIYHQLKRHNGFKTVHVDGLAASITSVIALVGDKVVIPAPAQFMIHKPWIALGN  142
            VF GLAIYHQL RHNGFKTVHVDGLAASI SVIALVGDKV+IPAPAQFMIHKPWIALGN
Sbjct  61   VFGGLAIYHQLNRHNGFKTVHVDGLAASIMSVIALVGDKVIIPAPAQFMIHKPWIALGN  120

Query  143  ANEFRKQANDLDIAEESILKIYEANLNEGVNIETIKQLMADETWMTGEKAAEYFNIEVEE  202
            ANEFRKQANDLDIAEESILKIYEANLNEGVNIETIKQLMADETWMTGEKAAEYFNIEVEE
Sbjct  121  ANEFRKQANDLDIAEESILKIYEANLNEGVNIETIKQLMADETWMTGEKAAEYFNIEVEE  180

Query  203  TQELVACTSQYFNEYKNMPKNLFAKSIKPKEPENKDNFIKDEEIESLIERVNNTLKFEEE  262
            TQELVACTSQYFNEYKNMPKNLFAKSIKPKE +NKD+FIKDEEIE+LI RVNNTLKFEEE
Sbjct  181  TQELVACTSQYFNEYKNMPKNLFAKSIKPKELKNKDHFIKDEEIEALISRVNNTLKFEEE  240

Query  263  RINE  266
            RINE
Sbjct  241  RINE  244
```

Figure 1. The BLAST result shows various data, such as the gene product, length of the gene, e-value, and organism, which is Clostridium botulinum, supporting the annotation of the gene product as a Clp Protease

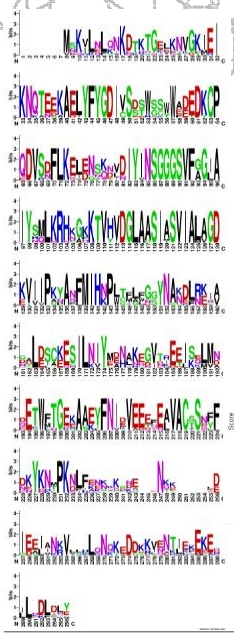

Figure 2 (at left). The WebLogo shows the amino acid sequence of CBF_3259 in a graph form. There is no predominant color in the sequence logo as black, green, red, and blue are represented approximately equally, The colors green, black, red, and blue represent hydrophilic, hydrophobic, acidic, and basic regions, respectively. The results are in line with the fact that Clp Protease is an enzyme, and is required to function in many medium of of either hydrophilic and hydrophobic environments in addition to environments of different acidities and pH's.

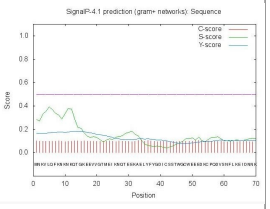

Figure 3. In the SignalP graph for the gene, due to the low values of the C, S, and Y-Score, it can be interpreted that there is no signal peptide predicted or cleavage site.

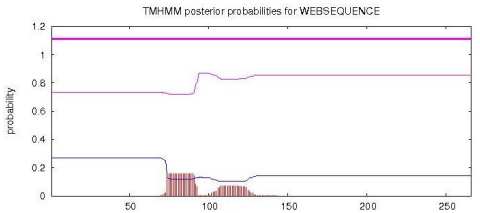

Figure 4. The TMHMM Transmembrane topology graph lacks distinct red regions with high probability, indicating that it is highly unlikely for this protein to be located in or on the cellular membrane of *Clostridium Botulinum*, due to no predicted transmembrane helices, which correlates with the fact that the protein is an enzyme, which do not function in relationship with the cellular membrane.

Figure 5. This image depicts the Pfam pairwise alignment of Clp Protease, with a low e-value of 3.4e-27, indicating that the match is significant.

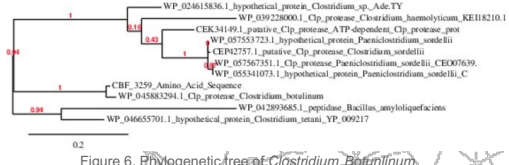

Figure 6. Phylogenetic tree of *Clostridium Botulinum*.

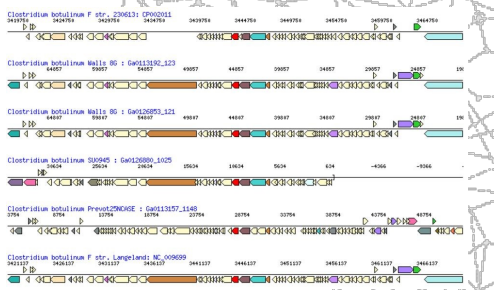

Figure 7. The ortholog neighborhood region from IMG/ER shows that this gene is present in many similar species. This indicates that this gene is coserved in the species *Clostridium Botulinum*.

Conclusion

The GENI-ACT proposed gene product did not differ significantly from the proposed gene annotation for the Clostridium Botulinum gene from the loci 3444248 to 3445148. Therefore, the gene appears to be correctly annotated by the Genebank computer database.

Gene Locus	Geni-Act Gene Products	Proposed Annotation
3259	putative Clp Protease	putative ATP-dependent Clp Protease

References

Erbguth, F. J. (2004). "Historical Notes on Botulism, Clostridium Botulinum, Botulinum Toxin, and the Idea of the Therapeutic Use of the Toxin." *Movement Disorders : Official Journal of the Movement Disorder Society*: U.S. National Library of Medicine, Mar. 2004.

Acknowledgments

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