# Annotation of the *Clostridium botulinum* Genome at Locus Tag CBF_3493

Nidhish Gokhale and Sangeeta Gokhale
Williamsville East High School, participating in the BEAM - Buffalo Engineering Awareness for Minorities Program and The Western New York Genetics in Research Partnership

## Abstract

The process of annotating or finding information on a gene's location, its product, evolutionary connections, pathways and enzymatic functions is necessary to study all species. GENI-ACT, is an annotation collaboration tool that provides resources to facilitate genomics research (www.geniscience.org/). Clostridium is known to produce a neurotoxin which deforms acetylcholinesterase to cause severe infections. The Genbank proposed gene product for the annotated gene (CBF_3493) was reviewed. The following steps were taken: the assessment of amino acid sequence similarity data by looking the redundancy in the genetic code, the structure-based evidence from the amino acid sequence to find similarity in the structures of functional domains, the cellular localization data to determine where in cell the protein is encoded, the potential alternative open reading frames, and the possibility of horizontal gene transfer. The data obtained manually matched the computer's data for the corresponding gene product names.

## Introduction

***Clostridium botulinum*** is a Gram-positive, rod-shaped, anaerobic, spore-forming, motile bacterium with the ability to produce the neurotoxin botulinum. The botulinum toxin can cause a severe flaccid paralytic disease in humans and other animals by inhibiting the enzyme, acetylcholinesterase and is the most potent toxin known to mankind. The toxin is a zinc metalloprotease that cleaves one or more of the fusion proteins by which neuronal vesicles release acetylcholine into the neuromuscular junction. (Montecucco C, *Curr Top Microbiol Immunol.* 1995;195:1-278). *C. botulinum* is commonly associated with bulging canned food and misshapen cans due to an internal increase in pressure caused by gas produced by the bacteria. The contamination is primarily from soil (Tucker J. B, Cambridge, Mass: MIT Press; 2000). Genome sequencing is often known as "decoding". In Genbank, there are over 61,582 readings involving the *Clostridium botulinum* gene, CBP_3493. It is suggested to code for a zinc metalloprotease which binds acetylcholinesterase and deforms it so that acetylcholine will not be able to send nerve signals. The toxin does not allow neurotransmitters to be released; thus, avoiding formation of the snare complex.

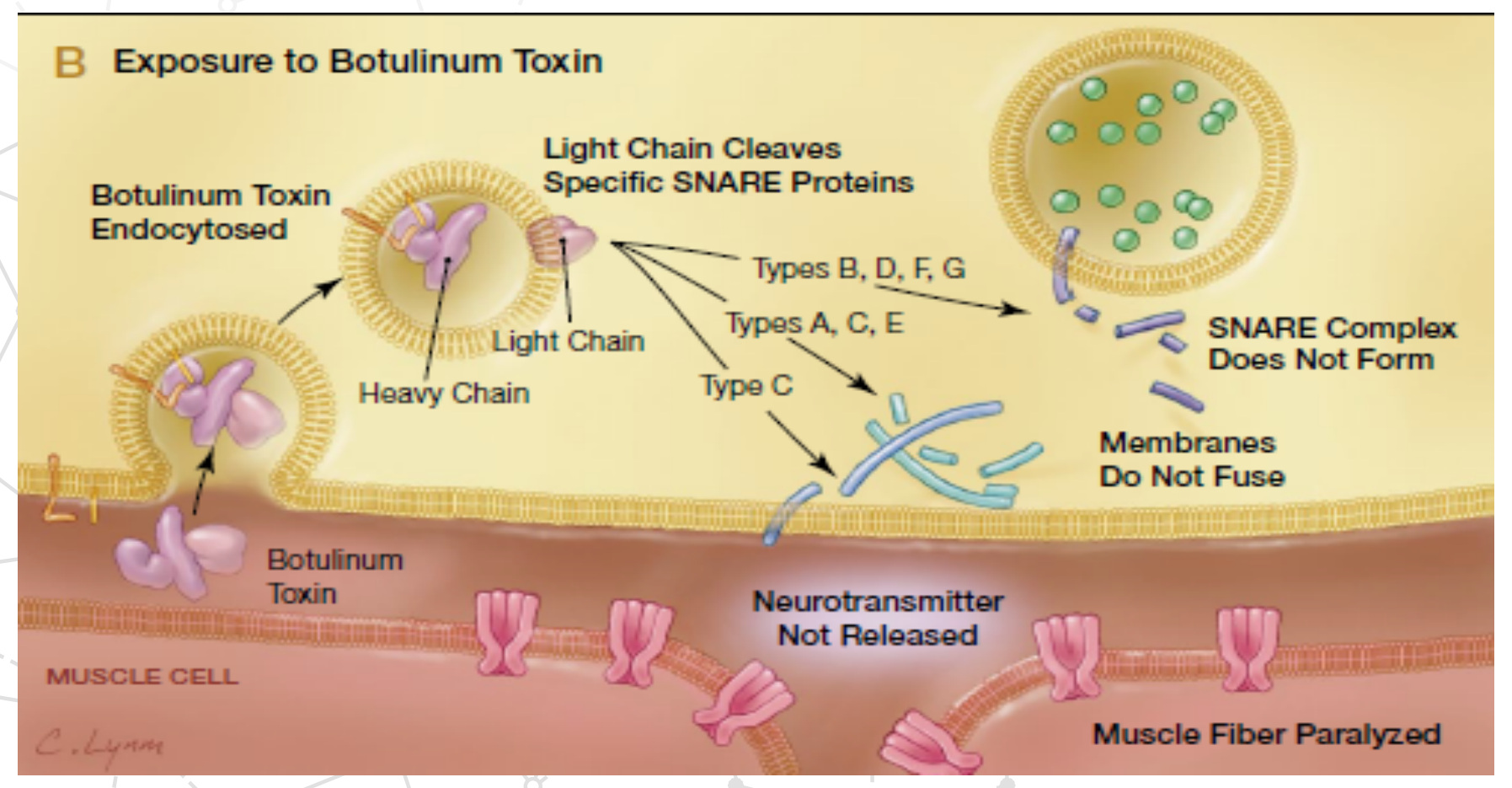


Figure I – Botulinum toxin (zinc metalloprotease) action

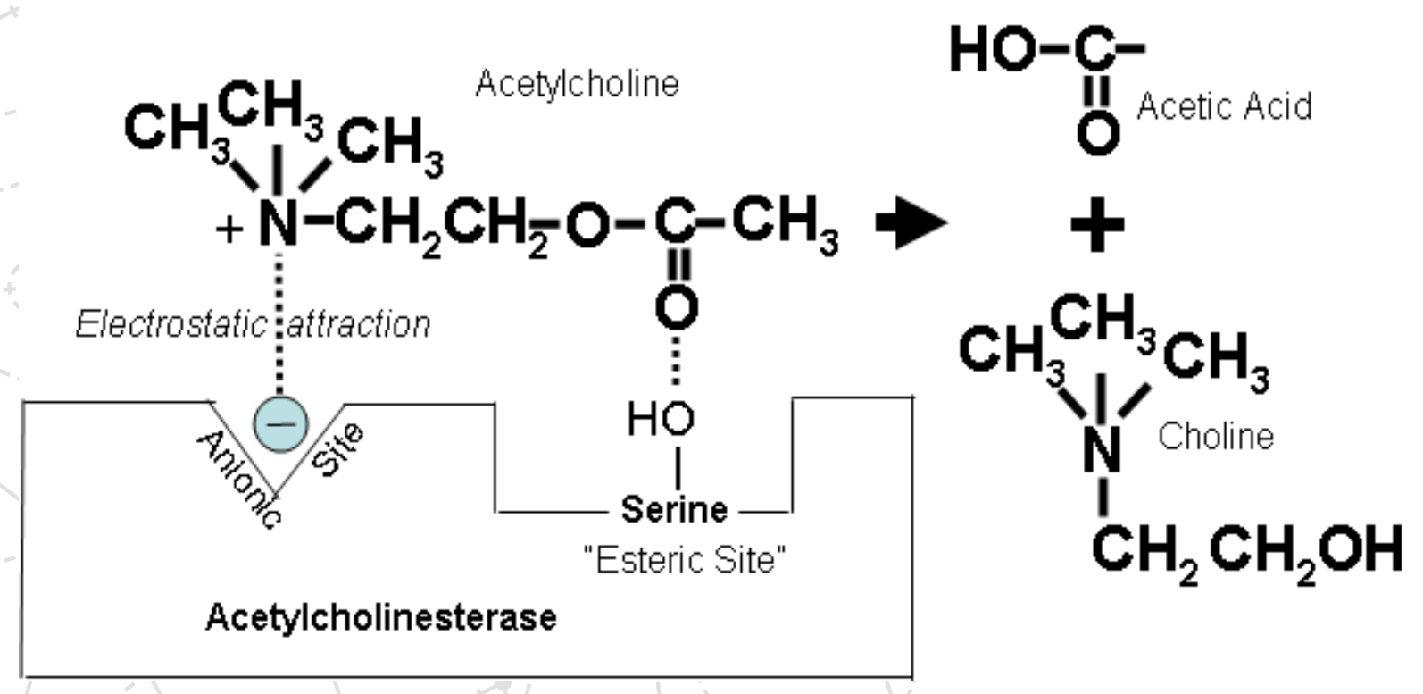


Figure II - Image of the "Esteric Site" where acetylcholinesterase acts

## Methods

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

*Clostridium botulinum* (CBF_3493):

The initial proposed product of this gene by GENI-ACT was a predicted zinc metalloprotease. This gene product proposal was supported by the top BLAST hits suggest the gene product to be aminopeptidase. The COG hit suggested it to be aspartylaminopeptidase.

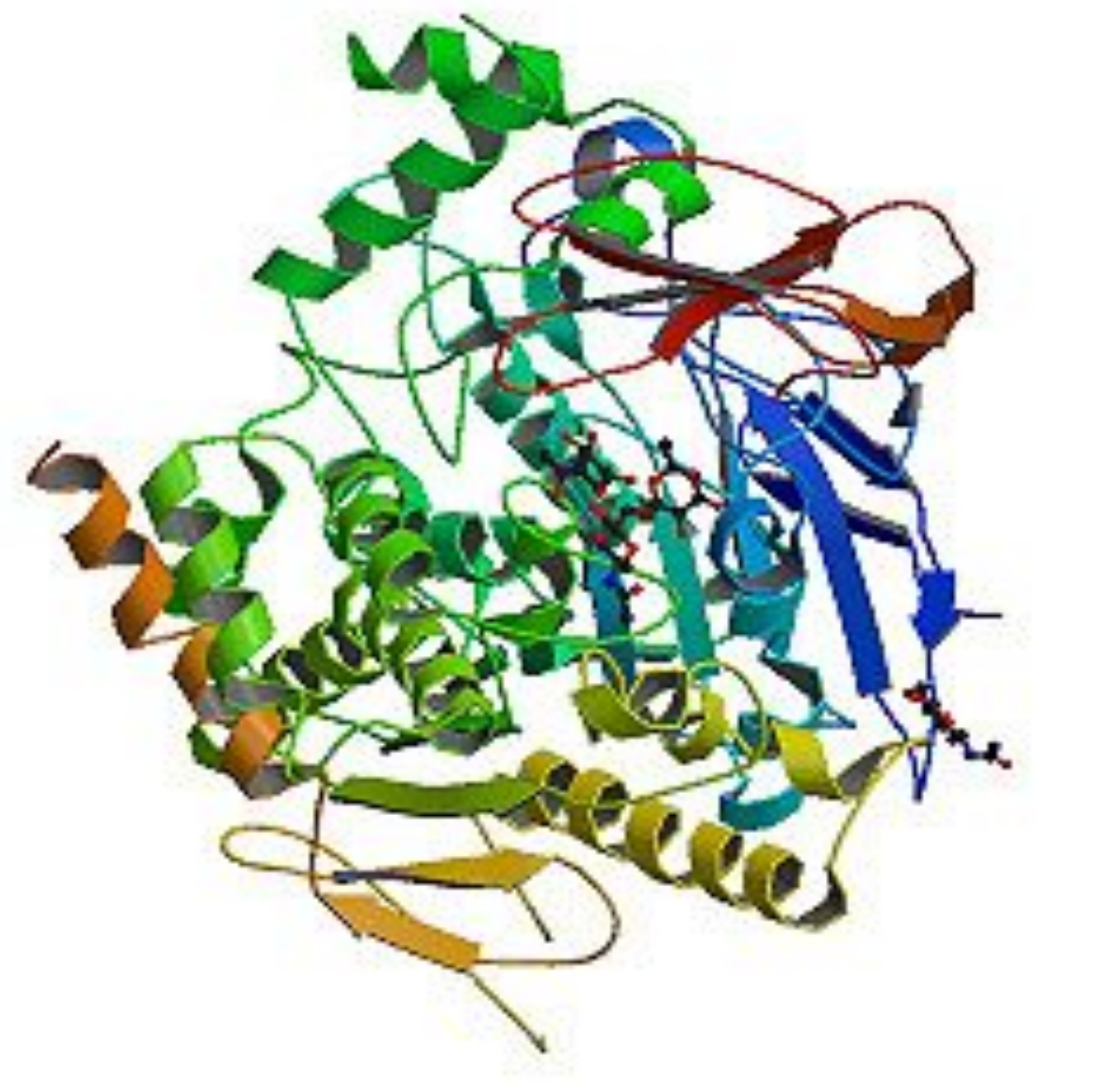
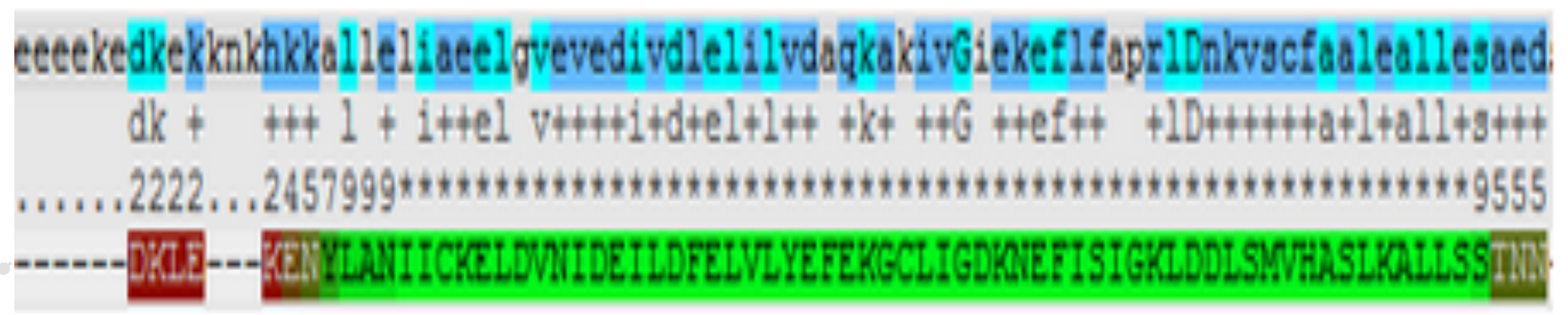
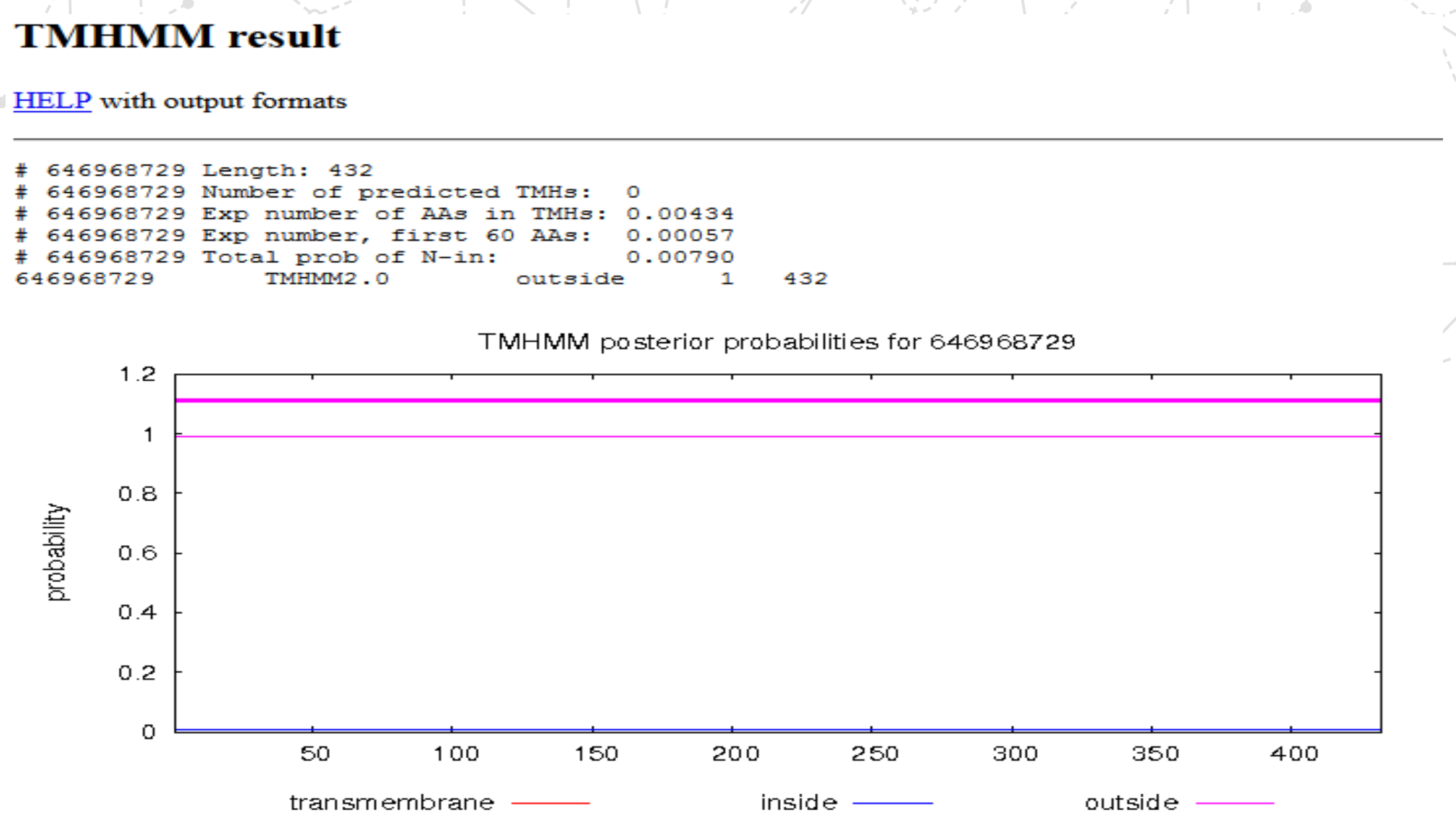


Figure III – Ribbon model for the protein/enzyme, acetylcholinesterase



```
eeeekedkekknkhkkalleliaeelgvevedivdlelilvdaqkakivGiekeflfaprlDnkvscfaaleallesaed
     dk +   +++ l + i++el v++++i+d+el+l++ +k+ ++G ++ef++  +lD++++++a+l+all+s+++
......2222...2457999*************************************************************9555
------DKLE---KENYLANIICKELDVNIDEILDFELVLYEFEKGCLIGDKNEFISIGKLDDLSMVHASLKALLSSTNN
```

Figure IV - Pfam pairwise alignment of the CBF_3493 gene product



**TMHMM result**

HELP with output formats

```
# 646968729 Length: 432
# 646968729 Number of predicted TMHs:  0
# 646968729 Exp number of AAs in TMHs: 0.00434
# 646968729 Exp number, first 60 AAs:  0.00057
# 646968729 Total prob of N-in:        0.00790
646968729       TMHMM2.0        outside      1   432
```

Figure V – The TMHMM shows the probability of the location of the protein. From this it is possible to conclude that the enzyme is located outside of the cytoplasmic region

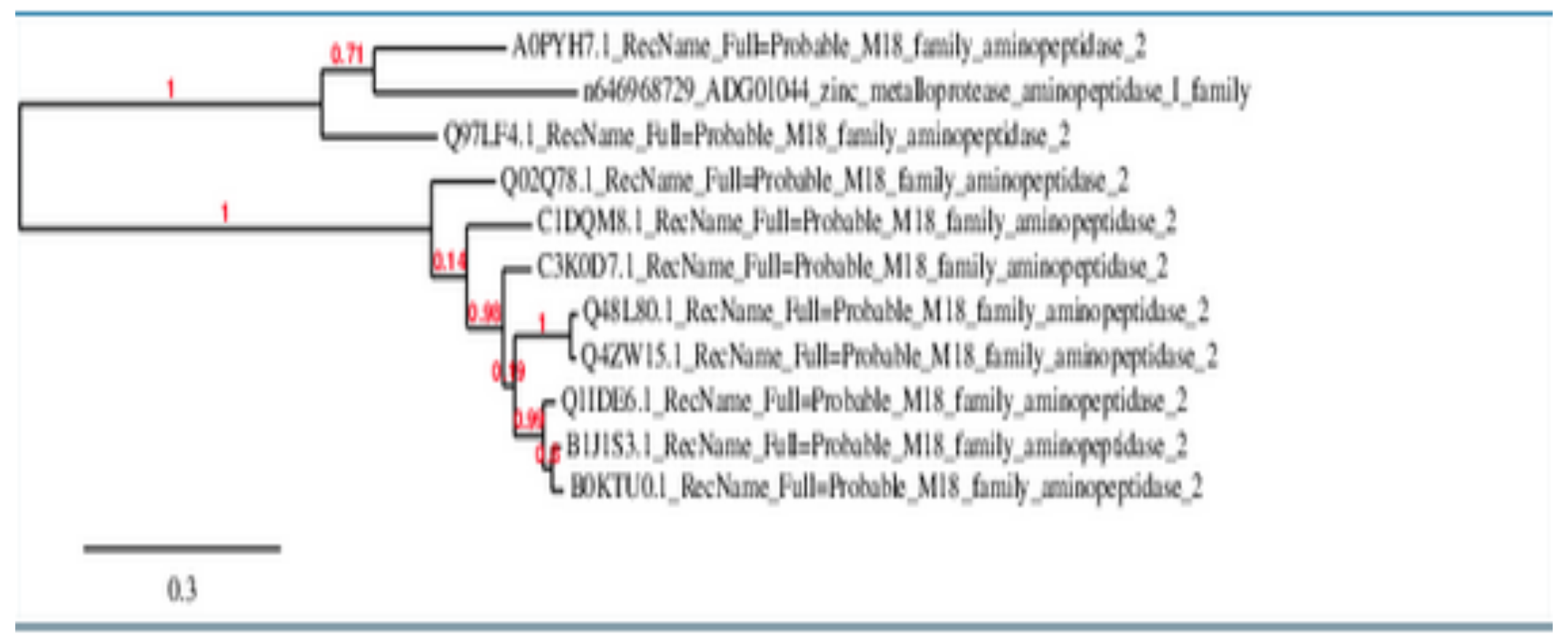


Figure VI - Phylogenic tree shows possible convergent or divergent evolution (relationship among similar organisms

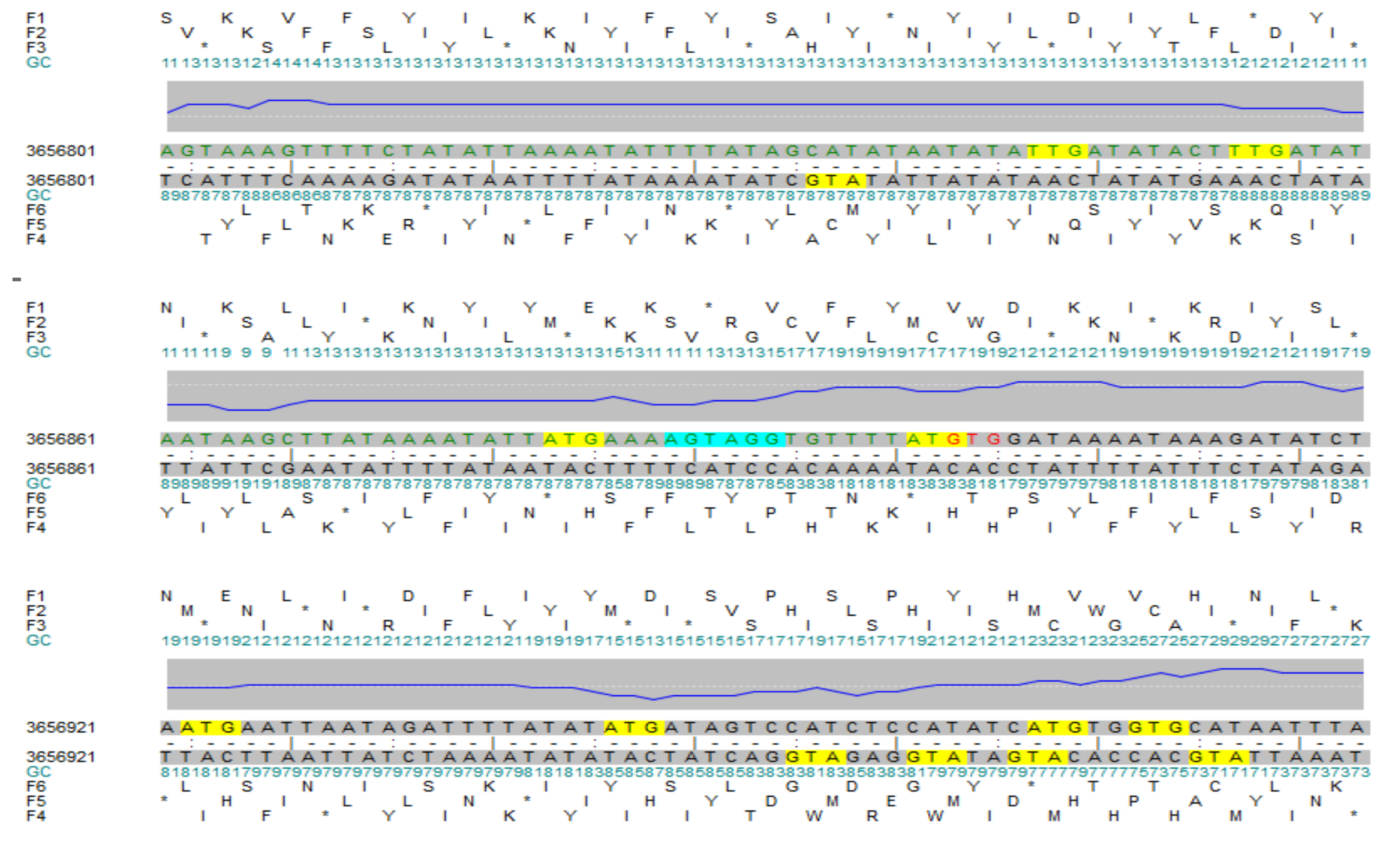


Figure VII – The image above shows possible alternate open ended reading frames for the gene CBF_3493. From the data observed, it can be concluded that there are no alternative reading frames, since there are not a significant amount of Shine-Delgarno sequences

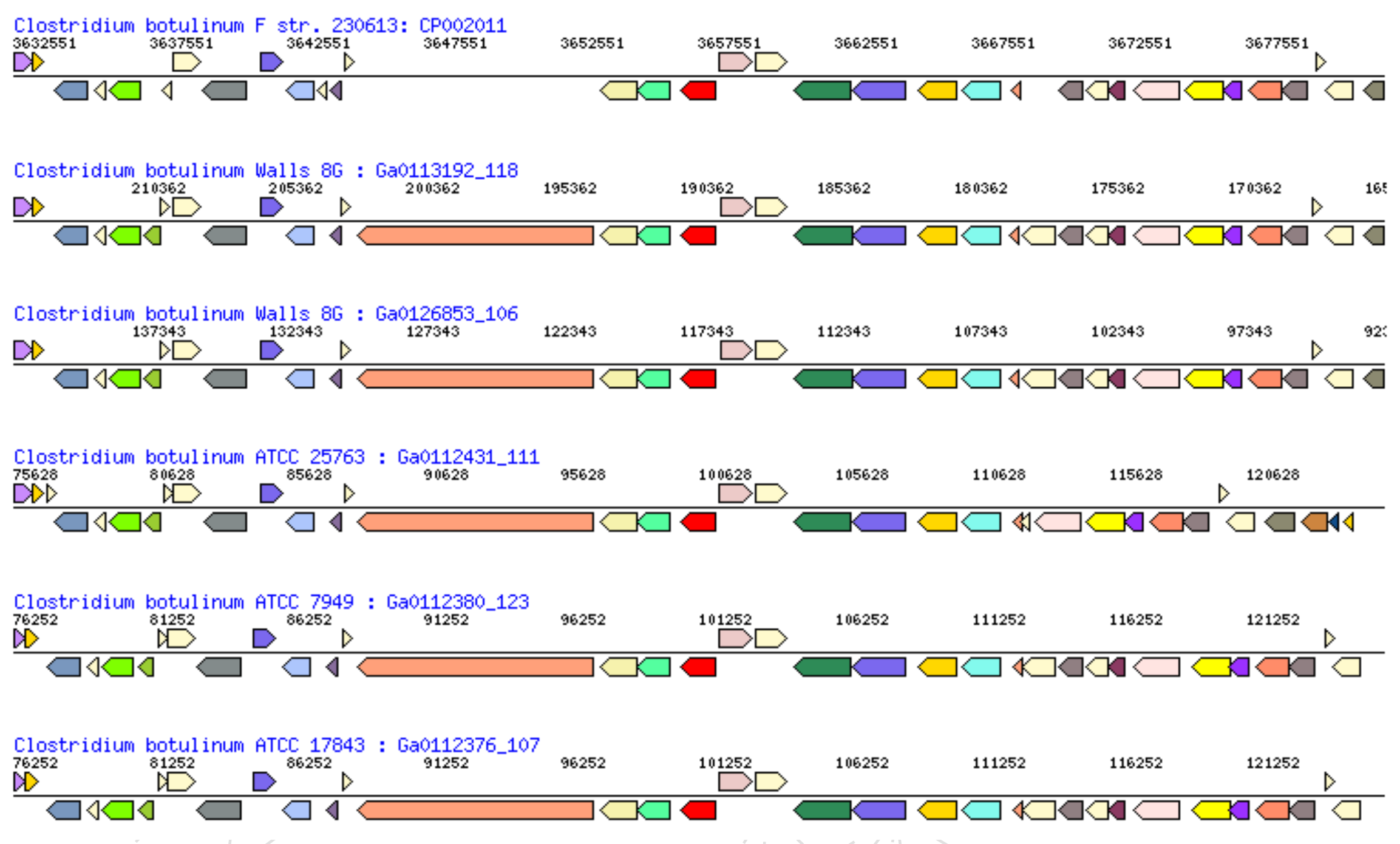


Figure VIII – The image shows the ortholog neighborhood region for the organism *Clostridium botulinum.* Similarities throughout the family of organisms can be seen

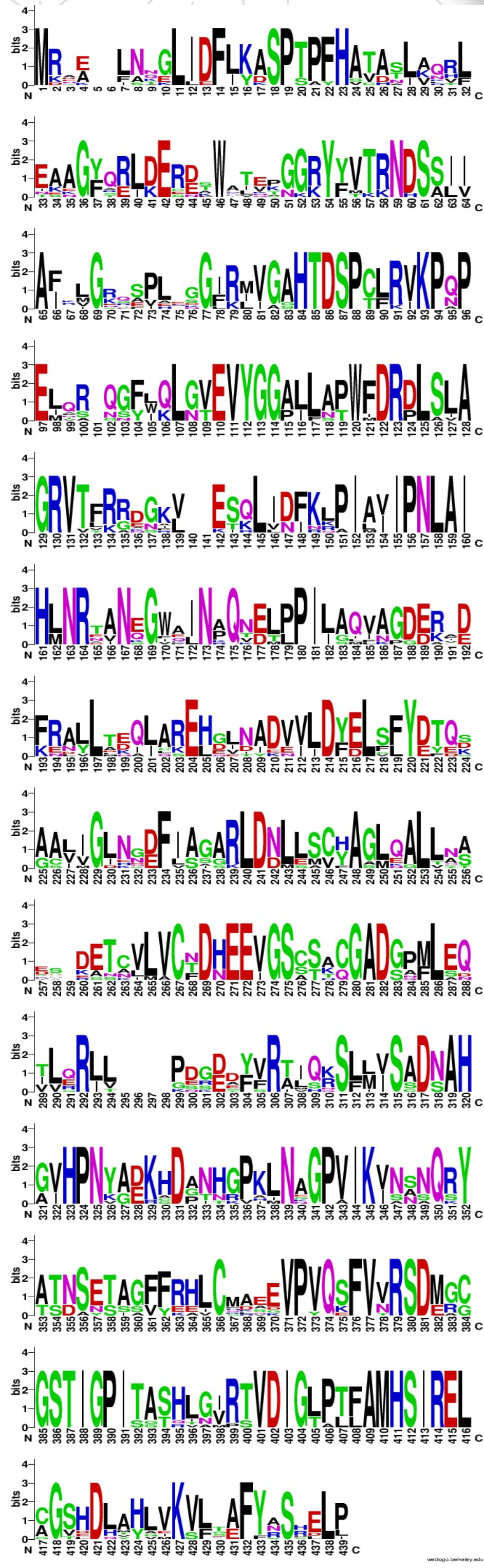
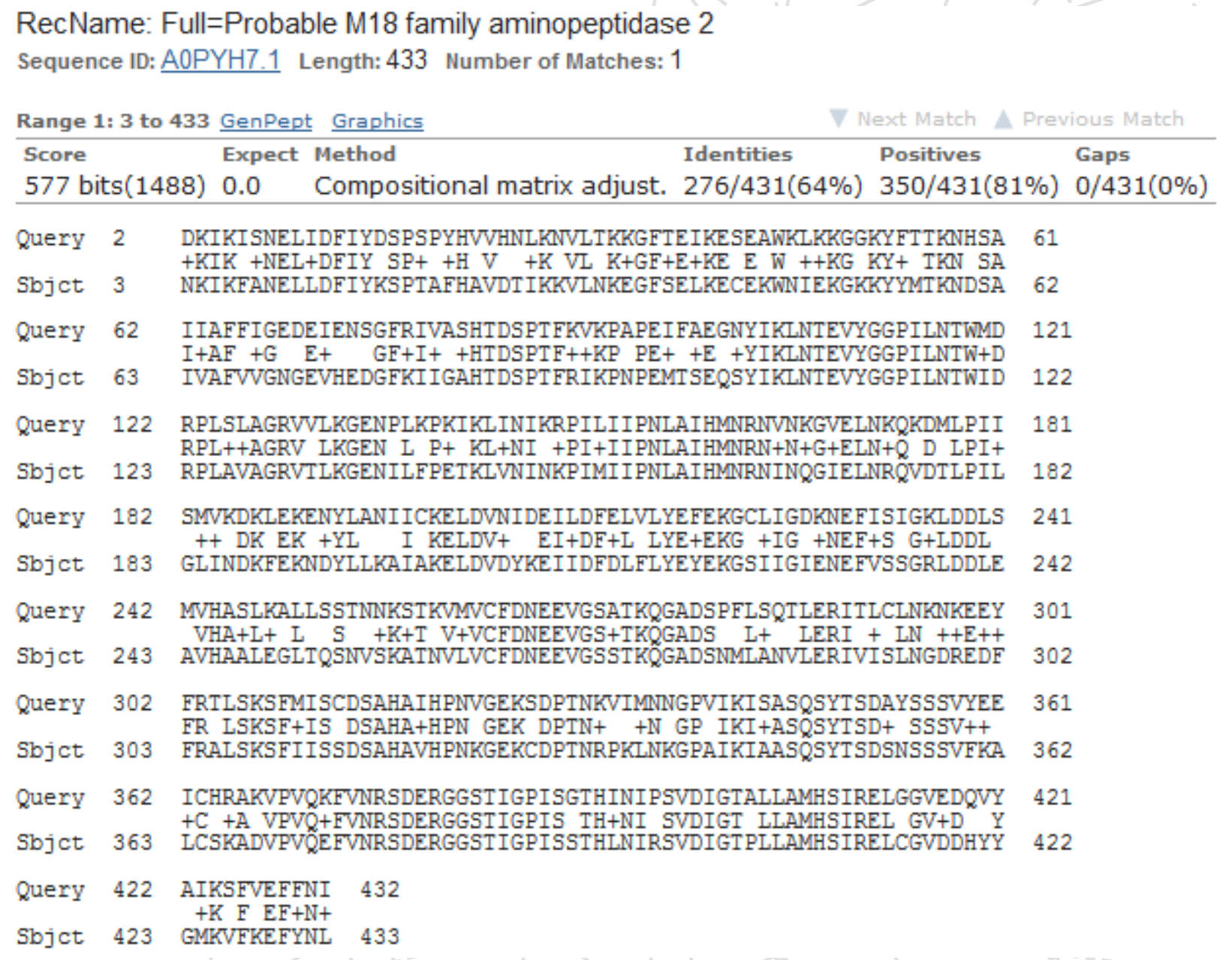


Figure IX – The WebLogo (left) compares the amino acid residues to other organisms' by using single-letter amino acid abbreviations as well as different sizes and colors to represent similarities and differences



RecName: Full=Probable M18 family aminopeptidase 2
Sequence ID: A0PYH7.1 Length: 433 Number of Matches: 1

Range 1: 3 to 433 GenPept Graphics ▼ Next Match ▲ Previous Match

| Score | Expect | Method | Identities | Positives | Gaps |
|---|---|---|---|---|---|
| 577 bits(1488) | 0.0 | Compositional matrix adjust. | 276/431(64%) | 350/431(81%) | 0/431(0%) |

```
Query  2    DKIKISNELIDFIYDSPSPYHVVHNLKNVLTKKGFTEIKESEAWKLKKGGKYFTTKNHSA  61
            +KIK +NEL+DFIY SP+ +H V  +K VL K+GF+E+KE E W ++KG KY+ TKN SA
Sbjct  3    NKIKFANELLDFIYKSPTAFHAVDTIKKVLNKEGFSELKECEKWNIEKGKKYYMTKNDSA  62

Query  62   IIAFFIGEDEIENSGFRIVASHTDSPTFKVKPAPEIFAEGNYIKLNTEVYGGPILNTWMD  121
            I+AF +G  E+   GF+I+ +HTDSPTF++KP PE+ +E +YIKLNTEVYGGPILNTW+D
Sbjct  63   IVAFVVGNGEVHEDGFKIIGAHTDSPTFRIKPNPEMTSEQSYIKLNTEVYGGPILNTWID  122

Query  122  RPLSLAGRVVLKGENPLKPKIKLINIKRPILIIPNLAIHMNRNVNKGVELNKQKDMLPII  181
            RPL++AGRV LKGEN L P+ KL+NI +PI+IIPNLAIHMNRN+N+G+ELN+Q D LPI+
Sbjct  123  RPLAVAGRVTLKGENILFPETKLVNINKPIMIIPNLAIHMNRNINQGIELNRQVDTLPIL  182

Query  182  SMVKDKLEKENYLANIICKELDVNIDEILDFELVLYEFEKGCLIGDKNEFISIGKLDDLS  241
             ++ DK EK +YL   I KELDV+   EI+DF+L LYE+EKG +IG +NEF+S G+LDDL
Sbjct  183  GLINDKFEKNDYLLKAIAKELDVDYKEIIDFDLFLYEYEKGSIIGIENEFVSSGRLDDLE  242

Query  242  MVHASLKALLSSTNNKSTKVMVCFDNEEVGSATKQGADSPFLSQTLERITLCLNKNKEEY  301
             VHA+L+ L  S  +K+T V+VCFDNEEVGS+TKQGADS  L+  LERI + LN ++E++
Sbjct  243  AVHAALEGLTQSNVSKATNVLVCFDNEEVGSSTKQGADSNMLANVLERIVISLNGDREDF  302

Query  302  FRTLSKSFMISCDSAHAIHPNVGEKSDPTNKVIMNNGPVIKISASQSYTSDAYSSSVYEE  361
            FR LSKSF+IS DSAHA+HPN GEK DPTN+  +N GP IKI+ASQSYTSD+ SSV++
Sbjct  303  FRALSKSFIISSDSAHAVHPNKGEKCDPTNRPKLNKGPAIKIAASQSYTSDSNSSSVFKA  362

Query  362  ICHRAKVPVQKFVNRSDERGGSTIGPISGTHINIPSVDIGTALLAMHSIRELGGVEDQVY  421
            +C +A VPVQ+FVNRSDERGGSTIGPIS TH+NI SVDIGT LLAMHSIREL GV+D  Y
Sbjct  363  LCSKADVPVQEFVNRSDERGGSTIGPISSTHLNIRSVDIGTPLLAMHSIRELCGVDDHYY  422

Query  422  AIKSFVEFFNI  432
             +K F EF+N+
Sbjct  423  GMKVFKEFYNL  433
```

Figure X – The query sequence (above) shows the alignment length, e-value, score (in bits) and much more information regarding the gene. The information was received from the swissprot database during a BLAST search

## Conclusion

The GENI-ACT proposed gene product did not differ significantly from the proposed gene annotation; thus, the genes appear to have been correctly annotated by the computer database. Overall the results were more accurate from the nr database.

| Gene Locus | Geni-Act Gene Products | Proposed Annotation |
|---|---|---|
| **3493** | Zinc metalloprotease | Zinc metalloprotease |

## References

Montecucco C, *Curr Top Microbiol Immunol.* 1995;195:1-278
Tucker J. B, Cambridge, Mass: MIT Press; 2000

## Acknowledgments

Supported by an NSF Innovative Technology Experiences for Students and Teachers (ITEST) Award – 1311902

**Special Thanks to**:
Dr, Rama Dey-Rao and Dr. Stephen Koury
Mrs. Sangeeta Gokhale

**www.buffalo.edu**