Annotation of the *Pseudomonas aeruginosa* Genome from DNA Coordinates 69788 to 86086(or Locus Tags T223_00295 to T223_00380)

Erin Muhleisen, Kayla Flaitz, *Sophia Carretto, *Erin Briggs, *Vivianne Spitulnik, Emma Flaitz, Kate Colucci, Madison Murphy, and Kathy Flint * indicates same effort

Hornell City School District, Hornell 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

The bacterium *Pseudomonas aeruginosa* (T223_00295 – T223_00380) was annotated using the collaborative genome annotation website GENI-ACT. The Genbank proposed gene productname 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 and cellular localization data. The numerous mutations and adaptability of this gene intrigues researchers to investigate the gene's amino acid sequence. The Genbank proposed gene product was supported by the proposed gene annotation in all but one gene in the group and as such, overall, the genes appear to be correctly annotated by the database.

Introduction

Pseudomonas aeruginosa is a gram-negative bacterium from the Pseudomonadaceae family that grows in soil, coastal marine habitats as well as plant and animal tissue. This organism is an omnipresent environmental bacterium that is a leading cause of opportunistic human infections and, due to its larger genome size and adaptability, resists antimicrobial therapy. There are over 1,200 mutations of the cystic fibrosis gene causing diagnosis and treatment to be difficult. *Pseudomonas aeruginosa* is a significant bacterium since it is the predominant cause of mortality in cystic fibrosis patients. Because of their abnormal airway epithelia, long-term colonization of *Pseudomonas aeruginosa* is allowed. Biofilm-growing mucoid strains in the lungs of patients with cystic fibrosis allowing for chronic Pseudomonas aeruginosa infections thus leading to respiratory failure (Friedrich, 2016). Only a very few successful antibiotics for the *P. aeruginosa* bacterium have been found definitively because of its high adaptability. Because cystic fibrosis has no known cure, researchers are interested in analyzing the *Pseudomonas aeruginosa* gene.

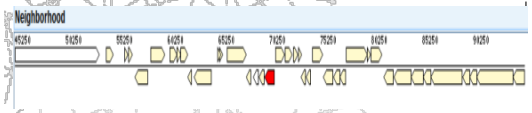


Figure 1. The locus tags and relative position of the genes under investigation in this research

Methods

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

Results

T223_00295:

The leading proposed product of this gene by GENI-ACT was a transcriptional regulator. The top BLAST hits for the amino acid sequence indicate well-curated protein functional domains within the amino acid sequence present, and the cellular location of the sequence support this gene product. As such, the proposed annotation is a transcriptional regulator. The database used was the non-redundant protein sequence database. The score of the hit was 416 bits and the e-value was 9e-144, thus proving the hit was significant. Refer to figure 2.

T223_00330:

The initial proposed product of this gene was an amino acid sequence hit was an organism called Enterobacter Cloacae (aminopeptidase).The e-value was zero indicating a significant match. There was only one major gap in the WebLogo sequence between 2 to 126, but otherwise the sequence was conserved throughout. Using sites such as : TMHMM, SignalP, LipoP, PSORT-B, and Phobius, it was indicated that T223_00330 could be located in the cytoplasm. When searching about the enzymatic function of the sequence no hit could be found on TIGRFAM, while one hit was found on Pfam. The Pfam hit had multiple predominant upper case letters. Another website used was PDB, which came up with a result identical to our first blast hit, Structure of glycosylated human aminopeptidase N, indicating the strong connection between T223_00330 and aminopeptidase.

T223_00350:

The initial product of this gene by GENI-ACT amino acid sequence was the top hit from Swiss-prot was with the organism Oligopeptidase A. With a score of 809 the hit was significant. The TMHMM hits for this gene showed there were no predicted transmembrane helices. No transmembrane helices indicates this protein is not an integral protein, as shown in figure 5. This gene also had no indication of Signal peptide from the prediction number of 0.097. To have a Signal peptide prediction the number has to be above 0.450.

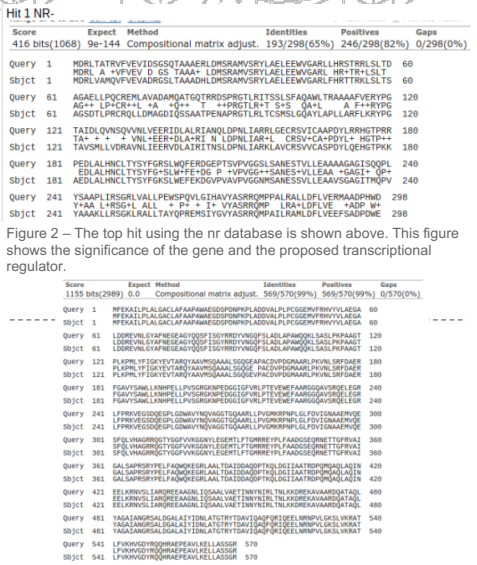


Figure 2 – The top hit using the nr database is shown above. This figure shows the significance of the gene and the proposed transcriptional regulator.

Figure 3 – T223_00370 Alignment of top nr database hit and query sequence for insignificant secretion protein with E-value of 0.0. (Refer to T223_00370 under Results section)

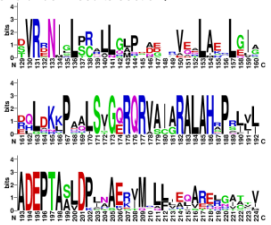


Figure 4 – T223_00380 A portion of the well-curated functional domains from 129-224 (WEbLogo).

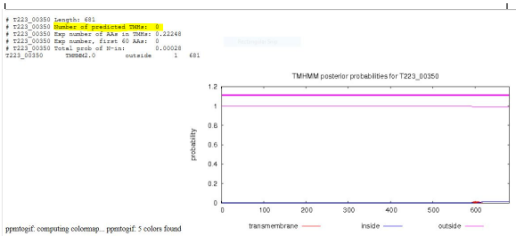


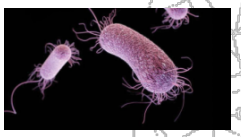
Figure 5 - As shown by the highlighted part there are no predicted transmembrane helices.

T223_00370:

The initial proposed product of this gene by GENI-ACT was a secretion protein. However the hits in the non-redundant database for this gene were either non-significant with an E-value of 0.0 or hypothetical proteins; this gene product proposal was not supported by the top BLAST hits for the amino acid sequence. As such, the proposed annotation was not supported by the non-redundant blast. The score of the top hit, which was a secretion protein was 1155 and the E-value was 0.0 and non-significant. Thus indicating no other sequence has homology to this protein because the insignificant E-value.

T223_00380:

The initial proposed product of this gene by GENI-ACT was an ATP binding protein. This gene product was supported by the top BLAST hits for the amino acid sequence in the Swiss-Prot, the enzymatic function of the amino acid sequence, and the location of the sequence. The e- value 1e-35 and the score was 129 bits, which validates the significance of the hit. Furthermore, the data found using the top hit in T-Coffee and WebLogo supports that T223_00380 is a well preserved and well-curated protein. See Figure 4.



Conclusion

The GENI-ACT proposed gene product supported the proposed gene annotation for all but one of the genes in the group. T223_00370 had a GENI-ACT gene product of a secretion protein, but a secretion protein doesn't resemble pseudomonas. Overall, the genes appear to be correctly annotated by the database.

Locus	Geni-Act Gene Product	Proposed Annotation	Number of modules completed
T223_00295	Transcriptional Regulator	Transcriptional Regulator	2
T223_00330	Aminopeptidase	Aminopeptidase	4
T223_00350	Oligopeptidase A	Oligopeptidase A	4
T223_00370	Secretion Protein	Secretion Protein doesn't resemble pseudomonas	2
T223_00380	ATP Binding protein	ATP Binding protein	2

References

Marcus Friedrich, 2016.

Acknowledgments

Supported by an NIH Science Education Partnership (SEPA) Award - R25ODO10536, Dr. Rama Dey-Rao, and Dr. Stephen Koury

www.buffalo.edu