University at Buffalo *The State University of New York* | REACHING OTHERS

Annotation of Locus Tags Ksed_09850 and Ksed_09830 of the *Kytococcus sedentarius* Genome

Matthew Spink, Joshua Mincer, Elizabeth Winters and Laura O’Donnell

Attica High School and the Western New York Genetics in Research Partnership

Abstract

Two separate genes from the microorganism *Kytococcus sedentarius* (Ksed_09850 and Ksed_09830) were annotated using the collaborative genome annotation website GENI-ACT. The Genbank proposed gene product name 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, cellular localization data, potential alternative open reading frames, structure-based evidence, enzymatic function, gene duplication and degradation, evidence for horizontal gene transfer, and RNA. The Genbank proposed gene product name did not differ significantly from the proposed gene annotation for each of the genes in the group, but we propose a new start codon for Ksed_09850.

Introduction

Kytococcus sedentarius is a strictly aerobic, non-motile, non-encapsulated, and non-endospore forming gram positive coccoid bacterium, found predominantly in tetrad formation. This organism is classified as a chemoheterotroph, as it requires methionine and several other amino acids for growth. Originally isolated from a microscope slide submerged in sea water in 1944, *Kytococcus sedentarius* grows well in sodium chloride at concentrations less than 10% (w/v).

According to Sims et al. (2009), *Kytococcus sedentarius* is a microorganism of interest for several reasons. This bacterium is a natural source of the oligoketide antibiotics monensin A and monensin B (Sims et al., 2009). *Kytococcus sedentarius* has been implemented as the etiological agent of a number of opportunistic infections including valve endocarditis, hemorrhagic pneumonia, and pitted keratolysis (Sims et al., 2009). Finally, the phylogeny of this microorganism is a source of interest, as it is a member of the family *Dermacoccaceae* within the actinobacterial suborder *Micrococcineae,* which has yet to have been thoroughly studied utilizing bioinformatics (Sims et al., 2009).

Methods and Materials

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

Ksed 09850		
Basic Information	**DNA Coordinates**	1017189..1017623
	DNA Length	435 Base Pairs
	Amino Acid Sequence Length	144 Amino Acids
Sequence Based Similarity	**COGs**	cdo5829 (Sortase_E) Pfam04203 (Sortase)
	TCOFFEE	Moderately well conserved
	WEBLOGO	Moderately well conserved
Cellular Localization	**Transmembrane Helices**	0
	Signal Peptide	1
	PSORTB Final Prediction	Extracellular
Alternative ORFs		1017181..1017623
Structure Based Evidence	**TIGRFAMs**	TIGR01076 (Sortase Family Protein)
	PFAMs	PF04203 (Sortase Family)
	HMM Logo Key Residues	H46, F54, T107, C108, R118
	PDB	4D70 (Clostridium perfringens Sortase D5 transpeptidase)

Figure 1.1- General information of *Ksed_09850*

Kytococcus sedentarius 09850: Annotation

For this gene, with and without the Alternate Open Reading Frame, all of the databases show the same result. All of the databases show that it produces a sortase family protein. Sortase_E is the most probable sortase family protein that is produced since it is the top COG hit and had a E-value of 2.26e-36. Despite the consensus regarding the family of protein it produces, what exact sortase family protein it produces is harder to predict because there is a much smaller consensus between the databases regarding the exact product. Never the less, the gene Ksed_09850 produces a sortase family protein based on this annotation.

The Enzymatic Function of Sortase:

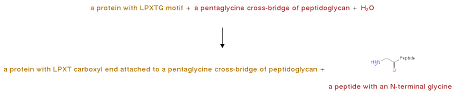

Figure 1.2- This is the MetaCyc pathway map for Sortase A. Although it is not certain that this gene produce Sortase A, all sortase family proteins function in a similar fashion. According to Wikipedia, sortase proteins “modify surface proteins by recognizing and cleaving a carboxyl-terminal.”

The Alternate Open Reading Frame for Ksed_09850:

As noted above this gene was called incorrectly and an Alternate Open Frame Reading Frame was found. Upon completing the Cellular Localization Data for this gene, as it was called by the gene caller, the location of the gene was unknown. According to the databases it was non cytoplasmic, and did contain any signal peptides or any transmembrane helices. These were odd results because it led to a great uncertainty regarding the location of the protein.

A possible Alternate Open Frame Reading was found. It was eight base pairs upstream from the called start codon by the gene caller. After testing the Cellular Localization data module again, a signal peptide was found by Signal P. This the presence of a signal peptide allowed the protein to be located outside the cell. This Alternate Open Reading Frame (AORF) was critical in confirming Ksed_09850’s annotation. Containing a signal peptide made Ksed_09850 predicted production of a sortase family protein logical. Figure 1.3 and 1.4 show the effect of the AORF on the SignalP results.

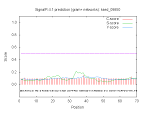

Figure 1.3 - Before AORF: # Measure Position Value Cutoff signal peptide? max. C 38 0.151 max Y 38 0.138 max. S 34 0.216 mean S 1-37 0.105 D 1-37 0.125 0.450 NO Name=ksed_09850 SP='NO' D=0.125 D-cutoff=0.450 Networks=SignalP-TM

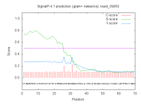

Figure 1.4 - After AORF: Measure Positon Value Cutoff signal peptide? max C 31 0.227 max Y 26 0.358 max. S 8 0.799 mean S 1-25 0.701 D 1-25 0.492 0.450 YES Name=ksed_09850 SP='YES' Cleavage site between pos. 25 and 26: RWA-SP D=0.492 D-cutoff=0.450 Networks=SignalP-TM

Kytococcus sedentarius 09830 Annotation:

The initial proposed product of this gene was an aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase subunit B, and the investigation of this gene has proved this to be the case. All of the hits in BLAST were of the same product or of just glutamyl-tRNA amidotransferase. This information, along with the results of the phylogenetic tree analysis and compared ortholog neighborhood regions of other organisms show that this gene product and this gene are not a result of horizontal gene transfer, but rather originate from a common ancestor. This data is also supported through the gene product’s Pfam family of GatB_N, which provides a means for the formation of glutaminyl- tRNA(Gln) through the transamidation of misacylated glutamyl-tRNA(Gln) in organisms which lack glutaminyl-tRNA synthetase., which correlates to the proposed annotation. As this product maybe observed in *Homo sapiens* as well, the common ancestor may be back chronologically far. This gene appears to be linked to gatA and gatC genes as well, in a gatCAB complex, as shown in the ortholog neighborhood regions depicted below in Figure 2.1

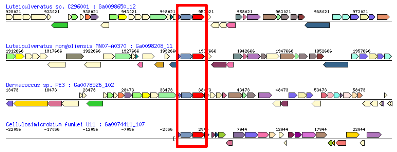

Figure 2.1 - As depicted above, the gene was found to be consistently linked with gatA and gatC to make gatCAB, with each of the subunits working together to create the amidotransferase.

The Function of aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase subunit B:

Amidotransferase are involved in transamidation, which is the transfer of an amide group from one compound to another. Aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase subunit B is involved in the transamidation of L-glutamyl-tRNA(Gln) and L-aspartyl-tRNA(Asn) to L-glutaminyl-tRNA(Gln) and L-asparaginyl-tRNA(Asn), respectively. The amide group is transferred from L-glutamine for both processes. This occurs in two steps: first, with glutamyl-tRNA(Gln) sythetase generating a misacylated L-glutamyl-tRNA(Gln) species, then with the glutamyl-tRNA(Gln) amidotransferase transferring the proper amide group. This will occur in organisms that lack either or both of asparaginyl-tRNA or glutaminyl-tRNA synthetases. The MetaCyc pathway is depicted below in Figure 2.2

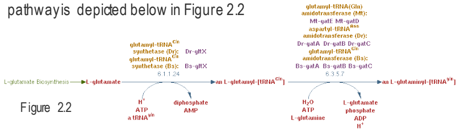

Figure 2.2

Overall, the aspartyl/glutamyl-tRNA (Asn/Gln) amidotransferase subunit B works with subunits A and C to metabolize L-Asparagine and L-glutamine and the formation of L-glutamyl-tRNA(Gln) and L-asparaginyl-tRNA(Asn), as is depicted in the KEGG Pathway map in Figure 2.3 below.

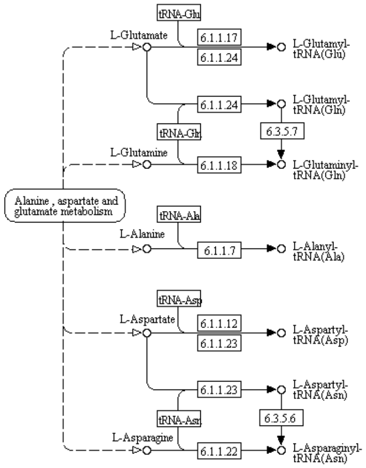

Figure 2.3. The KEGG Pathway Map of Alanine, Aspartate and Glutamate Metabolism.

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. These results support that *Ksed_09380* forms an aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase subunit B while *Ksed_09850* forms an uncertain sortase family protein.

Gene Locus	Proposed Annotation
09830	Aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase subunit B
09850	Sortase Protein

References

Sims et al. (2009). Complete genome sequence of Kytococcus sedentarius type strain (541T). *Standards in Genomic Sciences*, 12 - 20.

Acknowledgments

Supported by NSF ITEST Strategies Award Number 1311902