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

# Annotation of the *Kytococcus sedentarius* Genome from Locus Tags Ksed_08340 to Ksed_08370

**Adrian Bell, Jean Ji, Emma Fiorini, Alexandra DiTommaso and Betsy Vinton**

The Harley School Department of Biological Sciences and the Western New York Genetics in Research Partnership

## Abstract

A group of consecutive 4 genes from the microorganism *Kytococcus sedentarius* (Ksed_08340 -- Ksed_08350 -- Ksed_08360 – Ksed_08370) were annotated using the collaborative genome annotation website GENI-ACT. The Genbank proposed gene product name for each gene was evaluated 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. For these 4 genes, the Genbank proposed gene product name did not differ significantly from the proposed gene annotation, suggesting that the genes were correctly annotated by the database.

## Introduction

*Kytococcus sedentarius* is a strictly aerobic, non-motile, non-encapsulated, non-endospore forming, and gram positive coccoid bacterium, found predominantly in tetrad formation. This organism is classified a sa 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 *Micrococineae*, which has yet to have been thoroughly studied utilizing bioinformatics (Sims et al., 2009).

In addition, recent research has revealed that microbes play a vital role in humans, as well as many other organisms. Trillions of microbes combine together to create efficient microbiomes that significantly aid an organism's metabolic processes and immune defense. Microbes and organisms have coevolved to benefit each other, creating a mutually beneficial relationship. In fact, the human body hosts more than 100 trillions bacterial and fungal cells, while there are only 30 million human cells. Therefore, it is reasonable that human health is largely attributed to the diversity and well-being of these microbes. (Blaser, 2014)

In this study, students used a variety of online aids and databases to understand their gene's function and location in the cell.

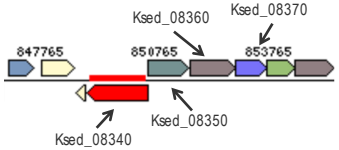


Figure I – *Kytococcus sedentarius* 08340 (red), 08350(gray), 08360(day), and 08370(purple) gene neighborhood.

## 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- Structure-Based Evidence | TIGRfam, Pfam, PDB | Are there functional domains in my protein? |
| Module 4- Cellular Localization Data | Gram Stain, TMHMM, SignalP, PSORT, Phobius | Is my protein in the cytoplasm, secreted or embedded in the membrane? |
| Module 5- 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? |
| Final Annotation | Review data from all modules | Does the student proposed name of the gene agree with that proposed by the automated computer annotation? Are any changes proposed to the pipeline annotation? |

## Results

***Kytococcus sedentarius* 08340:**

GENI-ACT proposed the product of this gene as a acetyl-CoA carboxylase, carboxyltransferase component (subunits alpha and beta). This proposal was supported by the top BLAST results (methylmalonyl-CoA carboxytransferase) of the NR database using the amino acid sequence. TIGRFAM predicted the gene as a methylmalonyl-CoA decarboxylase alpha subunit, which uses energy from decarboxylation of carboxylic acid substrates to extrude Na+ ions across cell membrane. The PBD predicted the gene as Acyl-CoA Carboxylase Beta Subunit. These results further support GENI-ACT's proposal. TMHMM indicated that there are no transmembrane helices, Signal IP revealed low probability for signal peptide, PSORT-B predicted Ksed_08340 to be a cytoplasmic protein. Phobius corroborates negative results for TMH and Signal peptide. Taking all of these results into account, the final prediction for localization is that Ksed_08340 is a cytoplasmic protein. To verify that the correct amino acid sequence was called by the computer, the IMG gene sequence viewer was implemented to search for other possible start codons. A BLAST search with the alternate sequences confirmed the computer had called the correct amino acid sequence.

***Kytococcus sedentarius* 08350:**

This locus included the DNA coordinates 851205 to 852248 on the forward strand as indicated by the GENI-ACT gene page. The gene product is 347 amino acids long. A BLAST query gave a top hit of sequence similarity with biotin--[acetyl-CoA-carboxylase] ligase from Luteipulveratus mongoliensis with an e-value of 1e-29. The second hit was the same gene product as the first hit but was from Phycicoccus sp. Soil802. The database used for the BLAST query was nr (Non-redundant protein sequences). The CDD database search resulted in one COG, COG0340, a biotin-(acetyl-CoA carboxylase) ligase used in coenzyme transport and metabolism. A TMHMM analysis did not detect any transmembrane helices. No signal peptide was predicted by the SignalP analysis.

Based on these results and the results from PSORT-B and Phobius, the protein is expected to be outside the cytoplasm. An alternate open reading frame was found that included the DNA coordinates 851244 to 852248 on the forward strand. The e-value found from the BLAST query was 1e-30 instead of 1e-29.

The following figures represent some interesting findings from the research into these 4 gene sequences:

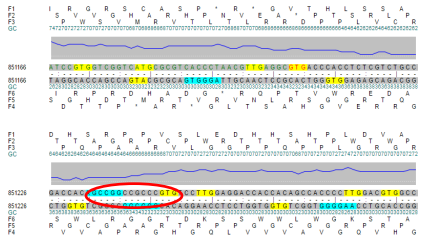

Figure II – An alternate reference frame search on *Kytococcus sedentarius* 08350 revealed another possible start codon, as indicated by the circled (red) Shine-Dalgarno region.


Figure III – A segment of the Web Logo for *Kytococcus sedentarius* 08340 reveals shows well conserved, hydrophobic and polar amino acid groupings.

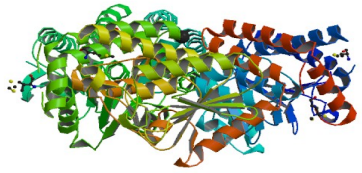

Figure IV – Generated biological assembly from Protein Data Bank of the top result for *Kytococcus sedentarius* 08370 (PBD Code: 3KQF).

***Kytococcus sedentarius* 08360:**

Ksed_08360 has the DNA coordinates of 852291 to 853505 for the forward strand. According to BLAST, Ksed_08360 matched two hits. The first hit was guanylate cyclase [Dermacoccus sp. PE3]; second hit was adenylyl cyclase [Knoellia aerolata]. CDD gave a COG name of Adenylate cyclase, class 3 [Signal transduction mechanisms] as the top hit.

T-Coffee was used to make a web logo. From amino acid positions 4-27, 147-174, 234-237, and 405-422 there were no similar amino acids. The Pfam number was PF16701 as the top hit. The Pfam name was Adenylate cyclase regulatory domain. There were no significant TIGRfam hits, but an alternative search could be using PDB. For the cellular localization data, the bacterium was found to be positive where it was originally isolated from a marine environment. There were no transmembrane helices or signal peptide detected, therefore this is not an integral membrane protein. Ksed_08360 was found to have a cytoplasmic score of 9.97 in PSORT-B. Lastly, the Alternative Open Reading Frame was processed. The proposed DNA coordinates were 852411 through 853505. The beginning proposed coordinates were different than the original ones, and the E-value was more favorable than the one identified by GENI-ACT.

***Kytococcus sedentarius* 08370:**

GENI-ACT proposed the product of this gene as short chain enoyl-CoA Hydratase. This proposal was supported by the top BLAST results (enoyl-CoA hydratase) of the NR database using the amino acid sequence. TIGRFAM, however, predicted the gene as badI: 2-ketocyclohexane carboxyl-CoA hydrola. TMHMM showed that there is no transmembrane helix. No signal peptide is predicted in the in the Signal IP. PSORT-B predicted Ksed_08370 to be a cytoplasmic protein. Based on the data above, the final prediction for localization is that Ksed_08370 is a cytoplasmic protein. The coordinates appear to be correct as called. A shorter alternative open reading frame in ORF1 with a potential Shine-Dalgarno sequence was investigated, but the BLAST E-value for the shorter ORF was increased (the computer predicted E-value was lower).

## 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.

| Gene Locus | Geni-Act Gene Products | Proposed Annotation |
|---|---|---|
| 08340 | Acetyl-CoA Carboxylase | Carboxyl Transferase Domain |
| 08350 | Biotin Ligase | Biotin protein Ligase C Terminal Domain |
| 08360 | Family 3 Adenylate Cyclase | Adenylate and Guanylate Cyclase Catalytic domain |
| 08370 | Short Chain Enoyl-CoA Hydratase | Enoyl-CoA Hydratase/Isomerase Family |

## References

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

Missing Microbes. ( 2014). Blaser, Martin MD.

## Acknowledgments

Supported by NSF ITEST Strategies Award Number 1311902

Thanks to Dr. Koury and Dr. Dey-Rao for their advice.