



Gene Annotation of Ksed_04890, a Hypothetical Protein of Unknown Function in *Kytococcus sedentarius*

Ariel Breads and Lon Knappenberger

Westfield Academy and Central School, Westfield, New York 14787 and The Western New York Genetics in Research Partnership

Abstract

Kytococcus sedentarius is an aerobic, gram positive organism found on a partially submerged slide. Students at Westfield Academy and Central School have annotated individual genes in order to further understand the function of this organism. Eight modules on GENI-ACT were completed in order to reach conclusions on the function of each gene. It is through these steps that the gene product Ksed_04890 was found to be a hypothetical protein of un-known function. However, a hypothesis was formulated that it can that Ksed_04890 may be a new transposase or a non-functional mutation of a transposase based on multiple sequence alignment of Ksed_04890 with a number of short transposase sequences found in the *Kytococcus* genome.

Introduction

Kytococcus sedentarius cells are spherical and occur predominantly in tetrads. *K. sedentarius* is Gram-positive, nonmotile, non-encapsulated, and not endospore-forming. *Kytococcus sedentarius* is strictly aerobic, requires methionine and grows well in NaCl at concentrations up to 10%¹

The genome of *Kytococcus sedentarius* has 2,785,024 base pairs coding for 2703 genes. 1948 have predicted functions¹. 64 are RNA genes. 2639 genes code for proteins. 1948 of the gene products have predicted functions. This means that almost 700 genes have no predicted function.¹

The coordinates of Ksed_04890 are (484140..484247), which contains 108 base pairs coding for 35 amino acids.

This gene is being studied in order to determine the function of each gene and the genome of *Kytococcus Sedentarius*. Having students study these genes not only gave each student real-life experience in gene annotation, but also verified information given by computers and subsequently contributed to a better understanding of *Kytococcus sedentarius*. Hopefully a further understanding of *Kytococcus sedentarius* could lead to the discovery of a gene product that could have potential medical uses.

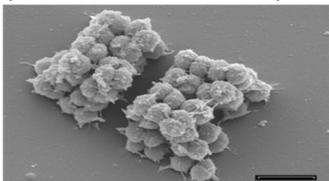


Figure 1. *K. sedentarius* (Sims, et al., 2009)

Materials and Methods

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

cytochrome C biogenesis protein [Arthrobacter sp. 35/47]
Sequence ID: [ref|WP_026543525.1](#) Length: 249 Number of Matches: 1

Score	Expect	Method	Identities	Positives	Gaps
34.7 bits(78)	1.6	Compositional matrix adjust.	15/19(79%)	18/19(94%)	0/19(0%)

Query 1 MGEYFAQTVQSGALLIATP 19
+GE+FA TVQSGALL+AIIP
Sbjct 3 IGEFFADTVQSGALLVAIP 21

transposase, partial [Kineospaera limosa]
Sequence ID: [ref|WP_006594951.1](#) Length: 70 Number of Matches: 1

Score	Expect	Method	Identities	Positives	Gaps
31.2 bits(69)	6.1	Compositional matrix adjust.	14/18(78%)	15/18(83%)	0/18(0%)

Query 18 IFFSTYEAARRAATLPMAA 35
+PSTYEA RAATLP AA
Sbjct 53 LSFSTYEAARRAATLPAA 70

Figure 1. The results of a protein BLAST using the Non-redundant database shows a Cytochrome C biogenesis protein as the top hit (non significant e-value of 1.6).

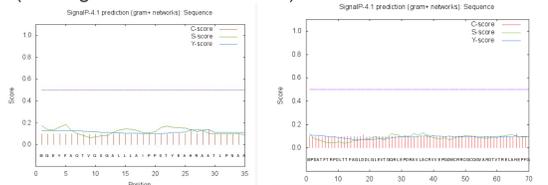


Figure 2. The results of a protein BLAST using the Non-redundant database shows a transposase as the fourth hit (non-significant e-value of 6.1).

Results

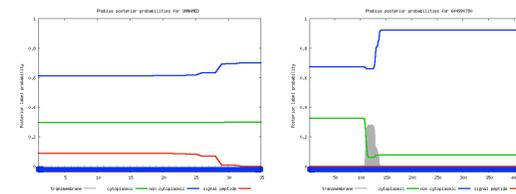


Figure 4. Left. Phobius of Ksed_04890 Hypothetical protein of unknown function. Right. Phobius graph of Ksed_04880, a known transposase.

SegID: ariel Breads	Analysis Report:	Unknown	Cytoplasmic	Extracellular	Final Prediction:
CNSVM+	Unknown	Unknown	Unknown	Unknown	Unknown
CNSVM+	Unknown	Unknown	Unknown	Unknown	Unknown
CNSVM+	Unknown	Unknown	Unknown	Unknown	Unknown
ECVM+	Unknown	Unknown	Unknown	Unknown	Unknown
ECVM+	Unknown	Unknown	Unknown	Unknown	Unknown
ECVM+	Unknown	Unknown	Unknown	Unknown	Unknown
ModHM+	Unknown	Unknown	Unknown	Unknown	Unknown
ModHM+	Unknown	Unknown	Unknown	Unknown	Unknown
ModHM+	Unknown	Unknown	Unknown	Unknown	Unknown
Profile+	Unknown	Unknown	Unknown	Unknown	Unknown
SCL-BLAST+	Unknown	Unknown	Unknown	Unknown	Unknown
SCL-BLAST+	Unknown	Unknown	Unknown	Unknown	Unknown
SCL-BLAST+	Unknown	Unknown	Unknown	Unknown	Unknown
Signal+	Unknown	Unknown	Unknown	Unknown	Unknown
Signal+	Unknown	Unknown	Unknown	Unknown	Unknown
Signal+	Unknown	Unknown	Unknown	Unknown	Unknown

Localization Scores: Cytoplasmic 2.50, CytoplasmicMembrane 2.50, Cellwall 2.50, Extracellular 2.50, Final Prediction: Unknown

Localization Scores: Cytoplasmic 7.50, CytoplasmicMembrane 1.15, Cellwall 0.62, Extracellular 0.73, Final Prediction: Cytoplasmic 7.50

Figure 5. PsortB of Ksed_04890 Hypothetical protein of unknown function (shown left) as compared to a PsortB of Ksed_04880, a known transposase (shown right).



Figure 6. T-Coffee of hypothetical protein 644990781 (Ksed_04890, shown with red arrow) and all known transposases of *Kytococcus* that have a short amino acid length. Short transposases and Ksed_04890 are conserved in the areas shown.

Results continued

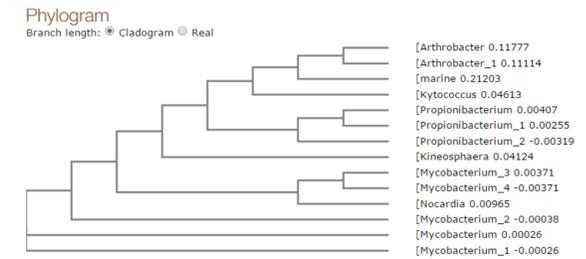


Figure 7. This graph is the phylogenetic tree for the . It shows Arthrobacter as the closest neighbor. *Kytococcus* and *Arthrobacter* are the same down to taxonomic family. *Kytococcus* and *Propionibacterium* are the same down to class. These indicate transmission by Heredity and not Lateral gene transfer.

Conclusions

According to the website used gene Ksed_04890 is a hypothetical protein of unknown name and unknown function. After looking closely at this gene product it was found that that may not be the case. After running a BLAST search of Ksed_04890, the first hit was a cytochrome C biogenesis protein. However, the gene product for Cytochrome C has already been found (Ksed_16240 & Ksed_16250). Ksed_04890 was then compared to several known transposase gene products. Ksed_04890 seemed to have some conservation of domains found in transposase gene products found in the *Kytococcus* genome. We hypothesize that Ksed_04890 could be a new transposase or one that has mutated and lost the functional part of the gene. Further research must be done before any conclusions can be made.

Literature Cited

1. Sims et al., 2009. Stand Genomic Sci. 2009 Jul 20; 1(1): 12–20.

Acknowledgements

This project was supported by a National Science Foundation Innovative Technology Experiences for Students and Teachers (ITEST) Award (#1311902).