

Annotation of the *Kyococcus sedentarius* Genome from DNA Coordinates 364717 to 370640

Nidhish Gokhale¹, Parlang², Jessica Phillips³ and Sangeeta Gokhale

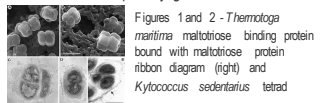
¹Williamsville East High School, ²McKinley High School, ³East Aurora High School (participating in the BEAM - Buffalo Engineering Awareness for Minorities Program) and the Western New York Genetics in Research Partnership

Abstract

Annotations are part of genome analysis that can be done by computational means before a genome sequence is deposited in Genbank (Koonin et al., 2003). The leading website, GENI-ACT, (www.geni-science.org) is an annotation collaboration tool that provides access to genomes, bioinformatics and project designing resources to facilitate genomics research was used to annotate 5 consecutive genes of *Kyococcus sedentarius* (Ksed_03790-Ksed_03830). (www.geni-science.org). The Genbank proposed gene product name for each gene was assessed. Following steps were taken. The assessment of amino acid sequence similarity data to 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

Kyococcus sedentarius is the only bacteria known to us that produces the antibiotic monesin A and B which is used as TMR (Total mixed ration) for increased milk production efficiency in dairy cows. (Whitman et al., 2012). But *Kyococcus* is an opportunistic pathogen. It has been isolated from varying environments, including human skin, groundwater, and even airline cabins. *Kyococcus sedentarius* can cause dermatological infections like pitted keratolysis (foot infection). It is a Gram positive organism that grows as spherical acid in tetrads which can be arranged in cubical packets. It is non-encapsulated and does not form endospores. It is strictly aerobic and chemorganotrophic, requires methionine and other amino acids for growth, and grows well in NaCl at concentrations at optimally grows at 25-37° C.



More studies needed to be done on the commercial uses of the antibiotics and the enzymes made by *Kyococcus*. Ex. Antibiotics Monesin A and B can be used as fodder, serine protease enzymes may be of commercial use in the biodegradation of a range of keratin polymers, biological washing powders and in the treatment of human skin. (Longshaw et al., 2002)



The goal of this study is to study the genome of *Kyococcus sedentarius*, to compare the proposed gene product given by the Genbank with the data given by the genome data base and to compare the results from the Genbank to other amino acid sequences found in other common organisms and bacterium as using different programs like Blast, CCD, T-Coffee, WebLogo, phylogenetic trees, KEGG, Metacyc pathway. The above gene products and sequences were also studied to find possibility of duplications, horizontal gene transfers or being a pseudogene. This project used crowdsourcing to determine protein structure and function. The crowdsourcing compares computer generated results to human based input. It allows for more collaboration and fast paced evidence gathering.

Methods and Materials

Modules of the GENI-ACT (<http://www.geni-act.org/>) were used to complete *Kyococcus sedentarius* genome annotation as 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

Kyococcus sedentarius 03790:

The top BLAST hits were *Escherichia Coli* O157 and *Serinococcus profundus* and the COG Names were Maltose binding periplasmic protein MalE [Carbohydrate transport and metabolism] and PoD: Spermidine/putrescine-binding periplasmic protein [Amino acid transport and metabolism]. The only Pfam name found was SBP_bac_8. Two accepted names, provided by BLAST, are maltose binding periplasmic protein and ABC transporter substrate-binding protein [Serinococcus profundus]. All of this data, research and information leads up to the conclusion: the sequence codes for a carbohydrate (specifically maltose) binding periplasmic protein.

According to the phylogenetic tree, it seems as though *Kyococcus sedentarius* has evolved independently from other bacteria. Thus, this bacterium is divergently evolving from the other species. Its closest relative is *Serinococcus marinus*. It is safe to assume that *Kyococcus sedentarius* has paralogs of the maltose binding ATPase periplasmic protein in other organisms. Other species such as *Ornithinimicrobium pekingsense* and *Serinococcus profundus* seem to be closely related as they have evolved from similar pathways. When aligning the raw translation of Ksed_03790 with the amino acid sequence for this gene in Genbank, it was observed that no frameshift occurred in the sequence as noted in the alignment above. These findings suggest that Ksed_03790 is most likely not a pseudogene, since the sequence is an identical, perfect match and doesn't have an other start and stop codon with its Shine Dalgarno's sequence. On the SignalP results, the probability for the appearance of a signal peptide is 0.616 > 0.45. This indicates a clear presence of a signal peptide with possibility of being present in the extracellular/non-cytoplasmic region.

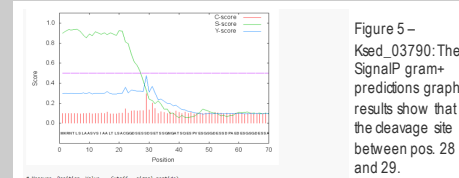
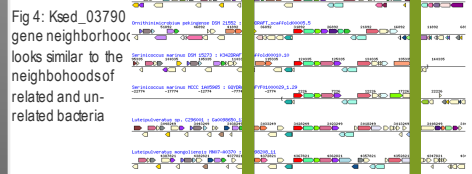


Figure 5 - Ksed_03790: The SignalP gram-predictions graph results show that the cleavage site is between pos. 28 and 29.

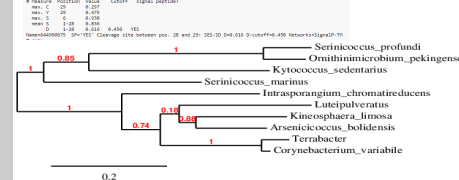


Figure 6 - Ksed_03790: The phylogenetic tree shows *Kyococcus sedentarius* has evolved independently due to some change at point 0.85, closest relative is *Serinococcus marinus*. *Kyococcus sedentarius* may have paralogs of the protein in other organisms, such as *Ornithinimicrobium pekingsense* and *Serinococcus profundus* and they may have evolved from similar pathways.

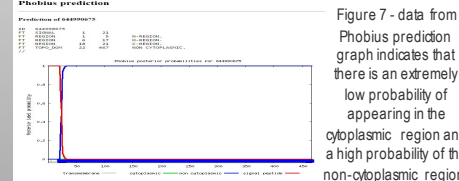


Figure 7 - data from Phobius prediction graph indicates that there is an extremely low probability of appearing in the cytoplasmic region and a high probability of the non-cytoplasmic region.

***Kyococcus sedentarius* 03800:** Genbank proposed gene product name is permease component of ABC-type sugar transporter. Top BLAST hits suggest ABC type sugar transporter proteins. Top COG hits are sugar transporter permease components confirm that protein product is a sugar transporter protein

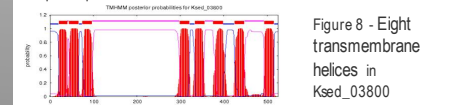


Figure 8 - Eight transmembrane helices in Ksed_03800

***Kyococcus sedentarius* 03810:** BLAST top hits: Maltose transport system permease protein MalG. Cog top hits: ABC-type maltose transport system, permease and ABC-type glycerol-3-phosphate transport system, permease component [Carbohydrate transport and metabolism]. TIGRFam name: BPD_transp_1: ABC transporter, permease. Pfam top hit: Binding-protein-dependent transport system inner membrane component. This concludes that the protein is a transporter membrane protein

***Kyococcus sedentarius* 03820:** Blast top hit HTH-type transcriptional regulator MalR; AltName: Full-Maltose operon transcriptional repressor. Second hit: Transcriptional regulator TIGRFam name: scpA catabolite control protein A. Pfam top hit name: Bacterial regulatory proteins, lacI family Pfam second hit Periplasmic binding protein-like domain. All of the above suggests that the protein is a transcriptional regulator

Kyococcus sedentarius 03830:

The initial proposed product of this gene by GENI-ACT was a Trehalose-phosphatase. This gene product proposal was supported by the top BLAST hits for the amino acid sequence, the presence of well-curated protein functional domains within the amino acid sequence, the cellular location of the amino acid sequence, and the enzymatic function of the amino acid sequence. Therefore, the proposed annotation is Trehalose-phosphatase.

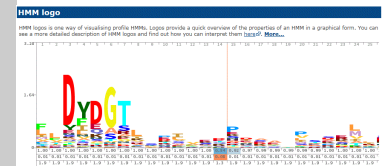


Figure 9 - HMM logo of Ksed_03830

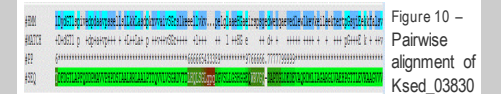


Figure 10 - Pairwise alignment of Ksed_03830

Conclusion

The GENI-ACT proposed gene product did not differ significantly from the proposed gene annotation for each of the genes in the group; therefore, the genes appear to be correctly annotated by the computer database.

Gene Locus	Geni-Act gene product	Proposed Annotations
03790	Maltose/maltodextrin ABC transporter, substrate binding periplasmic protein MalE	Maltose-binding periplasmic protein MalE
03800	Carbohydrate ABC transporter membrane protein 1	Permease component of ABC-type sugar transporter
03810	Carbohydrate ABC transporter membrane protein 2	Maltose transport system permease protein
03820	Transcriptional Regulator, LacI family	HTH-type transcriptional regulator MalR
03830	Trehalose 6 phosphatase	Trehalose-phosphatase

References

Bergey's Manual of Systematic Bacteriology. Volume 5: The Actinobacteria. Whitman et al (2012)
Kyococcus sedentarius, the organism associated with pitted keratolysis, produces two keratin-degrading enzymes. Longshaw et al (2002)
 GENI-ACT Guiding Education through Novel Investigation, Developing Next-generation Academic Scientists <http://www.geni-science.org/>
 Genome annotation: data flow and performance, 5.1.1. Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. Koonin et al (2003)
 Geni-act, <http://www.geni-act.org/>

Acknowledgments

Supported by NSF IT EST Strategies Award Number 1311902
 Special Thanks to: Dr. Rama Dey-Rao and Dr. Stephen Koury (WNYGIRP).