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

Annotation of the *Kytococcus sedentarius* Genome from DNA Coordinates 764347 to 773451

M. Abbadi, S. Ahmed, A. Alabadi, H. Alhajjaji, A. Ali and Entasar Saif

Global Concepts Charter High School and Western New York Genetics in Research Partnership

Abstract

Five genes (Ksed_07600, Ksed_07610, Ksed_07640, Ksed_07650, Ksed_07660) from the microorganism *Kytococcus sedentarius* were annotated using the collaborative genome annotation website GENI-ACT. The Genbank proposed gene product name for each gene was assessed in terms of some or all of the following: 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, and enzymatic function. The Genbank proposed gene product name did not differ significantly from the proposed annotation (based on Modules 1-6) for Ksed_07600. Ksed_07600 and the remaining genes mentioned above were not annotated to completion however. It is proposed, therefore, that further study of these genes is needed to determine their function.

Introduction

To annotate genes is to assign function to them. While supercomputers do this, a person, in contrast, has the unique ability to make connections between genes that a computer cannot. An investigator who manually annotates genes can also catch errors made by a computer. In this project, we participated in the manual annotation of genes from the bacteria, *Kytococcus sedentarius*. This experience has allowed us and others to contribute our findings to the scientific community regarding genes and their functions. Learning more about genes and their functions may lead to the treatment and prevention of disease.

According to Sims et al. (2009), *Kytococcus sedentarius* is a strictly aerobic, free-living, nonmotile, coccoid, non-spore forming, Gram positive bacteria originally isolated from a marine environment. *K. sedentarius* is known for the production of oligoketide antibiotics and for its role as an opportunistic pathogen causing endocarditis, hemorrhagic pneumonia, and pitted keratolysis. It has a genome of interest because of its location in *Dermacoccaceae*, a poorly studied family within the actinobacterial suborder *Micrococcineae* (Sim se tal., 2009). Our study of the genes of *K. sedentarius* comes after its complete genome sequencing as part of the Genomic Encyclopedia of Bacteria and Archea (GEBA) project. With the complete sequencing of this genome, we can look at genes in question and compare them to other genes that have had their function predicted or experimentally determined.

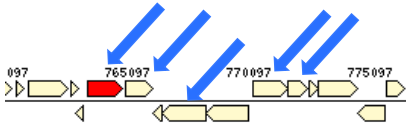


Figure 1: Neighborhood of the genes studied: From left to right- Ksed_07600, Ksed_07610, Ksed_07640, Ksed_07650, and Ksed_07660.

Methods and Materials

Modules of the GENI-ACT (http://www.geni-act.org/) that 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?
Module 6- Enzymatic Function	KEGG, MetaCyc, E.C. Number	In what process does my protein take part?

Results

*Kytococcus sedentarius*07600:

The initial proposed product of this gene by GENI-ACT was inosine-5'-monophosphate dehydrogenase. This gene product proposal was supported by the top BLAST (swiss-pro and non-redundant database) hits for the amino acid sequence where high scores and low e-values were observed. Additionally, one could interpret the presence of significant Clusters of Orthologous Groups (COGs) as an indication that this protein has functional domains in common with other characterized proteins. In addition, WEBLOGO results suggest this protein is very well conserved among various species. Cellular localization data suggest this protein contains no transmembrane helices or signal peptides. Results from the alternate ORF search suggest the proposed start codon for this protein is correct. Searching its amino acid sequence against the TIGRFAM database and finding a good hit also suggests that there is a functional relationship between the gene in question and the hit in the database. Work in the Enzymatic Function Module suggest the protein encoded by this gene is involved in purine metabolism. As such, the proposed annotation is inosine-5'-monophosphate dehydrogenase.

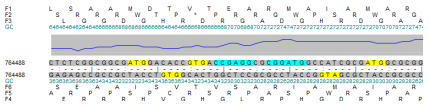


Figure 2: The gene coordinates for Ksed_07600 (764347..765843) appear to be correct as called by the gene caller although there is no applicable Shine-Dalgarno sequence upstream of the start codon (ATG) (not shown).

Figure 2 cont: Figure shows a potential SD sequence (cyan) downstream of the original start codon and it is located 9 nucleotides upstream of a different potential start codon (ATG) (highlighted in yellow at far right) at coordinate 764536. A blast search on the new nucleotide sequence using the swiss-pro database suggest the alternative ORF is not of a significant find due to same scores and e-values.

*Kytococcus sedentarius*07610:

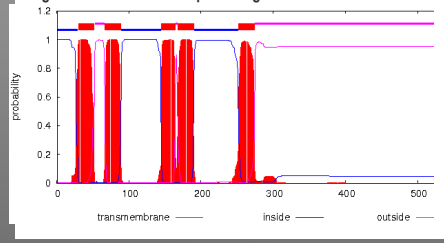
The initial proposed product of this gene by GENI-ACT was a IMP dehydrogenase family protein. The top BLAST hits for the amino acid sequence resulted in high scores and low e-value sand showed a name of uncharacterized oxidoreductase in the Swiss pro database and guanosine monophosphate reductase in the nr database. Significant COG hits showed a name of IMP dehydrogenase/GMP reductase (nucleotide transport and metabolism). No transmembrane or signal peptides were predicted. In addition, WEBLOGO results suggest this protein is very well conserved among various species especially in the stretch between amino acid residues 202-226. As such, the proposed annotation is IMP dehydrogenase family protein.

Figure 3: WEBLOGO results showing high degree of conservation between amino acid residues 202-226 (only shown to 224).

*Kytococcus sedentarius*07640:

The initial proposed product of this gene by GENI-ACT was ABC-type multidrug transport system, ATPase and permease component. This gene product proposal was supported by the top BLAST hits for the amino acid sequences in both databases. The swiss-pro COG top hit was COG1132 and was listed as ABC-type multidrug transport, ATPase and permease component as a defensive mechanism. The COG name was part of the Conserved Protein Domain Family (MdlB), with the e-value being 3.75e-58. WEBLOGO results showed some conservation amongst species with almost no conservation toward the C-terminus. All TIGRFAM hits resulted in negative scores therefore protein families with similar function could not be determined. A Pfam name however was found to be ABC_membrane. Results from TMHMM and Phobius predicted 5 transmembrane helices. There is no predicted signal peptide according to signalIP and PSORTb showed a cytoplasmic membrane score of 10.00. As such the proposed annotation is ABC-type multidrug transport system, ATPase and permease component.

Figure 4: TMHMM results predicting 5 transmembrane helices.



*Kytococcus sedentarius*07650:

The initial proposed product of this gene by GENI-ACT was a YVTN family beta-propeller repeat protein. This gene product proposal was supported by the top swiss-pro BLAST hits, CDD, TIGRFAM, and PFAM searches. WEBLOGO results suggest poor conservation especially in the N and C termini. TMHMM results suggest no TM helices while signalIP and Phobius results support the presence of a signal peptide.

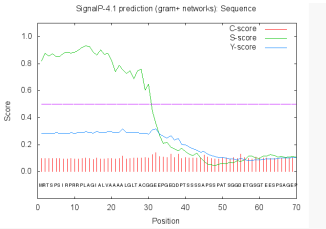
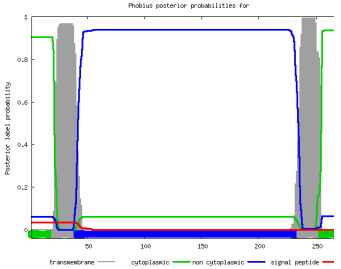


Figure 5: Signal IP graph suggesting presence of signal peptide.

*Kytococcus sedentarius*07660:

The initial proposed product of this gene by GENI-ACT was an uncharacterized protein. Results of the swiss-pro search were discounted due to high e-values, whereas nr results also suggest uncharacterized protein. A CCD search suggests a COG name of Shy1 (Cytochrome oxidase assembly protein ShyY1). A Pfam name of SURF1 was obtained. Phobius predicted this protein to have 2 transmembrane helices.

Figure 6: Phobius results depicted the presence of 2 TM helices.



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.

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