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

Global Concepts CHARTER SCHOOL

Manual Annotation of Four Genes in *Kytococcus sedentarius* Genome

A. Emran, A. Salihovic, S. Aggi, M. Wilczewski, and Christine Masiulionis

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

ABSTRACT

A group of four genes, three consecutive and one non-consecutive, from the microorganism *Kytococcus sedentarius* (Ksed_04830, Ksed_04840, Ksed_04850 and Ksed_04870) were manually annotated using the collaborative genome annotation website GENI-ACT (Genomics Educational National Initiative- Annotation Collaboration Toolkit). The Genbank proposed gene product name for each gene was assessed in terms of the basic genomic information, amino acid sequence-based similarity data, and cellular localization data. The GenBank proposed gene product name 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 in the GENI-ACT database (automated annotation). These genes were called by the computer correctly.

INTRODUCTION

Kytococcus sedentarius is a non-motile, non-endospore forming, non-encapsulated aerobic gram-positive spherical bacterium, found most often as a tetrad. This organism is classified as chemoorgantrophic, as it obtains energy from the oxidation of carbon compounds, including amino acids. Originally found on a microscope slide containing sea water in 1944, *Kytococcus sedentarius* grows well in sodium chloride at concentrations under 10% (w/v).

There are several reasons for choosing the genome from *Kytococcus sedentarius* for manual annotation. This bacterium has a potential to be the source of the natural oligoketide antibiotics (Sims et al., 2009). It has been found to be responsible for a variety of common infections including valve endocarditis, hemorrhagic pneumonia, and pitted keratolysis (Sims et al., 2009). Last but certainly not least, the position on the tree of life is a source of interest, as a member of the family *Dermacoccaceae* within the order *Micrococcales*, and class *Actinobacteria*. The genomic sequence has yet to be completely studied (Sims et al., 2009). It is essential to annotate the genome completely to be able to understand more about the underlying biology of the organism.

Gene neighborhoods of the four annotated genes.

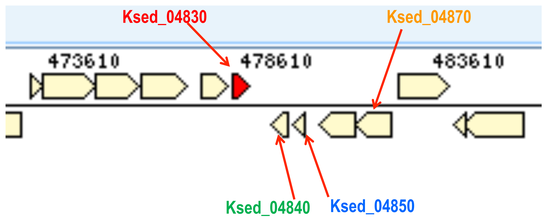


Figure 1 : Gene neighborhoods of:

(1) Ksed_04830 (**red**), DNA coordinates 478376-478840 (+) (465 bp), 154 amino acids; **(2)** Ksed_04840 (**green**), DNA coordinates 479376-479849 (-) 474bp, 157 amino acids; **(3)** Ksed_04850 (**blue**), DNA coordinates 479965-480180 (-) (216 bp), 71 amino acids; **(4)** Ksed_04870 (**yellow**), DNA coordinates 481563-482504 (-) (942 bp), 313 amino acids.

MATERIALS AND METHODS

Modules of the GENI-ACT (http://www.geni-act.org) 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?

Table 1

RESULTS

Ksed_04830 and 04850:

The initial proposed product of these genes by GENI-ACT were hypothetical proteins. BLAST searches using the nr database returns low-level matches with hypothetical proteins and insignificant e-values. No further information was revealed since multiple sequences were not available to go through with T-coffee or Weblogo.

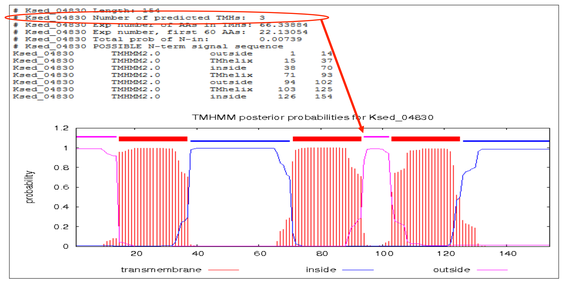


Figure 2: Significant evidence for Ksed_04830 to be a membrane protein with 3 predicted TMHs.

There is preliminary evidence to indicate that Ksed_04830 is a membrane protein. We propose further analyses to substantiate the claim.

Although PSORT-B predicts Ksed_04850 to be a cytoplasmic protein, we concluded that this was based on insufficient evidence.

Ksed_04840:

The initial proposed product of this gene by GENI-ACT was a DNA-binding protein, excisionase family. BLAST search finds a moderate to high match with excisionase from *Jiangella alkaliphila*.

Multiple alignment with the top 10 orthologs showed the presence of a reasonably well conserved amino acid sequence from position 56-172. The PSORT-B score of 7.5 indicates that it is a cytoplasmic protein.

RESULTS

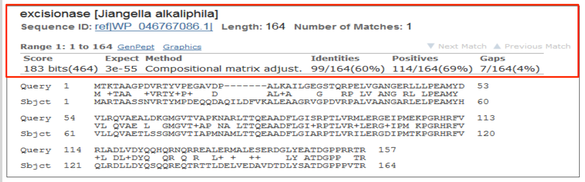


Figure 3: BLAST search using the amino acid sequence of Ksed_04840. Searches against the nr database returned a moderately high alignment with excisionase from *Jiangella alkaliphila* (e-value 3e-55, Score 183 bits with 69% positive) over a length of 157 amino acids

Ksed_04870:

The initial proposed product of this gene by GENI-ACT was a cation diffusion facilitator family transporter.

We performed BLAST searches with both the Swissprot database (SP database) as well as the non-redundant (nr) database.

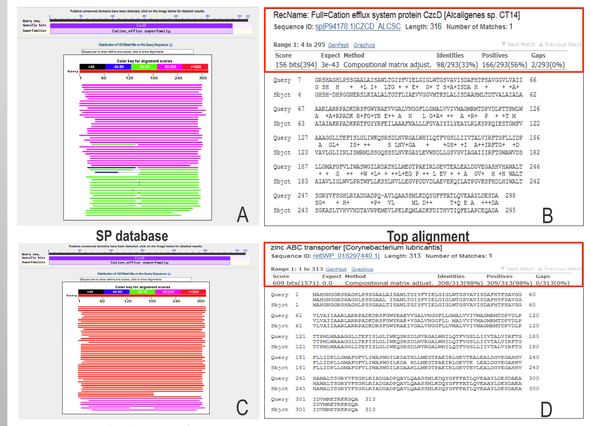
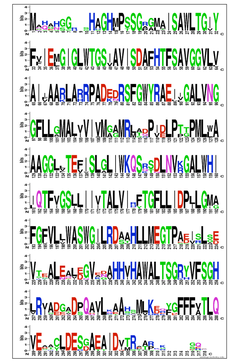


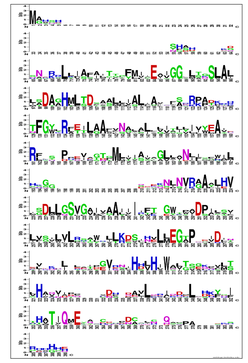
Figure 4: The results for Ksed_04870 show the top match by the curated Swissprot database to be a cation efflux system protein CzcD from *Alcaligenes sp* and zinc ABC transporter from *Corynebacterium lubricans.*

Additional evidence from the CDD search resulted in finding a COG-1230 Co/Zn/Cd efflux system component [Inorganic ion transport and metabolism] with an e-value of 3.73e-69. Using the top 10 orthologs from the nr database (**Figure 5C**) search we found a Weblogo that reflected a very well conserved protein which performs an essential function in *Kytococcus sedentarius*. Using the top 10 sequences from the Swissprot database **(Figure 5A)** search, we found a Weblogo with a moderate conservation which reflected the alignment results of BLAST from SP database. We therefore learned that it was useful to use the curated database of Swissprot for sequence matches to well studied proteins, even if the match was moderate to high. However, to study conservation of proteins, it might be a better choice to probe the nr database and get high quality matches from a larger number of orthologs (low e-values) to get a true picture of conservation of the protein. We clearly show that the protein encoded from the gene in locus tag Ksed_04870 is very well conserved.

RESULTS



A Ten orthologs used from nr database



B. Ten orthologs used from Swissprot database

Figure 5: Ksed_04870 Clear differences in Weblogo created from sequences chosen from nr and SP databases.

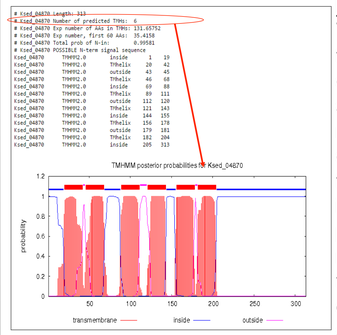


Figure 6: Ksed_04870 TMHMM clearly indicates 6 transmembrane helices containing hydrophobic amino acids. PSORT-B returns a score of 10.0 for cytoplasmic membrane, and there is no signal peptide indicated. This transporter protein is likely to be a membrane protein helping in transport of molecules in and out of the cell as the name suggests.

CONCLUSIONS

All the proposed gene products were the same as the automated gene annotation proposed in GENI-ACT.

Gene Locus Ksed	Geni-Act Product	Proposed Annotation
04830	hypothetical protein	hypothetical protein
04840	DNA-binding protein, excisionase family	DNA-binding protein, excisionase family
04850	hypothetical protein	hypothetical protein
04870	cation diffusion facilitator family transporter	cation diffusion facilitator family transporter

Interestingly, although Ksed_04830 is predicted to be a hypothetical protein we can speculate that it is a membrane protein that requires further analyses to substantiate. We also learned how to use BLAST and two databases as a tool to study protein alignments and conservation.

REFERENCES

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

ACKNOWLEDGMENT

We would like to thank Drs. Rama Dey-Rao and Stephen T Koury for assistance with training and teaching of the project.

Supported by NSF ITEST Strategies Award Number 1311902