Annotation of the *Kytococcus sedentarius* Genome Locus Tags Ksed_16650, Ksed_16700 and Ksed_16780

Gunnar Hammonds, Thomas Neumaier, Sam Reeder and Dr. Betsy Vinton
The Harley School, Brighton, New York and the Western New York Genetics in Research Partnership

Abstract

A group of 3 genes from the microorganism *Kytococcus sedentarius* (Ksed_16650, Ksed_16660, and Ksed_16780) were annotated using the collaborative genome annotation website GENI-ACT. The Genbank proposed gene productname for each gene was assessed 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, enzymatic function, presence or absence of gene duplication and degradation, the possibility of horizontal gene transfer, and the production of an RNA product. 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 r database.

Introduction

Kytococcus sedentarius (Ksed) is a gram positive, non-motal, stricly aerobic prokaryote that has the ability to grow only when several amino acids are present. It is a pathogen that is normally commensal but often takes advantage of individuals with weakened immune systems to cause several diseases including pitted keratolysis – a bacterial infection of the feet.

In general, it is important to understand how to study organisms such as *Kytococcus sedentarius* in silica because our current understanding and knowledge of prokaryotic organisms is extremely limited. One recent article suggests that less than 1% of all microbial organisms have been discovered, let alone researched, and therefore any information and research tecniques gathered in a study is valuable and could be applied to other investigations. The study of prokaryotes is not only important to pathogenic research but for that of humans as well. Gathering information on the organisms that are the root of all living things can shed a lot of light in dark corners.

With the entire DNA sequence of Ksed annotated, and it being part of the scarcely populated genus within actinobacterial family *Dermacoccaceae,* it is a valuable prokaryote to investigate.

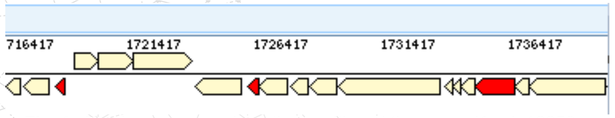

Fig. 1. The locus tags and relative position of the genes Ksed_16650, Ksed_16700, and Ksed_16780.

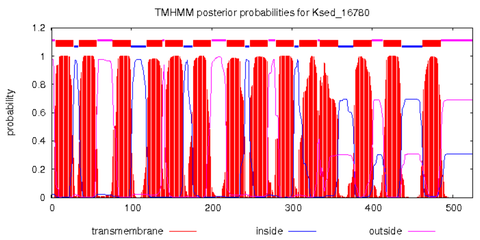

Fig. 2. TMHMM results of Ksed_16790 which show 15 transmembrane helices

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- 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?
Module 7- Gene Duplication/Gene Degradation	ParaloModulg, 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?
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

Ksed_16780:

Ksed_16780 is a 1578-nucleotide sequence located on the complementary strand on coordinates 1735112 to 1736689. BLAST results showed the closest similarities to the Na+/H+ antiporter subunit D in Bacillus pseudofirmus and Bacillus subtilis, with e-values of 2e-95 and 7e-79 respectively, and many other genuses have genes with sufficiently low e-values to be considered homologous. 14 transmembrane helices were detected by TMHMM, PSORTb and PHOBIUS, and PSORTb predicted location in the cytoplasmic membrane with a score of 10.00. However SignalP and PSORTb contradicted each other, with SignalP predicting a signal peptide with probability 0.510, but PSORTb predicting no signal peptide. The open reading frame predicted by the gene caller is most likely correct, and a Shine-Dalgarno sequence AGGAGG is located starting 12 nucleotides upstream of the start codon. Pfam classified the gene in the clan CL0425, and the rough 3D structure is given in PDB under the code 3RKO. As the protein encoded is not an enzyme, no enzymatic function data was collected. There was 28% identity to Na+/H+ antiporter subunit A, and there is no evidence that it is a pseudogene. There is no evidence that the gene was gained by horizontal transfer; phylogenetic analysis clusters Kytococcus with other members of the family Actinobacteria based on the gene, and these similarities are accounted for by vertical, not horizontal transfer. No non-coding RNA sequences were found by Rfam. Overall, the gene caller's prediction is backed very well by these results.

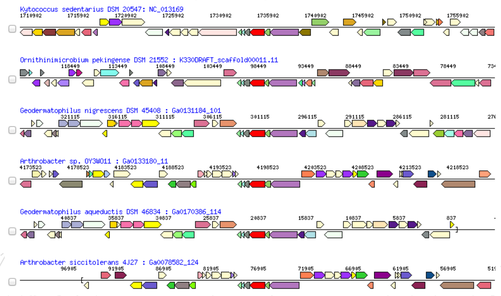

Fig. 3. Orthologs of Ksed_16780 and its neighborhood in related organisms

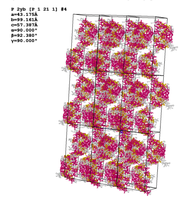

Fig. 4
3D crystal packing structure of UspA stress protein encoded by Ksed_16700

Fig. 5
3D structure of Mraz transcriptional regulator protein., encoded by Ksed_16650.

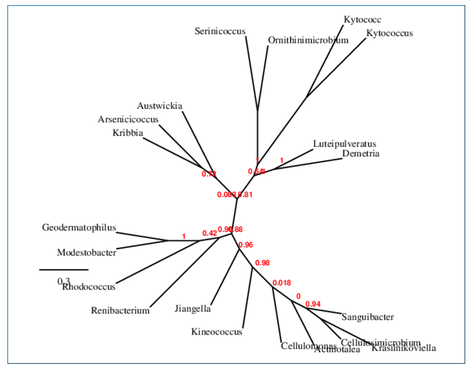

Fig. 6
Unrooted phylogenetic tree created by comparing orthologs of Ksed_16780 from related organisms. The clustering of K. sedentarius with other closely related organisms indicates that the similarity of the orthologs is due to vertical transfer, not horizontal.

Ksed_16650:

This locus, Ksed_16650, contains DNA coordinates 1718606 to 1719037 on the forward strand according to the GENI-ACT gene page. The nucleotide sequence is 432 units long, and the amino acid sequence is 143 units long. GENI-ACT proposed the product of this gene as an Mraz transcriptional regulator protein. Based on the top BLAST hit of the NR database, this proposed product is accurate with an e-value of 2e-69 from Frankia alni. The second hit had the same results but was present in the Thermobifida fusca organism. A search using the CDD database found one COG, COG2001, an Mraz DNA-binding transcriptional regulator and inhibitor of RsmH methyltransferase activity. This means the gene product is used in one of the DNA to RNA regulation processes, and also represses the ribosomal methylation of cytidine in RNA. Pfam also suggests that it could act as an antitoxin in the cell cycle. These results further expand the GENI-ACT proposal. TMHMM results indicated that there are no transmembrane helices, and SignalP predicted no signal peptides. Phobius also confirmed the absence of these structures. PSORT-B predicted that it is cytoplasmic (a score of 7.50 compared to the extracellular score of 0.73). The lack of transmembrane helices and signal peptides rules out the possibility of a trans membrane or secreted protein, so the gene product is most likely located in the cytoplasm. An IMG search was used to verify that the amino acid sequence originally proposed was correct. The results found that there was a Shine-Dalgarno sequence 10 units upstream of the proposed start codon. There were no other Shine-Dalgarno sequences that were near any other possible start codons, which verifies the accuracy of the original amino acid sequence.

Ksed_16700:

This gene is located from coordinates 1726210-1726620 in the protein. The gene has a total of 418 gene bases. The there are a total of 140 amino acids which are transcribed from the gene. Geni-act proposed the product of this gene to be a universal stress protein UspA-like protein. This was the supported by the top BLAST results (universal stress protein UspA) According to Pfam UspA is a structural protein which changes its function when the cell is exposed to stressful environments, giving the cell greater chances of survival. TIGRFAM however, proposed that the protein was phenylacetic acid degradation protein. However the e-value, for this prediction was too low to be taken into consideration. TMHMM determined that this protein had zero transmembrane helices. And that the protein was most likely located in the cytoplasm. This was supported by the PSORT-B results. IMG Sequence Viewer predicted that there were no Shine-Dalgarno sites, and proposed the same DNA coordinates as the Geni-Act Database.

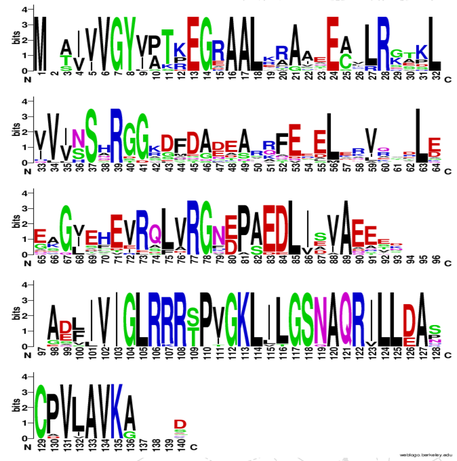

Fig. 7
WebLogo illustrating the conservation of Ksed_16700 and related genes.

Conclusion

The GENI-ACT proposed gene products 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.

Reference

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

Acknowledgments

Supported by an NSF Innovative Technology Experiences for Students and Teachers (ITEST) Award - 1311902

www.buffalo.edu