Annotation of the *Kytococcus sedentarius* Genome Locus Tag Ksed_16680, Ksed_16740 and Ksed_16860

Jacob LaDue, Audrey Scudder, Emma Fiorini and Betsy Vinton
The Harley School, Brighton, New York and The Western New York Genetics in Research Partnership

University at Buffalo

Abstract

A group of 3 genes from the microorganism *Kytococcus sedentarius* (Ksed_16680, Ksed_16740 and Ksed_16860) 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.

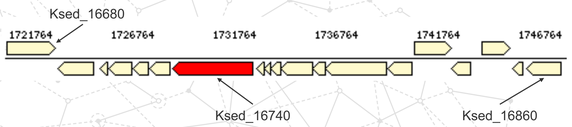

Figure 1. The locus tags and relative position of the genes under investigation in this research

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?
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_16740:

This locus tag refers to a gene made up of 3996 nucleotides oriented on the complementary strand from coordinates 1729765 to 1733760. This gene codes for 1331 amino acids. Using the Swissprot database in Genbank, two hits were found with e-values of 0 - Full=ATP-dependent RNA helicase HrpA and its homolog. These two genes yielded e-values of 0. From the distribution of BLAST hits, it can be observed that after the first two hits, which both have significantly higher overall scores, the alignment is truncated down to half of the length of the first two hits. From this, it can be inferred that the first two hits have a domain at the second half of their sequence that is significant. This domain, however, is a domain of unknown function. COG1643, HrpA-like RNA helicase [Translation, ribosomal structure and biogenesis], was found to be similar to Ksed_16740. Before moving to Weblogo, it could be seen that the sequence of Mobilicoccus had conserved residuals close to 150 amino acids before the alignment of the other 10 sequences. The weblogo made from the T-Coffee clustal alignment shows that the sequence is not well conserved from the amino end to around residue 175 based off of the gaps. The sequence is well conserved from about residue 175 to 1120 based off of the large relative heights of the letters and heights and widths of stacks.

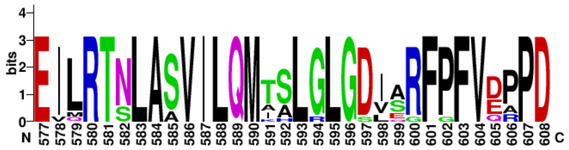

Figure 2 . A section of the Weblogo for Ksed_16740; from amino acid 577 to 608. This section of the sequence is included in the alignment of the OB_NTP_bind domain. These amino acids are very conserved which indicates some of the other organisms used in the T-Coffee multiple sequence alignements.

However, within this range, there are some small gaps with absence of amino acids or not well conserved amino acids (385-405, 630-645, 776-800). From residue 1120 to the carboxy end, there is reasonable conservation , however, they are not as consistent with or conserved as those from 175 to 1120. Output from the TMHMM, SignalP, PSORT-B, and Phobius databases are all consistent with the prodictions that the protein is located in the cytoplasm and tested negative for any Transmembrane Helices or signal peptides. The domains from the top hits of TIGRFAM are most regularly found in company of each other in the. DEAD-box helicases which can be found in most eukaryotes and prokaryotes. PDB resulted in a hit with a significantly low e-value, Crystal Structure of the PRP43P DEAH-box RNA Helicase in complex with ADP. There were no Shine Delgarno sequences.

Ksed_16680:

GENI-ACT predicted the product of this gene to be a transglutaminase-like enzyme, predicted cysteine protease. This was only partially supported by the top BLAST results, as most of the top hits were only hypothetical proteins. Pfam also predicted the gene to be in the transglutaminase-like superfamily. TIGRFAM and PDB searches returned no results with an E-value less than .1 . SignalP did not find any signal peptides, but TMHMM found 7 transmembrane helices. PSORT-B predicted that the protein would be located in the cell membrane as well. This does not support the BLAST prediction, since recent research has suggests that cysteine proteases are usually secreted. An alternate reading frame was investigated, but BLAST results had higher E-values, so the sequence was most likely called correctly by the gene caller.

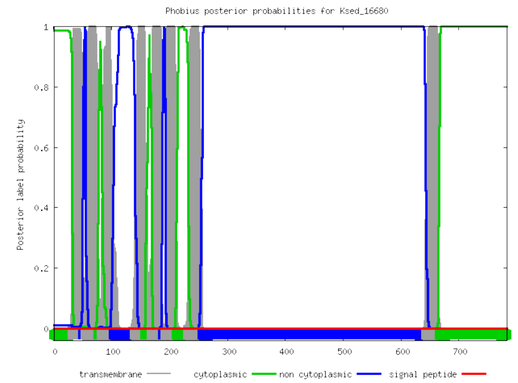

Figure 3. A Phobius output showing the presence of 7 transmembrane helices and no signal peptide.

Ksed_16860:

Ksed_16860 has a sequence length of 1707 base pairs and an amino acid length of 568 amino acids. After analyzing the gene with BLAST, the gene product name was confirmed to be Phosphoenolpyruvate-protein phosphotransferase (top hit). The e-value was found to be 7e-121. The protein contains many conserved domains in the middle of the sequence, but the homology is negligible at the amino and carboxy termini. In TMHMM, the results indicated that no transmembrane helixes were found, meaning that the proteins were not located on the cell membrane. SignalP scores indicate that no signal peptide is present and the peptide is likely not secreted. In PSORTb there was a large score (9.97) for the location "cytoplasmic", and extremely low scores for other locations. Since the microbe tested as gram-positive, the outer membrane and periplasmic scores do not apply. Phobius confirmed the TMHMM results, and no trans membrane helices were found. The findings of these four tools indicate that the protein is most likely located inside the cell. An Alternative Open Reading Frame analysis shows the presence of a Shine-Dalgarno sequence five amino acids upstream from the predicted start codon. No others were observed upstream.

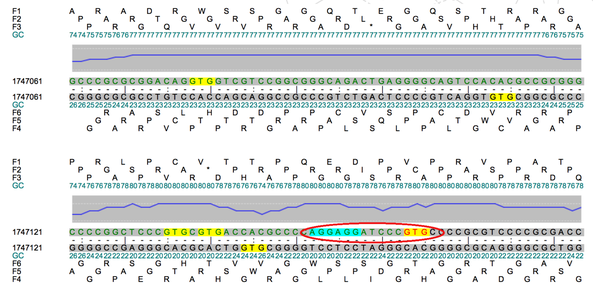

Figure 4. The graphical six frame translation of Ksed_16860 which shows a Shine Delgarno sequence.

Conclusion

The gene products for Ksed_16680, Ksed_16740 and Ksed_16860 were all confirmed by the first hit of BLAST. However, Ksed_16680's cellular localization data warrants further study

References

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