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

Annotation of the *Kytococcus sedentarius* Genome from DNA Coordinates 370637 to 376721

Sina Soltanieh[1], Robert Mentkowski[2], Niyat Berhe[3], Haben Berhe[3], Sangeeta Gokhale

[1] Williamsville East High School, [2] Williamsville North High School, [3] Amherst High School, (participating in the BEAM – Buffalo Engineering Awareness for Minorities Program)

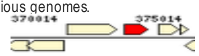
Abstract

Bioinformatics is a growing field of biology, helping scientists around the globe to realize, understand, and apply functioning of genes, regulation of cells, drug target selections, drug design, and numerous diseases in real world situations. Bioinformatics tools assist in comparing and interpreting laboratory generated scientific research data, and help understand evolutionary aspects of molecular biology and to catalogue and explain important biological processes integrated within systems biology. The objective of analyzing the four genes with consecutive loci in the *Kytococcus sedentarius* bacteria was exactly so, to identify and interpret information about proteins, to catalog correct information about the function of these proteins and their effect in biological processes, as identified in the proposed annotations. To check the validity and reliability of these annotations, we followed a nine-module method to create a final annotation from the evidence we obtained, from the basic information of the gene, which includes the nucleotide and amino acid sequences, to the RNA hits and alignment, eventually drawing a conclusion in our final annotations. The proposed gene products matched the results obtained with GENI-ACT research, with evidence to support all four of the final annotations, suggesting the genes were correctly annotated in the database.

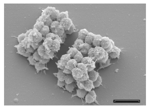
Introduction

Kytococcus sedentarius is a gram positive, coccus shape bacteria, formerly classified as a micrococcus. It is found in irregular clusters, tetrads, in groups of eight. It is a strict aerobe, testing catalase positive, as well as oxidase positive, and is chemoheterotrophic, found in marine environments with an optimal temperature of 25-27 degrees Celsius, and functioning at an optimal NaCl concentration of less than 10%. *K. sedentarius* is non-motile, non-encapsulated, and non-endospore, and it plays a key role in polyketide antibiotics in commercial use. Genomic sequencing was performed on the gene sequence of *K. sedentarius* to verify the annotations in the Genebank database using GENI-ACT, with assistance from programs such as BLAST, T-Coffee, and Weblogo. The goal and purpose of the project was the verification and evaluation on the gene and its annotation.

According to Good, et al, (2013), crowd sourcing increases the scientists' ability in all areas, in this case, to determine protein structure and function by comparing computer generated output to human input within a database. This information sourcing has allowed for scientists to increase communication and gather more information and evidence, for example, about the *K. sedentarius* gene. Sims et al. (2009), identifies why *K. sedentarius* is of particular interest to scientists. It is a natural source of oligoketide antibiotics, monensin A and monensinB specifically (Sims et al., 2009). Hemorrhagic pneumonia and pitted keratolysis are some diseases which are caused by this bacteria. (Sims et al., 2009). *K. sedentarius* is a member of the family Dermacoccaceae within the actinobacterial suborder Micrococcineae. Organisms in this phylogeny have yet to have been thoroughly studied utilizing bioinformatics (Sims et al., 2009). Therefore, the use of bioinformatics tools would greatly help scientists acquire more information about various genomes.



Gene Neighborhood in *K. sedentarius*. The highlighted gene is that of the locus tag Ksed_03860.



Methods and Materials

The gene annotations were done using the GENI-ACT (http://geni-act.org/) website with module instructions, and various other programs, such as BLAST and T-Coffee to assist in these annotations. Information about each of the nine individual modules are summarized in the table below. All modules were completed.

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

<u>*Kytococcus sedentarius*03840</u>: Trehalose-6-phosphate synthase are the names of the top blasts his. The COG hit was Trehalose-6-phosphate synthase. There was no second COG hit. This genes is located in the cytoplasm. Alpha-trehalose is the TIGRFAM name. Phosphate Glycosyltransferase family 20 was the PFAM name . There was no second PFAM hit. This evidence says that the name of the enzyme is trehalose-6-phosphate synthase.

<u>*Kytococcus sedentarius* 03850</u>: The top BLAST hits were all hypothetical proteins. COG hit #1 is "tRNAvngjv,gj,gj, A-37 threonylcarbamoyl transferase component Bud32 [Translation, ribosomal structure and biogenesis]" COG hit #2 is "Serine/threonine protein kinase [Signal transduction mechanisms]" The TIGRFAM name was associated with the protein kinase. The PFAM was found as the protein kinase family.

No alternate reading frame found and DNA coordinates were corresponding to the Gene Caller. Based on the results, the gene codes for serine /threonine protein kinase

<u>*Kytococcus sedentarius* 03860</u>: The need for a conclusion has arisen that states that the information and annotation of the gene under investigation as proposed by the computer, is indeed, correct. The top blast hits included transglycosylases, but the best results were resuscitation promoting factors, Rpf1, and RpfB. The blast search also indicates a transglycosylase with a lysozyme-like superfamily. The Pfam hit results were Transglycosylas and Trypan_PARP. The Transglycosylas is supported by various other sources, such as a blast conserved domain hit, and by the horizontal gene transfer phylogenetic tree evidence, which brings up the evidence for possible gene transfer through evolution of a common ancestor. The accepted enzyme name was peptidoglycan glycosyltransferase. This again supports the evidence for a transglycosylase-like protein/enzyme, as transglycosylases are a class of Glycoside Hydrolase enzymes which catalyze the transformation of one glycoside to another.

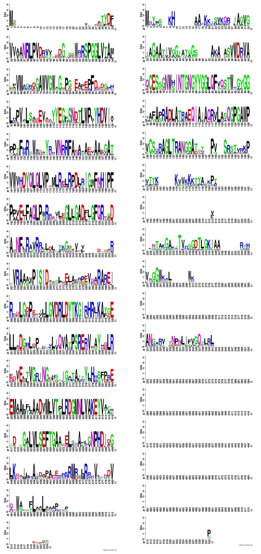


Figure 1- WebLogo of Ksed_03840 (left) and Ksed_03860 (right). The beginning of the Weblogo signifies the 5' end of the sequence and the end signifies the 3' end of the sequence. Relative stack height indicates percent conservation. Relative sizes of the letters indicate the frequency of the amino acid in the alignment. Ksed_03840 portrays a relatively high percent conservation throughout the sequence alignment. In contrast, Ksed_03860 indicates a well conserved sequence at the 5' end, but a poorly conserved sequence at the 3' end.

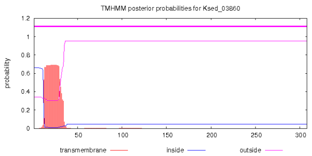


Figure 2- These results of using TMHMM for the Ksed_03860 gene indicate no transmembrane helices for this protein The program may incorrectly mark signal peptides as transmembrane helices, as shown by the probability at the beginning of the sequence.

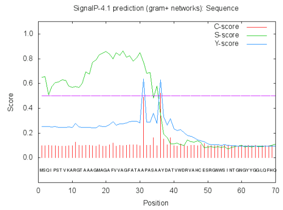


Figure 3- The SignalP results show that a Signal Peptide is present in the Ksed_03870 gene. Further tools, namely LipoP, were then used to further elaborate, indicating thee type of SP and its cleavage sites.

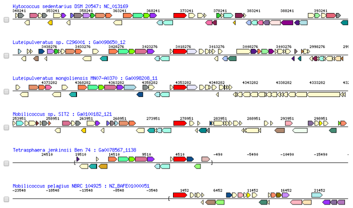

Kytococcus sedentarius DSM 20547: NC_013169
Luteipulveratus sp. C296001 : Ga0098650_12
Luteipulveratus mongoliensis MN07-A0370 : Ga0098280_11
Mobilicoccus sp. SIT2 : Ga0104182_121
Tetrasphaera jenkinsii Ben 74 : Ga0078567_1138
Mobilicoccus pelagius NBRC 104925 : NZ_BAFE01000051

Figure IV- The gene neighborhood of genes near Ksed_03850 is used to try to find evidence for horizontal gene transfer. If it looks similar in related species, this evidence would support vertical transfer; however, if it looks similar to unrelated species, this would support horizontal gene transfer.

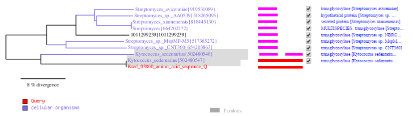

Streptomyces_avicenniae{919531089}
Streptomyces_sp._AA0539{518265095}
Streptomyces_xiamenensis{818445130}
Streptomyces{664202272}
1011299239{1011299239}
Streptomyces_sp._MspMP-M5{517365272}
Streptomyces_sp._CNT360{654203863}
Kytococcus_sedentarius{502480548}
Kytococcus_sedentarius{502480547}
Ksed_03860_amino_acid_sequence_Q
8 % divergence
Query
cellular organisms
Paralogs
transglycosylase [Streptomyces avicenniae]
hypothetical protein [Streptomyces sp. ...
secreted protein [Streptomyces xiamenensis]
MULTISPECIES: transglycosylase [Strepto...
transglycosylase [Streptomyces sp. NBRC...
transglycosylase [Streptomyces sp. MspM...
transglycosylase [Streptomyces sp. CNT360]
transglycosylase [Kytococcus sedentariu...
transglycosylase [Kytococcus sedentariu...

Figure V- Phylogenic Tree of Ksed_03860

<u>*Kytococcus sedentarius* 03870</u>*:* The top and second BLAST hits are suggestive of a transglycosylase. The names of Pfam hits are (PF06737) PFAM: Bacterial SH3 domain; Transglycosylase-like domain; this is also suggestive of a transglycosylase-like protein. One transmembrane helix was found and TIGRfam results are suggestive of a transglycosylase. Thus the protein is transglycosylase-like.

Conclusion

The GENI-ACT Gene Products were approximately similar and match the proposed annotations of the *Kytococcus sedentarius* genes from the loci 03840-03870. Therefore, the genes seem to be correctly annotated by the Genome Online Database, Genebank.

Gene Locus	GENI-ACT Gene Products	Proposed Annotation
Ksed_03840	Trehalose 6 Phosphatase Synthase	Trehalose 6 Phosphate Synthase
Ksed_03850	Protein Kinase	Serine / Threonine Protein Kinase
Ksed_03860	Transglycosylase-like protein	Transglycosylase-like protein
Ksed_03870	Transglycosylase family protein	Transglycosylase family protein

References

Good, Benjamin M., and Andrew I. Su. (2013) "Crowdsourcing for bioinformatics." Bioinformatics. btt333.

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
Special acknowledgements to Dr. Rama Dey-Rao and Dr. Stephen Koury, Research Associate Professor; WNYGiRP