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

Annotation of the *Kytococcus sedentarius* Genome from DNA Coordinates 05530 to 05550

Patrick Marrone*, Scott Rizzo*, Jacquelyn Weber* and Samantha Evans

Iroquois High School and The Western New York Genetics in Research Partnership *Indicates equal contribution

Abstract

A group of 3 consecutive genes from the microorganism *Kytococcus sedentarius* (Ksed_05530 – Ksed_05550) were annotated using the collaborative genome annotation website GENI-ACT. The Genbank proposed gene product name 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, and the possibility of horizontal gene transfer. 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. With using the phylogenetic tree we can see what organisms are related to *Kytococcus sedentarius*. Ksed_05530 has a close relative to a melon *Cucumis melo*. Ksed_05540 location according to SignalP and PSORTb both suggests that the protein is going to be extracellular or part of the cell wall since there was a signal peptide.

Introduction

There is a need for manual gene annotation for *Kytococcus sedentarius.* The computer had predicted what the gene products would be but there is a need for humans to check the computer. The computer can make errors and it is essential for humans to make sure the data is accurate. Module 4, the alternative open reading frame is where the most computer errors made. Humans need to check the computers work for all possible errors that may occur.

Previously *Kytococcus sedentarius* was only annotated by a computer, but now with Western New York Genetics Partnership, students are performing the strenuous task of double checking the computer. Our neighborhood Ksed_05530-05550 was correctly predicted by the computer, but there are many more genes to be double checked by students to check for possible errors.

Through our group's research, we have confirmed that our gene neighborhood does indeed match what the computer predicted and much more including information such as the locations of our proteins, their purpose, and their close relatives. All of this is important to science not just because a computer can make an error, but for students to learn the scientific process of annotating a gene.

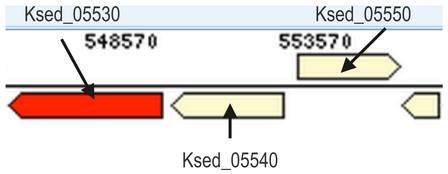


Figure I – Gene Neighborhood for Annotated Genes: Ksed_05530, Ksed_05540 and Ksed_05550

Methods and Materials

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- 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 8- Evidence for Horizontal Gene Transfer	Phylogenetic Tree	Has my gene co-evolved with other genes in the genome?

Results

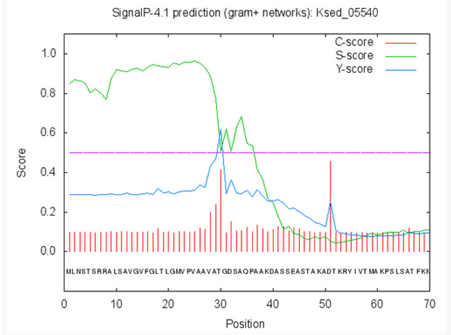
Ksed_05530- Ksed_05530, a cell wall binding protein located in the extracellular region of *Kytococcus sedentarius* is actually closely related to the Crystal structure of Cucumisin (Fig. II) by an E value of $1.3855e^{-65}$. Cucumisin is a glycoprotein located in *Cucumis melo L.* (Fig. III) which interestingly enough is a melon fruit native to the Middle East called Muskmelon.(Murayama, K., et. al, 2012)

Figure II – Crystal Structure of Cucumisin

Figure III – Muskmelon

Ksed_05540: There is a good probability that the protein has a signal peptide as predicted by the D score, 0.728. which is above the probability cut-off, 0.45. The SignalP program has predicted that the cleavage site is between position 29 and 30 which indicated that the protein is going to be in the outside of the cell membrane. (see Figure V)

The program PSORTb showed that protein was split between either being extracellular with a score of (4.71) or being in the cell wall with a score of (5.12). These results show that it is most likely going to be on the outside of the cell yet PSORTb wasn't able to pinpoint the exact cellular localization. The program SignalP also thought that the protein had a signal peptide which indicates the protein is excreted outside of the cell, agreeing with the results found with the program PSORTb.




```
# Measure  Position  Value   Cutoff   signal peptide?
  max. C    51       0.458
  max. Y    30       0.620
  max. S    25       0.965
  mean S     1-29    0.897
  D          1-29    0.728   0.450   YES
Name=Ksed_05540 SP='YES' Cleavage site between pos. 29 and 30: AVA-TG D=0.728 D-cutoff=0.450 Networks=SignalP-TM
```

Figure IV – SignalP-4.1 Prediction of Signal Peptide

PSORTb Results

```
SeqID: Ksed_05540 animo acid
  Analysis Report:
    CMSVM+              Unknown
    CWSVM+              Unknown
    CytoSVM+            Unknown
    ECSVM+              Extracellular
    ModHMM+             Unknown
    Motif+              Unknown
    Profile+            Unknown
    SCL-BLAST+          CytoplasmicMembrane, Cellwall
    SCL-BLASTe+         Unknown
    Signal+             Non-Cytoplasmic
  Localization Scores:
    Cytoplasmic         0.00
    CytoplasmicMembrane 0.02
    Cellwall            5.27
    Extracellular       4.71
  Final Prediction:
    Unknown (This protein may have multiple localization sites.)
```

Figure V – PSORTb Cellular Localization Predictions

Ksed_ 05550- This image is a phylogenetic tree of Ksed_05550.(fig VI) A phylogenetic tree is a diagram that shows the evolutionary relationships between species. The lengths of the lines on the image show how close a species is related to another species, and the numbers are a statistical estimate of significance of the branching . From this image we can see that Ksed_05550 is most closely related to *Streptomyces*.

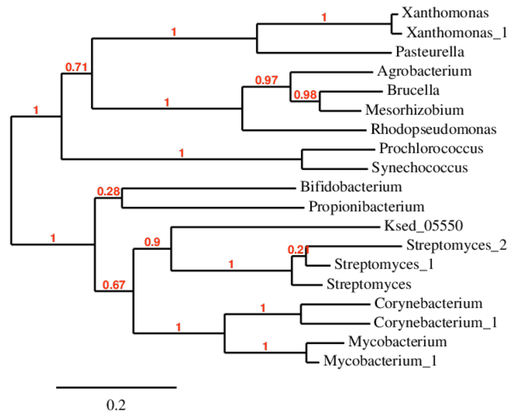


Figure VI– Phylogenetic Tree of Ksed_05550

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.

Gene Locus	Geni-Act Gene Products	Proposed Annotation
05530	cell wall-binding protein	cell wall-binding protein
05540	cell wall-binding protein	cell wall-binding protein
05550	ATPase with chaperone activity, ATP-bindingsubunit	ATPase with chaperone activity, ATP-bindingsubunit

References

Murayama, K., Kato-Murayama, M., Hosaka, T., Sotokawauchi, A., Yokoyama, S., Arima, K., & Shirouzu, M. (2012). Crystal Structure of Cucumisin, a Subtilisin-Like Endoprotease from Cucumis melo L. Journal of Molecular Biology, 386-396. doi:10.1016/j.jmb.2012.07.013

Acknowledgments

Special thanks to Dr. Stephen Koury and Dr .Rama Dey-Rao for their assistance with this project. This work was supported by the National Science Foundation ITEST Strategies Award Number 1311902.