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

Annotation of the *Kytococcus sedentarius* Genome from DNA Coordinates 889528 to 907866

Za Tei Thang, Flor Ree Na, Sabina Manuel, Gabriel Ngandu, Halimah Hassoon and Jeffery Besinger

Newcomer Academy at Lafayette High School and the Western New York Genetics in Research Partnership

Abstract

A group of consecutive 5 genes from the microorganism *Kytococcus sedentarius* (Ksed_08680 – Ksed_08770) were annotated using the collaborative genome annotation website GENI-ACT. The objective of this study was to introduce students to techniques involved with determining gene function by checking that the computer correctly annotated the *Kytococcus sedentarius* genome. The gene product name proposed by Genbank for each gene was assessed using various search tools (BLAST, Pfam, TIGRfam) against collaborative online databases (SwissProt and NR). The Genbank proposed gene product name did not differ significantly from the proposed gene annotation for each of the genes in the group.

Introduction

Since the conclusion of the Human Genome Project, technological improvements in DNA sequencing resulted in faster/cheaper sequencing. This has led to the establishment of large, online libraries of DNA sequences which are continually growing with new sequences. These databases can be used to look for patterns in genes. Sequence similarities between genes suggest similar structural shapes and therefore similar function. The wealth of scientific discoveries hidden in these databases is unprecedented. Increasing the number of people who understand these databases and are capable using them is a logical next step in making best use of this information. We attempted to familiarize ourselves with online genomic databases and associated search tools by manually annotating the function of 5 genes from the microbe *Kytococcus sedentarius*.

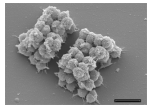

Figure 1 Electron microscope image of *Kytococcus sedentarius*. Photo credit: Dr. Manfred Rohde at Helmholtz Centre for Infection Research, Braunschweig, Scale = 2 µm.

Kytococcus sedentarius is a gram positive bacterium that lives in the ocean. It has been found to require oxygen (obligate aerobe) and several essential amino acids in order to live. It has medical relevance because it produces an antibiotic, and is an opportunistic pathogen causing pneumonia, heart valve infections, and pitted keratolysis in the soles of feet. Its entire genome was sequenced from an ocean water sample collected near San Diego in 1944, and published in 2009. The genome was found to be made up of 2,785,024 base pairs, (71% being G-C) and consisting of 2639 protein coding genes (Sims, D et al, 2009). All genes were automatically identified by a gene caller program and protein products were generated based on automated analysis. Our goal was to manually follow the automated steps so that we can understand how gene functions are determined, and to double check the process, looking for potential errors.

We were assigned 5 genes starting with base pair 889528 to base pair 907866. Our genes were identified as 08790, 08810, 08840, 08850, and 08880. Automated proposed annotations for each gene was listed in the corresponding Geni-Act online notebooks. Each student was responsible for manually annotating their assigned gene by following the modules in their Geni-Act notebook.

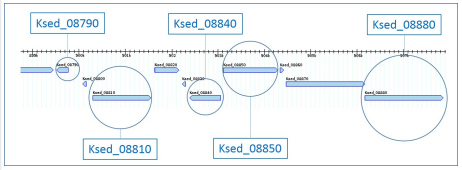

Figure 2 Gene neighborhood of *Kytococcus sedentarius* from nucleotide 889528 to 907866. Annotated genes are indicated by circles, sizes are relative to length of gene.

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?

Results

Kytococcus sedentarius 08810

Initial amino acid sequence BLAST search in the NR database indicated that the Ksed_08810 had a high degree of similarity (3e-104) with ribonuclease BN from *Jiangella gansuensis*. However, the additional search tools Pfam and TIGRfam indicate that the characterization of this gene sequence as producing a ribonudease is incorrect. With no clearly identifiable protein product, and evidence contradicting the BLAST match, we used TMHMM to identify trans-membrane helices which support the original annotation.

Figure 3 Pfam results showing that original prediction of ribonudease is incorrect

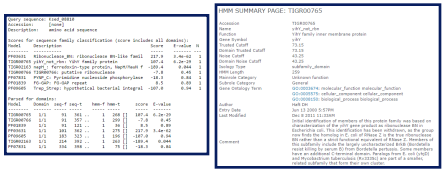

Figure 4 and 5 TIGRfam results and summary page for TIGR00765 with additional information regarding the misclassification as ribonudease.

Kytococcus sedentarius 08840

Amino acid sequence BLAST search in the NR database indicated that the Ksed_08840 had a high degree of similarity (9e-93) with oligoribonudease from *Konellia sinensis* as well as several other species which supports the original annotation.

Figure 6 BLAST NR search showing multiple species matches for oligoribonudease

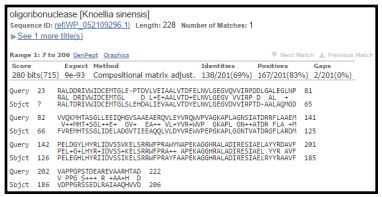

Figure 7 alignment of Ksed_08840 with an oligoribonudease showing conservation of amino acid sequence (69% of the sequence is conserved)

Kytococcus sedentarius 08850

Ksed_08850 was characterized as a predicted membrane protein. TMHMM confirmed that this gene contains 8 regions of the amino acid sequence that would result in the chemical composition of possible trans-membrane helices.

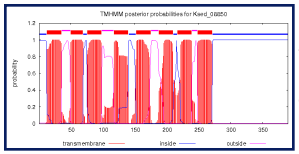

Figure 8 Results of TMHMM for Ksed_08850 indicting 8 possible trans-membrane helices

Kytococcus sedentarius 08880

While BLAST, COG, Pfam and TIGRfam all confirmed the original prediction of a GTPase, we did a check of possible alternate reading frames to confirm that the gene caller program correctly identified the start and stop codons. The proposed start codon (red nucleotides in Figure 9) was not preceded by an upstream Shine-Dalgarno sequence. We identified a valine start codon which is preceded by a Shine-Dalgarno sequence 66 base pairs upstream from the proposed start codon in frame 2. Reading frame 2 from that start codon terminated after 22 amino acids at a stop codon. We determined that 22 amino acids was too short to be a protein and therefore deferred to the original proposed start.

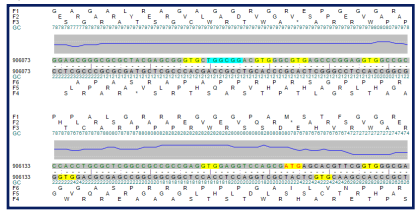

Figure 9 ORF search showing the proposed start codon in red and an additional 99 base pairs that precede it. All possible starts are yellow, and Shine-Dalgarno sequences are blue.

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.

Locusaa	Geni-Act Gene Product	Proposed Annotation
08790	transcription factor WhiB	transcription factor WhiB
08810	predicted membrane protein	predicted membrane protein
08840	oligoribonuclease (3'->5' exoribonuclease)	oligoribonuclease/exoribonuclrease
08850	predicted membrane protein	predicted membrane protein
08880	predicted GTPase	GTPase

References

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