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

Annotation of the *Kytococcus sedentarius* Genome from DNA Coordinates 886597 to 899459

Amina Adillahi, Amina Ali, Filmon Asmelash, Gospel Djongara, Radhika Chapagain and Jeffery Besinger

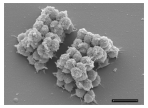
Newcomer Academy at Lafayette High School and the Western New York Genetics in Research Partnership

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

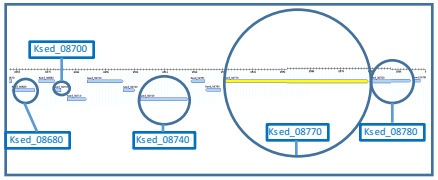
Since the conclusion of the Human Genome Project, technological improvements in DNA sequencing resulted in faster/cheaper sequencing. This has led to the establishment of large, online libraries of DNA sequences which are continually growing with new sequences. These databases can be used to look for 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*.



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, Det 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 886597 to base pair 899459. Our genes were identified as 08680, 08700, 08740, 08780, 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.



Gene neighborhood of *Kytococcus sedentarius* from nucleotide 886597 to 899459. 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 08680

Ksed_08680 was characterized as containing a LysM domain, which was confirmed by Pfam. TMHMM also revealed that this gene contains 3 regions of the amino acid sequence that would result in the chemical composition of possible trans-membrane helices.

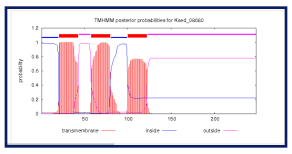


Figure 3 Results of TMHMM for Ksed_08680 indicting 3 possible trans-membrane helices

Kytococcus sedentarius 08700

Ksed_08700 was characterized as containing a DNA-binding domain, which was confirmed by BLAST, and TIGRfam. Pfam match characterized the sequence as containing a Helix-Turn-Helix domain. We did some additional research and determined that the Helix-Turn-Helix domain is a protein structure that binds to DNA. Despite the fact that the domains had different names, their functions are similar, therefore Pfam results could be considered supporting the original annotation.

Figure 4 Results of Pfam for Ksed_08700 showing a significant match to "Helix-Turn-Helix domain"

Figure 5 shows an example of a "Helix-Turn-Helix domain" binding to DNA

Kytococcus sedentarius 08740

Ksed_08740 was characterized as containing a delta-1-pyrroline-5-carboxylate dehydrogenase, group 1 which was confirmed by BLAST, Pfam, TIGRfam and PDB. BLAST results show e values as low as 0.0 showing an extremely high amount of sequence similarities.

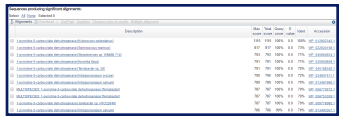


Figure 6 Results of BLAST for Ksed_08740 showing e values of 0.0

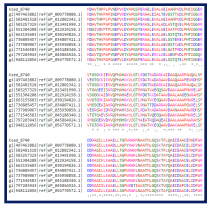


Figure 7 shows a protion of T-Coffee alignment for Ksed_08740 and 10 additional 1-pyrroline-5 carboxylate dehydrogenases with a high amount of conservation.

Kytococcus sedentarius 08770

Ksed_08770 was characterized as NAD-Glutamate Dehydrogenase which was confirmed by BLAST, and Pfam. TIGRfam indicated that the sequence was likely to have a hydrophobic domain, crystal structure suggests that glutamate dehydrogenase has a hydrophobic pocket which confirms the original annotation.

Figure 8 Results of Pfam for Ksed_08770 showing e values of 0.0

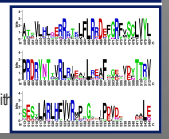


Figure 9 shows a protion of weblogo and 10 additional NAD-glutamate dehydrogenases with a high amount of conservation.

Kytococcus sedentarius 08780

Ksed_08780 was characterized as an ATPase, which was confirmed by BLAST, COG Pfam and TIGRfam. A search of the Protein Database yielded a significant match with an ATPase domain of histadine kinase.

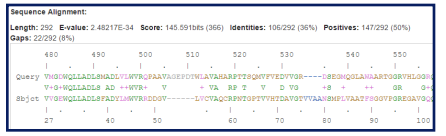
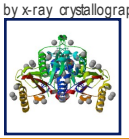


Figure 10 Results of PBD alignment for Ksed_08780 and the ATPase domain of histadine kinase. A portion of the full alignment is shown with a high degree of conservation at the start of the sequence.

Figure 11 is the 3 dimentional structure of ATPase domain of histadine kinase as determined by x-ray crystallography and described in the PDB article.



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. TMHMM provided addition information regarding cellular localization of Ksed_8680 and Ksed_08770

Locus	Geni-Act Gene Product	Proposed Annotation
Ksed_08680	LysM domain-containing protein	LysM domain-containing membrane protein
Ksed_08700	DNA-binding protein, excisionase family	DNA-binding protein
Ksed_08740	delta-1-pyrroline-5-carboxylate dehydrogenase,group 1	1-pyrroline-5-carboxylate dehydrogenase
Ksed_08770	NAD-specific glutamate dehydrogenase	NAD-specific glutamate dehydrogenase membrane protein
Ksed_08780	signal transduction histidine kinase	signal transduction histadine kinase

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