

Annotation of the *Kytococcus sedentarius* Genome from DNA Coordinates 567417 to 571832

Abbey Scheffler, Anna Marburger, Sydney Hensel, Fasha Antonio and Amanda Jasper

Holland Central School District and the Western New York Genetics in Research Partnership



Abstract

A group of consecutive 4 genes from the microorganism *Kytococcus sedentarius* (Ksed_5670 – Ksed_05700) were annotated using the collaborative genome annotation website GENI-ACT. The objective of this study was to prove that the computer correctly annotated the *Kytococcus sedentarius* genome. 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, and potential alternative open reading frames. By using crowd-sourcing databases, the process of annotation, the Genbank proposed gene product name did not differ significantly from the proposed gene annotation for each of the genes in the group. One noteworthy occurrence, was the BLAST results of Gene Locus tag 05700, which showed a transcriptional regulator, but all other databases indicated the proposed gene product.

Introduction

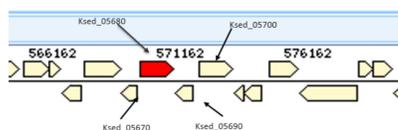
Kytococcus sedentarius is a marine dwelling Gram positive bacterium in the genus *Kytococcus*. It is known for the production of polyketide antibiotics as well as for its role as an opportunistic pathogen. This Gram positive coccus is found in tetrads, irregular clusters, and cubical packets of eight. It is catalase positive, oxidase positive, and exhibits strictly aerobic metabolism. Optimum growth temperature is 25-37°C. It is primarily isolated from human skin, and plays no known role in disease. It was once considered a species of the genus *Micrococcus*.

Genome sequencing is the collecting and finding of the order of nucleotides in a DNA sequence. Genome sequencing is often known as “decoding.” The gene sequence was taken from a Genome Online Database called GenBank. In Genebank, there are over 61,582 readings involving *Kytococcus sedentarius*. From the data base, computers decode the gene sequences and classify each gene within the DNA helix. The DNA strand was separated into bases of adenine (A), guanine (G), thymine (T), and cytosine (C). The bases, found by the computer, convert the DNA sequence into groups called amino acids.

Genomes of *Kytococcus sedentarius* were placed in the program Geni-Act for groups to confirm the computers results. They compared the results from the GenBank to other amino acid sequences found in other common organisms and bacterium. Types of program included Blast, CCD, T-Coffee and WebLogo. The goal of this project was to verify the data given the genome data base.

According to Good, et al, (2013) the use of crowd sourcing has greatly increased the ability of scientists within various scientific domains. Within this project, crowdsourcing was used to determine protein structure and function by comparing computer generated results to human based input. There are pros and cons to using this type of information sourcing, but it seems to allow for more collaboration and expedited evidence gathering.

Figure I – Locus tags and coordinates for *Kytococcus sedentarius* genes that were manually annotated.



Methods and Materials

Modules 1-5 of the GENI-ACT (<http://www.geni-act.org/>) were used to complete *Kytococcus sedentarius* genome annotation. All 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 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

Kytococcus sedentarius 05670:

The initial proposed product of this gene by GENI-ACT was a Predicted Permease. This gene product proposal was supported by the top BLAST hits for the amino acid sequence, the presence of well-curated protein functional domains within the amino acid sequence, the transmembrane topography of the amino acid sequence, and the cellular location of the amino acid sequence. As such, the proposed annotation is a permease.

Kytococcus sedentarius 05680:

The initial proposed product of these genes by GENI-ACT was an Acetyl-CoA Carboxylase Beta Subunit. This gene product proposal was supported by the top BLAST hits for the amino acid sequences, the presence of well-curated functional domains within the amino acid sequences and the cellular location data of the amino acid sequence,.

Kytococcus sedentarius 05690:

The initial proposed product of these genes by GENI-ACT was a Transcription Regulator. This gene product proposal was supported by the top BLAST hits for the amino acid sequences, the presence of well-curated functional domains within the amino acid sequences and the cellular location data of the amino acid sequences.

Kytococcus sedentarius 05700: The initial proposed product of this gene by GENI-ACT was a Na⁺/H⁺ antiporter NhaD-like permease. This gene product proposal was not supported by the top BLAST hits for the amino acid sequence in the nr database, rather, the top hits indicated a Transcriptional Regulator. All other modules indicate that the protein was a Na⁺/H⁺ antiporter NhaD-like permease. It thus appears that a sequence from an earlier genome sequencing project was incorrectly annotated as a transcriptional regulator in the nr database, resulting in the erroneous BLAST results (Brad Goodner, personal communication).

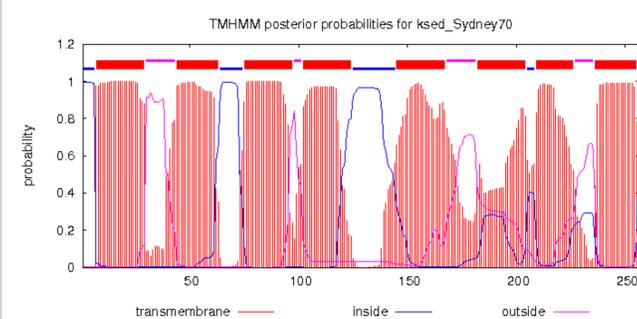


Figure I – *Kytococcus sedentarius* 05670 shows 8 transmembrane helices which supports the proposed annotation of a permease protein.



Figure II – *Kytococcus sedentarius* 05700 was originally called as a Antiporter, but Non-Redundant BLAST results show 05700 being called as a Transcriptional Regulator with good scores and low e values.

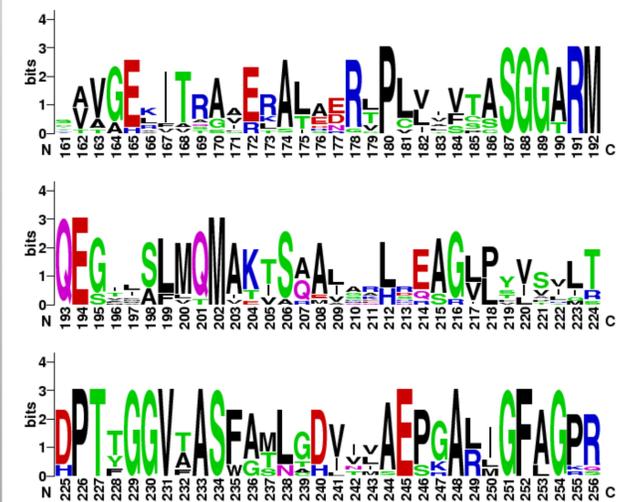


Figure III – *Kytococcus sedentarius* 05680 WEBLOGO results were well conserved between amino acids 162 and 256.

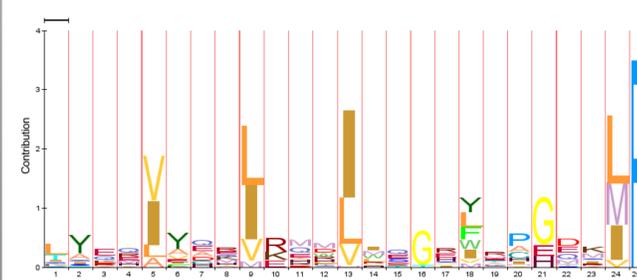


Figure IV – *Kytococcus sedentarius* 05690 HMM LOGO results show that were key functional residues at 5, 9, 13 and 25 within the amino acid sequence.

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. One exception was Gene Locus 05700, wherein n BLAST results supported a transcriptional regulator annotation, but other results, such as PDB, indicate the original call (Na⁺/H⁺ Antiporter) was supported. The erroneous BLAST results from the nr data base likely arose from an incorrect annotation in a genome sequenced in the past, which was perpetuated as new sequences were added to the nr data base (Brad Goodner, personal communication).

Gene Locus	Geni-Act Gene Products	Proposed Annotation
05670	Predicted Permease	Permease
05680	Acetyl-CoA Carboxylase Beta Subunit	Putative Acetyl-Coenzyme A Carboxylase Carboxyl Transferase Subunit Beta
05690	Transcription Regulator	GntR Family Transcriptional Regulator
05700	Na ⁺ /H ⁺ antiporter NhaD-like permease	*BLAST -Transcriptional Regulator *CDD- Na ⁺ /H ⁺ antiporter NhaD-like permease

References

Sims et al. (2009). *Standards in Genomic Sciences*, 12 – 20.
 Longshaw, C. M., et al. (2002) *Journal of Applied Microbiology*. 93.5: 810-816.
 Good, Benjamin M., and Andrew I. Su. (2013) "Crowdsourcing for bioinformatics." *Bioinformatics*. btt333.
 Becker, Karsten, et al. (2002) *International journal of systematic and evolutionary microbiology* 52.5: 1609-1614.
 Overbeek, Ross, et al. (2005) *Nucleic acids research* 33.17: 5691-5702.
 Schäffer, Alejandro A., et al. (2001) *Nucleic acids research* 29.14: 2994-3005.
 Tatusova, Tatiana A., and Thomas L. Madden. (1999). *FEMS microbiology letters* 174.2 : 247-250.

Acknowledgments

Supported by NSF ITEST Strategies Award Number 1311902