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

Annotation of the *Kytococcus sedentarius* Genome from DNA Coordinates 773902 to 781786

T.Almontaser, D. Al-Yafaei, D. DeJesus, M. Noman and Entasar Saif

Global Concepts Charter High School and Western New York Genetics in Research Partnership

Abstract

Four genes (Ksed_07680, Ksed_07690, Ksed_07700, Ksed_07730) from the microorganism *Kytococcus sedentarius* were annotated using the collaborative genome annotation website GENI-ACT. The Genbank proposed gene product name for each gene was assessed in terms of some or all of the following: general genomic information, amino acid sequence-based similarity data (BLAST, CDD, T-Coffee, WebLogo), structure-based evidence from the amino acid sequence (TIGRfam, Pfam, PDB), cellular localization data (TMHMM, SignalIP, PSORT, Phobius), potential alternative open reading frames (IMG Sequence Viewer For Alternate ORF Search), and enzymatic function (KEGG, MetaCyc, E.C. Number). The Genbank proposed gene product name did not differ significantly from the proposed annotation (based on Modules 1-6) for Ksed_07730 and the remaining genes mentioned above were not annotated to completion. It is proposed, therefore, that further study of these genes is needed to determine their function.

Introduction

Our project revolves around studying genes in order to assign function to them. Existing databases contain information on genes that have been studied either experimentally in a lab or through computer computation. Specifically, we are studying the genes of *Kytococcus sedentarius*, a micro-organism, which has had its genome recently sequenced. We are taking specific genes in question from this organism and comparing them to other genes of various organisms through computer programs and databases. Information found can tell us more about the function of the proteins that our genes code for. It is not enough to rely on computers to do this study or annotation of genes. Human beings can make observations and connections between genes that a computer program may not. By manually annotating genes, we can also catch mistakes that a computer program may miss. Understanding genes continues to be a key development for the scientific community because it may lead to treatment and prevention of disease.

According to Sims et al. (2009), *Kytococcus sedentarius* is a strictly aerobic, free-living, nonmotile, coccoid, non-spore forming, Gram positive bacteria originally isolated from a marine environment. *K. sedentarius* is known for the production of oligoketide antibiotics and for its role as an opportunistic pathogen causing endocarditis, hemorrhagic pneumonia, and pitted keratolysis. It has a genome of interest because of its location in *Dermacoccaceae*, a poorly studied family within the actinobacterial suborder *Micrococcineae* (Sims et al., 2009). Our study of the genes of *K. sedentarius* comes after its complete genome sequencing as part of the Genomic Encyclopedia of Bacteria and Archea (GEBA) project. With the complete sequencing of this genome, we can look at genes in question and compare them to other genes that have had their function predicted or experimentally determined.

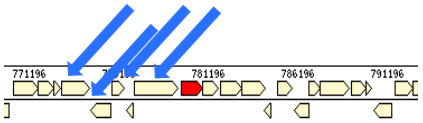

Figure 1: Neighborhood of the genes studied: From left to right- Ksed_07680, Ksed_07690, Ksed_07700, Ksed_07730

Methods and Materials

Modules of the GENI-ACT (http://www.geni-act.org/) that 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- Structure-Based Evidence	TIGRfam, Pfam, PDB	Are there functional domains in my protein?
Module 4- Cellular Localization Data	Gram Stain, TMHMM, SignalP, PSORT, Phobius	Is my protein in the cytoplasm, secreted or embedded in the membrane?
Module 5- 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 6- Enzymatic Function	KEGG, MetaCyc, E.C. Number	In what process does my protein take part?

Results

Kytococcus sedentarius 07680:

The initial proposed product of this gene by GENI-ACT was GMP synthase (glutamine-hydrolyzing). The top BLAST (swiss-pro and non-redundant database) and CDD search suggested hits for the amino acid sequence where high scores and low e-values were observed to be for chromosomal replication initiation protein DnaA. WEBLOGO results suggest this protein is very well conserved among various species but only toward the C-terminus (between residues 257 and 574). Cellular localization data in TIGRFAM suggest this protein contains no transmembrane helices. As such, the proposed annotation may not coincide with the results obtained.

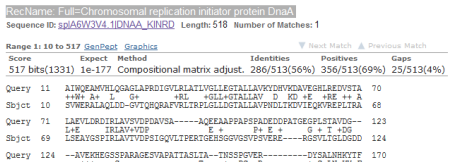
RecName: Full=Chromosomal replication initiator protein DnaA
Sequence ID: sp|A6W3V4.1|DNAA_KINRD Length: 518 Number of Matches: 1
Range 1: 10 to 517 GenPept Graphics
Score Expect Method Identities Positives Gaps
517 bits(1331) 1e-177 Compositional matrix adjust. 286/513(56%) 356/513(69%) 25/513(4%)
Query 11 AIWQEAWVHLQGAGLAPRDIGVLRLATLVGLLEGTALLAVKVDHVWDAVEGHLREDVSTA 70
Sbjct 10 SVWERALAQLDD-GVTQHQRAFVRLTRPLGLLDGTALLAVPNDLTKDVIEQKVREPLTRA 68
Query 71 LAEVLDRDIRLAVSVDPDAVSA-----AQEEAAPAPSPADEDDPATGEGPLSTAVDG-- 123
Sbjct 69 LSEAYGSPIRLAVTVDPSIGQVLTPERTGEHSGGVGSVPSVERE----RGSVLTGLDGDD 124
Query 124 --AVEKHEGSSPARAGESVAPATTASLIA--TNSSPGVER---------DYSALNHKYTF 170
Sbjct 125 GLHLDERRSGSLEEDSPLDDSDPDLLFTGYKVDRGPGTGRQPRRPTTRIENSRLNPKYIF 184

Figure 2: Portion of Swiss pro results showing high score and low significant e-value.

Kytococcus sedentarius 07690:

The initial proposed product of this gene by GENI-ACT is an integral membrane protein. Swiss-pro BLAST hits were discounted due to high expect values. Hypothetical protein hits were observed for the nr database. A CDD search suggested the following: Uncharacterized membrane protein YbhN, UPF0104 family [Function unknown].

WEBLOGO suggest regions of variable conservation at residues 293-352 and low areas of conservation at the N and C terminus. Results also suggest the protein coded by the gene ksed_7690 to be in the cytoplasmic membrane by P-Sort whereas results from TMHMM and Phobius suggest that there are 9 transmembrane helices with hydrophobic amino acids that most likely reside in the membrane. Signal IP and Phobius suggest no signal peptides present. A Pfam name of LPG_synthase_TM was also observed.

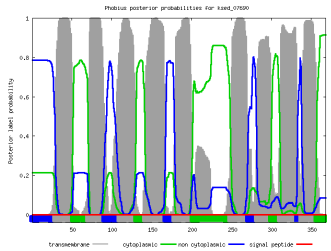

Figure 3: Phobius results predicting 9 transmembrane helices.

Kytococcus sedentarius 07700:

The initial proposed product of this gene by GENI-ACT was an NTP pyrophosphohydrolase. The top swiss-pro BLAST hit with a significant e-value corresponded to a Nudeoside diphosphate. The top two nr BLAST hits corresponded to a hypothetical protein and NUDIX domain-containing protein. COG hits resulted in two significant hits: (1) 8-oxo-dGTP pyrophosphatase MutT and related house-deaning NTP pyrophosphohydrolases, NUDIX family [Defense mechanisms](linked to 3-D structure and (2) NADH pyrophosphatase NudC, Nudix superfamily [Nudeotide transport and metabolism]. WEBLOGO results suggested the protein is not well conserved amongst species, except for areas towards the C-terminus. Although no TIGRFAM hits were found, a Pfam hit of NUDIX was also obtained. No transmembrane helices or signal peptides were identified. A KEGG analysis suggests this protein is involved in pyrimidine metabolism and the accepted name for the EC number (3.6.1.12) obtained is dCTP diphosphotase.

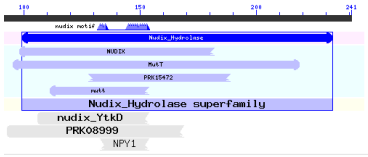

Figure 4: Conserved domains on NTP pyrophosphohydrolase.

Kytococcus sedentarius 07730:

The initial proposed product of this gene by GENI-ACT is succinyl-CoA synthetase (ADP-forming) betasubunit. This gene product proposal was supported by the top BLAST (swiss-pro and non-redundant database) hits for the amino acid sequence where high scores and low e-values were observed. A CDD search resulted in a significant COG hit named SucC (Succinyl-CoA synthetase, beta subunit [Energy production and conversion], linked to 3D-structure).

WEBLOGO results suggest this protein is very well conserved among various species throughout the entire sequence. As such, the proposed annotation is succinyl-CoA synthetase (ADP-forming) betasubunit. No transmembrane or signal peptides were predicted. A TIGRFAM search resulted in the hit sucCoAbeta with function listed as succinate-CoA ligase, beta subunit. A Pfam search resulted with Pfam names of ATP-grasp_2 and Ligase_CoA. A KEGG search suggests this protein is involved in the Citrate or TCA Cycle. An accepted name of Succinate--CoA ligase (GDP-forming) was noted with the E.C. number of 6.2.1.4.

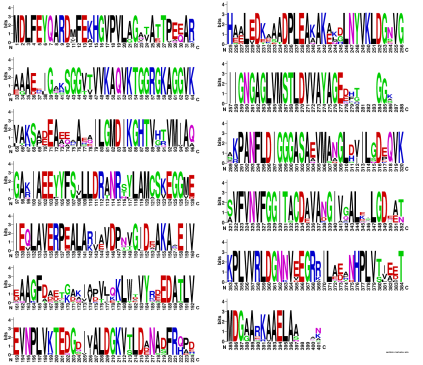

Figure 5: WEBLOGO showing high degree of conservation amongst various organisms.

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. In cases where not enough data supported a proposed annotation, further study is suggested.

Gene Locus	Geni-Act Product	Proposed Annotation
07680	GMP synthase (glutamine-hydrolyzing)	Further study is proposed
07690	integral membrane protein	Uncharacterized integral membrane protein
07700	NTP pyrophosphohydrolase	Further study is proposed
07730	Succinyl-CoA synthetase (ADP forming) betasubunit	Succinyl-CoA synthetase (ADP forming) betasubunit

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