Annotation of the *Kytococcus sedentarius* Genome from DNA Coordinates 06950 to 06960

***John Durrenberger, *Louisa Izydorczak and Paul Weber**

Amherst Central High School and the Western New York Genetics in Research Partnership

* indicates the same effort

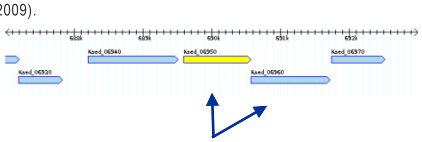
Abstract

A group of consecutive 7 genes from the microorganism *Kytococcus sedentarius* (Ksed_06950 – Ksed_06960) 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, enzymatic function, presence or absence of gene duplication and degradation, the possibility of horizontal gene transfer, and the production of an RNA product. 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 nr database.

Introduction

Kytococcus sedentarius is a strictly aerobic, non-motile, non-encapsulated, and non-endospore forming gram positive coccoid bacterium, found predominantly in tetrad formation. This organism is classified as a chemoheterotroph, as it requires methionine and several other amino acids for growth. Originally isolated from a microscope slide submerged in sea water in 1944, *Kytococcus sedentarius* grows well in sodium chloride at concentrations less than 10% (w/v).

According to Sims et al. (2009), *Kytococcus sedentarius* is a microorganism of interest for several reasons. This bacterium is a natural source of the oligoketide antibiotics monensin A and monensin B (Sims et al., 2009). *Kytococcus sedentarius* has been implemented as the etiological agent of a number of opportunistic infections including valve endocarditis, hemorrhagic pneumonia, and pitted keratolysis (Sims et al., 2009). Finally, the phylogeny of this microorganism is a source of interest, as it is a member of the family *Dermacoccaceae* within the actinobacterial suborder *Micrococcineae*, which has yet to have been thoroughly studied utilizing bioinformatics (Sims et al., 2009).



Genes used in our annotation.

This research poster contains information about two genes in *Kytococcus sedentarius*, Ksed_06950 and Ksed_06960. As seen above, they are next to each other in the gene sequence. Their proposed products are actually similar. They are both ABC-type transporters. Also, as seen above, the genes are about the same length and pointing in the same direction. If the arrows are pointing to the right, it means the genes are on the forward strand of the DNA. The products proposed by the computer for other surrounding genes are similar to these proposed products. This area of the genome seems to have to do with phosphate binding and transport.

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 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

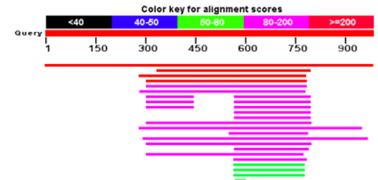
<u>*Kytococcus sedentarius* 06950 (John):</u>

The initial proposed product of this gene by GENI-ACT was a ABC-type molybdate transport system, permease component. This gene product proposal was supported by the top Blast hits for the amino acid sequence. With a relatively low e value making it the most similar to Ksed_06950. Here is the image

```
Score         Expect  Method                         Identities     Positives      Gaps
205 bits(522) 9e-62   Compositional matrix adjust.   134/303(44%)   188/303(62%)   24/303(7%)

Query  38   RLGDKIFSGLSIGSGALIAVTLALVAIFLLMEAWPALK----------QASEIGDEGFWM  87
            R G+++F   +  +G+ I + + L+AIFLL+ A P+L+          Q     DE
Sbjct  20   RRGERLFKLAASAAGSTIVIAILLIAIFLLVRAVPSLRANHANFFTSTQFDTSDDEQLAF  79

Query  88   LAANLLGGTLLASAIAMVIAVPVAVGIALFISHYAPPRLSKSIAFVIDLLAAIPSVIFGL  147
              +L    T L+S  A+V+AVPVAVGIA+F++HYAP RLS+     ++DLLAA+PS+IFGL
Sbjct  80   GVRDLFMVTALSSITALVLAVPVAVGIAVFLTHYAPRRLSRPFGAMVDLLAAVPSIIFGL  139

Query  148  WGYLTFAPFLVPLFEWLEDNLGFIPLFEDASRT---GRVLLTASIVLAIMILPIITSVAR  204
            WG    AP L P+  +L  NLG++ LF+    +   G  + TA IVL++MILPI+TS++R
Sbjct  140  WGIFVLAPKLEPIARFLNRNLGWLFLFKQGNVSLAGGGTIFTAGIVLSVMILPIVTSISR  199

Query  205  DVFQQVPRGHEEASLALGATKWEMIKQAVLPFGWPGVVGGAMLGLGRALGETMAVAMIL-  263
            +VF+Q P    EA+LALGATKWE+++  VLP+G  GVV  +MLGLGRALGET+AV +IL
Sbjct  200  EVFRQTPLIQIEAALALGATKWEVVRMTVLPYGRSGVVAASMLGLGRALGETVAVLVILR  259

Query  264  APKDEFQWNIIGSGNFPIPAYIA-----LNFPDATGLKVSTLIALGLVLFVLTLAVNLVA  318
            +     W++    G +    +IA       + P  TG  +S       G  LFVLT  VN  A
Sbjct  260  SAARPGTWSLF-DGGYTFASKIASAASEFSEPLPTGAYISA----GFALFVLTFLVNAAA  314

Query  319  RAI  321
            RAI
Sbjct  315  RAI  317
```

The top gene product was Phosphate transport protein from *Mycobacterium bovis*. With an alignment length of 284. And a E value of 9e-62 and a max score and a min score of 205.

Ksed_06950 graphical blast results



Ksed_06950 did not have a lot of blast hits in the Swis_Prot dabase (above). However, many hits with high scores and low E values corresponding to phosphate ABC transporter permease subunit PstC were found in the nr database.

<u>*Kytococcus sedentarius* 06960 (Louisa):</u>

The initial proposed product of this gene by GENI-ACT was a phosphate ABC transporter membrane protein. This gene product proposal was supported by the top BLAST hits for the amino acid sequence. The image below shows the amino acid sequence for this gene compared with the amino acid sequence for phosphate ABC transporter permease [*Citricoccus sp.* CH26A]. With a very low e-value, these proteins are very similar to one another, though they are found in different organisms.

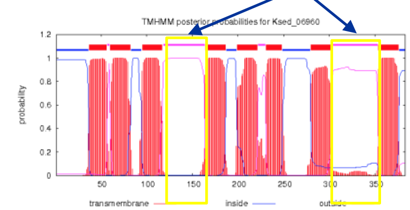
```
1st hit:
phosphate ABC transporter permease [Citricoccus sp. CH26A]
Sequence ID: ref|WP_010143601.1|  Length: 394  Number of Matches: 1

Range 1: 20 to 392  GenPept  Graphics                     ▼ Next Match  ▲ Previous Match
Score          Expect  Method                        Identities     Positives      Gaps
342 bits(876)  1e-110  Compositional matrix adjust.  189/373(51%)   255/373(68%)   15/373(4%)

Query  26   QRNTIHPASKT--LWMGLLAASAVVGGLITWLIG---WNIALFVVVTIVVFMLAVYIGYR  80
            ++N++    K    W+ +LA + VV    IT L+G    ++    ++V+  V+++L +Y+  R
Sbjct  20   KKNSLTAGQKPGWTWIAVLAGALVVAAGITALLGGGSFSWVGWLVIAAVLYVLGMYLATR  79

Query  81   VTAGHRQGLDQVVRALVYLAFALAVIPLIWLLWTVFSNAIPPLLGNLQVFTQDMYGISGD  140
            V    R+  D + R LV+ AF LA+IPLI +LWTV +N +P LL     Q+   +DM G++G
Sbjct  80   VIENSRRATDGLWRNLVWAAFLLALIPLISVLWTVAANGLPTLLETPQLLWRDMQGVTGS  139

Query  141  VDEARVTEGAEIVGGMKHALIGTLMITVIASIISIPVGLMAAIYLVEYGTNGWFSRTLRF  200
            D A   TEGA + GG  H  L+GTLMIT++A+IIS+P+GL+ +IYLVEYG  GWFSR + F
Sbjct  140  DDIATQTEGAPLAGGFLHGLVGTLMITLLATIISVPIGLLTSIYLVEYGRGGWFSRAIVF  199

Query  201  LVDVMTGIPSIVAGLFAYALFSMIF------GPGT----KFGFAGAIALSVLMIPIVVRN  250
            VDVMTGIPSIVAGLFA+A   ++          GP         K G + AIALSVLMIP+VVR+
Sbjct  200  FVDVMTGIPSIVAGLFAFAAVQLVITTFIGDGPRQLQMVKMGLSAAIALSVLMIPVVVRS  259

Query  251  SEEMLRVVSRDLRESALALGVPKWRVITKIVLPTAVAGLMSSVFLGVARVIGETAPLLVA  310
            +EEMLRVV   +LRE++  ALGV KWR I K+VLPTA++G+ S V L +ARV GETAP+LV
Sbjct  260  TEEMLRVVPNELREASYALGVRKWRTIFKVVLPTAMSGIASGVTLSIARVTGETAPILVT  319

Query  311  AGYNAATNWNPFEGPMNSLAFYAYQLFQYPLNPGYREVSDERSWIAVFVLVMLVLTLNLI  370
            AGY A TNWNPF+G M +L  Y Y+      P  P    + S ER+W A    VL+++V+ LNLI
Sbjct  320  AGYIATTNWNPFDGWMTALPVYIYRQLINPTAPAAADPSAERAWAAALVLIVIVMLLNLI  379

Query  371  ARLVAAALAPKKS  383
            AR+VA A APKK+
Sbjct  380  ARIVAKAFAPKKT  392
```

The following image is another BLAST hit with a very low e-value. This protein from another organism is similar to what the computer predicted for Ksed_06960's product, both in name and sequence.

```
phosphate ABC transporter permease [Serinicoccus marinus]
Sequence ID: ref|WP_022925372.1|  Length: 385  Number of Matches: 1

Range 1: 6 to 382  GenPept  Graphics                       ▼ Next Match  ▲ Previous Match
Score          Expect  Method                        Identities     Positives      Gaps
337 bits(864)  5e-109  Compositional matrix adjust.  195/382(51%)   260/382(68%)   6/382(1%)

Query  1    MTLSPVQADHGPTHEADRGPDMPELQRNTIHPASKTLWMGLLAASAVVGGLITWLIGWNI  60
            T SP ++DH PT +        +P     + + PA +   M + A + V+  L+  L+  ++
Sbjct  6    QTPSPQRSDHRPTQQEAVSGIVP----SPVRPAGRRPAMIIAAGAVVLALLLWQLLDLHV  61

Query  61   ALFVVVTIVVFMLAVYIGYRVTAGHRQGLDQVVRALVYLAFALAVIPLIWLLWTVFSNAI  120
            A ++++ + ++ A+Y  YR   G R  +D +VR+LVY AF  A++PL+ +LWTV +
Sbjct  62   AAALLLSFLAYLGAIYATYRQLGGTRVAVDHLVRSLVYFAFVCAIVPLVSVLWTVIAQGA  121

Query  121  PPLLGNLQVFTQDMYGISGDVDEARVTEGAEIVGGMKHALIGTLMITVIASIISIPVGLM  180
                 +   Q M G++G  D+       A+ VGG  HA++GTL+IT +A++IS+P+G+
Sbjct  122  GRAF-DFAFLNQTMNGVTGLNDQQYADGSADFVGGAYHAIMGTLLITGLATLISVPIGMF  180

Query  181  AAIYLVEYGTNGWFSRTLRFLVDVMTGIPSIVAGLFAYALFSMIFG-PGTKFGFAGAIAL  239
            AIYLVEYG     SR +RFLVDVMTGIPSIVAGLFA+ALF+ IFG     K G AGA+AL
Sbjct  181  TAIYLVEYGGRNQLSRGIRFLVDVMTGIPSIVAGLFAFALFTQIFGIRDAKMGLAGAVAL  240

Query  240  SVLMIPIVVRNSEEMLRVVSRDLRESALALGVPKWRVITKIVLPTAVAGLMSSVFLGVAR  299
            SVLMIP VVRNSEEMLR+V  +LRE+ALALGVPKWRVI K+VLPTA +GL S V L +AR
Sbjct  241  SVLMIPTVVRNSEEMLRIVPNELREAALALGVPKWRVIAKVVLPTAASGLASGVTLAIAR  300

Query  300  VIGETAPLLVAAGYNAATNWNPFEGPMNSLAFYAYQLFQYPLNPGYREVSDERSWIAVFV  359
            VIGETAPLLVA GY      NWN F+GPMN+LA YAY +F  PL+P  RE S ER+W A  +
Sbjct  301  VIGETAPLLVAIGYARGLNWNAFDGPMNTLAVYAYFMFTKPLSPAVREPSLERAWAAALL  360

Query  360  LVMLVLTLNLIARLVAAALAPK  381
            LV++V+ LNL+ARL+A   APK
Sbjct  361  LVLIVIALNLLARLIATMFAPK  382
```

Also, results were found in BLAST in the swissprot database with low e-values that support the proposed product, and swissprot has less information but more reliable information than nr.

This next image shows the sections of the amino acid sequence that are transmembrane, intracellular and extracellular. There are many places, seen in red, where the protein crosses the membrane. There are a few large gaps between these crossings, however, where the protein is outside of the membrane.



The CytoplasmicMembrane score in PSORT-B was also relatively high. It was 10.00, while Cytoplasmic, Cellwall and Extracellular scores were all zero. The PSORT-B final prediction was CytoplasmicMembrane 10.00. There is a Shine-Dalgarno sequence 5 to 15 bases upstream of the start codon. Also, names found in TIGRFAM, COG names and names of paralog gene products in Kytococcus sedentarius were found that are similar to the name of the proposed product.

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.

For Ksed_06960, the proposed product is the same as what the computer called: phosphate ABC transporter membrane protein. The blast hits with low e-values support this conclusion. The TMHMM posterior probabilities graph supports the idea that the protein is found in the membrane. The large gaps could have physiological functions or be places where the amino acids wrap around, putting certain transmembrane sections next to each other that aren't shown next to each other in the graphical representation. Also, the shine-dalgarno sequence 5 to 15 bases upstream of the start codon suggests that the amino acid sequence of the protein has been called correctly by the computer.

References

Sims et al. (2009). Complete genome sequence of Kytococcus sedentarius type strain (541T). *Standards in Genomic Sciences*, 12 - 20.

Acknowledgments

Dr. Rama Dey- Rao, Dr. Stephen Koury
Supported by NSF ITEST Strategies Award Number 1311902