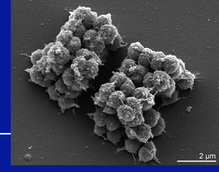
University at Buffalo *The State University of New York* | REACHING OTHERS



Annotation of the pathway for the biosynthesis of chorismate from erythrose-4-phosphate in *Kytococcus Sedentarius* using the IMG-ACT database

Ashlesha P. Odak, Patricia Masso-Welch, Rama Dey-Rao, Stephen Koury.

Department of Biotechnical and Clinical Laboratory Sciences, School of Medicine and Biomedical Sciences, State University of New York, University at Buffalo, 26 Cary Hall, 3435 Main Street, Buffalo, NY 14214

Abstract

Kytococcus sedentarius is an aerobic, gram positive, coccoid, non-endospore forming bacterium isolated from a marine environment. The genome of *Kytococcus sedentarius* is of great interest to annotate because of its biotechnological potential as a source of oligoketide antibiotics (monensin A and B), for its role in causing pitted keratolysis of the foot and in causing opportunistic infections. The genome has been sequenced and automated computer annotation of the genome has been performed. Manual annotation, however, holds great value because of documented inaccuracies involved in the computer annotation. The purpose of this study was to define and annotate the genes involved in the biosynthesis of chorismate in this organism. Chorismate is the direct precursor of many aromatic compounds and its biosynthesis occurs via the shikimate pathway only in bacteria, fungi, yeasts, algae, plants and certain parasites. Tools like BLAST, TMHMM, Phylogeny were used to annotate the genes with respect to the various characteristic features of the gene like amino acid sequence of the gene product, localization of the protein in the cell, structure, function, phylogenetic origin and pathways in which the gene product is involved. The annotation results suggested that gene for 3-dehydroquinate synthase was involved in 2 steps in the pathway. The gene encoding the enzyme 3-dehydroquinate dehydratase may have undergone a horizontal gene transfer from organisms like *Clostridium* based on results from the phylogenetic tree and gene neighborhood.

Introduction

Kytococccus sedentarius is a free-living, non-motile, gram positive bacterium [2] and has been generally considered to be non-fatal although there have been some studies that have found it to be associated with pitted keratolysis [3], dermatoses [4], peritoneal dialysis-associated peritonitis [5], nail infections [6], opportunistic infections, and fatal hemorrhagic pneumonia [7] especially in immunocompromised patients. Pitted keratolysis caused due to infection of *Kytococccus sedentarius* is associated with pits in the skin on the plantar surfaces of the feet and toes and production of a strong foot odor. Two extracellular proteases have been isolated from the organism that can degrade human keratin and thus contribute to pitting of skin [3]. The organism represents the scarcely populated genus *Kytococcus* (2 species) within the actinobacterial family *Dermacoccaceae* [2].

The genome of *Kytococcus sedentarius* has been sequenced and subjected to computer-based gene calling and annotation. However, there lies great importance in performing the manual annotations that we are performing here because many of the computer gene calls and annotations are incorrect [8].

Annotation, can be thought of as *in silico* biology as opposed to *in vivo* or *in vitro* biology. Using *in silico* tools to support or even to substitute wet laboratory work could help better focusing the laboratory experiments, resulting not only in considerable saving of resources but also increasing the number of molecules and scenarios investigated [1].

The biosynthesis of chorismate occurs via the shikimic acid pathway in seven enzyme-mediated steps. This pathway is a central pathway for the synthesis of various aromatic compounds in bacteria, fungi, yeasts, algae, plants and certain parasites[9]. The metabolic pathway for the biosynthesis of chorismate from erythrose-4-phosphate was annotated for *Kytococcus sedentarius* with regards to the specific genes involved in each step. Tools such as BLAST using Swiss-prot and non-redundant databases, T-COFEE, Web logo, TMHMM, PSORT-B, KEGG, COG, TIGRFam, PFam, Phylogeny, ExPasy and MetaCyc were utilized.

It is hoped that this annotation of the metabolic pathway for the biosynthesis of chorismate and comparison with that used in other organisms will greatly help in establishing the characteristics of the genes involved and thus the relation of *Kytococcus Sedentarius* to other organisms.

Materials and Methods

The following modules were looked at to annotate the genes using the IMG-ACT website:

1. Basic Information
2. Sequence-based Similarity Data
3. Cellular Localization Data
4. Alternative Open Reading Frame
5. Structure-based Evidence
6. Enzymatic Function
7. Duplication and Degradation
8. Horizontal Gene Transfer

Sequence based similarity-

This module was used to determine the homology of the protein sequence in K.sedentarius to similar sequences in other organisms. The protein sequence for the gene was entered as the query sequence in the NCBI based **Basic Local Alignment Search tool (BLAST)** (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The software gives the gene product name, organism, sequence, alignment length, score and E-value of similar sequences from other organisms and arranges them according to the best match of sequence. The sequences with the lowest e-value (closest to 0) and highest bit scores are the most similar and were noted down. **Conserved Domain Database (CDD)** accessed through the top part of the BLAST results page shows conserved putative domains. This was used to find the clusters of orthologous groups (COGs) based on similarity to the query sequence.

Materials and Methods

The sequences of the top 10 orthologs were obtained through the IMG listing of known orthologs for the gene. These were used to generate a multiple sequence alignment using the **Tree-based Consistency Objective Function For Alignment Evaluation (TCoffee)** software (http://www.ebi.ac.uk/Tools/msa/tcoffee). It shows the sequence similarity between the orthologs. The multiple sequence alignment was then used to generate a **Weblogo** (http://weblogo.berkeley.edu) that shows conserved residues using color coding and font variation for different amino acid residues.

Cellular localization: The location of the gene product in the cell was determined by combining the results obtained using the following tools- **1. Trans-membrane Hidden Markov Model (TMHMM)-** Predicts presence of transmembrane helices (http://www.cbs.dtu.dk/services/TMHMM) **2. SignalP-** Predicts presence of signal peptide (http://www.cbs.dtu.dk/services/SignalP) **3. Psortb-** Predicts subcellular location of protein (http://www.psort.org/psortb) **4. Phobius-** Combines TMHMM and Phobius results into one graphical output (http://phobius.sbc.su.se)

Determination of alternate open reading frame: Sequence viewer tool available on the IMG gene details page was used for viewing and verifying the proposed Open Reading Frame (ORF) and identifying any novel ORF. The start codon called by the gene caller was checked for the presence of a Shine Dalgarno sequence upto 15 base pairs upstream to it. Conclusions were then made as to whether the start codon called was correct or whether there was an alternate ORF that gave a protein with better BLAST hits.

Structure based evidence: The tools used in this module gave the probable function of the protein based on the structure of the protein. The tools were-

TIGRFAM (http://tigrblast.tigr.org/web-hmm)- Predicts name and function of the protein based on structural similarity to functionally understood, well conserved domains.

Pfam (http://pfam.sanger.ac.uk/search)- Identifies protein domains and clans based on domain function, sequence conservation and critical residues from multiple sequence alignment of common protein families. Generates pairwise alignment of similar domains and HMM (Hidden Markov Model) logo showing conserved residues in the domain sequence.

Protein Data Bank (PDB) (http://www.rcsb.org/pdb/home/home.do)- Gives 3d crystal structure of homologous proteins

Enzymatic function- The function of the gene product was confirmed by using the databases **Kyoto Encyclopedia of Genes and Genomes (KEGG)** http://www.genome.jp/kegg/pathway.html) and **MetaCyc** (http://metacyc.org/). The metabolic pathway in which the gene is involved was obtained through these tools. The EC number for each gene product was verified using the **ExPASY** (Expert Protein Analysis System) (http://www.expasy.ch/enzyme/enzyme-search-ec.html) database.

Duplication and degradation: The possibility of the gene to be a pseudogene was tested by checking for the presence of paralogs on the homolog listing in the IMG database. If a paralog was identified, the gene product was analyzed using the BLAST software. The BLAST hits were then compared to the gene being annotated.

Horizontal gene transfer: The possibility of horizontal gene transfer having occurred was evaluated based on 3 tools: **1. Phylogeny.fr (**http://www.phylogeny.fr/)- Creates phylogenetic tree as Radial Drawtree and cladogram images to determine phylogenetic origin of the gene

2. Ortholog neighbourhood map- Shows neighborhood of gene and enables comparison between homologs **3. GC content of genome** obtained through the IMG database- GC content of the gene is compared with the genome of K.Sedentarius.

Results

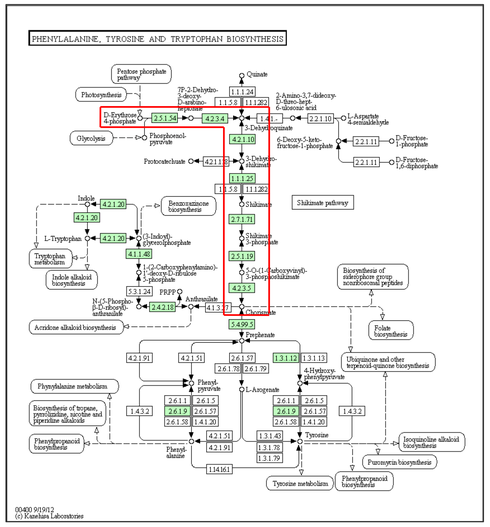


Figure 1: KEGG Pathway map for biosynthesis of chorismate from erythrose-4-phosphate. Genes highlighted in green are the genes used by *K. Sedentarius*

Results

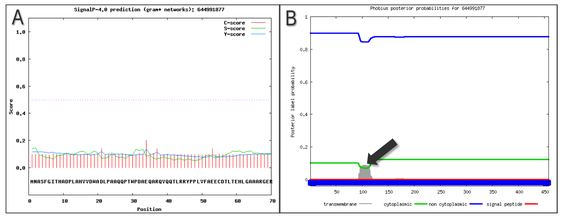


Figure 2: A) Signal peptide probability graph for 3-deoxy-D-arabinoheptulosonate-7-phosphate synthase illustrating absence of a peak indicating that the protein lacks a signal peptide. Dotted line shows minimum probability limit necessary for predicting presence of signal peptide. B) Phobius prediction graph for 3-deoxy-D-arabinoheptulosonate-7-phosphate synthase showing absence of transmembrane helices. Cut-off value of probability for prediction of location is 0.75

Figure 3: Pfam domain graphic for gene OID 644991533 showing two domains. The green domain is shikimate kinase and the pink is dehydroquinate synthase

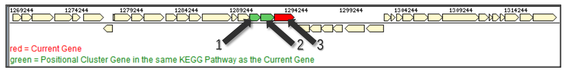


Figure 4: Gene neighbourhood map for gene OID 644991533 illustrates 1) Gene encoding shikimate-5-dehydrogenase 2) Gene encoding chorismate synthase 3) Gene encoding shikimate kinase and 3-dehydroquinate synthase .

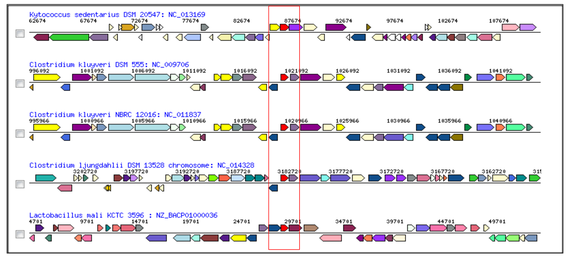


Figure 5: Gene ortholog neighborhood map for gene OID 644990396 encoding 3-dehydroquinate dehydratase that shows dissimilarity between gene neighborhoods indicating the possibility of horizontal gene transfer. The gene in red is the gene of interest.

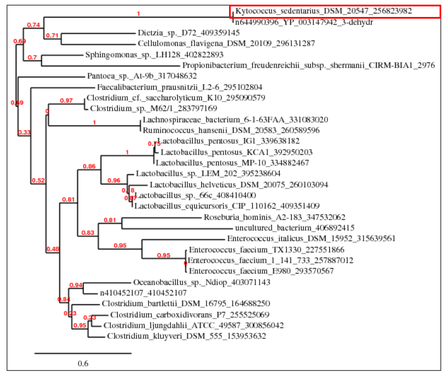


Figure 6: Phylogenetic tree in the form of a cladogram for gene OID 644990396 encoding 3-dehydroquinate dehydratase. *K. sedentarius* can be seen to be clustered with organisms of a different family indicating that the gene may have undergone horizontal gene transfer.

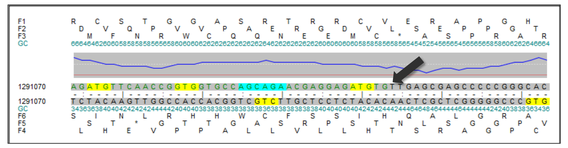


Figure 7: Sequence viewer for ORF search. The arrow points to the first nucleotide of the sequence. No red highlighted nucleotides indicates no start codon has been called by the gene caller.

Results

Figure 8: Pfam domain graphic for gene OID 644992153 encoding 5-O-(1-Carboxyvinyl)-3-phosphoshikimate synthase depicting presence of two domains having the same function arranged on the gene.

>sp|Q47QY9.2|AROC_THEFY RecName: Full=Chorismate synthase; AltName: Full=5-enolpyruvylshikimate-3-phosphate phospholyase
Length=393

GENE ID: 3579156 Tfu_1090 | chorismate synthase [Thermobifida fusca YX]
(10 or fewer PubMed links)

Score = 508 bits (1307), Expect = 1e-177, Method: Compositional matrix adjust.
Identities = 275/402 (68%), Positives = 313/402 (78%), Gaps = 17/402 (4%)

```
Query  1    MLRWLTAGESHGEALTGVLEGVPAGVPIEDLAGALGRRRLGYGRGARMSFEADRVTLI  60
            MLRWLTAGESHG AL +LEG+PAGV  V++D+A AL RRR GYGRGARM FE D V+I
Sbjct  1    MLRWLTAGESHGPALVAILEGLPAGVSVSDDIAAALLRRRAGYGRGARMKFEKDEVSII  60

Query  61   GGVPRGVTMGSPVALRIDNTEWPKWQAVMSADPVPPESLQVDAGRGDEREVRARNPLTRP  120
            GG+RRG T+G PVA+++ NTEWPKWQ VMS DPVP E L+          VARN PLTRP
Sbjct  61   GGIRHGRTLGGPVAIQVGNTEWPKWQKVMSFDPVPAEELE---------GVARNAPLTRP  111

Query  121  RPGHADLAGMQKYGLDEARPVLERASARETAARVALGEVARQAL-GALGIELGSHVVALG  179
            RPGHADL GMQKYG DEARP+LERASARETAARVALGEVARQ L  ALG+E+ SHVVA+G
Sbjct  112  RPGHADLVGMQKYGTDEARPILERASARETAARVALGEVARQFLRQALDVEIVSHVVAMG  171

Query  180  GVHVPDGVLDRAALGDLITTDGGRARIDADPVRCGDPATAATMVERVDEAKRAGDTLGGVV  239
             V VPD    A  DL          AR+DADPVRC DP T+A MV  VD+ ++ GDTLGGVV
Sbjct  172  SVDVPDDA-PIPAPELA------ARVDADPVRCTDPETSARMVARVDTRRSGDTLGGVV  224

Query  240  EVVVAGLPSGLGSHVHGDRRLDAQLAGALMSIQAIKGVELGDGFDVADQPGSAAHDEMEW  299
            EV+  GLP GLGSHVH DRRLDA+LAGALM IQAIKGVE GDGF  A + GSAAHDE++
Sbjct  225  EVIAYGLPPGLGSHVHWDRRLDARLAGALMGIQAIKGVEFGDGFRTAARRGSAAHDEIDV  284

Query  300  DEGAIRRVTDRSGGVEGGMSTGELRVRAAMKPISIVPRALRIVDVATGEAATAHHQRSD  359
             IRR ++R+GGVEGGMSTG+ LRVRAAMKPI+IVPRAL T+DV+TG+ A AHHQRSD
Sbjct  285  GPDGIRRRSNRAGGVEGGMSTGDPLRVRAAMKPIATVPRALDTIDVSTGQPAQAHHQRSD  344

Query  360  VCAVPAAGVVAEAEVALVLLRAVLEKFGGDHLEDTRAAVEHY  401
            V AVPAAGVVAEA VALVL  A LEKFGGD + +T    V
Sbjct  345  VIAVPAAGVVAEAMVALVLADALEKFGGDSVAETVRNLRGY  386
```

Fig 9: Alignment of the top BLAST hit for gene OID 644991532 encoding chorismate synthase with the red highlighted area showing the statistics for the alignment.

Conclusions

Step	Gene OID	Product name	E.C. Number	Reaction product
1.	644991877	3-deoxy-D-arabinoheptulosonate -7-phosphate synthase	2.5.1.54	7-phospho-2-dehydro-3-deoxy-D-arabinoheptonate
2.	644991533	3-dehydroquinate synthase	4.2.3.4	3-dehydroquinate
3.	644990396	3-dehydroquinate dehydratase	4.2.1.10	3-dehydroshikimate
4.	644991531	shikimate 5-dehydrogenase	1.1.1.25	shikimate
5.	644991533	shikimate kinase	2.7.1.71	shikimate-3-phosphate
6.	644992153	5-O-(1-Carboxyvinyl)-3-phosphoshikimate synthase	2.5.1.19	5-O-(1-Carboxyvinyl)-3-phosphoshikimate
7.	644991532	chorismate synthase	4.2.3.5	chorismate

Annotation of all the 7 genes involved in the biosynthesis of chorismate from erythrose-4-phosphate verified the data on the IMG website for the gene details. The gene OID 644991533 was found to be involved in 2 steps in the pathway- 1) Conversion of 2-keto-3-deoxy-D-arabinoheptulosonate-7-phosphate to 3-dehydroquinate 2) Conversion of shikimate to shikimate-3-phosphate. The gene OID 644990396 encoding the enzyme 3-dehydroquinate dehydratase responsible for conversion of dehydroquinate to 3-dehydroshikimate may have undergone a horizontal gene transfer from organisms like *Clostridium* based on results from the phylogenetic tree and gene neighborhood. The functions of all the genes in the pathway were confirmed by the KEGG and Metacyc results.

References

1. Terstyanszky, G., et al. (2012). "Application repository and science gateway for running molecular docking and dynamics simulations." Stud Health Technol Inform **175**: 152-161.
2. Stackebrandt, E., et al. (1995). "Taxonomic dissection of the genus Micrococcus: Kocuria gen. nov., Nesterenkonia gen. nov., Kytococcus gen. nov., Dermacoccus gen. nov., and Micrococcus Cohn 1872 gen. emend." Int J Syst Bacteriol **45**(4): 682-692.
3. Longshaw, C. M., et al. (2002). "Kytococcus sedentarius, the organism associated with pitted keratolysis, produces two keratin-degrading enzymes." Journal of Applied Microbiology **93**(5): 810-816.
4. Hsu, A. R. and J. W. Hsu (2012). "Topical review: skin infections in the foot and ankle patient." Foot Ankle Int **33**(7): 612-619.
5. Chaudhary, D. and S. N. Finkle (2010). "Peritoneal dialysis-associated peritonitis due to Kytococcus Sedentarius." Peritoneal Dialysis International **30**(2): 252-253.
6. Towersey, L., et al. (2008). "Kytococcus sedentarius nail infection." Journal of the American Academy of Dermatology **58**(2): AB88-AB88.
7. Levenga, H., et al. (2004). "Fatal hemorrhagic pneumonia caused by infection due to Kytococcus sedentarius - a pathogen or passenger?" Annals of Hematology **83**(7): 447-449.
8. Florea, L., et al., (2005). "Gene and alternative splicing annotation with AIR." Genome Research **15**(1): p. 54-66.
9. Zucko, J., et al., (2010). "Global genome analysis of the shikimic acid pathway reveals greater gene loss in host-associated than in free-living bacteria" Bmc Genomics **11**.