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

Annotation of genes responsible for synthesis of uroporphyrinogen III from L-glutamate, the first half of of the pathway of heme biosynthetic pathway, in *Kytococcus sedentarius*.

Vignesh M Iyer, Patricia Masso-Welch, Ph.D., Rama Dey-Rao, Ph.D, Stephen Koury, Ph.D.

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

Genome annotation is a set of algorithmic software tools that allows complete sequencing of an organism's genome. *Kytococcus sedentarius* is a gram positive bacterium, with callus degrading activity, closely associated with pitted keratolysis. The objective of this study was to annotate the genes in *Kytococcus sedentarius* strain 541 (DSM 20547) responsible for heme biosynthesis, involving the formation of uroporphyrinogen III from L-Glutamate. Metabolic functions of individual genes were characterized using the Integrated Microbial Genome Annotation Collaboration Toolkit (IMG-ACT) developed by the Joint Genome Institute (University of California, CA). Sequence based similarity was identified using BLAST and T-COFFEE. Cellular localization was determined using TMHMM and SignalP. Functional amino acids were assessed using WebLogo and protein functions were analyzed using Pfam. TIGRFAM was used to gather structure based evidence. KEGG and MetaCyc were used to verify metabolic functions of respective genes in the pathway. Significant importance was also given to the novel open reading frames, conserved domains, gene duplication, degradation and phylogenetic evolution. Genes were localized mostly in the cytoplasmic region and Weblogo indicated highly conserved amino acids. The proposed DNA coordinates could not be verified for glutamyl tRNA reductase (gene object id: 644992667). There was no evidence of any gene being a pseudo gene and phylogenetic analysis did not indicate any horizontal gene transfer. Heme is important in a number of microbial metabolic processes. Annotation reports indicated that the genes leading to the synthesis of uroporphyrinogen III were functionally active and highly conserved in *Kytococcus sedentarius*.

Introduction

One of the most important prerequisites to analyze a gene is to characterize the unknown sequence by *in silico* methods, to establish a skeletal framework based on which further studies can be carried out (Lio, Angelini et al., 2012). Genome annotation uses a set of algorithms and a statistical based approach to identify the biological characteristics and functions of a given query (Sims, Brettin et al., 2009). A series of computational software tools are used to match the query sequence with known databases; the gene, its genomic structure, protein function and almost every single feature of the query sequence is compared, analyzed and determined by a variety of sophisticated analytical software algorithms (Collins, Goward et al., 2003). A completely annotated sequence would include analyzing the sequence at the nucleotide level to predict its structure, comparing mRNA with other genomes for evidence of expression and finally identification of the conserved genes in vertebrates, and constructing a phylogenetic tree to predict the evolutionary map of the gene in context (Venter, Adams et al., 2001).

A gram-positive bacterium, *Kytococcus sedentarius* is a strain of species in the family *Dermacoccacea* in the suborder *Micrococcineae* (Sims, Brettin et al., 2009). A common skin organism in the *Kytococcus* genus, it is closely associated with pitted keratolysis, peritonitis and at times fatal pneumonia in immunosuppressed patients (Chaudhary and Finkle, 2010). Formerly known as *Micrococcus sedentarius*, *Kytococcus spp* to penicillin G and methicillin. While susceptible to bacitracin and erythromycin these grow more slowly than the other members of *Micrococcus* (Ertam, Aytimur et al., 2005; Sims, Brettin et al., 2009).

The purpose of this study was to annotate the genes in *Kytococcus sedentarius* responsible for coding the enzymes in heme biosynthesis, especially the metabolic pathway that converts L-glutamate to uroporphyrinogen III . In plants and most bacteria, uroporphyrinogen is usually produced from glutamyl transferase (Johansson and Hederstedt, 1999). However in *Kytococcus sedentarius* there is a mutation in the gene responsible for the conversion of glutamate 1 semi-aldehyde to 5-amino-levulinate. While a number of statistical tools were put to use to find specific information on the genes, significant information was extracted with respect to the related protein families using Pfam (Bateman, Coin et al., 2004).

Cellular localization for each of the genes was determined using TMHMM, SignalIP and P-SortB (Yu, Wagner et al., 2010). The enzymatic function for individual genes was majorly determined by KEGG and Metacyc (Ogata, Goto et al., 1999). Characteristic features and biological significance with respect to the heme synthesis were determined using a motley of statistical and analytical tools and all significant findings and conclusions were recorded in the Integrated Microbial Genomes Annotation Collaboration Toolkit (*IMG-ACT*) and submitted to the JGI's Genomics and Bioinformatics Education Program developed by the Joint Genome Institute(University of California, CA).

Materials and Methods

Basic information module

The gene detail page in this module was used to gather information about the DNA coordinates, protein and DNA sequence, isoelectric point and the gene ID for the assigned genes .

Sequence based similarity module

Gene homology and conservation of protein, using the amino acid sequence was assessed using an online tool from **NCBI** called **BLAST** (Basic Local Alignment Search Tool) (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

T-COFFEE (http://www.ebi.ac.uk/Tools/msa/tcoffee/). was used to retrieve the multiple sequence alignment for the genes.

The sequence was then used to create a **Web Logo** (http://weblogo.berkeley.edu/) to look at the amino acid conservation based on height of the single letter amino acid codes stacked at both the N-terminal and the C-terminal of the amino acid sequence..

Materials and Methods

Cellular localization data module

For this module amino acid sequences were used in **TMHMM** (Transmembrane Helices Hidden Markov Model) (http://www.cbs.dtu.dk/services/TMHMM/) to determine the presence of transmembrane helices in the protein.

SignalP (http://www.cbs.dtu.dk/services/SignalP) was used to identify signal peptides in the query gene.

Results obtained from the **PSORT-B** (http://www.psort.org/psortb) analysis tool and **Phobius** (http://phobius.sbc.su.se/) were used to assert the localization of the protein, based on specific cellular component scores.

Alternative open reading frame

The **Sequence viewer for alternative ORF** was used to validate the proposed DNA coordinates called by IMG database. Alternative start and stop codons were determined and BLAST was performed for the textual output of the reading frame to find supporting evidence for alternative novel open reading frames.

Structure based evidence module

TIGRFAM (http://blast.jcvi.org/web-hmm/) was used to identify protein families and **Pfam** (http://pfam.sanger.ac.uk/search) was used to identify similarities and conserved domains between organisms. The structure of the protein was obtained using results from **PDB** (Protein Database) (http://pfam.sanger.ac.uk/search).

Enzymatic function module

This module was used to determine the gene product name and its specific role in the assigned metabolic pathway. **KEGG (**Kyoto Encyclopedia of Genes and Genomes) (http://www.genome.jp/kegg/pathway.html) was used to find the metabolic pathway and **MetaCyc** (http://metacyc.org/) was used to asses enzymatic functions based on experimentally determined metabolic pathways.

Duplication and degradation module

The gene homolog section on the details page was used to look for paralogs. Results from **Pfam** (http://pfam.sanger.ac.uk/search) were used to look for truncated proteins in the query gene. Pair wise alignment for the genes and significant hit scores were obtained and **Prosite** (http://prosite.expasy.org/) was used to determine the functionality of the open reading frame. Reports from all the three tools were then used to identify non-functional genes if present, due to degradation and duplication. Jalview was used to find the functional residues and the HMM logo was used to identify and validate the key functional residues of induvidual genes.

Horizontal gene transfer module

This module was used to find evidence for horizontal gene transfer over the course of evolution. Phylogenic distribution tree was constructed using **Phylogeny** (http://www.phylogeny.fr/).

Results

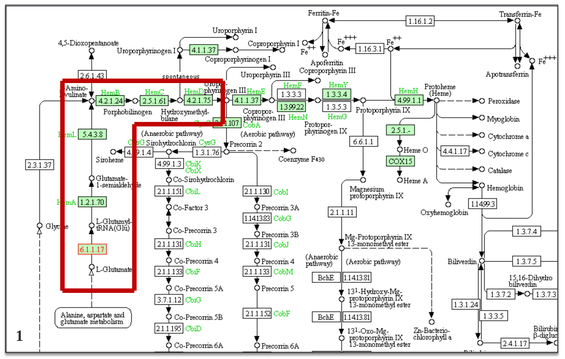

Figure 1. Pathway schematic in Kytococcus sedentarius tracing the production of uroporphyrinogen III. E.C. numbers and enzymes responsible for each step are mentioned along the lines for quick reference.

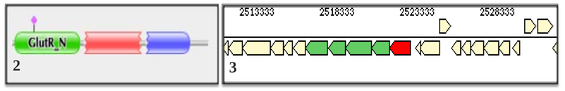

Figure 2. Pfam schematic for gene ID 644992667.

Figure 3. Sequential alignment of the proteins in the gene neighborhood responsible for uroporphyrinogen III synthesis in *Kytococcus sedentarius.*

Results

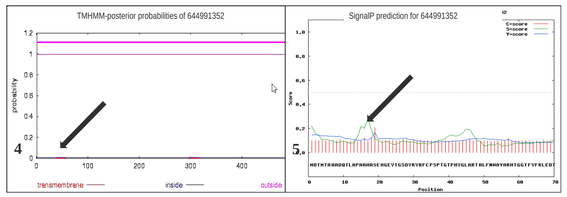

Figure 4. Results of TMHMM show zero probability(arrow) for the presence of transmembrane helices.

Figure 5. SignalP prediction shows peaks less than 0.75 which indicates very low probability of signal peptides in the gene .

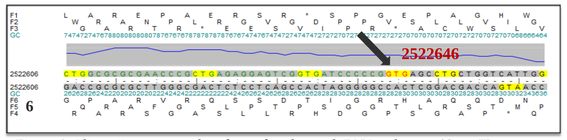

Figure 6. Alternative open reading frame for glutamyl tRNA reductase (Gene ID 644992667) shows a start codon at the proposed DNA coordinates but does not indicate a shine dalgarno sequence within 10 base pairs ,neither upstream nor down stream.

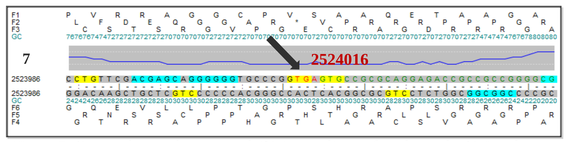

Figure 7. Stop codon at the proposed DNA coordinates for glutamyl tRNA reductase (Gene ID 644992667) .

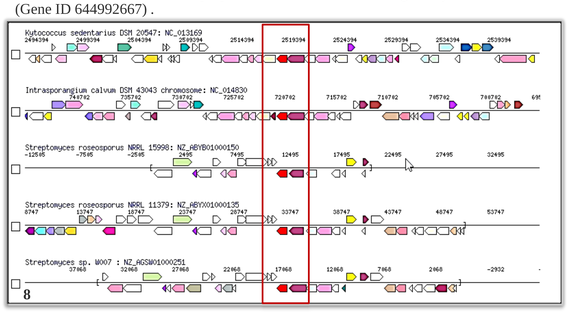

Figure 8. Gene neighborhood of the query gene with top COG hits with porphobilinogen synthase (Gene ID 644992664). Highlighted region in the box show similar enzymes for all species indicating highly conserved functions among organisms. Enzymatic functions for the purple clusters next to the genes similar to the query gene showed putative uroporphyrinogen III synthase which indicate a high degree of conservation.

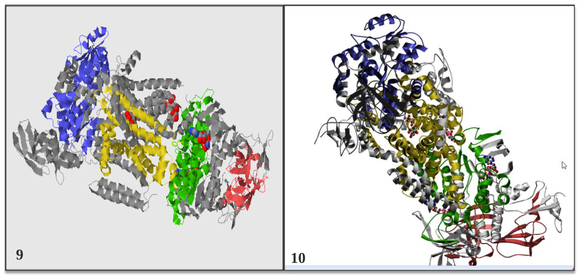

Figure 9. 3-dimensional protein structure for glutamate-1-semialdehyde2,1-aminomutase (gene bank ID 644992663) retrieved from the protein databank.

Figure 10. Jmol view of the crystal structure for the glutamate-1-semialdehyde2,1-aminomutase (gene bank ID 644992663) gene.

Results

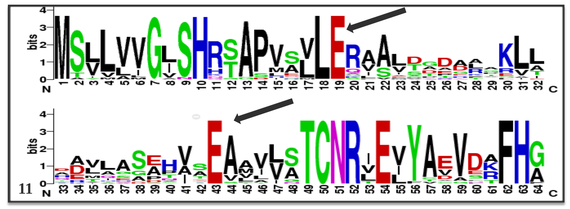

Figure 11. The WebLogo shows the conserved amino acids in the multiple aligned sequences from T-COFFEE. Stacked height and width for different single lettered amino acids is directly proportional to their degree of conservation. of amino acids.
The arrows indicate the amino acid that is most highly conserved in the sequence.

Conclusions

Gene object id	Product name	Amino acid length	Description	Pathways	E.C. Number
644991352	Glutamyl-tRNA synthetase	532 aa	Glutamyl-tRNA synthetase, bacterial family	Amino-acyl tRNA biosynthesis	6.1.1.17
644992667	Glutamyl-tRNA reductase	456 aa	Glutamyl-tRNA reductase	HemA glutamyl-tRNA reductase	1.2.1.70
644992663	Glutamate-1-semialdehyde 2,1-aminomutase	448 aa	Aminotransferase class-III	HemL glutamate-1-semialdehyde 2,1-aminomutase	5.4.3.8
644992664	Porphobilinogen synthase	340 aa	Delta-aminolevulinic acid dehydratase	Porphobilinogen synthase	4.2.1.24
644992666	Hydroxymethylbilane synthase	358 aa	Porphobilinogen deaminase	HemC, HMBS hydroxymethylbilane synthase	2.5.1.61
644992665 12	Uroporphyrinogen-III synthase	558 aa	Tetrapyrrole (Corrin/Porphyrin) Methylases	Uroporphyrinogen-III C-methyltransferase	4.2.1.75

Figure 12 : Table showing the genes involved in the synthesis of uroporphyrinogen III in *Kytococcus sedentarius*, the gene bank ID, specific enzymatic functions and the enzyme classification number.

Annotating the six genes in Kytococcus sedentarius responsible for the synthesis of uroporphyrinogen III using different modules form the IMG-ACT database suggested that the genes were functionally active . Results from the sequence based similarity module provided evidential information indicating that the genes were remarkably conserved among similar organisms. DNA coordinates for all the genes were actual except for glutamyl tRNA reductase (Gene bank ID 644992667) for which the start codon proposed by the gene caller could not be validated. KEGG and MetaCyc substantiated the metabolic functions of all genes. Proposed annotations for all the genes were reasserted and their metabolic functions in uroporphyrinogen III synthesis were confirmed.

References

1. Bateman, A., et al. (2004). "The Pfam protein families database." Nucleic acids research 32(suppl 1): D138-D141.
2. Chaudhary, D. and S. N. Finkle (2010). "Peritoneal dialysis-associated peritonitis due to Kytococcus sedentarius." Perit Dial Int 30(2): 252-253.
3. Collins, J. E., et al. (2003). "Reevaluating human gene annotation: a second-generation analysis of chromosome 22." Genome Res 13(1): 27-36.
4. Ertam, İ., et al. (2005). "Isolation of Kytococcus sedentarius from a case of pitted keratolysis." Ege Tıp Dergisi 44: 117-118.
5. Johansson, P. and L. Hederstedt (1999). "Organization of genes for tetrapyrrole biosynthesis in gram-positive bacteria." Microbiology 145(3): 529-538.
6. Lio, P., et al. (2012). "Statistical approaches to use a model organism for regulatory sequences annotation of newly sequenced species." PLoS One 7(9): e42489.
7. Ogata, H., et al. (1999). "KEGG: Kyoto encyclopedia of genes and genomes." Nucleic acids research 27(1): 29-34.
8. Sims, D., et al. (2009). "Complete genome sequence of Kytococcus sedentarius type strain (541)." Stand Genomic Sci 1(1): 12-20.
9. Venter, J. C., et al. (2001). "The sequence of the human genome." Science 291(5507): 1304-1351.
10. Yu, N. Y., et al. (2010). "PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes." Bioinformatics 26(13): 1608-1615.