Characterization of TPP-binding proteins in *Methanococci* archaeal species

Laura K. Harris¹,²*

¹Department of Science, Davenport University, Lansing, Michigan, United States of America; ²Department of Health Informatics, Rutgers School of Health Professions, Newark, New Jersey, United States of America; Laura K. Harris - E-mail: laura.harris@davenport.edu; *Corresponding author

Received August 22, 2016; Accepted November 5, 2016; Published November 22, 2016

Abstract:
Acetolactate synthase (ALS) is a highly conserved protein family responsible for producing branched chain amino acids. In *Methanocaldococcus jannaschii*, two ALS proteins, MJ0277 and MJ0663 exist though variations in features between them are noted. Researchers are quick to examine MJ0277 homologs due to their increased function and close relationship, but few have characterized MJ0663 homologs. This study identified homologs for both MJ0277 and MJ0663 in all 15 *Methanococci* species with fully sequenced genomes. EggNOG database does not define four of the MJ0663 homologs, JH146_1236, WP_004591614, WP_018154400, and EHP89635. BLASTP comparisons suggest these four proteins had around 30% identity to MJ0277 homologs, close to the identity similarities between other MJ0663 homologs to the MJ0277 homologous group. ExPASY physiochemical characterization shows a statistically significant difference in molecular weight and grand average hydropathy between homologous groups. CDD-BLAST showed distinct domains between homologous groups. MJ0277 homologs had TPP_AHAS and PRL06276 while MJ0663 homologs had TPP_enzymes super family and IlvB domains instead. Multiple sequence alignment using PROMALS3D showed the MJ0277 homologs a tighter group than MJ0663 and its homologs. PHYLIP showed these homologous groups as evolutionarily distinct yet equal distance from bacterial ALS proteins of established structure. The four proteins EggNOG did not define had the same features as other MJ0663 homologs. This indicates that JH146_1236, WP_004591614, WP_018154400, and EHP89635, should be included in EggNOG database cluster arCOG02000 with the other MJ0663 homologs.

Background:
The *Methanococci* class of archaeal organisms currently consists of 15 coccoid methanogens according to the National Center for Biotechnology Information (NCBI). *Methanococcus jannaschii* was the first fully sequenced archaea, leading to interesting revelations on the similarities across domains [1]. *M. jannaschii*, a hyperthermophilic organism isolated from the base of a hydrothermal vent on the East Pacific Rise that emits lighter-hued minerals containing barium, silicon, and calcium, grows best at 85°C with high pressure [2]. The National Center of Biotechnology Information (NCBI) re-classified it to the *Methanocaldococcus* genus alongside six other *Methanococcus* organisms due to their ability to thrive at high temperatures alongside a low 16S rRNA sequence similarity with their five-mesophilic relatives that retained the *Methanococcus* name. According to the UCSC Genome Browser, *M. jannaschii* and *M. fervens* are close, followed by *M. vulcanius*, *M. infernus*, and then *M. igneus* [3]. *M. aerolicus* forms a separate branch that includes *M. vanneilli* and *M. maripaludis*. *M. okinawensis* fits in between *M. jannaschii* and *M. aerolicus* while *M. voluae* groups with *M. aerolicus*. Interestingly, *M. thermolithotrophicus* has similar 16S rRNA sequence similarity to Methanococcus organisms, except it is thermophilic. Together they make up the *Methanococcaceae* family. According to NCBI, the thermophilic nature of *M. thermolithotrophicus* connects the mesophilic *Methanococcus* genus to their two thermophilic *Methanotorris* relatives.
When *M. jannaschii* was sequenced, open reading frame numbers 0277 and 0663 corresponding to locations relative to the ori, were assigned as genes encoding large sub-units of acetohydroxy acid synthase (EC 4.1.3.18, AHS) based on the algorithm by NCBI called the Basic Local Alignment Search Tool (BLAST). AHS assists with the production of branched chain amino acids: leucine, isoleucine, and valine [4-5]. Currently for *M. jannaschii* DSM2661, NCBI currently states that MJ0277 and MJ0663 are acetolactate synthase (ALS) large subunits. The Gene Ontology Consortium shows AHS and ALS (EC 2.2.1.6) to be synonymous [6]. AHS belongs to a superfamily of thiamine pyrophosphate (TPP)-dependent enzymes capable of catalyzing a variety of reactions. No one has determined the structure of archaenal ALS, but x-ray crystal structures are available for Klebsiella pneumoniae (1OZF) and Bacillus subtilis (4RJI) in the Protein Data Bank.

ALS is highly conserved across domains. Bowen showed through phylogenetic analysis that AHS (ALS) diverged from the other TPP-binding enzymes prior to the split between archaenal and bacterial lineages [7]. Therefore, it is not surprising that researchers detect AHS (ALS) activity in the cell extracts from several Methanococci species including Methanococcus aeolicus, Methanococcus maripaludis, and Methanococcus voltae [8-10]. However, data from several researchers suggest that MJ0277 and MJ0663 are different from each other. Phylogenetic studies show MJ0277 and an AHS (ALS) from Methanococcus aeolicus related to AHS (ALS) proteins from bacterial and eukaryotic species more closely than to MJ0663 and MJ0663 did not look related to other bacterial or eukaryotic TPP-binding proteins like AHS (ALS) or pyruvate oxidase [7]. Gardner showed that the amino acid sequence for MJ0277 was more similar than MJ0663 when compared to ilvB in Methanococcus maripaludis, 72.9% and 31.4%, respectively [10]. Because of these differences, Universal Protein Resource (Uniprot) currently calls MJ0663 an uncharacterized protein whereas MJ0277 reads as ilvB.

The MJ0277 protein and its homologs have received much attention due to their clear membership in the ALS protein family. Because of its differences, MJ0663 has not received the same focus so to date there are no studies on the homologs of MJ0663 in the literature. However, the EggNOG 4.5 database of orthologous groups and functional annotation shows MJ0663 as belonging to two clusters of archaenal orthologous groups (COG): COG0028 and arCOG02000 (TPP-binding proteins). COG0028 has over 5000 ALS proteins from more than 1700 species across domains whereas arCOG02000 had proteins from 11 Methanococci species. NCBI taxonomy currently lists 15 Methanococci species. If MJ0663 and its homologs are part of a conserved protein family, as EggNOG suggests, there should be identifiable homologs in the four species not currently included in the EggNOG database.

The purpose of this study is to use in silico methods to identify and characterize homologs of either MJ0277 or MJ0663 in Methanococci species. Since there are notable differences between MJ0277 and MJ0663 and prior research suggests they belong to two different, yet related, protein families, any new homologs should have similar observable differences. These analyses would confirm current information about these protein sub-families, further the understanding of the relatedness of ALS-related TPP-binding proteins in Methanococci archaenal species, and improve public database accuracy.

**Methodology:**
Both protein sequences for MJ0277 and MJ0663 underwent a NCBI protein-protein BLAST (BLASTP) with each individual Methanococci species to identify homologous groups. Table 1 lists the identified homologs from each organism. The sequences for all Table 1 proteins plus ALS proteins from Klebsiella pneumoniae and Bacillus subtilis, Protein Data Bank entries 1OZF and 4RJI, respectively were downloaded from NCBI.

The Expasy Protparam server calculated several physicochemical characteristics for each protein including number of amino acids, amino acid composition and frequencies, molecular weight, and the total number of charged residues (aspartic acid plus glutamic acid for positively charged and the sum of arginine and lysine for negatively charged) [11]. From that, the program calculates the theoretical isoelectric point, which is the pH where a molecule carries no net electrical charge. The algorithm also determines the amount of light a protein absorbs at a 280nm wavelength also known as the extinction coefficient, which is helpful for purification procedures [12]. ExPASy calculated the relative volume of a protein occupied by open side chain amino acids as the aliphatic index [13]. The grand average hydropathy (GRAVY) is the sum of hydropathy values of all amino acids in the protein divided by the number of residues [14]. Therefore, GRAVY relates to the extent of hydrophobicity for a given molecule. Minitab calculated the statistical significance using the Chi-squared (c²) Goodness of Fit (one variable) analyses.

Both Pfam and the conserved domain database (CDD) identified domains. Pfam is a comprehensive collection of multiple sequence alignments and Hidden Markov Models (HMMs) that represent protein domains and families [15-16]. PfamA is a set of manually curated and annotated models each based on a seed alignment and an automatically created full alignment. The seed alignment contains a group of proteins in the same family while the full alignment contains all noticeable protein sequences belonging to the family as defined by HMMs searches of primary sequence databases. Within NCBI lies another complimentary program for domain identification. The CDD is searchable using a protein query via the CD-Search interface. This algorithm uses Reversed Position Specific BLAST (RPS-BLAST), a Position-Specific Iterative (PSI)-BLAST variant, to establish position-specific scoring matrices.
with the protein sequence [17]. Together, Pfam and CDD-BLAST examine protein domains.

PROMALS3D used whole protein sequences in FASTA format for three multiple sequence alignments, one with MJ0277 homologs only, one with MJ0663 homologs only, and one with all proteins in Table 1 plus the *Klebsiella pneumoniae* (1OZF) and *Bacillus subtilis* (4RJI). The analyses used PROMALS3D’s default settings [18].

Similarly, ClustalW aligned all Table 1 proteins plus 1OZF and 4RJI whole protein sequences in FASTA format for input into the PHYLIP package Protdist program to produce a distance matrix using default settings such as the Jones-Taylor-Thornton matrix distance model [19-20]. Neighbor, another program in the PHYLIP suite, used this matrix to construct a neighbor joining and unweighted pair group method with arithmetic mean trees. The Fitch-Margoliash and Least-Squares Distance method, another phylogenetic tree building approach, verified the results. The program DrawTree illustrated all phylogenetic trees.

**Table 1:** A dataset of Methanothermococcus archaeal proteins studied

<table>
<thead>
<tr>
<th>Name of organism including strain</th>
<th>NCBI Protein Locus Tag</th>
<th>NCBI Assigned Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methanothermococcus jannaschii DSM 2661</td>
<td>Not applicable</td>
<td>Not applicable</td>
</tr>
<tr>
<td>Methanothermococcus bethoarodescens JH146</td>
<td>JH146_RS01170</td>
<td>JH146_RS03630</td>
</tr>
<tr>
<td>Methanothermococcus sp. FS406-22</td>
<td>MFS40622_RS01340</td>
<td>MFS40622_RS07655</td>
</tr>
<tr>
<td>Methanothermococcus fervens AG86</td>
<td>MEFER_RS00680</td>
<td>MEFER_RS04855</td>
</tr>
<tr>
<td>Methanothermococcus vulcanius M7</td>
<td>METIN_RS01430</td>
<td>METIN_RS00550</td>
</tr>
<tr>
<td>Methanothermococcus infernus ME</td>
<td>MAEO_RS07420</td>
<td>MAEO_RS01405</td>
</tr>
<tr>
<td>Methanothermococcus aelolicus Nankai-3</td>
<td>MVOL_RS06100</td>
<td>MMARC7_RS08600</td>
</tr>
<tr>
<td>Methanothermococcus voltae A3</td>
<td>MEVAN_RS07925</td>
<td>MEVAN_RS05905</td>
</tr>
<tr>
<td>Methanothermococcus maripaludis C7</td>
<td>WP_018154052</td>
<td>WP_004591614</td>
</tr>
<tr>
<td>Methanothermococcus maripaludis SB</td>
<td>WP_0188835</td>
<td>WP_018154052</td>
</tr>
<tr>
<td>Methanothermococcus okinawensis IH1</td>
<td>METOK_RS01215</td>
<td>METOK_RS01215</td>
</tr>
<tr>
<td>Methanothermococcus igneus Kol 5</td>
<td>METIC_RS03470</td>
<td>METIC_RS00805</td>
</tr>
</tbody>
</table>

**Note:** ALS, acetolactate synthase; NCBI, National Center for Biotechnology Information; TPP, thiamine pyrophosphate.
Figure 1: Alignment of MJ0277 homologs aligned by PROMALS3D. Magenta names are representative sequences colored red to identify predicted alpha-helix secondary structures. The black names belonging to the same alignment group as the magenta name above it, indicating a strong relationship between the two. Consensus_aa, consensus amino acid sequence; Consensus_ss, consensus predicted secondary structures; h, consensus predicted secondary structure alpha-helix.
Methanococcus species, every species, listed in Table 1. The MJ0663 homologs for four proteins were new proteins related to MJ0663, they averaged 97% query coverage (SD ±1%) with 29% identity (SD ±0.4%). Alternatively, when MJ0663 was compared with its homologs, they achieved an average 97% query coverage (SD ±3%) with 64% identity (SD ±13%) to the target protein sequence. When MJ0663 and its homologs were compared to MJ0663, they averaged 97% query coverage (SD ±1%) with 29% identity (SD ±0.4%). Alternatively, when MJ0663 was compared with its homologs, they achieved an average 97% query coverage (SD ±3%) with 64% identity (SD ±13%) to the target protein sequence. When MJ0663 and its homologs were compared to MJ0663, they averaged 97% query coverage (SD ±1%) with 29% identity (SD ±0.4%). Alternatively, when MJ0663 was compared with its homologs, they achieved an average 97% query coverage (SD ±3%) with 64% identity (SD ±13%) to the target protein sequence.
Figure 2: Alignment of MJ0663 homologs aligned by PROMALS3D. Magenta names are representative sequences colored red to identify predicted alpha-helix secondary structures. The black names belonging to the same alignment group as the magenta name above it, indicating a strong relationship between the two. Consensus_aa, consensus amino acid sequence; Consensus_ss, consensus predicted secondary structures; h, consensus predicted secondary structure alpha-helix.
Figure 3: Representative phylogenetic tree of proteins produced via PHYLIP package programs showing members of the acetolactate synthase (ALS) family. Nomenclature of genes is consistent with Table 1. The solid arrow highlights MJ0277 while the dashed arrow points to MJ0663. This illustrates that MJ0277 and MJ0663 are closely related to their respective homologs from other Methanococci species, but are different from each other. Both groups are equally distant from experimentally established bacterial ALS proteins.

**Physiochemical Characterization**
Table 2 summarizes several physiochemical characterizations. MJ0277 and homologs averaged 595 amino acids (SD ±8) while MJ0663 and homologs averaged 506 amino acids (SD ±26, p=0.298). Molecular weight reflected similar findings (64951 ±941 versus 56498 ±2653 for the MJ0277 and MJ0663 groups, respectively, p=0.000). The groups had similar theoretical isoelectric point averages (p=1.000). This was not surprising because the MJ0277 and MJ0663 groups had an average amino acid composition of 12.4% and 12.1% for negatively charged amino acids alongside...
10.8% and 11.4% for positively charged residues. Extinction coefficients and aliphatic index between the groups were unremarkable (p=1.000), but there was a difference in hydropophicity as seen in the average GRAVY results (-0.05 ±0.03 versus -0.23 ±0.08 for the MJ0277 and MJ0663 groups, respectively, p=0.000). These results indicate that the two protein families are similar in physiochemical properties but have some identifiable differences in molecular weight and hydroporphicity.

**Domain Characterization**
Both Pfam and CDD-BLAST algorithms identified domains for these TPP-binding proteins. Pfam-A results did not show a difference between the groups. All proteins except METIN_RS00550 had Thiamine pyrophosphate enzyme N-terminal binding, central, and C-terminal binding domains corresponding to clans CL0254, CL085, and CL0254, respectively. Protein METIN_RS00550 was missing the TPP enzyme N-terminal domain, but had the other two domains.

The CDD-BLAST database identified a difference between the groups. While CDD-BLAST assigned TPP_PYR_POX_like and TPP_enzyme_M domains to all proteins regardless of group, MJ0277 and its homologs had TPP_AHAS and PRK06276 domains whereas MJ0663 and its homologs had TPP_enzymes super family and IlvB domains instead. The TPP_enzyme_M domain came from the Pfam database, so it is interesting that Pfam itself did not assign this domain to METIN_RS00550, yet CDD-BLAST did. Both of the two domains specific to MJ0277 and its homologs were ALS related. The TPP_AHAS domain referred to the AHS (ALS) subfamily of TPP-binding proteins, comprised of proteins similar to the large catalytic subunit of AHS [21]. NCBI defines PRK06276 as an ALS catalytic subunit. The domains specific to MJ0663 and its homologs were not as function specific. The TPP_enzymes super family domain simply referred to these proteins having TPP-binding module found in many key metabolic enzymes that use TPP as a cofactor. The IlvB domain indicated an ALS large subunit or other TPP-requiring enzyme related to amino acid or coenzyme transport and metabolism.

**Sequence Alignment**
MJ0277 and its homologs were aligned together and separate from MJ0663 and its homologs. Figures 1 and 2 shows the multiple alignment data for separate homologous group alignments. MJ0277 and its homologs are more conserved than MJ0663 is with its homologs as seen by the magenta color illustrating representative alignment sequences. This is further illustrated comparing consensus sequences between groups. In both groups, the N-terminus is more closely conserved than the rest of the protein as seen by the consensus sequence. The consensus sequence for the MJ0663 homologous group becomes ill defined for the central and C-terminus whereas the consensus sequence for MJ0277 homologous group is well defined throughout the protein.

**Phylogeny Characterization**
To examine phylogenetic relatedness, various PHYLIP programs analyzed a CLUSTALW alignment of all 30 proteins and two bacterial ALS proteins with established structure to produce phylogenetic trees. Different algorithms with ProtDist were used, as were different tree building methods. All trees looked similar to Figure 3, illustrating how related MJ0277 and its homologs are to each other yet are distant to MJ0663 at its homologs with both groups equal distance to bacterial ALS proteins. These results support those from PROMALS3D.

**Conclusion**
For each of 15 Methanococcaceae species with genomes available on NCBI, there was an identifiable homolog for both MJ0277 and MJ0663. Four MJ0663 homologs JH146_1236, WP_004591614, WP_018154400, and EHP89635 from species Methanocaldococcus bethoarsdenscens, Methanocaldococcus villosus, Methanothermococcus thermolithotrophicus, and Methanotorriss formicicus, respectively, are proteins not included in EggNOG database cluster arCOG02000. BLASTP comparisons suggest these homologs had a 30% identity to MJ0277 homologs, similar to identity similarities between other MJ0663 homologs to the MJ0277 homologous group. ExPASy characterization showed the physiochemical chemical properties such as molecular weight and GRAVY are significantly similar among MJ0663 homologs but not MJ0277 homologs. CDD-BLAST identified two domains common among all MJ0663 homologs that are not present in MJ0277 homologs and vice versa. MJ0277 homologs had TPP_AHAS and PRK06276 while MJ0663 homologs had TPP_enzymes super family and IlvB domains instead. Multiple sequence alignment analysis showed all MJ0277 homologs as closely related but there are subtle differences among MJ0663 homologs. The consensus sequence for the MJ0663 homologous group becomes ill defined for the central and C-terminus whereas the consensus sequence for MJ0277 homologous group is well defined throughout the protein. PHYLIP illustrated MJ0277 and its homologs as phylogenetically related but the group is separate from their conserved MJ0663 relatives. Both homologous groups are equally distant to bacterial ALS proteins of established structure. These results support those from multiple sequence alignment. Ergo, the four MJ0663 homologs identified here, JH146_1236, WP_004591614, WP_018154400, and EHP89635, should be included in EggNOG database cluster arCOG02000.

**References**
Herring PA and Jackson JH. Microb Comp Genomics. 2000 5:2 [PMID: 11087175]


Edited by PKangueane

Citation: Harris, Bioinformation 12(8) 359-367 (2016)

License statement: This is an Open Access article which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. This is distributed under the terms of the Creative Commons Attribution License