Ibn Al Haitham Journal for Pure and Applied Science

Journal homepage: http://jih.uobaghdad.edu.iq/index.php/j/index

Automated Reconstruction and Manual Curation of Amino Acid Biosynthesis Pathways in *Sulfolobus solfataricus P2*

Barzan Ibrahim Khayatt

Department of Natural Resources, College of Agricultural Sciences Engineering, University of Sulaimani, Sulaimani, Kurdistan-IRAQ

barzan.khayatt@univsul.edu.iq

Article history: Received 10 June 2019, Accepted 10 July 2019, Publish September 2019.

Doi:10.30526/32.3.2271

Abstract

The efficient sequencing techniques have significantly increased the number of genomes that are now available, including the Crenarchaeon *Sulfolobus solfataricus P2* genome. The genome-scale metabolic pathways in *Sulfolobus solfataricus P2* were predicted by implementing the "Pathway Tools" software using MetaCyc database as reference knowledge base. A Pathway/Genome Data Base (PGDB) specific for *Sulfolobus solfataricus P2* was created. A curation approach was carried out regarding all the amino acids biosynthetic pathways. Experimental literatures as well as homology-, orthology- and context-based protein function prediction methods were followed for the curation process. The "PathoLogic" component of the "Pathway Tools" programme was able to predict many amino acid biosynthetic metabolic pathways. The total number of the metabolic pathways was modified to 168 pathways by adding extra pathways that have not been detected by the "PathoLogic". Amino acid biosynthetic pathways such as alpha-aminoadipic acid (AAA) pathway of Lysine biosynthesis and Alanine biosynthesis as well as the super-pathway of Phenylalanine, Tyrosine and Tryptophan biosynthesis variation II were added to the Pathway/Genome data base of *Sulfolobus solfataricus P2*. Discovery of the missing enzymes that have to fill in the metabolic holes in the pathways under study was the main curation task. This approach and the curated amino acid biosynthetic pathways in the PGDB of *Sulfolobus solfataricus P2* can be used for genomic annotations and metabolic pathway reconstructions of closely related Bacteria and Archaea.

Keywords: Crenarchaeon, Pathway Tools, Amino acid biosynthesis, Metabolic pathways reconstruction, Protein function prediction. (Supplementary Figures can be found at: https://figshare.com/s/8bea91e91cb6ce193fee)

1. Introduction

Ribosomal RNA phylogenetic analysis, translation, transcription, and replication-involving protein studies positioned Archaea as a distinct domain with Bacteria and Eukarya. Being distanced from either Bacteria or Eukaryotes does not mean the absence of common features among these life domains. It has been noted that Archaea have a eubacterial form and a

eukaryotic content [1]. Most archaeal proteins participating in DNA replication are more similar in sequence to those in Eukarya than to analogous replication proteins in Bacteria [2]. The absence of a nucleus and the genome organization especially the operonic organization of many genes in a single circular chromosome are the important common features with Bacteria. The complete genome of *Sulfolobus solfataricus P2* was released in 2001 with a single chromosome of 2,992,245 bp, holding 3032 identified genes that code for 2,997 proteins [3]. (for updates see NCBI RefSeq: NC_002754.1). Approximately 1/3 of the proteins has no homologs in other sequenced genomes, 40% appear to be archaeal-specific and only 12% and 2.3% are shared exclusively with Bacteria and Eukaryotes, respectively (same reference). *Sulfolobus solfataricus* as an important model organism belongs to the phylum Crenarchaeota of the life domain Archaea. It inhabits extreme soil and water environmental niches with acidic (pH 2 - 4) and high temperature conditions (around 80^{o} C). *Sulfolobus solfataricus* as a thermophilic Archaea, its proteins remain functional in high temperatures, which can be beneficial in food, synthetic industry, biotechnology, soil and water environmental researches.There is so far limited experimental work regarding the archaeal-specific metabolic pathways and the relating genes characterization. Variant metabolic pathways have evolved in Archaea, whereas the central metabolic fluxes of Bacteria and Eukaryotes are generally well conserved [4].

Automated prediction of metabolic pathways via the entire genome sequence of an organism is referred to as metabolic reconstruction [5]. Reference knowledge bases (KBs), especially metabolic pathway-specific databases such as MetaCyc [6-9]. are efficiently adopted in metabolic reconstructions of target organisms. Pathway/Genome databases (PGDBs) specific for different organisms are created through the use of “Pathway Tools” software [10,11]. These databases are collections of the annotated genome sequences of the target organisms as well as their reconstructed metabolic pathways including information on compounds, intermediates, cofactors, reactions, genes and their products. The recent record of the PGDBs that are now collected in the BioCyc database is 14560 [12]. In general, the initial non-curated PGDBs lack many chemical pathways that are actually present in the corresponding organism or conversely consist of false positives. The addition of the absent pathways and removing those that are really not exist in an individual organism help in the quality improvement of the genome annotation. If the corresponding genes encoding the majority enzymes in a metabolic pathway have been identified in the annotated genome, the missing steps are likely to be present amongst the unidentified genes and are worth to be identified and assigned. If a protein has not been assigned a specific function during the annotation process (annotations fail to assign function to 40 – 60% of the newly sequenced proteins), any reaction catalyzed by that protein would appear as a missing enzyme (pathway hole) in a PGDB [13].

Few available research works regarding metabolic reconstructions for the model thermoacidophilic archaeon *Sulfolobus solfataricus*, focus on the metabolic networks response to different carbon sources and central carbohydrate metabolism [14-16].

The majority of the genes that are known to be involved in different amino acid biosynthetic pathways were detected in *Sulfolobus solfataricus P2* genome (http://www-archbac.u-psud.fr/projects/sulfolobus/Amino_acids_biosynth.html). Depending on these detected genes, one cannot assume that *Sulfolobus solfataricus* is able to synthesize all amino acids, despite the signals of that ability (see [17]). There are very limited lab results so far to get complete

insight into the amino acids biosynthesis in *Sulfolobus solfataricus.* They do not exceed identification and characterization of some individual enzymes catalyzing reactions in some amino acid biosynthetic pathways. There is diversity in amino acid biosynthetic pathways even among the archaeal species. For example the cysteine biosynthesis pathway in *Methanosarcina barkeri* is achieved by a bacteria-like *cysK* gene encoding O-acetylserine(thiol)-lyase-A and an adjacent to *cysK* gene known as *cysE* gene (serine transacetylase) [18]. Another bacteria-like *cysM* gene encoding O-acetylserine(thiol)-lyase-B is detected in *Sulfolobus solfataricus* [3]. Whereas no orthologs for this gene can be detected in *Methanocaldococcus jannaschii, Methanobacterium thermoautotrophicum and Archaeoglobus fulgidus*. This can be interpreted by the existence of unrecognizable genes catalyzing the same function (Non-orthologous gene displacement) or a different cysteine biosynthesis pathway. The genes encoding histidine biosynthesis pathway, were detected in *Sulfolobus solfataricus* genome (Sso0592– 0600 and Sso6227), they show a novel organization in histidine operon, *hisCGABdFDEHI* [3]. In the Enterobacteria, the two reactions are associated with a single bifunctional polypeptide encoded by *hisB*, whereas the two reactions in Archaea are encoded by two separate genes [19].

The automated reconstruction of the metabolic pathways regarding amino acids biosynthesis in *Sulfolobus solfataricus P2* showed many enzymatic gaps (pathway holes). The current research intends filling in these gaps by the discovery of the genes and the enzymes they code for the enzymatic reactions in the amino acid synthetic pathways in *Sulfolobus solfataricus P2* strain.

2. Materials and Methods

2.1. *Sulfolobus solfataricus* PGDB Creation

The “Pathway Tools” software [10]. was used to create a Pathway/Genome Database specific for *Sulfolobus solfataricus P2*, and was saved in a local server (a non-curated version is available at BioCyc (https://biocyc.org/organism-summary?object=SSOL273057). The initial metabolic pathways were predicted for the target organism via the use of its genome annotation file and the entire genome sequence file (NCBI Reference Sequence: NC_002754.1) as input files for the “PathoLogic” component of the “Pathways Tools” software. The Pathway/Genome Navigator component provides query, visualization, and Web-publishing services for PGDBs. The Pathway/Genome “Editors” were used for curation and interactive updating of the created PGDB. The editors are pathway editor, reaction editor, protein editor ...etc. The “Pathway Tools” used MetaCyc [7]. as a reference database in the metabolic reconstruction process. MetaCyc is a multi-organism database of experimentally elucidated metabolic pathways and the associated enzymes as well as their commission (EC) numbers. Experimental literature in metabolic pathways of *Sulfolobus solfataricus* and other closely related Archaea was used in the PGDB curation process. Metabolic pathway-specific databases such as KEGG [20,21]. were also used. PubMed, Blast and Clusters of Orthologous Groups (COGs) from (NCBI, (http://www.ncbi.nlm.nih.gov) were also used in the curation. Other databases such as the enzyme database BRENDA (https://www.brenda-enzymes.org), UniProt (https://www.uniprot.org) and the Protein Information Resource (PIR) (http://pir.georgetown.edu) were used as well.

2.2. Missing Enzymes Detection

The manual curation of the initial *Sulfolobus solfataricus P2* PGDB was initiated with the literature search for amino acid biosynthetic pathways, their reactions, reactants and the involving enzymes. By comparing the reaction sets, reactants and the enzymes, identified by wet-lab experimental literature to those in MetaCyc and the created *Sulfolobus solfataricus P2* PGDB, enabled identifying the missing steps and enzymes in the pathways (pathway holes). Known EC numbers of the missing reactions helped in filling the missing steps depending on the available experimental literature in metabolic pathways of *Sulfolobus solfataricus,* and then the corresponding proteins and their encoding genes were assigned to these reactions. Adopting one or more of the protein function prediction methods (see below), was the alternative for the lack of experimental literature. Homology search using BLASTP programme and PSI-BLAST [22]. for a specific enzyme with an assigned EC-number (its sequence has been retrieved from closely related Archaea) in *Sulfolobus solfataricus P2* genome was the main strategy in this step. Clusters of Orthologous Groups of proteins (COGs) (http://www.ncbi.nlm.nih.gov/COG/) were used to identify the potential candidates for the pathway holes (see [23]). The study of corresponding pathways in phylogenetically close-related species to *Sulfolobus solfataricus* helped in filling the gaps and also creating some other pathway variants in the *Sulfolobus solfataricus P2* PGDB. Identification of the subunits of the enzyme complexes was also included in the curation to give a complete view of some reactions that are catalyzed by complexes. The predicted missing enzymes and the encoding genes were assigned to the corresponding reactions via the "Editors" component of "Pathway Tools" software.

2.3. Protein Function Prediction

In order to assign function to an enzyme, one or a combination of the following approaches were applied during the reconstruction and curation of the amino acid biosynthetic pathways (see [24]). a) *Homology-, orthology-based function prediction;* Homology search using a protein sequence of an experimentally validated function as a query to find the homologs in the target organism genome helped in finding the missing enzymes for a specific pathway. The detected homologs likely and not necessarily have the same function as the query, that is why the orthology is adopted to be the primary step for function prediction. The orthologs being originally one gene before speciation, they likely possess the similar function. b) *Gene neighbourhood (Context-based function prediction);* in case to select which paralog to be assigned the specific function, the physical position of the gene in the genome was taken in consideration. The association of a specific paralog or a gene to a set of genes in a specific operon performing same biological process is likely to be an indicator for the performance of a specific function. The evolutionary conservation of that association in different species confirms the approach. c) *Gene fusion;* Detecting subunits of experimentally known enzyme complexes via homology search in the target genome gives strong evidence that these sequences in the target organism perform gene fusion in order to create a specific enzyme complex likely to catalyze the same reaction. In some cases a bifunctional enzyme catalyzes two different steps in a specific pathway in an organism, while two different genes catalyze the two different steps in another organism. It is assumed that the bifunctional enzyme is a result of gene fusion between the two homologs of the other organism. d) *Phylogenetic distribution (gene co-occurrence);* the occurrence of genes of unknown function with other functionally known genes in the same phylogenetic pattern in different species gives

indication to the role that gene may act. e) *Conservation of co-expression;* Hierarchical clustering in the gene expression profile studies organized genes in specific clusters. The occurrence of genes of unknown functions with genes of known functions in same cluster may help in assigning functions to the genes of unknown function. The conservation of this co-expression pattern in different species validates this approach.

3. Results and Discussion

3.1. Initial Non-curated PGDB of *Sulfolobus Solfataricus* P2

The Pathway/Genome Database (PGDB) of *Sulfolobus Solfataricus* P2 that was created by the PathoLogic component of the Pathway Tools software is a collection of the genomic information and the reconstructed metabolic pathways of *S. Solfataricus P2 strain.* The PGDB is visualized, queried and can be edited via the Pathway Tools. The created PGDB is an object-oriented database since the software creates such types of databases and not simple rational databases. All components of the database are represented as frames in the graphical display of the programme. The programme compared all the genomic information of *S. solfataricus* especially the annotated EC numbers and the enzyme names with those stored in the reference database MetaCyc. Via PathoLogic algorithm, the programme has predicted all reactions that are probably been catalyzed by these enzymes and then the inferring of the corresponding candidate metabolic pathways for *S. solfataricus.* According to the programme algorithm, a pathway is added to the database when half or more reactions of that pathway could be detected by the PathoLogic. The Home screen of the Pathway Tools shows the list of the PGDBs, including the extra copies for manually curation in order to compare the initial PGDB with the so far curated one. Initially a total of 159 pathways and 787 enzymatic reactions were predicted. From 3009 polypeptides, 582 enzymes were detected in the genome. In this approach the PGDB was manually curated for the amino acid biosynthetic pathways. The curation is not only addition of extra pathways, which are not detected by the Pathway Tools but the user may also eliminate pathways for which there is little evidence [25].

3.2. Curation of Amino Acid Biosynthetic Pathways

PathoLogic could detect many amino acid biosynthetic pathways including some super pathways. The super pathways list consists of many amino acids groups such as the aromatic and branched chain amino acids. On the other hand the individual amino acids list consists of the majority but not all amino acid biosynthetic pathways. Both the super and individual amino acids pathways contain many metabolic holes (missing enzymes) to be filled in during the curation.

3.2.1. Aromatic Amino Acid Biosynthetic Pathways (Phenylalanine, Tyrosine and Tryptophan)

In order to give more complete display to the super pathway of aromatic amino acids biosynthesis, a new version of the super pathway of phenylalanine, tyrosine and tryptophan biosynthesis was created. **Figure 1.** shows the involvement of the three amino acids in the super pathway, whereas the super pathway of the initial non-curated PGDB contains tryptophan only. From chorismate, the new super pathway version shows the production of the three amino acids. In the aromatic amino acids biosynthetic pathways the metabolic holes were identified to be: EC 2.7.1.71 (in chorismate synthesis), EC 2.6.1.5 (in phenylalanine pathway) and EC 1.3.1.12 (in tyrosine pathway) **Table 1.**

The enzyme (EC 2.7.1.71) catalyzing the fifth reaction of the seven-step chorismate synthesis pathway is an archaeal shikimate kinase (member of the GHMP-kinase family). Chorismate is the precursor for the three aromatic amino acids. Chromosomal clustering of the chorismate biosynthetic genes helped in identifying the candidate gene encoding this enzyme in *Methanocaldococcus jannaschii* and then later the enzyme was experimentally verified [26]. The archaeal shikimate kinases were found to have no sequence homology to their bacterial counterparts. In *S. solfataricus*, the enzyme is not experimentally characterized so far. The homology search with the characterized shikimate kinase from *Methanocaldococcus jannaschii* against *S. solfataricus* genome helped in finding the candidate gene for this enzyme to be Sso0308 (COG 1685).

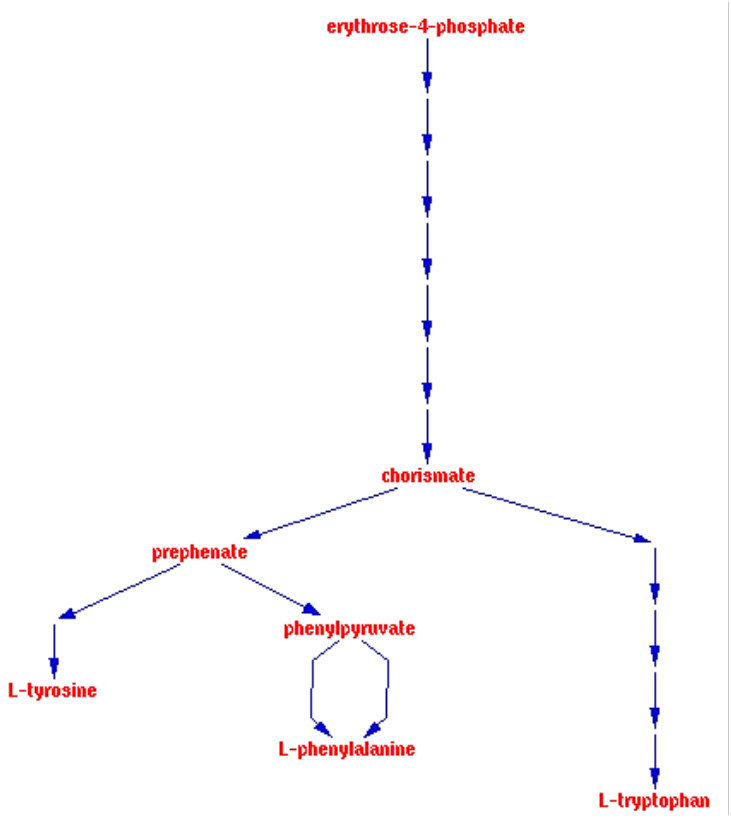

Figure 1. A new version (less-detailed display) complete super pathway of phenylalanine, tyrosine and tryptophan biosynthesis.

Table 1. Metabolic holes (missing enzymes) in the predicted super biosynthetic pathways of the amino acids groups and the pathways of the individual amino acids in the non-curated PGDB of *Sulfolobus solfataricus* P2.

Aromatic amino acids biosynthesis (phenylalanine, tyrosine and tryptophan):				
EC 2.7.1.71	Archaeal shikimate kinase		Sso0308	COG1685
EC 2.6.1.5 (Homologous to EC 2.6.1.1)	Aspartate aminotransferase	*aspB-1* *aspB-2* *aspB-3* *aspB-4*	Sso0104 Sso0897 Sso1177 Sso3245	COG1167 COG0436 COG0436 COG1167
	Tryptophan synthase beta chain 2	*trpB*	Sso1145	COG0133
EC 1.3.1.12	Archaeal prephenate dehydrogenase	*tyrA*	Sso0302	COG0287
Branched-chain amino acids biosynthesis (leucine, isoleucine and valine):				
EC 2.6.1.42	Aspartate aminotransferase	as EC 2.6.1.5		
Individual amino acids biosynthesis:				
Methionine:				
EC 2.3.1.46	Homoserine O-succinyltransferase	?	?	COG1897

EC 2.5.1.48 (formerly 4.2.99.9)	Cystathionine beta-lyase or O-succinylhomoserine (thiol)-lyase (cystathionine gamma-synthase)	*metB*	Sso2368	COG0626
EC 4.4.1.8	Cystathionine beta-lyase or O-succinylhomoserine (thiol)-lyase (cystathionine gamma-synthase)	*metB*	Sso2368	COG0626
EC 2.1.1.13	Methionine Synthase	?	?	
Arginine:				
EC 2.3.1.1	LysX/Glutamate N-acetyltransferase	*lysX /rimK-1*	Sso0159	COG0189
EC 2.6.1.13	Acetylornithine/acetyl lysine aminotransferase(ArgD/LysJ)	*argD/lysJ*	Sso0160	COG4992
Alanine: Three pathway variations were created (I, II and III)				
EC 2.6.1.66	Valine-pyruvate transaminase	*aspB-1*	Sso0104	COG1167
EC 2.6.1.2	Alanine aminotransferase (alanine aminotransaminase)	*aspB-2*	Sso0897	COG0436
Cysteine:				
EC 2.3.1.30	Serine O-acetyltransferase (CysE)	*cysE*	Sso0372	COG1045
Serine:				
EC 2.6.1.52	Phosphoserine transaminase/Serine-pyruvate aminotransferase (AgxT)	*agxT*	Sso2597	COG0075
EC 3.1.3.3	Phosphoserine phosphatase (PSP)		Sso0094	COG0560
Glycine:				
EC 3.1.3.18	Phosphoglycolate phosphatase 1 (PGP 1) Phosphoglycolate phosphatase 2 (PGP 2)		Sso0094 Sso2157	COG0561 COG0561
EC 1.1.99.14 replaced by EC 1.1.1.26	Glycolate reductase/D-3-phosphoglycerate dehydrogenase (serA-1)	*serA-1*	Sso0905	COG0111
EC 2.6.1.4	Glycine transaminase /Aspartate aminotransferase (EC 2.6.1.1)	*aspB-2*	Sso0897	COG0436
Glutamate:				
EC 1.4.1.4	NAD specific glutamate dehydrogenase (gdhA-1) (gdhA-2) (gdhA-3) (gdhA-4)	 *gdhA-1* *gdhA-2* *gdhA-3* *gdhA-4*	 Sso1457 Sso1907 Sso1930 Sso2044	 COG0334 COG0334 COG0334 COG0334
Proline:				
Unknown EC: Found to be EC 1.2.1.41	Aldehyde dehydrogenase (aldhT) glutamate-5-semialdehyde dehydrogenase	*aldhT* *?*	Sso3117 ?	COG0014 ?
Histidine:				
EC 3.6.1.31	Phosphoribosyl-ATP pyrophosphatase	*hisE*	Sso6223	COG0140
EC 3.1.3.15	Histidinol-phosphatase	?	c0849?	COG1387
Lysine: alpha-aminoadipic acid (AAA) pathway was completely created in *Sulfolobus solfataricus* PGDB				

The chorismate pathway was curated by adding the gene and its product to the corresponding EC number **Figure 2.**

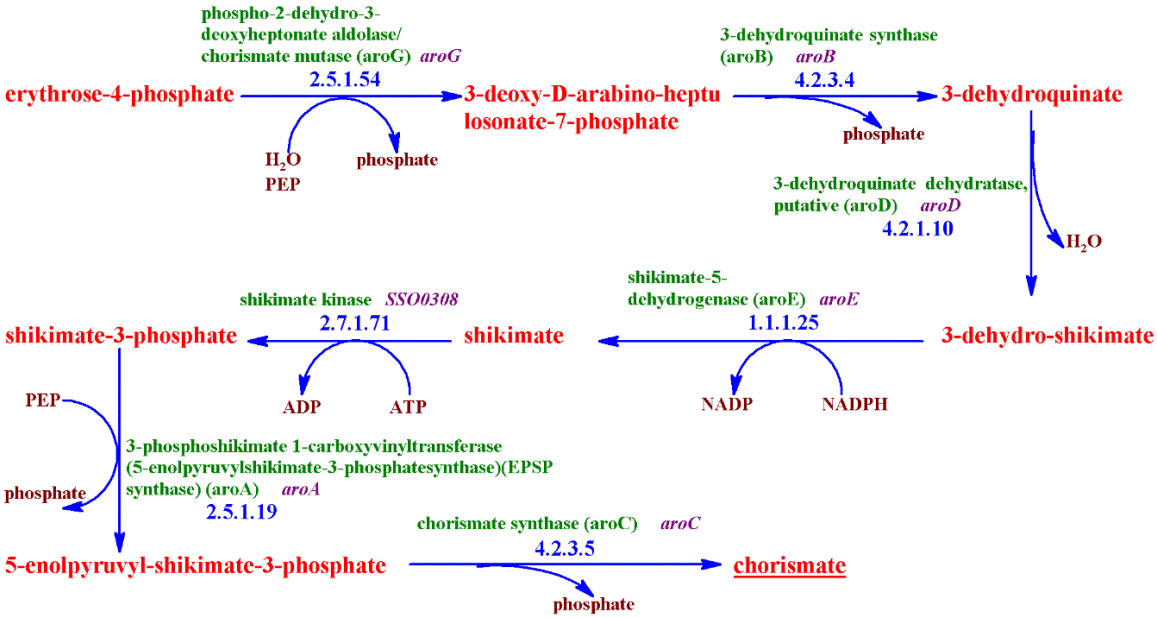

Figure 2. Curated chorismate biosynthesis.

The last step in the phenylalanine and tyrosine synthetic pathways in Eukarya and Bacrteria is catalyzed by tyrosine aminotransferase (known also as phenylalanine aminotransferase and also as L-tyrosine aminotransferase (EC 2.6.1.5)). The enzyme is not identified in *S. solfataricus*, but the PathoLogic suggested the reaction for both final steps in tyrosine and phenylalanine pathways. In order to detect the candidate gene that encodes this enzyme in *S. solfataricus*, the results of [27]. were taken in consideration. They showed besides the broad specificity of aspartate aminotransferase in an related Archaea (*Pyrococcus furiosus*), that the gene encoding the enzyme was expressed as a polycistronic message as part of the aro operon, in which genes of aromatic amino acids pathways are clustered close to the other aromatic gene cluster (*trp* operon). The homology search for tyrosine aminotransferase in *S. solfataricus P2* genome identified aspartate aminotransferase as the candidate to fill in the gap in the reaction with the corresponding EC 2.6.1.5 **Figure 3.** The gene coding for aspartate aminotransferase (EC 2.6.1.1) was already cloned and characterized in *S. solfataricus* [28]. and was found to be homologous to tyrosine aminotransferase (EC 2.6.1.5). The other homologous genes were also listed **Table 1.**

By homology search for an archaeal prephenate dehydrogenase (from *Pyrococcus furiosus* DSM 3638) in the *S. solfataricus P2* genome, the candidate gene encoding this enzyme was identified to be *tyrA* (Sso0302) in *S. solfataricus P2* (COG0287 and COG1605). In different databases such as BRENDA, KEGG and NiceZyme, this enzyme especially in the enteric Bacteria possesses EC 5.4.99.5 activity and converts chorismate into prephenate just a step before the reaction (EC 1.3.1.12) catalyzed by this enzyme in tyrosine biosynthesis pathway [29]. **Figure 3.**

3.2.2. Branched-Chain Amino Acid Biosynthetic Pathways (Leucine, Isoleucine and Valine)

The final step in leucine, isoleucine and valine biosynthesis is catalyzed by the enzyme branched-chain aminotransferase (BCAT, EC 2.6.1.42), which is specified as *ilvE*. The enzyme and its coding gene are not characterized in *S. solfataricus* due to lack of homology

even with the branched-chain aminotransferase (*ilvE*) from the Crenarchaeota *Pyrobaculum aerophilum*. Dual substrate recognition by BCATs is implemented via the "lock and key" mechanism without side-chain rearrangements of the active site residues [30]. To find the candidate gene encoding this enzyme in *S. solfataricus*, the relationship among the pathways lysine-AAA, leucine and arginine was taken in consideration. It was found that lysine-AAA gene cluster was analogous in part to the leucine and arginine biosynthetic pathways [31]. and they found that the genes evolutionarily related to lysine-AAA biosynthetic genes in *T. thermophilus* were all present in the archaea *Pyrococcus horikoshii*. From this point of view, it was found that the analogous gene to *ilvE* in leucine pathway was *lysN* (LysN, Alpha-aminoadipate aminotransferase (AAAAT), EC 2.6.1.39) in lysine-AAA pathway [32]. LysN recognizes not only 2-oxoadipate, an intermediate of lysine biosynthesis, but also 2-oxoisocaproate, 2-oxoisovalerate and 2-oxo-3-methylvalerate, intermediates of leucine, valine and isoleucine biosyntheses [33]. They found also that LysN and aspartate aminotransferase (AspAT) share the same common ancestor. When AAAAT (ST1411) and AspAT (ST1225) in S*ulfolobus tokodai* were searched in the *S. solfataricus P2* genome for their homologous genes, the both searches retrieved again the aspartate aminotransferase (*aspB*-1 "Sso0104", *aspB*-2 "Sso0897", *aspB*-3 "Sso1177", *aspB*-4 "Sso3245"). The curated super pathway of branched-chain amino acids biosynthesis (leucine, isoleucine and valine) is shown in **Figure 4.** The EC 4.1.3.12 is exchanged by EC 2.3.3.13.

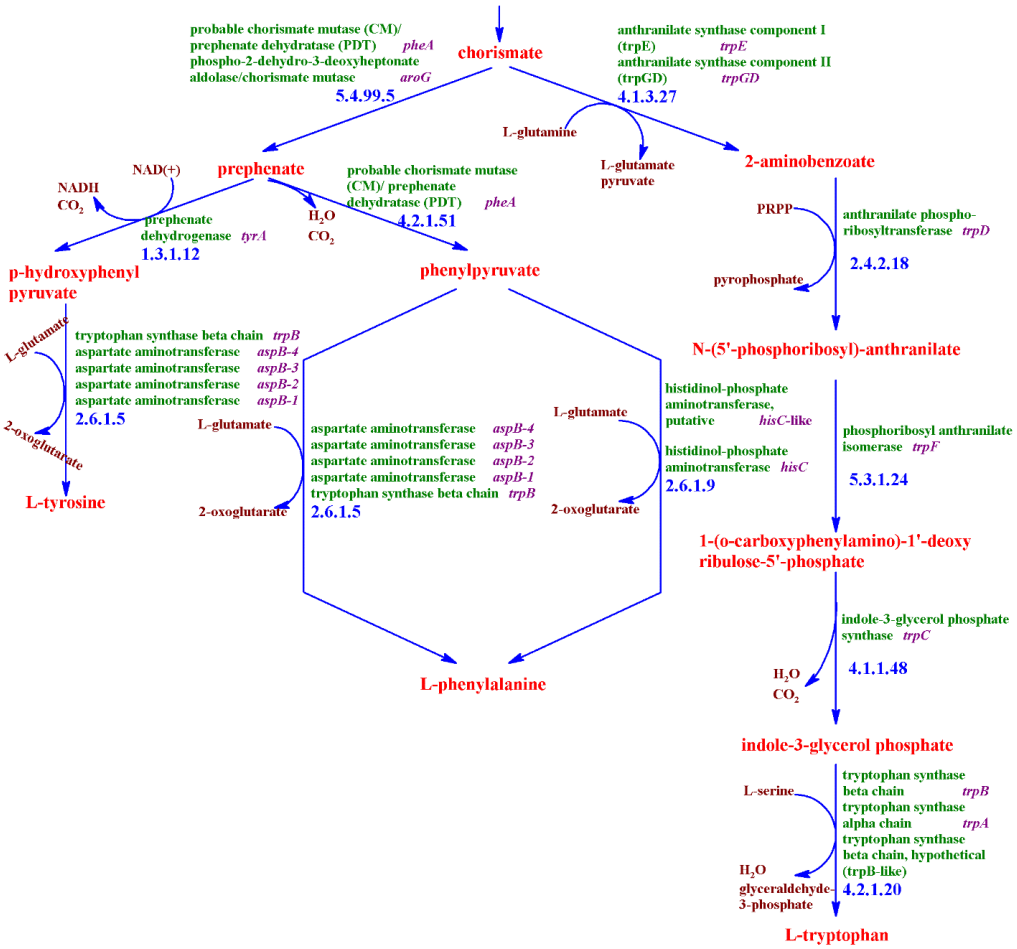

Figure 3. Curated super pathway of phenylalanine, tyrosine and tryptophan biosynthesis from chorismate (for chorismate biosynthesis, see **Figure 2**).

3.2.3. Individual Amino Acid Biosynthetic Pathways

Nishida [31]. proposed a modified variant of alpha-aminoadipic acid (AAA) pathway of lysine biosynthesis for the hyperthermophilic archaeon *Pyrococcus horikoshii*. Although the diaminopimelic acid (DAP) pathway of lysine biosynthesis was already detected by the PathoLogic of the Pathway Tools, the AAA pathway was not created in the PGDB of *S. solfataricus P2* because many catalyzing enzymes of this pathway are absent in *S. solfataricus P2* genome.

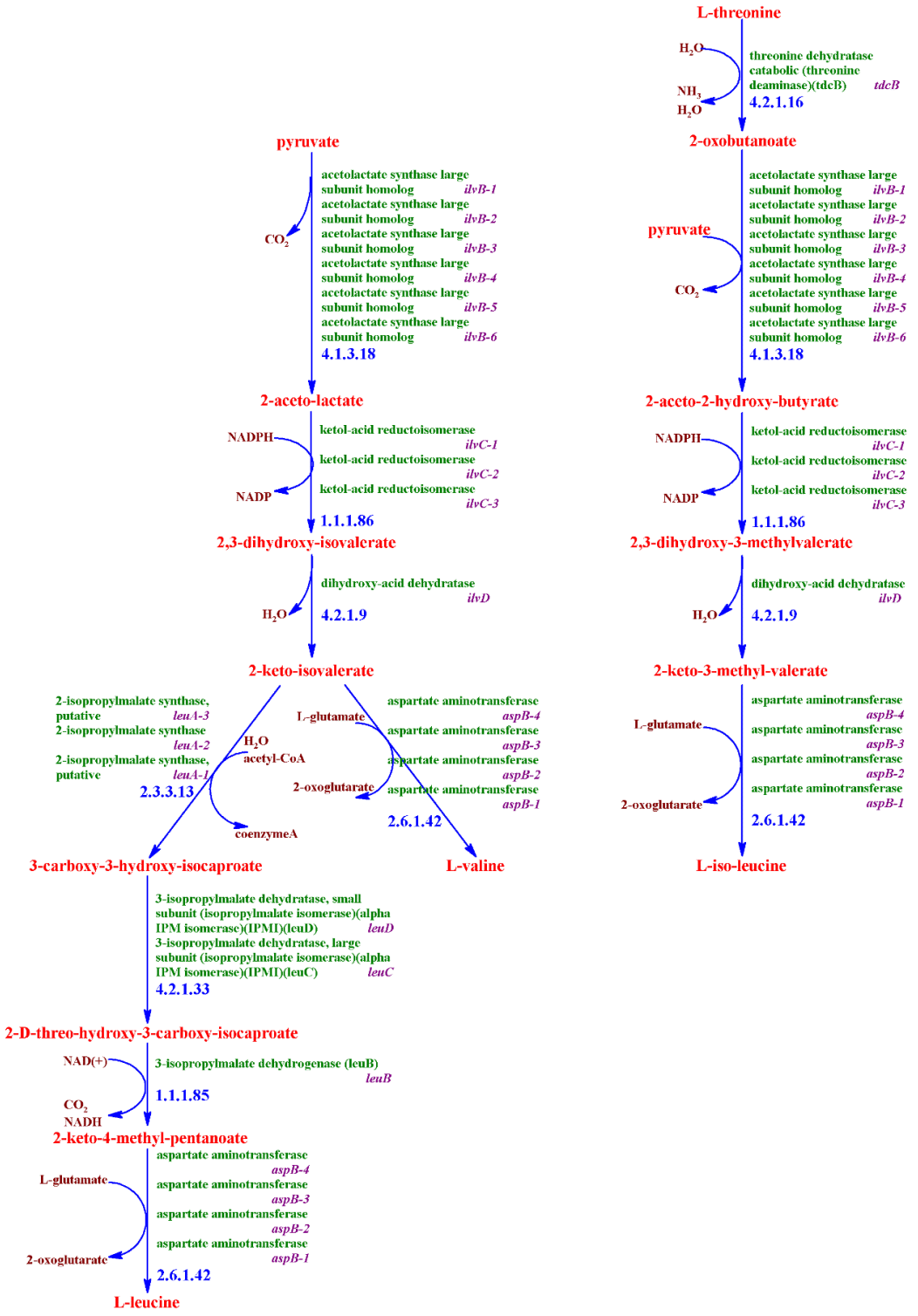

Figure 4. Curated super pathway of branched-chain amino acids biosynthesis (leucine, isoleucine and valine).

That's why a putative AAA pathway was created and added to the *S. solfataricus* PGDB **Figure 5.** The clear relationship among AAA-lysine, leucine and arginine pathways [31,32]. and the identified AAA pathway enzymes from *Pyrococcus horikoshii* helped in the identification of the AAA pathway enzymes for *S. solfataricus P2*.

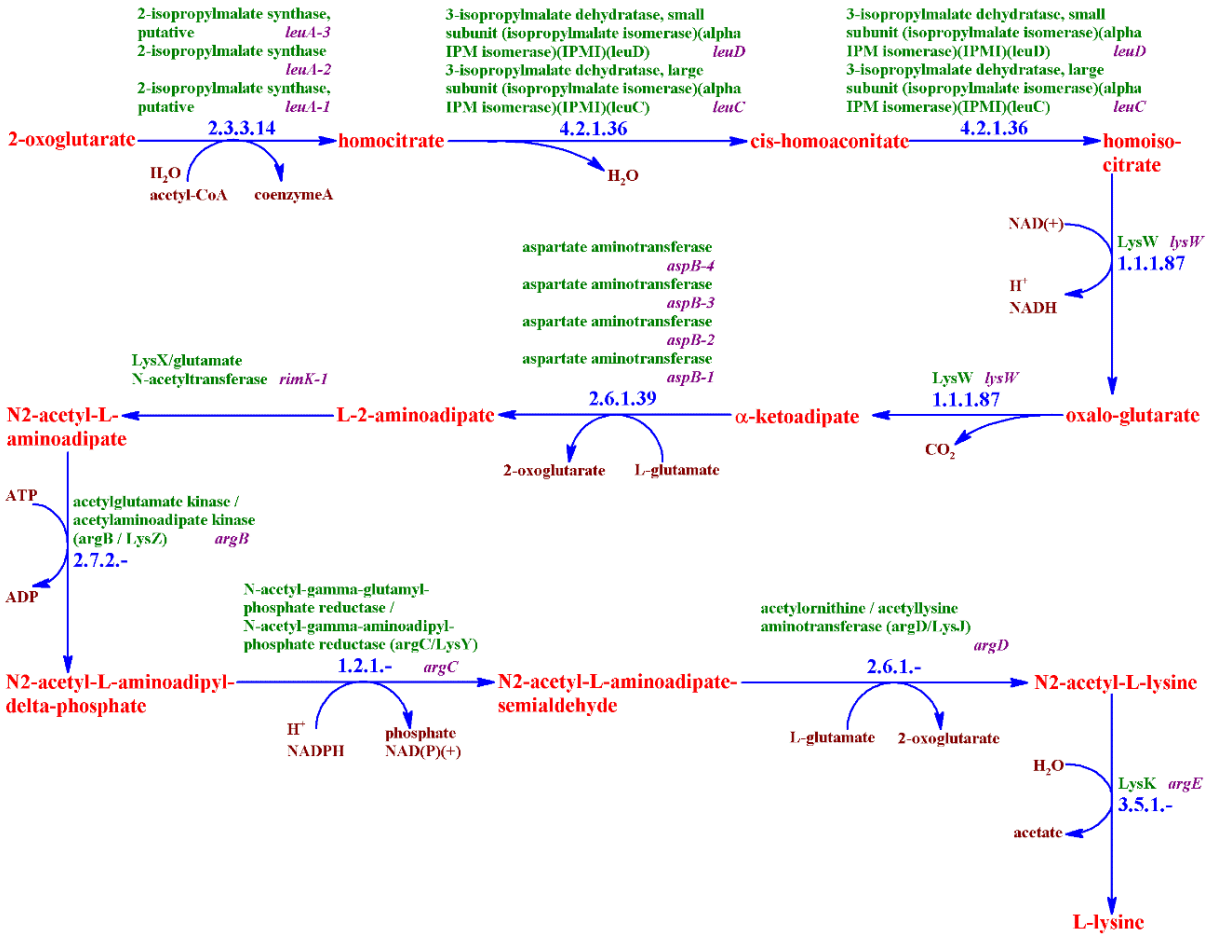

Figure 5. The **c**reated alpha-aminoadipic acid (AAA) pathway of lysine biosynthesis in *S. solfataricus*.

The broad substrate specificity of aspartate aminotransferase [34,35]. allowed the assignment of its enzymatic function to the reactions (EC 2.6.1.66) and (EC 2.6.1.2) in alanine biosynthetic pathway (three variations were created) **Figure 6.** and to (EC 2.6.1.4) in glycine biosynthesis **Figure 7.** and also to the EC-numbers in other missing steps of the aromatic and branched-chain amino acids biosynthesis **Table 1.** The homology search of such identified enzymes from the closely related Archaea in the *S. solfataricus* genome has always retrieved aspartate aminotransferase (AspB-1 to 4) and no other specific sequences.The previous mentioned analogy between lysine biosynthesis (from alpha-aminoadepate until lysine production) and arginine biosynthesis (from glutamate to arginine) and the close relation among the enzymes involved in both pathways led to the statement that ArgJ is LysX (RimK), (LysX and RimK are already annotated in *S. solfataricus P2* genome). And also ArgD is LysJ, and then the both proteins LysX/RimK and ArgD/LysJ were assigned to EC 2.3.1.1 and EC 2.6.1.13, respectively in arginine biosynthesis pathway **Figure 8.** None of MetA (EC 2.3.1.46) and methionine synthase (EC 2.1.1.13) was discovered in *S. solfataricus P2* genome, thus the route homoserine and methionine biosynthesis is the most reliable

pathway for methionine biosynthesis, in which MetB is responsible for the both reactions EC 2.5.1.48 (formerly 4.2.99.9) and EC 4.4.1.8 (Supplementary **Figure 1.**) (see https://www.ncbi.nlm.nih.gov/Structure/cdd/COG0626).

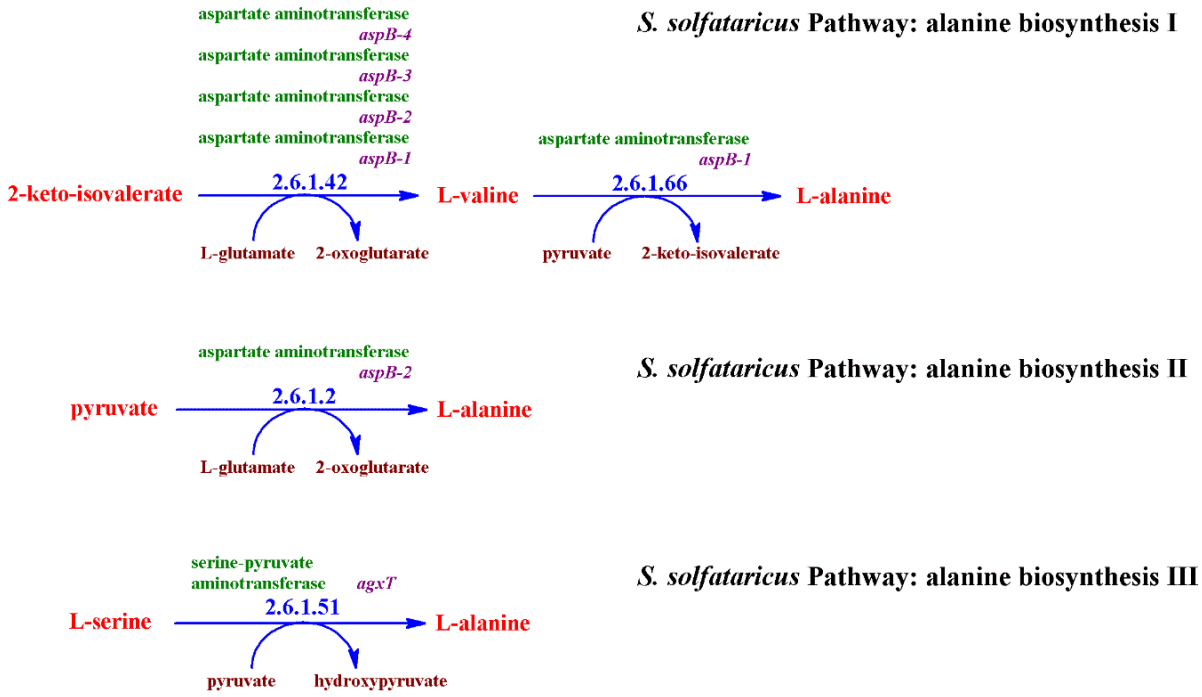

Figure 6. Three pathway variants of alanine biosynthesis (I, II and III).

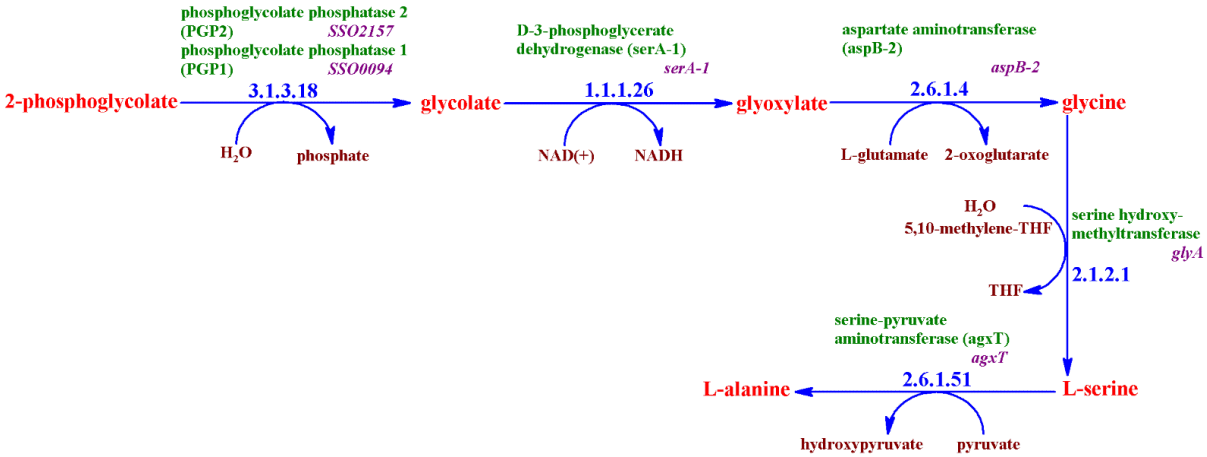

Figure 7. Super pathway of glycine, serine and alanine biosynthesis.

Cysteine is produced from serine via two steps (Supplementary **Figure 2).** The second step, which is catalyzed by CysM (EC 4.2.99.8 changed to EC 2.5.1.47), was already detected by PathoLogic. Whereas it is difficult to recognize CysE (catalyzes first step) in *S. solfataricus* due to the lack of homology with the known CysE of other Archaea such as *Methanosarcina* sp., *Pyrococcus abyssi* and *Haloarcula marismortui*. Sso0372 can be taken as a candidate after domain alignments (COG1045) and its position near to CysM (Sso0360).

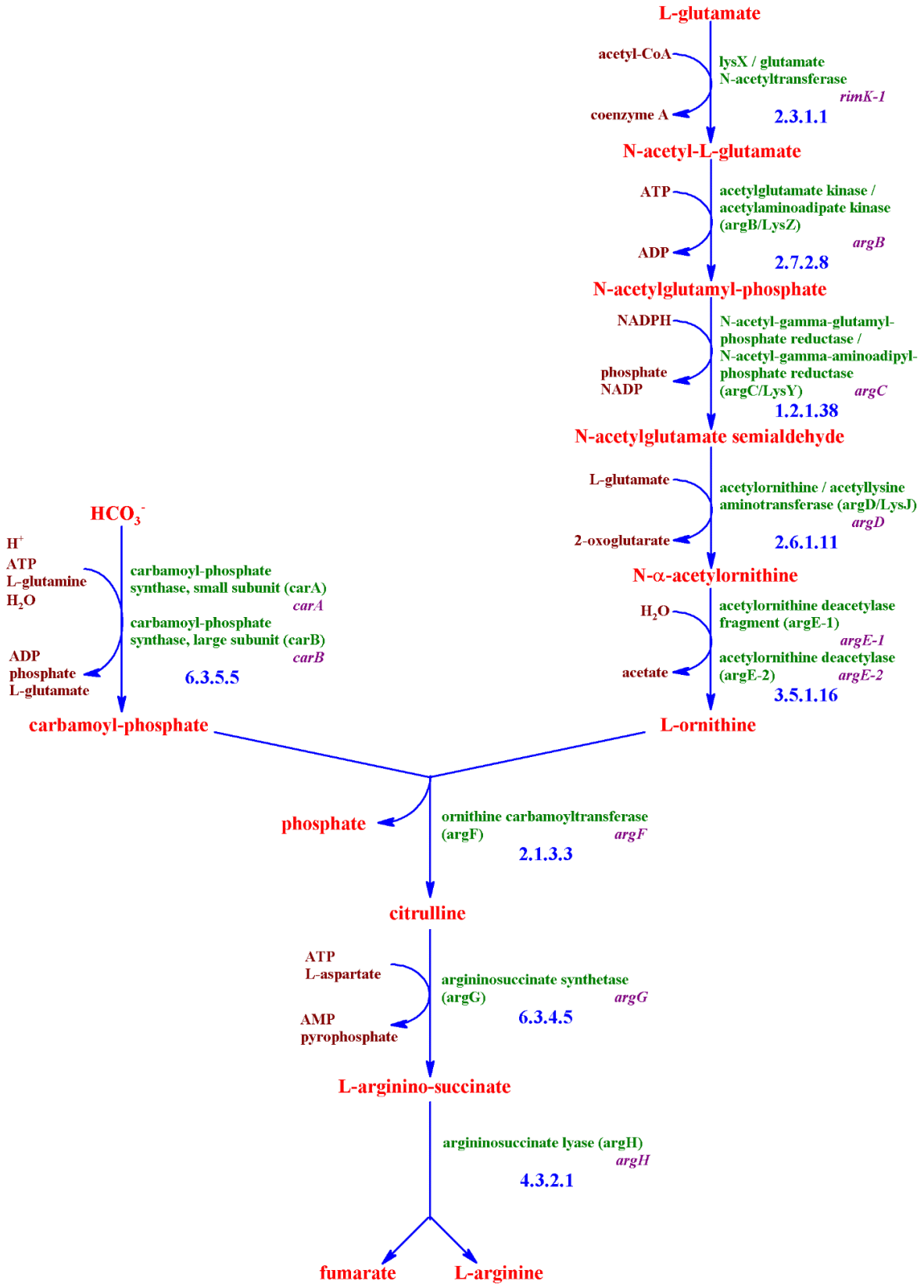

Figure 8. Arginine biosynthesis from glutamate.

Two super pathways (serine to glycine and reverse) were curated and the identified candidate enzymes **Table 1.** were assigned to the corresponding EC-numbers (Supplementary **Figures. 3 ,7**).

The unknown EC-number in proline biosynthesis pathway was recognized from the proline biosynthesis pathway of other organisms, and found to be EC 1.2.1.41 (COG0014). Searching for gamma-glutamyl phosphate reductase (ProA) from the archaeon *Methanosarcina acetivorans* in *S. solfataricus P2* genome retrieved aldehyde dehydrogenase (aldhT) (Supplementary **Figure 4**). NAD specific glutamate dehydrogenase (*gdhA*) from *Sulfolobus shibatae* was searched in the *S. solfataricus P2* genome retrieving (*gdhA-1"*Sso1457"), (*gdhA-2* "Sso1907"), (*gdhA-3* "Sso1930") and (*gdhA-4* "Sso2044"). The enzyme and the genes were assigned to EC 1.4.1.4 in glutamate biosynthesis pathway (Supplementary **Figure 5**). Phosphoribosyl-ATP diphosphatase (EC 3.6.1.31) in histidine biosynthesis pathway was already annotated in the *S. solfataricus P2* genome (*hisE "*Sso6223). In *S. solfataricus* the *hisBpx* gene encoding histidinol-phosphatase (EC 3.1.3.15) is either located at a different position in the genome, or it cannot be recognized due to lack of sufficient sequence similarity. The only candidate ORF in the *S. solfataricus* his operon is c0849 [36]. The both enzymes were assigned to the corresponding EC-numbers in the histidine biosynthesis pathway (Supplementary **Figure 6).**

4. Conclusions

Using the annotated genome, automated metabolic reconstruction created the Pathway/Genome Data Base (PGDB) specific for *Sulfolobus solfataricus*. Manual curation based on literature and protein function prediction methods improved the PGDB and consequently with continuous curation, the genomic annotations can also be improved. Curation process in the current study focused on a part of the metabolic pathways specific for amino acids biosynthesis. The curation should expand to cover the other metabolic pathways especially those which are considered to be essential for *Sulfolobus solfataricus* such as sulfur metabolism pathways. The advantage of using Pathway Tools software for automated metabolic reconstruction is the bird's eye view of the created data base components. The object-oriented nature of these (PGDBs) enables the users to easily visualize the pathways and get detailed information over the reactions, cofactors, genes and the encoding proteins and even the chemical structures of all intermediates. The complete curation of the *Sulfolobus solfataricus* PGDB should include detailed refining by removing non-existing pathways that have been false-detected by the PathoLogic. More detailed and complete annotations and citations should be added to get as a complete PGDB as possible. A complete-curated *Sulfolobus solfataricus* PGDB same as the other good curated PGDBs can serve as a reference knowledge base for genomic annotations and metabolic reconstructions for other organisms especially the closely related Archaea.

Acknowledgments

I would like to thank Dr Harmen van de Werken for his valuable advices during the implementation of this work. Thanks also to Dr Frank H. J. van Enckevort, who created the initial PGDB for *Sulfolobus solfataricus* P2. Also the author would like to thank and appreciate the great support from Professor John van der Oost.

References

1. Keeling, P.J.; Charlebois, R.L.; Doolittle, W.F. Archaebacterial genomes: eubacterial form and eukaryotic content. *Curr Opin Genet Dev*.**1994**, *4*, 816-22, URL: http://www.ncbi.nlm.nih.gov/pubmed/7888750.
2. Grabowski, B.; Kelman, Z. Archeal DNA replication: eukaryal proteins in a bacterial context. *Annu Rev Microbiol*. **2003**, *57*, 487-516, doi: 10.1146/annurev.micro.57.030502.090709.
3. She, Q.; Singh, R.K.; Confalonieri, F.; Zivanovic, Y.; Allard, G.; et al. The complete genome of the crenarchaeon Sulfolobus solfataricus P2. *Proc Natl Acad Sci U S A*. **2001**, *98*, 7835-40, doi: 10.1073/pnas.141222098.
4. Verhees, C.H.; Kengen, S.W.; Tuininga, J.E.; Schut, G.J.; Adams, M.W.; et al. The unique features of glycolytic pathways in Archaea. *Biochem J*. **2003**, *375*, 231-46, doi: 10.1042/BJ20021472.
5. Karp, P.D.; Paley, S.M. Representations of metabolic knowledge: pathways. *Proc Int Conf Intell Syst Mol Biol*. **1994**, *2*, 203-11, URL: http://www.ncbi.nlm.nih.gov/pubmed/7584392.
6. Caspi, R.; Billington, R.; Ferrer, L.; Foerster, H.; Fulcher, C.A.; et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*.**2016**, *44*, 471-80, doi: 10.1093/nar/gkv1164.
7. Caspi, R.; Billington, R.; Fulcher, C.A.; Keseler, I.M.; Kothari, A.; et al. The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res*.**2018**, *46*, 633-9, doi: 10.1093/nar/gkx935.
8. Karp, P.D.; Paley, S.; Altman, T. Data mining in the MetaCyc family of pathway databases. *Methods Mol Biol*.**2013**, *939*, 183-200, doi: 10.1007/978-1-62703-107-3_12.
9. Karp, P.D.; Riley, M.; Paley, S.M.; Pellegrini-Toole, A. The MetaCyc Database. *Nucleic Acids Res*.**2002**, *30*, 59-61, URL: http://www.ncbi.nlm.nih.gov/pubmed/11752254.
10. Karp, P.D.; Latendresse, M.; Paley, S.M.; Krummenacker, M.; Ong, Q.D.; et al. Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. *Brief Bioinform*. **2016**, *17*, 877-90, doi: 10.1093/bib/bbv079.
11. Karp, P.D.; Paley, S.; Romero, P. The Pathway Tools software.*Bioinformatics*.**2002**, *18* , 225-32, URL: http://www.ncbi.nlm.nih.gov/pubmed/12169551.
12. Karp, P.D.; Billington, R.; Caspi, R.; Fulcher, C.A.; Latendresse, M.; et al. The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform*.**2017**, bbx085, doi:10.1093/bib/bbx085.
13. Green, M.L.; Karp, P.D. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*. **2004**, *5*, 76-91, doi: 10.1186/1471-2105-5-76.
14. Khayatt, B.I. Bioinformatics Curation of the Carbohydrate Metabolism Pathways (Entner-Doudoroff and TCA Cycle) in *Sulfolobus solfataricus* P2. *Proceeding, World Agriculture Congress(4th World Research Journals Congress)*. **2018**, Timisoara; Romania.
15. Ulas, T.; Riemer, S.A.; Zaparty, M.; Siebers, B.; Schomburg, D. Genome-scale reconstruction and analysis of the metabolic network in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *PLoS One*.**2012**, *7*, e43401, doi: 10.1371/journal.pone.0043401.
16. Wolf, J.; Stark, H.; Fafenrot, K.; Albersmeier, A.; Pham, T.K.; et al. A systems biology approach reveals major metabolic changes in the thermoacidophilic archaeon *Sulfolobus solfataricus* in response to the carbon source L-fucose versus D-glucose. *Mol Microbiol*. **2016**, *102*, 882-908, doi: 10.1111/mmi.13498.

17. Grogan, D.W. Phenotypic characterization of the archaebacterial genus *Sulfolobus*: comparison of five wild-type strains. *J. Bacteriol.* **1989**, *171*, 6710-9, URL: http://www.ncbi.nlm.nih.gov/pubmed/2512283.
18. Kitabatake, M.; So, M.W.; Tumbula, D.L.; Soll, D. Cysteine biosynthesis pathway in the archaeon *Methanosarcina barkeri* encoded by acquired bacterial genes. *J Bacteriol*.**2000**, *182*, 143-5, URL: http://www.ncbi.nlm.nih.gov/pubmed/10613873.
19. Brilli, M.; Fani, R. Molecular evolution of hisB genes. *J Mol Evol*. **2004**, *58*, 225-37, doi: 10.1007/s00239-003-2547-x.
20. Kanehisa, M.; Goto, S.; Kawashima, S.; Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **2002**, *30*, 42-6, URL: http://www.ncbi.nlm.nih.gov/pubmed/11752249.
21. Kanehisa, M.; Sato, Y.; Furumichi, M.; Morishima, K.; Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.***2018**, *47*, D590–D595, doi: 10.1093/nar/gky962.
22. Johnson, M.; Zaretskaya, I.; Raytselis, Y.; Merezhuk, Y.; McGinnis, S.; Madden, T.L. NCBI BLAST: a better web interface. *Nucleic Acids Res*. **2008**, *36*, 5-9, doi: 10.1093/nar/gkn201.
23. Makarova, K.S.; Wolf, Y.I.; Koonin, E.V. Archaeal Clusters of Orthologous Genes (arCOGs): An Update and Application for Analysis of Shared Features between Thermococcales, Methanococcales, and Methanobacteriales. *Life (Basel)*.**2015**, *5*, 818-40, doi: 10.3390/life5010818.
24. Gabaldon, T.; Huynen, M.A. Prediction of protein function and pathways in the genome era. *Cell Mol Life Sci* **2004**, *61*, 930-44, doi: 10.1007/s00018-003-3387-y.
25. Tsoka, S.; Simon, D.; Ouzounis, C.A. Automated metabolic reconstruction for *Methanococcus jannaschii*. *Archaea*. **2004**, *1*, 223-9, URL: http://www.ncbi.nlm.nih.gov/pubmed/15810431.
26. Daugherty, M.; Vonstein, V.; Overbeek, R.; Osterman, A. Archaeal shikimate kinase, a new member of the GHMP-kinase family. *J Bacteriol*. **2001**, *183*, 292-300, doi: 10.1128/JB.183.1.292-300.2001.
27. Ward, D.E.; de Vos, W.M.; van der Oost, J. Molecular analysis of the role of two aromatic aminotransferases and a broad-specificity aspartate aminotransferase in the aromatic amino acid metabolism of *Pyrococcus furiosus*. *Archaea*.**2002**, *1*, 133-41, URL: http://www.ncbi.nlm.nih.gov/pubmed/15803651.
28. Cubellis, M.V.; Rozzo, C.; Nitti, G.; Arnone, M.I.; Marino, G.; Sannia, G. Cloning and sequencing of the gene coding for aspartate aminotransferase from the thermoacidophilic archaebacterium *Sulfolobus solfataricus*. *Eur J Biochem*.**1989**, *186*, 375-81, URL: http://www.ncbi.nlm.nih.gov/pubmed/2513189.
29. Koch, G.L.; Shaw, D.C.; Gibson, F. Tyrosine biosynthesis in *Aerobacter aerogenes*. Purification and properties of chorismate mutase-prephenate dehydrogenase. *Biochim Biophys Acta*.**1970**, *212*, 375-86, URL: http://www.ncbi.nlm.nih.gov/pubmed/5456988.
30. Bezsudnova, E.Y.; Boyko, K.M.; Popov, V.O. Properties of Bacterial and Archaeal Branched-Chain Amino Acid Aminotransferases. *Biochemistry (Mosc)*.**2017**, *82*, 1572-91, doi: 10.1134/S0006297917130028.
31. Nishida, H.; Nishiyama, M.; Kobashi, N.; Kosuge, T.; Hoshino, T.; Yamane, H. A prokaryotic gene cluster involved in synthesis of lysine through the amino adipate pathway: a key to the evolution of amino acid biosynthesis. *Genome Res*.**1999**, *9*, 1175-83, URL: http://www.ncbi.nlm.nih.gov/pubmed/10613839.

32. Lombo, T.; Takaya, N.; Miyazaki, J.; Gotoh, K.; Nishiyama, M.; et al. Functional analysis of the small subunit of the putative homoaconitase from *Pyrococcus horikoshii* in the Thermus lysine biosynthetic pathway. *FEMS Microbiol Lett*.**2004**, *233*, 315-24, doi: 10.1016/j.femsle.2004.02.026.
33. Miyazaki, T.; Miyazaki, J.; Yamane, H.; Nishiyama, M. alpha-Aminoadipate aminotransferase from an extremely thermophilic bacterium, *Thermus thermophilus*. *Microbiology*.**2004**, *150*, 2327-34, doi: 10.1099/mic.0.27037-0.
34. Marino, G.; Nitti, G.; Arnone, M.I.; Sannia, G.; Gambacorta, A.; De Rosa, M. Purification and characterization of aspartate aminotransferase from the thermoacidophilic archaebacterium *Sulfolobus solfataricus*. *J Biol Chem*.**1988**, *263*, 12305-9, URL: http://www.ncbi.nlm.nih.gov/pubmed/3137225.
35. Xing, R.Y.; Whitman, W.B. Characterization of amino acid aminotransferases of *Methanococcus aeolicus*. *J Bacteriol*. **1992**, *174*, 541-8, URL: http://www.ncbi.nlm.nih.gov/pubmed/1729242.
36. Charlebois, R.L.; Sensen, C.W.; Doolittle, W.F.; Brown, J.R. Evolutionary analysis of the hisCGABdFDEHI gene cluster from the archaeon *Sulfolobus solfataricus* P2.*J Bacteriol*.**1997**, *179*, 4429-32, URL: http://www.ncbi.nlm.nih.gov/pubmed/9209067.