Bioinformatics Approach for Metabolism Pathways Curation: Carbohydrate Metabolism and TCA Cycle in the Archaeon *Sulfolobus solfataricus* P2

Barzan Ibrahim Khayatt

Affiliation ; Department of Natural Resources, College of Agricultural Engineering Sciences, University of Sulaimani, Bakrajo, Sulaimani, Kurdistan-Iraq

***** Correspondence: *Email:* barzan.khayatt@univsul.edu.iq

Received: 23/2/2023

Accepted: 22/3/2023

Published: 8/5/2023

Abstract

Background: Metabolic and genomic informatics integrations in organism-specific databases require comprehensive and intensive efforts. PathoLogic, a component of the Pathway Tools software package can create complete Pathway/Genome Databases (PGDBs) from genomic sequence and annotation files for any organism. This tool can predict the metabolic pathways using MetaCyc as a reference knowledge base. This work aimed to apply a bioinformatics approach to curate a PGDB created for the Crenarchaeon *Sulfolobus solfataricus* P2. This archaeon grows optimally at 80^{o} C and pH 2-4. The complete genome of *S. solfataricus* P2 was released in 2001. Created PGDBs often need manual curations to fill in the metabolic gaps that the software failed to detect.

Methods: We used Pathway Tools to create the PGDB for the *Sulfolobus solfataricus* P2. Bioinformatics curation for carbohydrate metabolism pathway (Entner-Doudoroff "ED") and TCA cycle was carried out. Literature search as well as homology-, orthology- and context-based protein function prediction methods were followed for this curation using the Editors component of the Pathway Tools program.

Results: Curation modified the number of the pathways in the database by adding extra pathways that have not been detected by the PathoLogic. New pathways such as semi-phosphorylated ED and a new variation of the TCA cycle were added to the PGDB of *S. solfataricus* P2. Filling in the metabolic holes (missing enzymes) in the pathways under study was also involved in the curation process.

Conclusion: The bioinformatics curation of the PGDB of *S. solfataricus* P2 improved the database that can serve as a reference knowledge base for genomic annotations and metabolic pathway reconstructions of other organisms especially the closely related Archaea.

Keywords: Bioinformatics curation, Crenarchaeota, Pathway Tools, ED pathway, TCA cycle, *In silico study*.

***Supplementary Figures Link:** https://figshare.com/s/feb6ae05fb2e4935f571

https://doi.org/10.24126/jobrc.2023.17.1.692

1. Introduction

Automated prediction of metabolic pathways is referred to as metabolic reconstruction. Using entire genome sequences and genomic data, many metabolic reconstruction methods and algorithms have been published so far (1-6). Reference knowledge bases (KBs), especially metabolic pathway-specific databases such as MetaCyc (https://metacyc.org) (7-12) facilitate the achievement of metabolic reconstruction of the target organism. The so-called Pathway/Genome databases (PGDBs) for different organisms can be created by using Pathway Tools software (2, 4). These databases are collections of the annotated entire genome sequences of the target organisms as well as the reconstructed metabolic pathways of the organism including information on compounds, intermediates, cofactors, reactions, genes and their products. Many research groups have used the Pathway Tools software package, and a significant number of created PGDBs were collected in BioCyc (https://biocyc.org), which contains a collection of 19495 PGDBs for model eukaryotes and thousands of microorganisms. The initial non-curated PGDBs may lack specific chemical pathways that are actually present in the organism or conversely consist of false positives. The addition of the absent pathways and removing some pathways that are not exist naturally in that individual organism help in improving the quality of genome annotation. If the majority enzymes in a pathway have corresponding genes 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 new sequences), any reaction catalyzed by that protein would appear as a missing enzyme (pathway hole) in a PGDB (13).

Sulfolobus solfataricus P2 belongs to the phylum Crenarchaeota of the life domain Archaea. Comparative genomics has revealed a conserved core of 313 genes that are represented in all completely sequenced archaeal genomes, plus a variable 'shell' that is prone to lineage specific gene loss and horizontal gene exchange (14). There is limited experimental work available regarding the characterization of archaeal genes and prediction of the archaeal specific pathways. Whereas the central metabolic routes of bacteria and eukaryotes are generally well conserved, variant pathways have developed in Archaea (15). *S. solfataricus* P2 is widely adopted as a model organism, grows optimally at 80^{o} C and pH 2-4, aerobically metabolizes sulfur for energy and uses organic compounds as C source (chemoorganotroph) (16). The complete genome of *S. solfataricus* P2 was released in 2001, corresponding to a single chromosome of 2,992,245 bp. The identified number of the genes is 3032 genes. The number of proteins is 2,997 of which 1/3 have no detectable homologs in other sequenced genomes, and 40% appear to be archaeal specific and only 12% and 2.3% are shared exclusively with bacteria and eukaryotes, respectively (17).

Using ^{14}C-glucose, it was found that the oxidative breakdown of glucose to pyruvate in *S. solfataricus* differs from the classical Entner-Doudoroff pattern (18), lacking first steps phosphorylations. Metabolic network analysis by (19), revealed that a significant amount of glucose is metabolized to pyruvate via the non-phosphorylative rout, followed by the tricarboxylic acid cycle. The non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase (GAPN) was suggested to act a key role in the regulation of carbon degradation via modifications of the Emden-Meyerhof-Parnas (EMP) and the branched Entner-Doudoroff (ED) pathway in hyperthermophilic Archaea (20, 21). This raised the concept "non-phosphorylated Entner-Doudoroff (nED)" pathway. The anticipated central metabolic pathways (glycolytic pathway, pentose phosphate pathway and the citric acid cycle) were constructed for the *S. solfataricus* P2 (17). The construction was based on the prediction of the majority of the encoding genes. This came in agreement with the assumption that *S. solfataricus* possesses the non-phosphorylated Entner–Doudoroff (ED) pathway because all but one of the involving genes encoding non-phosphorylated ED enzymes were identified (2-keto-3-deoxygluconate [KDG]; glyceraldehyde-3P [GAP]; phosphoenolpyruvate [PEP]; dihydroxyacetone-P [DHAP]).

Tricarboxylic Acid Cycle (TCA Cycle), known as Krebs cycle and also as citric acid cycle is the metabolic pathway, its complete set comprises eight reactions. The cycle links to the ED pathway via pyruvate or PEP (22, 23). The cycle also links to amino acids biosynthesis pathways via providing precursors for some amino acids like L-lysine biosynthesis (24). Most organisms with completely sequenced genomes encode only a certain subset of TCA cycle enzymes, utilizing only fragments of the cycle (25). Via a comparative genomic study, (26) showed the diversity of TCA cycle pathway in different species, as well as detecting the non-orthologous gene displacement for some enzymes within the cycle. Although it was mentioned by (17) that all TCA cycle genes

https://doi.org/10.24126/jobrc.2023.17.1.692

were predicted in *S. solfataricus* genome (Sso1077, 1095, 2182, 2356 to 2359, 2482, 2483, 2585, 2589, 2815, 2816, 2863), there are still gaps in the TCA cycle to be filled in.
This bioinformatics study comprises an approach that allies the other ongoing attempts worldwide of computational metabolic reconstructions of metabolic pathways for many organisms. The study intended reconstruction of the metabolic pathways in *S. solfataricus* P2 focusing on carbohydrate metabolism (modified Entner-Doudoroff) and TCA cycle pathways and introducing an approach to curate the Pathway/Genome Database specific for *S. solfataricus* P2.

2. Materials and Methods

2.1. Creating PGDB specific for *Sulfolobus solfataricus* P2

To create the Pathway/Genome Database specific for *S. solfataricus* P2, the following tools were used: **1)** Pathway Tools software (2, 4), this package is also useful for querying, visualization and curation of MetaCyc and other pathway databases. By using this software, initial metabolic pathways were predicted for the target organism (*S. solfataricus* P2) via its entire annotated genome sequence. The main components of the Pathway Tools software are: a) The PathoLogic, b) The Pathway/Genome Navigator c) The Pathway/Genome Editors.
2) The reference database MetaCyc (10), which is a multi-organism database of experimentally elucidated metabolic pathways and the associated enzyme commission (EC) numbers and enzyme names. **3)** The annotated genome of *S. solfataricus* P2 (17).
4) Experiment-based literature of *S. solfataricus* P2 pathways and pathways of other closely related Archaea. **5)** Metabolic pathway-specific databases such as KEGG (27, 28).
6) The National Center for Biotechnology Information (NCBI) site for PubMed, Blast and Clusters of Orthologous Groups (COGs) (https://www.ncbi.nlm.nih.gov). **7)** Other databases such as: Enzyme database BRENDA (https://www.brenda-enzymes.org) (29), and the UniProt consortium members such as the Protein Database Swiss-Prot and Protein Information Resource (PIR) (30).

2.2. Detection of the Missing Enzymes

In order to achieve the metabolic reconstruction of *S. solfataricus* P2, the following steps were carried out: **1)** by using PathoLogic (a component of Pathway Tools software), the initial non-curated PGDB for *S. solfataricus* P2 was created. The annotated genome of *S. solfataricus* P2 has been used as input and the software would use MetaCyc as a reference database. From the enzymes annotated in the genome the PathoLogic could predict the set of the reactions in their corresponding pathways in *S. solfataricus* P2. **2)** The manual curation process of the initial PGDB was initiated with literature search for the mentioned pathways in *Sulfolobus solfataricus*, their reactions, reactants and the involving enzymes. **3a)** Checking was carried out for the metabolic pathways and their components in the reference database MetaCyc (for *S. solfataricus* P2). **3b)** Checking was also done for the mentioned pathways and the corresponding reactants and enzymes in other metabolic pathway-specific databases such as KEGG. **4)** By comparing the reaction sets, reactants and the enzymes, identified by experimental literature to those in MetaCyc and the created PGDB of *S. solfataricus* P2, enabled identifying the missing steps and enzymes in the pathways (pathway holes).
5) 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.
6) Adopting one or more of the protein function prediction methods [Details in (31)] was the alternative for the lack of experimental literature. Homology search using BLASTP and PSI-BLAST (32) of a specific enzyme with an assigned EC-number (formerly its sequence has been retrieved in closely related Archaea) in *S. solfataricus* P2 genome was the main strategy in this step. Clusters of Orthologous Groups of proteins (COGs) were used to identify the potential candidates for the pathway holes. **7)** Studying the relating pathways in evolutionarily close species to *S. solfataricus* P2 not just helped in filling the gaps but also creating some new pathway variations in the PGDB. **8)** Identification of the subunits of the enzyme complexes to give a complete view of some reactions that are catalyzed by complexes. **9)** The predicted missing enzymes and the encoding genes were assigned to the corresponding reactions via the Editors component of Pathway Tools. **10)** Citations were also added to the curated reactions automatically by the software via entering the PubMed ID of the

https://doi.org/10.24126/jobrc.2023.17.1.692

corresponding literature. **11)** Finally, partial curated PGDB was achieved including all modifications in the Entner-Doudoroff and TCA Cycle pathways.

3. Results

The Pathway/Genome Database (PGDB) of *S. 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. The PGDB is visualized, queried and can be edited via the Pathway Tools. The created PGDB is not a simple rational database but an object-oriented database. All components of the database are represented as frames in the graphical display of the program. The program compared all the genomic information of *S. solfataricus* P2 especially the annotated EC numbers and the enzyme names with those stored in the reference database MetaCyc. Via PathoLogic algorithm, the program 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* P2. The genomic information about the total gene number and the genes encoding proteins as well as the number of enzymes, enzymatic reactions and the predicted pathways are present in the PGDB. The number of predicted pathways is 159 and the enzymatic reactions are 787 reactions. From 3009 polypeptides, 582 enzymes were detected in the genome. The PGDB was manually curated for the carbohydrate metabolic pathways (Entner-Doudoroff) and TCA cycle.

3.1. Curation of Carbohydrate Metabolism Pathways

3.1.1. Semi-Phosphorylated ED Pathway

Non-phosphorylated ED pathway was constructed by (17) for *Sulfolobus solfataricus* P2 depending on the detection of the majority of the encoding genes in the genome. Because the genome of *Sulolobus solfataricus* P2 containing these enzymes was the input to the Pathway Tools software, the program could also detect the non-phosphorylated ED pathway. As it is depicted in **Figure 1** (ignoring the first reaction), from eight reactions only four reactions were assigned an enzyme in the predicted non-phosphorylated glucose degradation.

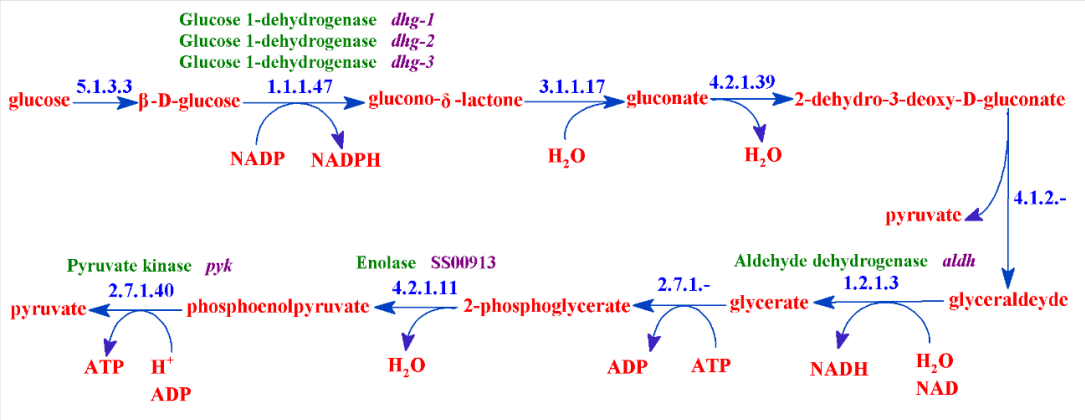

Figure 1 Detected non-phosphorylated carbohydrate metabolic pathway in the PGDB of Sulfolobus solfataricus P2 with five metabolic holes (missing enzymes). In the metabolic holes the EC* numbers are shown in blue. Enzyme names are in green and the gene names are in violet color.

*: EC number is Enzyme Commission number (classification scheme for enzymes). EC followed by four digits, the first digit defines the general type of reaction catalyzed by the enzyme (ranges from one to six). The second digit indicates the subclass. The third gives the sub-subclass and the fourth digit is the serial number of the enzyme in its sub-subclass.

The initial PGDB did not contain the semi-phosphorylated ED pathway. Using the Editors component of Pathway Tools software, the semi-phosphorylated ED pathway was also created. **Figure 2** shows the initial semi-phosphorylated ED pathway with five metabolic holes. The curation should fill in these gaps (**Figure 2**).

https://doi.org/10.24126/jobrc.2023.17.1.692

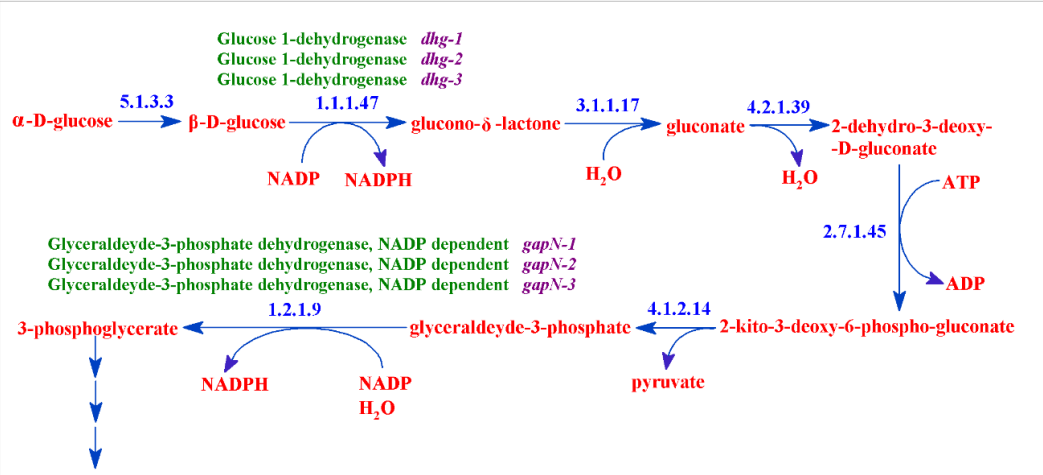

Figure 2 Initial semi-phosphorylated ED pathway with five metabolic holes. In the metabolic holes, the EC numbers are shown in blue above the reaction arrows. Enzyme names are in green and the gene names are in violet color.

3.1.1.1. Identifying Enzymes Assigned to the EC Numbers in the Metabolic Holes of Semi-Phosphorylated ED Pathway

In *S. solfataricus* P2 the both enzymes, glucose dehydrogenase and 2-keto-3-deoxygluconate aldolase (KDG aldolase) are responsible for the catabolism of glucose and galactose (33). Their results assume the promiscuity of the carbohydrate metabolic pathway. Depending on these findings and the orthology-based function prediction, the enzyme aldose 1-epimerase as well as the encoding gene (COG2017) were assigned to the first metabolic gap (EC 5.1.3.3) in the semi-phosphorylated ED pathway (**Figure 3**). Following the bioinformatics approach, the rest metabolic gaps were filled (see Discussion).

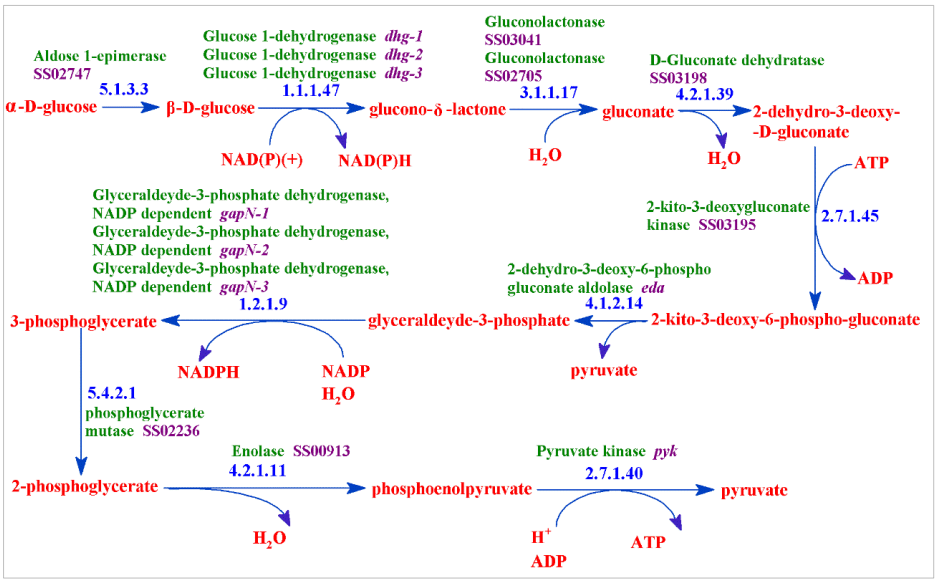

Figure 3 Complete curated semi-phosphorylated ED pathway in Sulfolobus solfataricus P2. All enzyme names and genes were assigned to the EC numbers above the reaction arrows. The EC numbers are shown in blue, enzyme names are in green and the gene names are in violet color.

3.1.2. TCA Cycle Pathway

The prediction of all citric acid cycle genes in *S. solfataricus* P2 genome (17) allowed the PathoLogic to detect and add different variations of TCA cycle to the created PGDB of *S. solfataricus* P2. In order to display a better view of the TCA cycle pathway in the PGDB, a new variation of the pathway was added as the first curation achieved for the initial PGDB. The complete set of the TCA cycle EC-numbers was used as input to the

https://doi.org/10.24126/jobrc.2023.17.1.692

Editor component of the Pathway Tools software to create the new variation of TCA cycle. The initially created TCA cycle variations contain many metabolic holes (missing enzymes) that have to be filled in during the curation process. *Supplementary **SFigure 1** shows a part of the TCA cycle in which some gaps with known EC-numbers are present. Literature search, homology-, orthology- and neighborhood-based protein function prediction helped in the identification of the missing enzymes to form the most likely complete TCA cycle in *S. solfataricus* P2 (**Table 1, Figure 4**).
It was mentioned that the prediction of the citrate synthase enzyme seems to be a good indicator for the presence of a (nearly) complete TCA cycle in a given organism (25). The detection of citrate synthase (Sso2589) was already achieved in the *S. solfataricus* P2 genome.

3.1.2.1. Identifying and Assign Enzymes to the EC Numbers in the Metabolic Holes of TCA Cycle

Table 1 shows the identified missing enzymes for each EC-number in the missing steps of the TCA cycle, which was created an added to the PGDB. The identified enzymes also include the missing enzymes of the linked reactions to the TCA cycle such as the link to pyruvate and phosphoenolpyruvate (PEP).
Malate dehydrogenase ((S)-malate:NAD^+ oxidoreductase) catalyzes the reaction [(S)-malate + NAD^+ = oxaloacetate + NADH + H^+] (EC 1.1.1.37). Malate dehydrogenase can be either dimeric or tetrameric (34). Crystallographic analysis by the same reference of MalDH from the halophilic archaeon *Haloarcula marismortui* has shown a tetramer (of two dimers) interacting mainly via complex salt bridge clusters. Putative malate dehydrogenase was detected in the *S. solfataricus* P2 genome to be encoded by the gene *mdh* (named also as *sqdB*).

Table 1 The identified missing enzymes for each EC-number in the missing steps of the TCA cycle, with the encoding genes and the COGs they belong.
*****: Clusters of the orthologous groups

EC-Number	Enzyme	Gene-Name	Gene#	COG*
EC 1.1.1.37	**Malate dehydrogenase**	***mdh(sqdB)***	***Sso2585***	**COG0039**
EC 1.1.1.38	**Malic enzyme**	**-**	***Sso2869***	**COG0281**
EC 1.1.1.40	**Malic enzyme**	**-**	***Sso2869***	**COG0281**
EC 1.2.7.1	**Pyruvate synthase (pyruvic-ferredoxin oxidoreductase)**			
	--alpha chain	***porA-1***	***Sso1207***	**COG0674**
		porA-2	***Sso2757***	
		porA-like	***Sso2129***	
	--beta chain	***porB-1***	***Sso1206***	**COG1013**
		porB-2	***Sso2756***	
		porB-like	***Sso2130***	
	--delta chain	***porD-1***	***Sso7412***	**COG1144**
		porD-2	***Sso11071***	
		porD-like	***Sso2128***	
	--gamma chain	***porG-1***	***Sso1208***	**COG1014**
		porG-2	***Sso2758***	
EC 1.2.7.3	**2-oxoglutarate synthase (2-oxoglutarate-ferredoxin oxidoreductase)**			
	--alpha chain		***Sso2815***	**COG0674**
	--beta chain		***Sso2816***	**COG1013**
EC 6.4.1.1	**Pyruvate carboxylase**	***AccC?***	***Sso2466***	**COG0439**
EC 4.1.1.31	**Archaeal type phosphoenolpyruvate carboxylase (atPEPC)**		***Sso2256***	**COG1892**
EC 4.1.1.32	**Archaeal GTP-dependent phosphoenolpyruvate carboxykinase (PEPCK)**	***pckg***	***Sso2537***	**COG1274**

https://doi.org/10.24126/jobrc.2023.17.1.692

The gene Sso2869 in *S. solfataricus* P2 produces malic enzyme (oxaloacetate decarboxylating). It was assigned to the reaction EC 1.1.1.38, that also catalyzes the reaction EC 1.1.1.40 (**Figure 4**). In the Discussion section, the strategies and argues are shown for filling the majority of the missing metabolic enzymes and the corresponding genes.

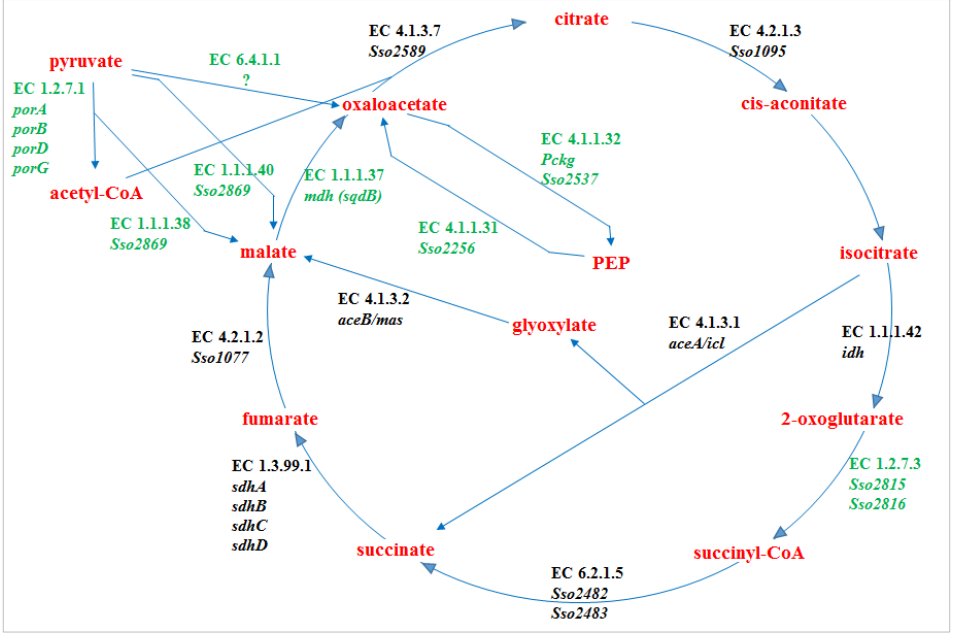

Figure 4 Curated TCA cycle of Sulfolobus solfataricus P2. The genes encoding the identified missing enzymes were assigned to the corresponding EC-numbers (in green) (The enzyme names are listed in **Table 1**). The software already has assigned enzymes and the encoding genes to the EC numbers in black.

4. Discussion

4.1. Semi-Phosphorylated ED Pathway

Identification of the key phosphorylation reaction (KDG kinase) for the modified ED pathway in an experimental work (35) lead to add the concept for a semi-phosphorylated ED pathway in *S. solfataricus* P2. This finding helped in the curation process of the ED metabolic pathway in the PGDB of *S. solfataricus* P2.

In vitro reconstruction experiments, (35) found that no gluconolactonase (catalyses the reaction: D-glucono-1,5-lactone + H_2O = D-gluconate) was needed for a functional pathway. However, they suggested a candidate gene (Sso3041), which located close to the glucose dehydrogenase (Sso3042) to fill the second metabolic-hole (EC 3.1.1.17) in the semi-phosphorylated ED pathway (**Figure 3**). This gene due to the above-mentioned neighborhood is most likely to be the gene that catalyzes the reaction rather than another gene (Sso2705), which is also found in *S. solfataricus* (COG3386).

Conversion of D-gluconate to 2-dehydro-3-deoxy-D-gluconate (KDG) is catalyzed by gluconate dehydratase (an octameric protein). N-terminal amino acid sequencing of the purified D-gluconate dehydratase from *S. solfataricus* was done and found to be in exact agreement with Sso3198, which has been annotated in the genome as muconate cycloisomerase (36). The enzyme and the encoding gene were added to the EC 4.2.1.39 in the third missing enzyme in the semi-phosphorylated ED pathway (**Figure 3**). This gene resides in a cluster with KDG aldolase (Sso3197), KDG kinase (Sso3195) and non-phosphorylating GAP dehydrogenase (Sso3194) (35, 36).

The identification and biochemical characterization of KDG kinase from *S. solfataricus* P2 added an evidence for the presence of the semi-phosphorylated ED pathway. The enzyme catalyzes the phosphorylation of 2-

https://doi.org/10.24126/jobrc.2023.17.1.692

dehydro-3-deoxy-D-gluconate (KDG) to form 6-phospho-2-dehydro-3-deoxy-D-gluconate (KDPG). The gene (Sso3195) that encodes this enzyme is present in the Ed-gene cluster. The enzyme belongs to sugar kinases, ribokinase family (COG0524). Within this COG other three genes were found in *S. solfataricus* but based on the above mentioned clustering the strong candidate is (Sso3195).The enzyme KDG kinase and the encoding gene were added to the EC 2.7.1.45 in the fourth missing enzyme in the semi-phosphorylated ED pathway (**Figure 3**).
The KD(P)G aldolase (a tetrameric protein**)** has been purified and characterized from *S. solfataricus*. This enzyme converts KD(P)G via non-phosphorylated ED pathway (35, 37) and also via semi-phosphorylated ED pathway (35) to pyruvate and glyceraldehyde(-3-phosphate). The bifunctional KD(P)G aldolase characterization was the extra evidence for the presence of semi-phosphorylated ED pathway in *S. solfataricus*. The gene *kdgA* known also as *eda* (Sso3197) was identified and cloned by (35) to be the gene encoding KD(P)G aldolase and it resides in the ED cluster as mentioned previously. The gene and its product were assigned to the EC 4.1.2.14 in the fifth missing enzyme in the semi-phosphorylated ED pathway (**Figure 3**).
Recent evidences showed that all enzymes of this pathway exhibit catalytic promiscuity that enables them to catalyze the metabolism of galactose as well (38).

4.2. TCA Cycle Pathway

Malic enzyme (Pyruvic-malic carboxylase) catalyzes the reaction [(S)-malate + NAD(P)(+) <=> pyruvate + CO(2) + NAD(P)H] (EC 1.1.1.40), acting on the CH-OH group of donors with NAD+ or NADP+ as acceptor, and also decarboxylates the oxaloacetate. The enzyme was purified by (39) from *S. solfataricus* P2 and its molecular weight was determined to be dimer of Mr 105,000 +/- 2,000 with apparently identical Mr 49,000 +/- 1,500 subunits.
Pyruvate synthase (Pyruvate:ferredoxin oxidoreductase) complex catalyzes the conversion of pyruvate to acetyl-CoA (EC 1.2.7.1), which enters the TCA cycle and both with oxaloacetate produce citrate [Pyruvate + CoA + oxidized ferredoxin <=> acetyl-CoA + CO(2) + reduced ferredoxin]. In general, Archaea utilize a pyruvate ferredoxin oxidoreductase (EC 1.2.7.1) instead of pyruvate dehydrogenase complex, which is already isolated and characterized from some Archaea such as *S. solfataricus* (40). The enzyme consists of four subunits that have already been detected in *S. solfataricus* genome (**Table 1**) with a molecular mass of 260 kDa (same reference). The cofactors, which activate the complex, are thiamine diphosphate and iron-sulfur clusters. Although the genes that encode the enzyme complex have been detected in the genome, the Pathway Tools software could not assign the enzyme complex, the genes and even the EC number to the reaction of converting pyruvate to acetyl-CoA. It suggested dehydrogenation reaction with NAD as acceptor (*Supplementary **SFigure 2**).
A reaction similar to EC 1.2.7.1 is catalyzed by the enzyme complex 2-oxoglutarate synthase (2-oxoglutarate-ferredoxin oxidoreductase) [2-oxoglutarate + CoA + oxidized ferredoxin <=> succinyl-CoA + CO(2) + reduced ferredoxin] (EC 1.2.7.3). The enzyme from *Sulfolobus sp*. is a heterodimer consists of two subunits, alpha (632 amino acids) and beta (305 amino acids) and it shows a broad specificity for 2-oxoacids such as pyruvate and 2-oxoglutarate (41). The genes encoding the both subunits were detected in the *S. solfataricus* genome, but the software same as in the previous reaction suggested a dehydrogenation reaction with NAD as acceptor (*Supplementary **SFigure 3**).
Pyruvate carboxylase catalyzes the conversion of pyruvate to oxaloacetate [pyruvate +ATP + HCO(3)(-) <=> oxaloacetate + ADP + phosphate]. There is no literature available regarding the experimental characterization of this enzyme in *S. solfataricus* P2. Although in the pathway-based database KEGG, the enzyme pyruvate carboxylase and its encoding gene Sso2466 (*accC*) were assigned to the EC 6.4.1.1, surprisingly neither the gene Sso2466 nor any archaeal genes exist within the COG1038 (pyruvate carboxylase (PycA)), but they exist within COG0439 (biotin carboxylase (AccC)). Within this COG, archaeal pyruvate carboxylase as well as archaeal biotin carboxylase are present.
Another TCA cycle link to glycolysis is via phosphoenolpyruvate, which is converted to oxaloacetate by an archaeal type phosphoenolpyruvate carboxylase (atPEPC) (EC 4.1.1.31). The mass of a typical bacterial and eukaryal PEPC ranges from 90 to110 kDa, whereas the calculated molecular masses of atPEPC subunits (found to be tetramer) are approximately half this size, ranging from 55 to 60 kDa and the enzyme complex requires Mg^2 for its activity (42). The enzyme atPEPC in *S. solfataricus* was found to be encoded by Sso2256 (42). Using Bioinformatical tools, they could also find significant similarity between the archaeal proteins from

https://doi.org/10.24126/jobrc.2023.17.1.692

COG1892 (contains Sso2256) and the bacterial, eukaryal PEPC (BE-PEPC) family (COG2352) [for details see (42)].
Preferably, the reverse conversion of oxaloacetate to phosphoenolpyruvate as the first step of gluconeogenesis is catalyzed by an archaeal GTP-dependent phosphoenolpyruvate carboxykinase (PEPCK) (EC 4.1.1.32). Among the completely sequenced Archaea, (43) found **a)** highly conserved homologues in the closely related genus *Pyrococcus* to the studied archaeal (*Thermococcus kodakaraensis*) PEPCK (84 to 86% identical in amino acid sequences), **b)** moderately (50 to 51% identical) and **c)** weakly (33 to 35% identical) homologous genes in *Sulfolobus* and *Thermoplasma*, respectively (COG1274). The orthologous gene in *S. solfataricus* within this COG is Sso2537. There is no lab biochemical characterization of the enzyme from *S. solfataricus* so far, thus it is not known to be homotetrameric such as PEPCK from *Thermococcus kodakaraensis* or monomeric such as all other known GTP-PEPCKs (43). The identified enzymes (**Table 1**) and the encoding genes were added to the gaps in the TCA cycle, each to the corresponding EC-number (**Figure 4**).

5. Conclusion

The bioinformatics approach implemented in the curation of the PGDB specific for *S solfataricus* P2 improved the metabolic pathways that can subsequently serve as a reference for genomic annotations and metabolic pathway reconstruction of other organisms especially the closely related Archaea. As mentioned previously, many research groups worldwide created a huge number of PGDBs (19495 PGDBs) that have been collected in BioCyc (https://biocyc.org) (44) for many model eukaryotes and thousands of microorganisms. The majority of the initial non-curated PGDBs may lack many chemical pathways that have not been detected by the software packages used to build these metabolic databases, or conversely consist of false positives that need comprehensive curations. The current bioinformatics approach, which utilizes sufficient techniques, can efficiently contribute in such curations. The approach mainly aimed to identify and assign enzymes and their encoding genes to the metabolic holes, curating specific metabolisms in the database. For the complete set metabolism curation of the whole PGDB database, the same bioinformatics approach can be applied.

6. References

1. Faria JP, Rocha M, Rocha I, Henry CS. Methods for automated genome-scale metabolic model reconstruction. Biochemical Society transactions. 2018;46(4):931-6.
2. Karp PD, Midford PE, Billington R, Kothari A, Krummenacker M, Latendresse M, et al. Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. Brief Bioinform. 2021;22(1):109-26.
3. Karp PD, Paley SM. Representations of metabolic knowledge: pathways. Proc Int Conf Intell Syst Mol Biol. 1994;2:203-11.
4. Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, Lee TJ, et al. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief Bioinform. 2009;11(1):40-79.
5. Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Res. 2018;46(15):7542-53.
6. Pilalis E, Koutsandreas T, Valavanis I, Athanasiadis E, Spyrou G, Chatziioannou A. KENeV: A web-application for the automated reconstruction and visualization of the enriched metabolic and signaling super-pathways deriving from genomic experiments. Computational and structural biotechnology journal. 2015;13:248-55.
7. Altman T, Travers M, Kothari A, Caspi R, Karp PD. A systematic comparison of the MetaCyc and KEGG pathway databases. BMC Bioinformatics. 2013;14(112).
8. Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 2012;40(Database issue):D742-53.

https://doi.org/10.24126/jobrc.2023.17.1.692

9. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 2016;44(D1):D471-80.
10. Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, et al. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. Nucleic Acids Res. 2020;48(D1):D445-D53.
11. Karp PD, Riley M, Paley SM, Pellegrini-Toole A. The MetaCyc Database. Nucleic Acids Res. 2002;30(1):59-61.
12. Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, et al. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res. 2004;32(Database issue):D438-42.
13. Green ML, Karp PD. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. BMC Bioinformatics. 2004;5:76.
14. Makarova KS, Koonin EV. Comparative genomics of Archaea: how much have we learned in six years, and what's next? Genome Biol. 2003;4(8):115.
15. Verhees CH, Kengen SW, Tuininga JE, Schut GJ, Adams MW, De Vos WM, et al. The unique features of glycolytic pathways in Archaea. Biochem J. 2003;375(Pt 2):231-46.
16. Zillig W, Stetter KO, Wunderl S, Schulz W, Priess H, Scholz I. The Sulfolobus-"Caldariella" group: Taxonomy on the basis of the structure of DNA- dependent RNA polymerases. Arch Microbiol. 1980;125:259-69.
17. She Q, Singh RK, Confalonieri F, Zivanovic Y, Allard G, Awayez MJ, et al. The complete genome of the crenarchaeon Sulfolobus solfataricus P2. Proc Natl Acad Sci U S A. 2001;98(14):7835-40.
18. De Rosa M, Gambacorta A, Nicolaus B, Giardina P, Poerio E, Buonocore V. Glucose metabolism in the extreme thermoacidophilic archaebacterium Sulfolobus solfataricus. Biochem J. 1984;224(2):407-14.
19. Ulas T, Riemer SA, 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(8):e43401.
20. Ettema TJ, Ahmed H, Geerling AC, van der Oost J, Siebers B. The non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase (GAPN) of Sulfolobus solfataricus: a key-enzyme of the semi-phosphorylative branch of the Entner-Doudoroff pathway. Extremophiles. 2008;12(1):75-88.
21. Figueiredo AS, Kouril T, Esser D, Haferkamp P, Wieloch P, Schomburg D, et al. Systems biology of the modified branched Entner-Doudoroff pathway in Sulfolobus solfataricus. PLoS One. 2017;12(7):e0180331.
22. Haferkamp P, Tjaden B, Shen L, Brasen C, Kouril T, Siebers B. The Carbon Switch at the Level of Pyruvate and Phosphoenolpyruvate in Sulfolobus solfataricus P2. Frontiers in microbiology. 2019;10:757.
23. Koendjbiharie JG, van Kranenburg R, Kengen SWM. The PEP-pyruvate-oxaloacetate node: Variation at the heart of metabolism. FEMS microbiology reviews. 2020.
24. Khayatt BI. Automated Reconstruction and Manual Curation of Amino Acid Biosynthesis Pathways in Sulfolobus solfataricus P2. Ibn Al Haitham J for Pure and Appl Sci. 2019;32(3):1-18.
25. Koonin EV, Galperin MY. Sequence-Evolution-Function. Computational Approaches in Comparative Genomics: Kluwer Academic; 2002.
26. Huynen MA, Dandekar T, Bork P. Variation and evolution of the citric-acid cycle: a genomic perspective. Trends Microbiol. 1999;7(7):281-91.
27. Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. Nucleic Acids Res. 2002;30(1):42-6.

https://doi.org/10.24126/jobrc.2023.17.1.692

28. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. Nucleic Acids Res. 2019;47(D1):D590-D5.
29. Schomburg I, Chang A, Placzek S, Sohngen C, Rother M, Lang M, et al. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. Nucleic Acids Res. 2013;41(Database issue):D764-72.
30. Magrane M, Consortium U. UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford). 2011;2011:bar009.
31. Gabaldon T, Huynen MA. Prediction of protein function and pathways in the genome era. Cell Mol Life Sci. 2004;61(7-8):930-44.
32. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389-402.
33. Lamble HJ, Heyer NI, Bull SD, Hough DW, Danson MJ. Metabolic pathway promiscuity in the archaeon Sulfolobus solfataricus revealed by studies on glucose dehydrogenase and 2-keto-3-deoxygluconate aldolase. J Biol Chem. 2003;278(36):34066-72.
34. Madern D, Ebel C, Mevarech M, Richard SB, Pfister C, Zaccai G. Insights into the molecular relationships between malate and lactate dehydrogenases: structural and biochemical properties of monomeric and dimeric intermediates of a mutant of tetrameric L-[LDH-like] malate dehydrogenase from the halophilic archaeon Haloarcula marismortui. Biochemistry. 2000;39(5):1001-10.
35. Ahmed H, Ettema TJ, Tjaden B, Geerling AC, van der Oost J, Siebers B. The semi-phosphorylative Entner-Doudoroff pathway in hyperthermophilic archaea: a re-evaluation. Biochem J. 2005;390(Pt 2):529-40.
36. Kim S, Lee SB. Identification and characterization of Sulfolobus solfataricus D-gluconate dehydratase: a key enzyme in the non-phosphorylated Entner-Doudoroff pathway. Biochem J. 2005;387(Pt 1):271-80.
37. Buchanan CL, Connaris H, Danson MJ, Reeve CD, Hough DW. An extremely thermostable aldolase from Sulfolobus solfataricus with specificity for non-phosphorylated substrates. Biochem J. 1999;343 Pt 3:563-70.
38. Theodossis A, Walden H, Westwick EJ, Connaris H, Lamble HJ, Hough DW, et al. The structural basis for substrate promiscuity in 2-keto-3-deoxygluconate aldolase from the Entner-Doudoroff pathway in Sulfolobus solfataricus. J Biol Chem. 2004;279(42):43886-92.
39. Bartolucci S, Rella R, Guagliardi A, Raia CA, Gambacorta A, De Rosa M, et al. Malic enzyme from archaebacterium Sulfolobus solfataricus. Purification, structure, and kinetic properties. J Biol Chem. 1987;262(16):7725-31.
40. Witzmann S, Bisswanger H. The pyruvate dehydrogenase complex from thermophilic organisms: thermal stability and re-association from the enzyme components. Biochim Biophys Acta. 1998;1385(2):341-52.
41. Fukuda E, Wakagi T. Substrate recognition by 2-oxoacid:ferredoxin oxidoreductase from Sulfolobus sp. strain 7. Biochim Biophys Acta. 2002;1597(1):74-80.
42. Ettema TJ, Makarova KS, Jellema GL, Gierman HJ, Koonin EV, Huynen MA, et al. Identification and functional verification of archaeal-type phosphoenolpyruvate carboxylase, a missing link in archaeal central carbohydrate metabolism. J Bacteriol. 2004;186(22):7754-62.
43. Fukuda W, Fukui T, Atomi H, Imanaka T. First characterization of an archaeal GTP-dependent phosphoenolpyruvate carboxykinase from the hyperthermophilic archaeon Thermococcus kodakaraensis KOD1. J Bacteriol. 2004;186(14):4620-7.

https://doi.org/10.24126/jobrc.2023.17.1.692

44. Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, et al. The BioCyc collection of microbial genomes and metabolic pathways. Brief Bioinform. 2019;20(4):1085-93.

استخدام المعلوماتية الحيوية في معالجة و تنظيم المسارات الأيضية: أيض الكربوهيدرات ودورة كريبس في الآركيون *Sulfolobus solfataricus* P2

برزان ابراهيم خياط

قسم الموارد الطبيعية، كلية علوم الهندسة الزراعية، جامعة السليمانية
بكرجو، السليمانية، كردستان-العراق
البريد الإلكتروني: barzan.khayatt@univsul.edu.iq

الخلاصة

خلفية عن الموضوع: تتطلب عمليات تكامل المعلوماتية الأيضية والجينومية في قواعد بيانات خاصة بكائنات حية جهودا شاملة ومكثفة. باثولوجيك (PathoLogic) ؛ احدى مكونات باقة برنامج أدوات المسار (Pathway Tools) يمكنه إنشاء قواعد بيانات مسار/جينوم كاملة (Pathway/Genome Databases) (PGDBs) من التسلسل الجينومي وملفات الشروح التوضيحية الجينومية (Genome Annotation Files) لأي كائن حي. يستخدم باثولوجيك MetaCyc كقاعدة معارف مرجعية في تنبؤه و تأسيسه للمسارات الأيضية. الهدف من هذه الدراسة هو تطبيق المعلوماتية الحيوية (Bioinformatics) لإنشاء وتنظيم قاعدة بيانات مسار/جينوم (PGDB) خاص بالكرين آركيون (Crenarchaeon) *Sulfolobus solfataricus* P2. ينمو هذا الآركيون بشكل مثالي في بيئة درجة حرارتها $80C^{o}$ و رقم حموضتها 4-2. تم إطلاق الجينوم الكامل لهذا الآركيون في 2001. كثيرا ما تحتاج قواعد البيانات هذه إلى عمليات تنسيق وتنظيم يدوية لملأ الفجوات الأيضية (Missing Enzymes) التي فشل البرنامج في الكشف عنها.

المواد و طرق العمل: إستخدمنا أدوات المسار (Pathway Tools) ضمن منهج متكامل في مجال المعلوماتية الحيوية من أجل إنشاء تنظيم واكتمال المسارات الأيضية للكربوهيدرات Entner-Doudoroff (ED) ودورة كريبس (TCA cycle). واتبعت الدراسة أساليب للبحث ضمن المؤلفات المنشورة في هذا المجال، فضلا عن أساليب للتنبؤ الخاصة بوظائف البروتين و القائمة على التماثل أو التنادد (homology) و التشابه الاورثولوجي (orthology) و السياق الموقعي للجينات (gene context). ومن أجل تعديل وتنظيم و اضافة الأنزيمات في مواقعها ضمن المسارات الأيضية، استخدم مكون المحرر Editors من برنامج أدوات المسار.

النتائج: حقق نهج التنظيم والمعالجة باتباع المعلوماتية الحيوية (Bioinformatics Curation) في هذه الدراسة تعديلاً في عدد المسارات الأيضية في قاعدة البيانات بإضافة مسارات إضافية لم يتم اكتشافها بواسطة PathoLogic حيث اضيفت مسارات ايضية جديدة مثل ED شبه الفوسفوريل (semi-phosphorylated ED) ومتغيير جديد لدورة TCA إلى قاعدة البيانات (PGDB) الخاصة بالآركيون *S. solfataricus* P2. عملية المعالجة تضمنت أيضا ملأ الثقوب الأيضية (الأنزيمات المفقودة) في المسارات قيد الدراسة.

الاستنتاج: أدت هذه المعالجة باتباع المعلوماتية الحيوية إلى تحسين قاعدة البيانات PGDB الخاصة ب *S. solfataricus* P2 و التي يمكن أن تستخدم كقاعدة معارف مرجعية للشروح الجينومية (Genome Annotations) وعمليات إعادة بناء و معالجة المسارات الأيضية للكائنات الأخرى، ولا سيما الآركيا (Archaea) الوثيقة الصلة.

الكلمات المفتاحية: المعالجة باتباع المعلوماتية الحيوية، كرين آركيوتا، دورة كريبس، مسار ED، أدوات المسار، دراسة باستخدام الحاسوب.

https://doi.org/10.24126/jobrc.2023.17.1.692