Functional annotation of an ecologically important protein from *Chloroflexus aurantiacus* involved in polyhydroxyalkanoates (PHA) biosynthetic pathway

Atikur Rahman¹ · Tasmina Ferdous Susmi¹ · Farzana Yasmin¹ · Md. Ekramul Karim² · Mohammad Uzzal Hossain³

Received: 4 July 2020 / Accepted: 28 September 2020 / Published online: 10 October 2020
© Springer Nature Switzerland AG 2020

Abstract

Chloroflexus aurantiacus J-10-fl strain is a thermophilic gram-negative bacterium that possesses many proteins in its genome; some are considered as hypothetical proteins. The use of bioinformatics tools can assist in understanding this organism through structural and functional annotation. Our study aimed to assign structure and function to an ecologically important hypothetical protein present in the bacterial genome. To analyze the hypothetical protein (WP_012259469.1), we used an in silico approach to find out various properties like physiochemical characteristics, subcellular localization, 3D structure, protein–protein interaction and functional annotation. Protein–protein interactions were obtained from the STRING database. In silico analysis revealed that the protein is a soluble protein with predominantly alpha-helices in its secondary structure. The 3D model of the protein has been found to be novel and possessed expected quality as assessed by several quality assessment tools. Functional annotation indicated that the protein acted like a (R)-specific enoyl-CoA hydratase which is linked with PHA synthesis. Protein–protein interactions also showed with high confidence that the protein interacted with a protein synthesizer of enoyl-CoA hydratase involved in PHA biosynthesis. Polyhydroxyalkanoate (PHA) is a novel polyester used as a biodegradable thermoplastic and plays a crucial role in environmental biodegradability and biocompatibility. An extensive variety of microorganisms produces PHA for intracellular carbon and energy storage purposes. In the present investigation, we bioinformatically confirmed that the WP_012259469.1 is associated with the PHA biosynthesis pathway. From our analyses, we also predict that polyhydroxyalkanoate (PHAs) has the potential to become an alternative source of renewable and biodegradable polyesters.

Keywords *Chloroflexus aurantiacus* · Hypothetical protein · Homology modeling · Functional annotation · Protein–protein interaction · PHA synthesis

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s42452-020-03598-x>) contains supplementary material, which is available to authorized users.

✉ Mohammad Uzzal Hossain, uzzalbge10044@gmail.com; Atikur Rahman, atikur.just@gmail.com; Tasmina Ferdous Susmi, tsusmi7@gmail.com; Farzana Yasmin, farzana.just@gmail.com; Md. Ekramul Karim, ekrammbio@gmail.com | ¹Department of Genetic Engineering and Biotechnology, Faculty of Biological Science and Technology, Jashore University of Science and Technology, Jashore, Bangladesh. ²Department of Microbiology, Faculty of Life and Earth Sciences, Jagannath University, Dhaka 1100, Bangladesh. ³Bioinformatics Division, National Institute of Biotechnology, Ganakbari, AshuliaSavar, Dhaka 1349, Bangladesh.

SN Applied Sciences (2020) 2:1810 | <https://doi.org/10.1007/s42452-020-03598-x>

1 Introduction

Chloroflexus aurantiacus is a well-known gram-negative bacterium having remarkable features like anoxygenic, thermophilic, filamentous, gliding and phototrophic properties [1]. Excluding other anoxygenic phototrophs, *C. aurantiacus* flourish in regions with intense heat, and their optimum growth temperature ranges from 50–60 °C [2]. Predominantly, they can acclimate in a diverse array of environmental conditions like hot springs, lakes, river water and sediments with high sulfide conditions [3]. Interestingly, the bacterial species share some common features with green sulfur bacteria and purple photosynthetic bacteria especially chimeric photosystem [4]. Considering the impressive photosynthetic and thermophilic characteristics, the bacterium prompted researchers to investigate several proteins associated in heat adaptation, synthesizing industrially valuable enzymes such as maltotetraose producing amylase (EC number 3.2.1.60), propionyl coenzyme A synthase (EC number 6.2.1.17), malonyl-CoA reductase (EC number 1.2.1.75), metalloprotease-I (EC number 3.4.24.80) and so on in recent years [5–7]. Besides, finding novel characters in its genome has gained much interest in recent times through analyzing genome databases.

Due to an immense evolution in computational biology, different servers and tools have been developed which helps predict protein function, identify sequence similarity, conduct phylogenetic analysis, evaluate active site residue similarity, protein–protein interaction, protein–ligand interaction, gene expression profiling, conserved domains, motif phosphorylation regions, etc. [8–13]. According to Muhammad Naveed et al., 50–60% of genes has a known function in most of the completely sequenced genomes. The number of genes having unknown functions called hypothetical proteins are present in each organism's genome [14]. In other words, hypothetical proteins are a category of the protein whose existence is not proved by any experimental evidence but can be portended to be expressed from an open reading frame (ORF) [15]. The hypothetical proteins can be classified as either uncharacterized protein families (UPF) or domain of unknown functions (DUF) [16]. UPFs are predicted and experimentally proven to exist but have not been characterized or linked to a known gene, whereas DUFs are experimentally identified proteins in the absences of known functional or structural domains [17, 18]. Those proteins having unknown characters can be attributed to homology-based annotation due to compatibility with known proteins [16]. Most of these hypothetical protein analyses can decipher the knowledge about new structures, functions, pathways

that are believed to play a crucial role in the cellular functions.

Analysis of hypothetical proteins using bioinformatics tools offers several advantages such as determination of 3D structural conformation, identification of new domains and motifs, assessment of new cascades and pathways, phylogenetic profiling as well as functional annotation [19]. This achieved information may also confer efficient pharmacological targets and help in potential drug design against several diseases [20–22].

Due to the relevance of the properties of *C. aurantiacus* J-10-fl strain, this work aimed to assign functions to the hypothetical protein (accession number WP_012259469.1) present on the genome of this species. The outcome of the present study can contribute to understanding the mechanism of this bacterium's survival in the acute environment and its role in PHA synthesis pathway. Among the natural polymers, polyhydroxyalkanoates (PHAs) comprise a family of biodegradable polyesters that are produced by an extensive variety of microorganisms for intracellular carbon and energy storage purposes [23]. PHAs are known as a renewable biodegradable alternative source to petrochemical polymers [24]. The ever-increasing environmental pollution due to various organic and inorganic contaminants is a major global concern. Among the organic pollutants, petroleum-based polymers such as synthetic plastics are the cause of significant environmental degradation due to their xenobiotic nature and recalcitrant properties [25–27]. The main advantages of PHAs involve its environmental biodegradability and biocompatibility [28]. Scientists have focused on bioproduction of PHAs by using various microorganisms [29]. This aims to produce industrially acceptable environment-friendly biopolymers by PHAs in bioreactors [30]. However, the production costs of PHAs compared to conventional non-biodegradable plastics are still much higher than what it would take to compete in an industrial scale [31].

Moreover, this study can help to identify new biotechnological targets by adopting an integrated workflow that contains conventional annotation programs allied to PPI analysis, functional annotation and three-dimensional protein modeling.

2 Methods and materials

2.1 Hypothetical protein selection

The amino acid sequence of the hypothetical protein (WP_012259469.1) from *C. aurantiacus* J-10-fl was retrieved from the NCBI database (<https://www.ncbi.nlm.nih.gov>). The complete framework which shows all the

Fig. 1 Complete framework which shows all the tools used for structural and functional annotation of the hypothetical proteins

tools used for the structural and functional annotation of the query protein is shown in Fig. 1.

2.2 Characterization of the physicochemical properties

The physicochemical properties of the hypothetical protein were predicted using the ProtParam tool in the ExPASy server (<https://www.expasy.org/>) [32]. The server predicts physicochemical properties including molecular weight, theoretical pI, amino acid composition, the total number of positive and negative residues, instability index, aliphatic index and grand average of hydropathicity (GRAVY) [33–35].

2.3 Evaluation of subcellular localization

The subcellular localization was predicted by the using PSORTb server (<https://www.psort.org/psortb/>) that contains information through the association of both laboratory experimentations and computational prediction for the bacterial and archaeal sequences [36]. STRING (<https://string-db.org/>) was also used for cross-checking the result derived from PSORTb [37]. The solubility of the protein was identified using the SOSUI server (<https://bio.tools/sosui>) [38].

2.4 Analysis of the secondary structure

To predict the secondary structure of the hypothetical protein, the server SOPMA (https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopma.html) was implicated. The result derived from the SOPMA server was validated by exploiting PRSIPRED (<https://bioinf.cs.ucl.ac.uk/psipred/>), predict protein and SYMPRED (<https://www.ibi.vu.nl/programs/sympredwww/>) servers [39, 40].

2.5 3D Structure modeling of the hypothetical protein

The FASTA sequence of the hypothetical protein was subjected to possible 3D structure model prediction through SWISS-MODEL (<https://swissmodel.expasy.org/>) server by automated comparative modeling [41]. The confidence of the predicted structure was also validated by Robetta (<https://rosetta.bakerlab.org/>) Web server that provides automated tools for protein structure prediction and analysis [42].

2.6 Quality assessment of built model

Several quality assessment tools were used to estimate the reliability of the predicted 3D structure model of the hypothetical protein. For assessment of the 3D model quality SAVES (<https://servicesn.mbi.ucla.edu/SAVES/>), RAMPAGE (<https://mordred.bioc.cam.ac.uk/~rapper/rampage.php>) and PROFUNC (<https://www.hs-ls.pitt.edu/obrc/index.php?page=URL1125088260>), servers were used [43]. The Ramachandran plot for the model was built using the RAMPAGE server to visualize the backbone dihedral angles ψ against ϕ of amino acid residues in protein structure. The protein structure was cross-checked with the help of the ERRAT server. Verify 3D (<https://servicesn.mbi.ucla.edu/Verify3d/>) server was used to determine the compatibility of an atomic model (3D) with its amino acid sequence as well as comparing the results to standard structures [44, 45].

2.7 Functional annotation of the hypothetical protein

Several programs and databases such as InterPro (<https://www.ebi.ac.uk/interpro/>), Pfam (<https://pfam.xfam.org/>),

smart (<https://smart.embl-heidelberg.de/>), cdart, superfamily (<https://supfam.org/>), hhpred (<https://toolkit.tuebingen.mpg.de/tools/hhpred>) were applied to decipher the function of the selected hypothetical protein [46–48]. These analyses laid out the information about the protein domain as well as protein family based on homology search on multiple databases.

Manual annotation of the homologous protein from related organisms using blast (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) was performed against non-redundant Uniprot/Swissprot sequences where several parameters like the e-value, query cover and maximum score were calculated [49]. Hhpred server was used to implement pairwise comparison to search various databases (PDB, SCOP, Pfam, SMART and CDD) which select template manually [50].

Interpro and Pfam database describes protein domain and family based on a group of proteins that share a common evolutionary origin, reflected by their related functions and similarities in sequence or structure [51, 52]. Protein families are often arranged into hierarchies with proteins that share a common ancestor are subdivided into smaller, more closely related groups [53]. CDART and SMART were used to search conserved domains using the Conserved Domain Database (CDD) and protein function based on domain architecture, respectively [47, 48]. CATH-Gene3D (<https://www.cathdb.info/>) server was also used to identify protein families to predict the function of an unknown protein. CATH is a classification of protein structures and Gene3D uses the information in CATH to predict the locations of structural domains to predict function [54].

2.8 Protein–protein interaction

STRING database (<https://string-db.org/>) is known for the prediction of protein–protein interactions. It aims to collect and integrate the information by direct (physical) and

indirect (functional) associations of data for a large number of organisms [55]. STRING database search was performed to find out the plausible interactors with the hypothetical protein. The validity of protein–protein interaction was checked by the cytoscape program. Cytoscape (<https://cytoscape.org/>) is an open-source software platform for visualizing molecular interaction networks and biological pathways that allows the user to add an arbitrary node, edge and network information, integrating these networks with annotations, gene expression profiles and other state data [56].

3 Results

3.1 Retrieval of hypothetical protein

We selected a hypothetical protein (WP_012259469.1) from the organism *C. aurantiacus* J-10-fl as it is a photosynthetic, thermophilic, heat-stable bacterium. This hypothetical protein contains 119 amino acids. Further information collected from the NCBI database regarding this hypothetical protein is given in Table 1.

3.2 Physicochemical properties of the protein

The physicochemical properties of a protein can be characterized by analysis of the analogous properties of the amino acids. The physicochemical characteristics of the hypothetical protein were obtained from the ExPASy ProtParam server. The molecular weight of the hypothetical protein was 13,011.15 Dalton, and the isoelectric point was 5.47. The instability index (II) was computed to be 24.72. The GRAVY index was 0.181. The most abundant amino acid residue was observed to be leucine (11.8%), followed by threonine and valine, both of which were in equal percentage (10.9%). The lowest number of amino acids was

Table 1 Retrieval of the hypothetical protein

Protein individualities	Hypothetical protein information
Locus	WP_012259469
amino acid	119 aa
Definition	hypothetical protein [<i>Chloroflexus aurantiacus</i>]
Accession	WP_012259469
Version	WP_012259469.1
GI	501,216,451
Keywords	RefSeq
source strain	<i>Chloroflexus aurantiacus</i> J-10-fl
Organism	<i>Chloroflexus aurantiacus</i>
Fasta sequence	>WP_012259469.1 hypothetical protein [<i>Chloroflexus aurantiacus</i>] MTALTVMQATVQRVIDQALIDEYARLCDDQPQKTVPEPLIGALFSHVLG- VKLPAGGSVYLQELIFHDTARIGEMLTATVTITHRTDKPLVTLRTQCTGS- DQRLIVDGEALVMVPQV

of cysteine, phenyl-alanine and tyrosine (1.7%). The amino acid sequence had 13 negatively charged residues and 10 positively charged residues. The molecular formula of hypothetical protein is expressed as $C_{573}H_{948}N_{158}O_{173}S_6$. All the physicochemical properties are given in Table 2.

3.3 Subcellular localization

Prediction of subcellular localization of a protein involves determining where a protein resides in a cell. The functional properties, determination of interaction, genome annotation of a protein are highly influenced by their subcellular location. Protein subcellular localization predicted from the PSORTb server indicated that this protein belongs to the cytoplasmic portion. The localization score exhibited that this protein contains the cytoplasmic score is 2.00, cytoplasmic-membrane is 2.00 and the periplasmic score is 2.00. The subcellular localization score from the PSORTb server summarized that the hypothetical protein is transmembrane in nature. Phyre2 server also showed that the hypothetical protein is a transmembrane protein which is shown in supplementary Fig. 1. The result of the SOSUI server revealed that the Average hydrophobicity of the hypothetical protein is 0.180672. Thus, this amino acid sequence of WP_012259469.1 is a soluble protein.

3.4 Secondary structure analysis of WP_012259469.1

Protein structure is closely related to its function. The secondary structure, viz. helix, sheet, turn and the coil of any protein have a great association with the structure, function and interaction of the protein [57–60]. The secondary structure of the hypothetical protein resulted from the SOPMA server. From the result from the server, alpha helices were found to be most predominant 36.97%, random coil 35.29%, extended strand 21.85% and beta-turn 5.88%. The results obtained from this database were also cross-checked from SYMPRED and PRISPPRED server which provides a similar result. The representative secondary structure of the protein WP_012259469.1 is given in Fig. 2.

3.5 Three-dimensional structure exploration

For the prediction of protein structure, homology modeling is an useful tool when only amino acid sequence data are available. Determining protein function, structural information is often more valuable than the sequence alone. Homology modeling or comparative modeling is based on the biological fact that when two sequences share structural similarity and their respective functions are also similar. Sequence identity over 30% to a known structure can often be predicted with an accuracy equivalent to a low-resolution X-ray structure. Conversely, sequence identity below 30% could never achieve the minimum acceptable accuracy in homology-based structure prediction [61]. The structure of the protein WP_012259469.1 was obtained by homology modeling using the SWISS-MODEL server. The SWISS-MODEL server is based on the analysis of the similar structure templates which is consequently performed by the BLASTP search to identify similar templates to build a 3D model structure. We set a sequence identity cutoff threshold as $\geq 30\%$. The highest template identity for the WP_012259469.1 was 37.61% with (R)-specific enoyl-CoA hydratase protein (PDB id 1IQ6), and the oligo-state was homo-dimer. The crystallographic resolution of the template used to model this protein was 1.50 Å by adopting the X-ray diffraction method [62]. Global quality estimate, local quality estimate, comparison of the protein size residue, model template alignment were also explored from this server. Structure prediction from Robetta by comparative modeling also gave high confidence (0.89) of the predicted model where 1.0 good, 0.0 bad. According to Ramachandran plot, amino acids in the most favored regions were found to be 93.8%. Figure 3 contains the figures obtained from swiss-model including global quality estimate, local-quality estimate. Ramachandran plot is in Fig. 4, and a possible three-dimensional structure is given in Fig. 5. Model template alignment can be found in Supplementary Fig. 2. For further assessment of the structure and establishing the compatibility of the protein model, the SAVES server was

Table 2 Physicochemical properties of the hypothetical protein WP_012259469

Properties	Value
Molecular weight	13,011.15
Theoretical pl	5.47
Total number of negatively charged residues (Asp + Glu):	13
Total number of positively charged residues (Arg + Lys):	10
Formula:	$C_{573}H_{948}N_{158}O_{173}S_6$
Instability index (II)	24.72
Aliphatic index:	111.34
Grand average of hydropathicity (GRAVY):	0.181

Fig. 2 Secondary structural model, **a** secondary structure predicted from multiple alignments by SOPMA server, **b** SYMPRED server predicted structure by using the PSI-BLAST algorithm under their

corresponding amino acid. **c** Simultaneous analyses of secondary structure from the PSIPRED server

Fig. 3 The swiss-model template library was searched with BLAST and HHBlits for evolutionary-related structures matching the target sequence to predict the three-dimensional structure of the protein

WP_012259469.1, **a** global quality estimate, **b** local quality estimate, **c** comparison of the protein size residue

used which provided ERRAT value overall quality factor A: 96.1165 and B: 86.7925 that is presented in Fig. 6. The three-dimensional figure of the protein was also cross-checked by using varify3d of the SAVES server and Pro-Func server.

3.6 Functional annotation of WP_012259469.1

Comparing the amino acid sequence of a protein against all functionally related sequences in the databases is the conventional pathway to annotate and describe the

Fig. 4 Ramachandran plot for the 3D model of the studied hypothetical protein of *C. aurantiacus* by the SWISS-MODEL server

function of a hypothetical protein. Searching basic local alignment search tool (BLAST) against non-redundant Uniprot/Swissprot sequences is the option for obtaining similar protein to annotate function. Table 3 designates the database of BLASTP search result against non-redundant Uniprot sequence including maximum score, total score, query cover, *E*-value, percent identity and accession number of the similar protein. Hhpred server result in supplementary Fig. 3 also shows the analogs data like the BLAST. Both the BLAST and Hhpred server narrate that this protein acts like (R)-specific enoyl-CoA hydratase which is linked with PHA synthase. Pfam server describes the protein family of the hypothetical protein of *C. aurantiacus* which is the MaoC dehydrat_N family (N-terminal half of MaoC dehydratase). CATH-Gene3D server is a core data resource that shows the superfamily of the protein which is Hotdog Thioesterase along with the functional family which is (R)-specific enoyl-CoA hydratase [(R)-specific enoyl-CoA hydratase from *Aeromonas caviae* involved in polyhydroxyalkanoate biosynthesis]. MaoC like hydratase (maoc) belongs to (R) enoyl-CoA hydratase group which

is involved in beta-oxidation and PHA biosynthetic pathways. For the efficient production of PHAs synthesis, (R)-hydratase catalysis mechanism plays a very critical role. SMART database describes the family of WP_012259469.1 is MaoC-like dehydratase in the taxonomy of Bacteria (superkingdom).

3.7 Protein–protein interaction

Search tool for the retrieval of interacting genes/proteins (STRING) is a Web-based database of known and predicted protein–protein interactions that include direct (physical) and indirect (functional) associations and form interactions aggregated from other (primary) databases. The function of a target protein and drug ability of molecules can be predicted by analyzing protein–protein interaction. Protein–protein interaction network analysis obtained from the STRING database revealed that our query protein Caur_3633 interacts with other proteins whose functions are experimentally known, and some of these interactions were mostly hypothetical. The protein Caur_3633 showed a high

Fig. 5 **a** Models are built based on the target-template alignment by coordinates which are conserved between the target, and the template is copied from the template to the model from SWISS-MODEL; **b** Possible 3D structure modeling structure from SAVES server; **c** Structure verified from Phyre2 server and **d** Robetta webtool

Fig. 6 ERRAT value overall quality factor A: 96.1165 and B: 86.7925 from the SAVES server

confidence interaction with a protein Caur_3631 producers of MaoC domain protein dehydratase (score 0.989). Our protein interacts with the proteins (Caur_3632, Caur_3221, Caur_0613, Caur_3630, Caur_3638, Caur_3629, Caur_1085, Caur_3634, Caur_0232) that their functional patterns were predicted from Pfam database. The network stats are the number of nodes-31, the number of edges-162, average node degree-10.5, average local clustering coefficient-0.722, PPI enrichment p-value < 1.0e-16 that represent this network

has significantly more interactions than expected. Figure 7 displays the protein–protein interaction of the protein WP_012259469.1.

Table 3 Data from BLASTP result against non-redundant uniprot/swissprot sequences

No	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
1	(R)-specific enoyl-CoA hydratase [Aeromonas caviae]	66.2	66.2	65%	6e-14	44.87%	O32472.1
2	(R)-specific enoyl-CoA hydratase; AltName: (R)-specific trans-2,3-enoylacyl-CoA hydratase [Rhodospirillum rubrum ATCC 11170]	60.5	60.5	93%	1e-11	36.51%	Q2RQ36.1
3	3-hydroxybutyryl-CoA dehydratase; AltName: R-specific crotonase [Methylobacterium extorquens AM1]	55.8	55.8	63%	1e-09	38.16%	C5AZ74.1
4	Methionyl-tRNA formyltransferase [Magnetospirillum magnetotacticum AMB-1]	29.3	29.3	25%	8.0	36.67%	Q2WAS9.1

Fig. 7 Protein-protein interaction network of the hypothetical protein of *C. aurantiacus*; data from STRING server. The color nodes represent query proteins and first shell of interactors, white nodes

represent the second shell of interactors, empty nodes represent proteins of unknown 3D structure, filled nodes some 3D structure is known or predicted

4 Discussion

Polyhydroxyalkanoates (PHA) have long been generated by certain microbes as bioplastics [23]. However, their commercialization is still in transit. A couple of issues have been identified as obstacles to PHA commercialization: significant expense and dangers on sub-atomic loads (Mw) and structures, subsequently precariousness on thermo-mechanical properties [63]. Further development of PHA-producing microorganisms ought to be centered around exploiting suitable microscopic organisms, particularly extremophiles. Increment on carbon substrates to PHA transformation and maintaining Mw of PHA are important considerations in the regard [64].

In this study, we tried to identify a novel protein involved in the PHA biosynthetic pathway. Since *Chloroflexus aurantiacus* bacteria has already been recognized for the evolution of photosynthesis, we targeted a hypothetical protein from this bacterium to examine its involvement in the PHA biosynthetic pathway. Therefore, we retrieved the amino acid sequence of the hypothetical protein (WP_012259469.1) from *C. aurantiacus* for further analysis.

In case of protein 3D model prediction using SWISS-MODEL, WP_012259469.1 showed highest template identity of 37.61% with an entry from the PDB database. Sequence coverage was also very high (98%). Hypothetical proteins are novel, and usually no similar template is present in any databases. Quality assessment of the

modeling process, the GMQE, the QSQE provides estimates of the expected quality of the resulting model at the tertiary and quaternary structure level [65]. QMEAN is a composite estimator based on different geometrical properties employed statistical potentials of mean force to generate quality estimates on a global and local scales. Interpretation of the obtained model quality the QMEAN global score is transformed into a Z-score, indicating whether the model scores as it would be expected from experimentally determined structures of similar size [66]. Robetta server utilizes a comparative modeling method to predict 3D structure of the hypothetical protein [42]. The results obtained from this server have high confidence which validates the predicted protein structure. 3D structure assessment of the hypothetical protein by the ab initio method through the Phyre2 server gives 100% confidence in the model by the single highest scoring template. Crystal structure from *A. caviae* which involved polyhydroxyalkanoate (PHA) biosynthesis was used as a template to generate three-dimensional coordinates for the query protein [62]. Ramachandran plot analysis showed an overall G-factor value 0.36 which indicates that 93.8% residues belonged in most favored regions. Residues in additional allowed regions and generously allowed regions were 5.3% and 0.4%, respectively, which evaluated the quality of the model to be good and reliable as it is generally accepted that if 90% residues are in the most favored regions, it is likely to be a reliable model [67]. The data regarding model quality is given in Fig. 4. The model is compatible with its sequence as Verify 3D analysis revealed that 78.63% of the residues had an average 3D–1D score of ≥ 0.2 . ERRAT is a so-called “overall quality factor” used to evaluate the amino acid environment for non-bonded atomic interactions. Higher scores indicate higher quality, and the generally accepted range is > 50 for a high-quality model [68]. The ERRAT server provided the overall quality factor to be 90.9828 for the current 3D model. The ERRAT output data is given in Fig. 6.

There is a relationship between three-dimensional structures and their biological functions of a protein where the relationship can be identified by using various bioinformatics databases to understand protein structure and its function. Therefore, the functional prediction has identified the function of proteins by adopting the sequence to function method. *C. aurantiacus* hypothetical protein WP_012259469.1 belongs to the hotdog thioesterase superfamily with MaoC like domain predicted by utilizing Web tools based on the homology-based method and structure-based methods. According to the data of Pfam (<https://pfam.xfam.org/>), the MaoC protein is found to share similarities with a wide variety of enzymes. Several bacterial proteins that are composed solely of this domain have (R)-specific enoyl-CoA hydratase activity [69, 70]. So

the functional family is (R)-specific enoyl-CoA hydratase which involves polyhydroxyalkanoate (PHA) biosynthesis from the β -oxidation cycle (Supplementary Fig. 3). The potential usage of bacterial polyhydroxyalkanoate (PHA) which is a novel polyester is as a biodegradable thermoplastic, and a wide variety of bacteria accumulate PHA within their cells as a source of carbon and energy under nutrient-starved conditions [71].

Protein–protein interaction (PPI) network analysis was performed for the query protein WP_012259469.1 to evaluate the interactions with other proteins of *C. aurantiacus*. We obtained the interaction network within this group of proteins given in Fig. 7. The color of the nodes evaluates the degree of interaction on the network; the darker the green the greater the interaction. The solid and discontinuous lines represent the physical and non-physical interactions, respectively. The IS value represents the thickness of the lines, such that thin lines have lower IS and thick lines have greater IS. The color of the lines represents the confidence of the interaction; a higher level of confidence is represented by yellow/green lines while experimental confidence is represented by the blue lines [72]. The proteins Caur_3633, Caur_3632 and Caur_3631 represented interactions with high confidence; they are also the neighborhood in the genome. The intergenic distance between Caur_3633, Caur_3632 and Caur_3632, Caur_3631 is 0 bp (score 0.845). Besides, homologous genes are neighbors in other genomes (score 0.076). Our protein (Caur_3633) also interacts with other proteins and showed a high confidence interaction with a protein producer of enoyl-CoA hydratase/isomerase which involved in PHA biosynthesis in bacteria (Fig. 7). Protein cellular functions are determined through their interaction with other proteins that led to the development of various experimental methods for determination [73]. Thus, protein–protein interaction studies are important to infer the function of unidentified proteins based on their interaction with the known proteome of a given organism. PPI network analysis helped in the development of computational methods to predict protein function and provides crucial information about biological pathways [74].

Overall, the protein 3D model was predicted, and quality assessment of the modeling process such as the GMQE, the QSQE, the QMEAN provided the expected quality of the resulting model. Ramachandran plot analysis indicated that 93.8% of residues were in most favored regions which concluded that the model is good and reliable. The protein functional family is (R)-specific enoyl-CoA hydratase which involves in polyhydroxyalkanoate (PHA) biosynthesis. Polyhydroxyalkanoates (PHAs) comprise a family of biodegradable polyesters that are produced by an extensive variety of microorganisms for intracellular carbon and energy storage purposes. Protein–protein interaction also showed

a high confidence interaction with a protein involved in PHA biosynthesis in bacteria.

5 Conclusion

Proteins play a crucial role in biological processes, and this hypothetical protein is from such a bacterial strain that possesses some unique properties. The present investigation was performed to find out a novel protein that had a beneficial role in the survival of *C. aurantiacus* in harsh environmental conditions and can also be used as a tool in genetic engineering for commercial production of PHA. The effective utilization of various bioinformatics tools has facilitated successful structural and functional annotation of the hypothetical protein found in *C. aurantiacus*. This hypothetical protein helps the bacteria to synthesize polyhydroxyalkanoates (PHA) which is an essential polyester to survive in nutrient strived conditions. This finding may encourage researchers to work on PHA delivering microorganisms which can be further utilized for biofilter production. Hence, the outcomes of the analysis indicate the reliability of bioinformatics tools used in this study as well as the scope of further experimental research on the hypothetical protein.

Acknowledgements There is no acknowledgement for this study.

Authors' contributions A.R. and M.U.H helped in conceptualization; A.R. and T.F.S were involved in formal analysis; A.R. contributed to methodology; A.R, T.F.S. and F.Y contributed to original draft; M.E.K was involved in writing, review and editing; M.U.H contributed to conceptualization, writing, review and editing. All authors have read and agreed to the manuscript.

Data availability All the data will be available for this article.

Compliance with Ethical Standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Consent for publication Not applicable.

Human or Animal Rights Not applicable.

References

- Pierson BK, Castenholz RW (1974) A phototrophic gliding filamentous bacterium of hot springs, *Chloroflexus aurantiacus*, gen and sp, nov. Arch Microbiol 100(1):5–24
- Feick RG, Fuller RC (1984) Topography of the photosynthetic apparatus of *Chloroflexus aurantiacus*. Biochemistry 23(16):3693–3700
- Oelze J, Fuller R (1983) Temperature dependence of growth and membrane-bound activities of *Chloroflexus aurantiacus* energy metabolism. J Bact 155(1):90–96
- Tang K-H, Barry K, Chertkov O, Dalin E, Han CS, Hauser LJ, Honchak BM, Karbach LE, Land ML, Lapidus A (2011) Complete genome sequence of the filamentous anoxygenic phototrophic bacterium *Chloroflexus aurantiacus*. BMC genomics 12(1):334
- Alber BE, Fuchs G (2002) Propionyl-coenzyme A synthase from *Chloroflexus aurantiacus*, a key enzyme of the 3-hydroxypropionate cycle for autotrophic CO₂ fixation. J Biol Chem 277(14):12137–12143
- Hügler M, Menendez C, Schägger H, Fuchs G (2002) Malonyl-coenzyme A reductase from *chloroflexus aurantiacus*, a key enzyme of the 3-hydroxypropionate cycle for autotrophic CO₂ fixation. J Bact 184(9):2404–2410
- Ratanakhanokchai K, Kaneko J, Kamio Y, Izaki K (1992) Purification and properties of a maltotetraose- and maltotriose-producing amylase from *Chloroflexus aurantiacus*. Appl Environ Microbiol 58(8):2490–2494
- Díaz DA, Barreto GE, Santos JG (2014) Structural and functional prediction of the hypothetical protein Pa2481 in *Pseudomonas aeruginosa* Pao1. Adv Comput Biol 232:47–55
- Canduri F, Fadel V, Basso LA, Palma MS, Santos DS, de Azevedo Jr WF (2005) New catalytic mechanism for human purine nucleoside phosphorylase. Biochem Biophys Res Commun 327(3):646–649
- Pereira JH, Canduri F, de Oliveira JS, da Silveira NJF, Basso LA, Palma MS et al (2003) Structural bioinformatics study of EPSP synthase from mycobacterium tuberculosis. Biochem Biophys Res Commun 312(3):608–614
- Canduri F, Cardoso Perez P, Caceres RA, de Azevedo WF (2007) Protein kinases as targets for antiparasitic chemotherapy drugs. Curr Drug Targets 8(3):389–398
- Gong J, Chen Y, Pu F, Sun P, He F, Zhang L et al (2019) Understanding membrane protein drug targets in computational perspective. Curr Drug Targets 20(5):551–564
- Tan J-X, Lv H, Wang F, Dao F-Y, Chen W, Ding H (2019) A survey for predicting enzyme family classes using machine learning methods. Curr Drug Targets 20(5):540–550
- Naveed M, Tehreem S, Usman M, Chaudhry Z, Abbas G (2017) Structural and functional annotation of hypothetical proteins of human adenovirus: prioritizing the novel drug targets. BMC Res Notes 10(1):706
- Bashir Z, Rizwan M, Mushtaq K, Munir A, Ali I (2017) In silico structural and functional prediction of Phaseolus vulgaris hypothetical protein PHA VU_004G136400g. J Proteom Bioinform 10:206–207
- Jaroszewski L, Li Z, Krishna SS, Bakolitsa C, Wooley J, Deacon AM, Wilson IA, Godzik A (2009) Exploration of uncharted regions of the protein universe. PLoS Biol 7(9):e1000205
- Varma PBS, Adimulam YB, Kodukula S (2015) In silico functional annotation of a hypothetical protein from staphylococcus aureus. J Infect Publ Health 8(6):526–532
- Mudgal R, Sandhya S, Chandra N, Srinivasan N (2015) De-DUF-ing the DUFs: deciphering distant evolutionary relationships of domains of unknown function using sensitive homology detection methods. Biol Dir 10(1):38
- Mills CL, Beuning PJ, Ondrechen MJ (2015) Biochemical functional predictions for protein structures of unknown or uncertain function. Comput Struct Biotechnol J 13:182–191
- de Azevedo WF (2011) Molecular dynamics simulations of protein targets identified in mycobacterium tuberculosis. Curr Med Chem 18(9):1353–1366
- de Azevedo WF, dos Santos GC, dos Santos DM, Olivieri JR, Canduri F, Silva RG et al (2003) Docking and small angle X-ray

- scattering studies of purine nucleoside phosphorylase. *Biochem Biophys Res Commun* 309(4):923–928
22. de Azevedo WF, Canduri F, Fadel V, Teodoro LG, Hial V, Gomes RA (2001) Molecular model for the binary complex of uropepsin and pepstatin. *Biochem Biophys Res Commun* 287(1):277–281
 23. Li Z, Yang J, Loh XJ (2016) Polyhydroxyalkanoates: opening doors for a sustainable future. *NPG Asia Mater* 8(4):e265
 24. Kovalcik A, Obruca S, Fritz I, Marova I (2019) Polyhydroxyalkanoates: their importance and future. *BioResources* 14(2):2468–2471
 25. Rajasekharreddy P, Huang C, Busi S, Rajkumari J, Tai M-H, Liu G (2019) Green synthesized nanomaterials as theranostic platforms for cancer treatment: principles, challenges and the road ahead. *Curr Med Chem* 26(8):1311–1327
 26. Moltó J, López-Sánchez B, Domene-López D, Moreno A, Font R, Montalbán M (2020) Pollutant emissions during the pyrolysis and combustion of starch/poly (vinyl alcohol) biodegradable films. *Chemosphere* 256:127107
 27. Gedanken A, Perkas N, Perelshtein I, Lipovsky A (2018) Imparting pharmaceutical applications to the surface of fabrics for wound and skin care by ultrasonic waves. *Curr Med Chem* 25(41):5739–5754
 28. Raza ZA, Abid S, Banat IM (2018) Polyhydroxyalkanoates: Characteristics, production, recent developments and applications. *Int Biodeterior Biodegrad* 126:45–56
 29. Steinbüchel A, Lütke-Eversloh T (2003) Metabolic engineering and pathway construction for biotechnological production of relevant polyhydroxyalkanoates in microorganisms. *Biochem Eng J* 16(2):81–96
 30. Koller M, Maršálek L, de Sousa Dias MM, Braunneg G (2017) Producing microbial polyhydroxyalkanoate (PHA) biopolyesters in a sustainable manner. *New Biotechnol* 37:24–38
 31. Muhammadi S, Afzal M, Hameed S (2015) Bacterial polyhydroxyalkanoates-eco-friendly next generation plastic: production, biocompatibility, biodegradation, physical properties and applications. *Green Chem Lett Rev* 8(3–4):56–77
 32. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *N Acids Res* 31(13):3784–3788
 33. Ikai A (1980) Thermostability and aliphatic index of globular proteins. *J Biochem* 88(6):1895–1898
 34. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 157(1):105–132
 35. Guruprasad K, Reddy BB, Pandit MW (1990) Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng, Des Sel* 4(2):155–161
 36. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 26(13):1608–1615
 37. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, Von Mering C (2012) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *N Acids Res* 41(D1):D808–D815
 38. Hirokawa T, Boon-Chieng S, Mitaku S (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* 14(4):378–379
 39. Lin K, Simossis VA, Taylor WR, Heringa J (2005) A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics* 21(2):152–159
 40. Geourjon C, Deleage G (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Bioinformatics* 11(6):681–684
 41. Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. *N Acids Res* 31(13):3381–3385
 42. Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the Robetta server. *N Acids Res* 32 (suppl_2): W526–W531
 43. Laskowski RA, Watson JD, Thornton JM (2005) ProFunc: a server for predicting protein function from 3D structure. *N Acids Res* 33 (suppl_2):W89–W93
 44. Kihara D, Chen H, Yang YD (2009) Quality assessment of protein structure models. *Curr Protein Pept Sci* 10(3):216–228
 45. Lüthy R, Bowie JU, Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. *Nature* 356(6364):83–85
 46. Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* 313(4):903–919
 47. Geer LY, Domrachev M, Lipman DJ, Bryant SH (2002) CDART: protein homology by domain architecture. *Genome Res* 12(10):1619–1623
 48. Letunic I, Doerks T, Bork P (2012) SMART 7: recent updates to the protein domain annotation resource. *N Acids Res* 40(D1):D302–D305
 49. Johnson M, Zaretskaya I, Raytselis Y, Merezukh Y, McGinnis S, Madden TL (2008) NCBI BLAST: a better web interface. *N Acids Res* 36 (suppl_2):W5–W9
 50. Söding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. *N Acids Res* 33 (suppl_2):W244–W248
 51. Mitchell A, Chang H-Y, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S (2015) The InterPro protein families database: the classification resource after 15 years. *N Acids Res* 43(D1):D213–D221
 52. Finn RD, Bateman A, Clements J, Coghill P, Eberhardt RY, Eddy SR, Heeger A, Hetherington K, Holm L, Mistry J (2014) Pfam: the protein families database. *N Acids Res* 42(D1):D222–D230
 53. Diplaris S, Tsoumakas G, Mitkas PA, Vlahavas I (2005) Protein classification with multiple algorithms. In *panhellenic conference on informatics*: Springer: 448–456
 54. Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, Lees JG, Lewis TE, Studer RA, Rentzsch R (2012) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *N Acids Res* 41(D1):D490–D498
 55. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P (2010) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *N Acids Res* 39 (suppl_1):D561–D568.
 56. Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27(3):431–432
 57. de Azevedo Jr WF, Canduri F, de Oliveira JS, Basso LA, Palma MS, Pereira JH et al (2002) Molecular model of shikimate kinase from *Mycobacterium tuberculosis*. *Biochem Biophys Res Commun* 295(1):142–148
 58. Uchôa HB, Jorge GE, Da Silveira NJF, Camera JC Jr, Canduri F, De Azevedo Jr WF (2004) Parmodel: a web server for automated comparative modeling of proteins. *Biochem Biophys Res Commun* 325(4):1481–1486
 59. Canduri F, Peres PC, Caceres RA, de Azevedo J, Filgueira W (2008) CDK9 a potential target for drug development. *Med Chem* 4(3):210–218
 60. de Azevedo WF, Canduri F, da Silveira NJF (2002) Structural basis for inhibition of cyclin-dependent kinase 9 by flavopiridol. *Biochem Biophys Res Commun* 293(1):566–571

61. Xiang Z (2006) Advances in homology protein structure modeling. *Curr Protein Pept Sci* 7(3):217–227
62. Hisano T, Tsuge T, Fukui T, Iwata T, Miki K, Doi Y (2003) Crystal structure of the (R)-specific enoyl-CoA hydratase from *Aeromonas caviae* involved in polyhydroxyalkanoate biosynthesis. *J Biol Chem* 278(1):617–624
63. Chen G-Q, Jiang X-R (2017) Engineering bacteria for enhanced polyhydroxyalkanoates (PHA) biosynthesis. *Synth Syst Biotechnol* 2:192–197
64. Numata K, Doi Y (2012) Biosynthesis of polyhydroxyalkanoates by a novel facultatively anaerobic *Vibrio* sp under marine conditions. *Mar Biotechnol* 14(3):323–331
65. Guex N, Peitsch MC, Schwede T (2009) Automated comparative protein structure modeling with SWISS-MODEL and Swiss-Pdbviewer: a historical perspective. *Electrophoresis* 30(S1):S162–S173
66. Benkert P, Biasini M, Schwede T (2011) Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* 27(3):343–350
67. Hooda V, Gundala babuChinthala PP (2012) Sequence analysis and homology modeling of peroxidase from *Medicago sativa*. *Bioinformation* 8(20):974
68. Messaoudi A, Belguith H, Hamida JB (2011) Three-dimensional structure of *Arabidopsis thaliana* lipase predicted by homology modeling method. *Evolut Bioinform* 7:EBO. S7122.
69. Sugino H, Sasaki M, Azakami H, Yamashita M, Murooka Y (1992) A monoamine-regulated *Klebsiella aerogenes* operon containing the monoamine oxidase structural gene (*maoA*) and the *maoC* gene. *J Bact* 174(8):2485–2492
70. Fukui T, Shiomi N, Doi Y (1998) Expression and characterization of (R)-specific enoyl coenzyme A hydratase involved in polyhydroxyalkanoate biosynthesis by *Aeromonas caviae*. *J Bact* 180(3):667–673
71. Kourmentza C, Plácido J, Venetsaneas N, Burniol-Figols A, Varone C, Gavala HN, Reis MA (2017) Recent advances and challenges towards sustainable polyhydroxyalkanoate (PHA) production. *Bioengineering* 4(2):55
72. Folador EL, Hassan SS, Lemke N, Barh D, Silva A, Ferreira RS, Azevedo V (2014) An improved interolog mapping-based computational prediction of protein–protein interactions with increased network coverage. *Integr Biol* 6(11):1080–1087
73. De Las RJ, Fontanillo C (2012) Protein–protein interaction networks: unraveling the wiring of molecular machines within the cell. *Brief Func Genom* 11(6):489–496
74. Ding Z, Kihara D (2019) Computational identification of protein–protein interactions in model plant proteomes. *Sci Rep* 9(1):1–13

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.