Computational Biology and Chemistry 61 (2016) 226–228

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research article

Comparative genomics for understanding the structure, function and sub-cellular localization of hypothetical proteins in *Thermanerovibrio acidaminovorans* DSM 6589 (tai)

Hitesh S. Thakare[a,b,*], Dilip B. Meshram[a], Chandrakant M. Jangam[a], Pawan Labhasetwar[a], Kunal Roychoudhary[b], Arun B. Ingle[b]

[a] National Environmental Engineering Research Institute, Nehru Marg, Nagpur 440020, Maharashtra, India
[b] Department of Microbiology, S. K. Porwal College, Kamptee 441001, Maharashtra, India

ARTICLE INFO

Article history:
Received 13 March 2015
Received in revised form 25 January 2016
Accepted 23 February 2016
Available online 24 February 2016

Keywords:
Anaerobic sludge bed reactor
Sugar refinery
Bioinformatics
Annotation
Hypothetical
Interproscan

ABSTRACT

The *Thermanerovibrio acidaminovorans* DSM 6589 (tai) is a unique bacterium isolated from anaerobic sludge bed reactor from sugar refinery in Netherland. The comparative genomic studies for understanding the hypothetical proteins in *T. acidaminovorans* DSM 6589 (tai) were carried out using different bioinformatic tools and web servers. In all 320 hypothetical proteins were screened from the total available genome. The Insilico function prediction for 320 hypothetical proteins was achieved by using different online servers like CDD-Blast, Interproscan and pfam whereas, the structure prediction for 202 hypothetical proteins were deciphered by using protein structure prediction server (PS2 server). The sub-cellular localization for the identified proteins was predicted by the use of cello v2.5 for 320. The study carried out has helped us to understand the structures and functions of unknown proteins available in *T. acidaminovorans* DSM 6589 (tai) through comparative genomic approach.

1. Introduction

Thermanerovibrio acidaminovorans DSM 6589 is a recently discovered bacterium isolated from anaerobic sludge bed reactor from sugar refinery and has been sequenced recently. There are very few publications and literature references available till date related to *T. acidaminovorans* DSM 6589. As the bacterium is included in the class of extremophils i.e. thermopiles, there is very rare amount of data for understanding their genetic makeup and metabolisms involved during their life cycles. However, the recent paper by Guangsheng et al. (1992) showed that very few enzymes have been identified and many others are still not discovered in *T. acidaminovorans* DSM 6589. This gives us a very good opportunity to study the available unknown proteins and hypothetical proteins in *T. acidaminovorans* DSM 6589, which will help us to know this bacterium in more depth and the data generated will prove very helpful for the ongoing research of on *T. acidaminovorans* DSM 6589.

Thermanerovibrio acidaminovorans DSM 6589 classification:

Domain: Bacteria	
Thermanerovibrio acidaminovorans DSM 6589	Phylum: Synergistete Class: Synergistia Order: Synergistales Family: Synergistaceae Genus: Thermanaerovibrio Species: *Thermanaerovibrio acidamonovorans*

The enzymes like succinate thiokinase, fumarate reductase, succinate dehydrogenase, β-methylaspartase, hydroxyglutarate dehydroge-nase, isocitrate dehydrogenase and formate dehydrogenase were not detected as reported from previous publications (Guangsheng et al., 1992).

The hypothetical proteins are unclassified proteins with denied or unknown functions in system biology. The Insilico methods can be applicable to model the structures of such hypothetical proteins which help in experimental designs for identifying their functions. Considering the majority of hypothetical proteins, pseudogenes and lack of various enzymes and proteins, has created interest for further study and understanding of unknown and unclassified proteins in *T. acidaminovorans* DSM 6589.

The developing online bioinformatics tools and servers have facilitated the annotation of genome for revealing the function of a particular gene (protein), determine the presence of the enzymatic

* Corresponding author.
E-mail address: hiteshthakre7@gmail.com (H.S. Thakare).

http://dx.doi.org/10.1016/j.compbiolchem.2016.02.018
1476-9271/

conserved domains in the sequences (which may assist in the categorizing protein into specific family) and three dimensional structures for protein sequences (Paunikar et al., 2011; Sanmukh and Paunikar, 2011, 2012a,b,c,d; Sanmukh et al., 2010, 2011a,b,c,d,e, f, 2012). Online web tools like CDD-BLAST (Alejandro et al., 2001; Altschul et al., 1997; Aron et al., 2006; Cédric et al., 2000), INTERPROSCAN (Zdobnov and Rolf, 2001), and PFAM (Alex et al., 2004) can be very useful for understanding the function of targeted protein sequence and Cello v2.5 can be used to find the location of protein or enzyme within the cell (Paunikar et al., 2011; Sanmukh and Paunikar, 2011, 2012a,b,c,d; Sanmukh et al., 2010, 2011a,b,c,d,e, f, 2012; Yu and Lin et al., 2004). The 3-D structure prediction of such proteins can be carried out by using Protein Structure Prediction Server (PS2 server) (Chih-Chieh et al., 2006; Edward et al., 2000; Zafer et al., 2006).

Examples of in silico methods for determination of structures and functions of hypothetical proteins are as follows:

1. In biological system *Mycoplasma hyopneumoniae*, 716 coding sequences are annotated as Hypothetical proteins. The structures of seven proteins were predicted using Insilico methods (Marbella Maria da Fonsêca et al., 2012).
2. *Natrialba magadii* is metabolically versatile, required alkaline and hypersaline conditions for growth and development. In this organism about 36% hypothetical proteins were predicted (Shivakumar Siddaramappa et al., 2012).
3. In *Mycobacterium tuberculosis* protein-protein interactions were explained using in vitro and in silico methods (Venkatesan et al., 2015).

Our main objective was to identify the functional properties of unidentified hypothetical proteins of *T. acidaminovorans* DSM 6589 using bioinformatics tools and online servers.

2. Methodology

2.1. Sequence retrieval

The whole genome sequences for *T. acidaminovorans* DSM 6589 was retrieved from the KEGG database (http://www.genome.jp/kegg/) (Mansi et al., 2009).

2.2. Functional annotation and categorization

The *T. acidaminovorans* DSM 6589 hypothetical proteins were screened and sorted out from the genome and were individually analysed for the presence of conserved functional domains by using computational biology tools like CDD-BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) (Alejandro et al., 2001; Altschul et al., 1997; Aron et al., 2006; Cédric et al., 2000), INTERPROSCAN (http://www.abi.ac.uk/interpro) (Venkatesan et al., 2015) Pfam (http://www.pfam.sanger.ac.uk/) (Alex et al., 2004). The sub-cellular localization of hypothetical proteins was determined by Cello (Yu and Lin et al., 2004). Moreover, the 3-D structures of the functionally active domain of the identified hypothetical proteins were dechiphered. These web tools have the ability to search the defined conserved domains in the sequences available in the online servers or databases and assist in the classification of proteins in appropriate family.

2.3. Protein structure prediction

The in-silico structure predictions of the hypothetical protein sequences showing functional properties were carried out by using PS2 protein structure prediction server (http://www.ps2.life.nctu.edu.tw/) (Chih-Chieh et al., 2006; Edward et al., 2000; Zafer et al., 2006). The online server helps to generate the 3D structures of the hypothetical proteins. The server accepts the sequences in FASTA format as a query to generate resultant proteins 3D structures. The structure determination is completely based upon the conserved template regions detected during functional annotations.

3. Results and discussion

The computational genomic studies for understanding 320 hypothetical protein genes were carried out. The functional annotation of the hypothetical proteins was made through different servers for getting maximum confidence level. The function prediction was carried out through CDD-Blast, pfam and interproscan. The probable function annotation, characterization, subcellular localization and 3-D structure prediction was successfully for 202 gene sequences of protein. The online-automated bioinfromatics tools like CDD- Blast, Interproscan, Pfam, Cello and PS^2 server were used for structurural and functional characterization of screened hypothetical proteins. The structures, functions, sub-cellular localizations of the hypothetical proteins from *T. acidaminovorans* DSM 6589 (tai) were obtained and represented. Also the 3-D structures generated after using the best scoring templates were represented in as Template ID in structure column of respective hypothetical proteins (supplementary Table 1). As the functions of hypothetical proteins denied in system biology however the structures for such proteins can be verified through biological processes or experiments.

4. Conclusion

The intensive bioinformatic study has helped to functionally annote 320 hypothetical, proteins, which have showed and help us, understand many functional proteins are available in the *T. acidaminovorans* DSM 6589. The sub-cellular localization of the 320 sorted hypothetical proteins was also carried out by using which further help us understand the localization of identified enzyme or proteins. We have successful characterized the 277 unknown proteins from 320 screened hypothetical protein sequences from *T. acidaminovorans* DSM 6589 for verifying their structure and functions of the gene products. This predicted functions and three dimensional structures may assist in establishing their role in the life cycle of the bacterium. This computationally generated data can also be used for developing new protocols for transgenic plant production or for modification of the existing protocols through computational docking studies (Altschul et al., 1997).

Authors contributions

Hitesh Thakre did the preparation of manuscript and referencing. Dilip Meshram and Chandrakant Jangum did the critical editing.

Acknowledgments

H.T. wants to thanks Swapnil Sanmukh, post graduate researcher, Sau Paulo state university, Brazil for his tremendous support and help during this study. Hitesh Thakre carried out extensive bioinformatics analysis as a part of his M.Sc. Microbiology dissertation. The authors want to thanks Senior Principal Scientist and Head, Water Technology and Management Division, CSIR-NEERI, Nagpur for his valuable guidance. Authors are also thankful to Associate Professor and Head, Department of