Structural Analysis of Hypothetical Proteins of Salmonella Typhi CT18

Anila Farid [1], Saadia Sadiq [1], Beenish Haider [1], Adnan Arshad [2]

ABSTRACT

Background: Salmonella *typhi* is a gram negative, flagellated, rod shaped facultative, anaerobic bacterium belonging to family enterobacteriace. It is etiological agent of typhoid fever. It spreads through contamination of water and under cooked food. It has caused death in many countries where sanitation is poor. It has evolved the ability to spread in deeper tissues of human body..
Objective: To identify the function and structure of hypothetical proteins of Salmonella *typhi* CT18 by using bioinformatic programs.
Material and Methods: The 3D structure of these proteins was carried out by the help of Bioinformatics programs such as CMR, Interproscan, BLAST, Modellor 9.10, Procheck and Prosa. Functional annotation was carried out by using Interproscan and structure analysis was carried by homology modeling using Modeller.
Results: It was found that there are 4934 proteins are present in Salmonella *typhi* out of which, 12.1 % were hypothetical proteins. Among hypothetical proteins, results of 463 proteins were shown by interproscan,6.3% were functionally detected proteins, 2.6% were uncharacterized proteins and for rest no hits were found. For homology modeling, 5 proteins were selected. The three dimensional structure of proteins was constructed by Modeller 9.10.
Conclusion: Structural bioinformatics is an effective method to find structure of hypothetical protein which provide good solution for drug discovery.
Keywords: Homology Modelling, Hypothetical proteins, Typhoid fever,Modellor, Structural Bioinformatics.

This article may be cited as: Farid A, Sadiq S, Haider B, Arshad A. Structural Analysis of Hypothetical Proteins of Salmonella Typhi Ct18. J Saidu Med Coll Swat. 2022;12(1):41-46.DOI:https://doi.org/10.52206/jsmc.2022.12.1.704

INTRODUCTION

Food borne infections are quite common and widely distributed worldwide. Typhoid is one such common disease by Salmonella enteric typhi[1]. Salmonella *typhi* is gram negative, flagellated, rod shaped and facultative anaerobic bacterium[2]. Salmonella enterica serovar Typhi is the causative agent of ty-phoid fever. S. Typhi does not have an animal reservoir and can be transmitted from a typhoid carrier only through contaminated water or food[3]. Its suppression seems doubtful due to latest emergency of multiple drug resistant strains [4].

The complete genome for *Salmonella typhi* has been sequenced. *Salmonella typhi* encodes for 204 genes. Out of 204 genes, twenty seven (27) are remnants of insert sequences, seventy five (75) are occupied in housekeeping function and 46 of the genes mutation are involved in host interaction. Protein sequence-structure analysis (PSSA) is fundamental in a wide range of biomedical research fields, especially in protein structure prediction and modeling[5].

The strain of Salmonella *typhi*CT18 has 4934 proteins. A large circular chromosome of 4.809, 037bp which is nearly equal to 4.8Mb in length and two plasmids, pHCM1 and pHCM2, pHCM1 plasmid is involved in drug resistant. Recently in bacterial genome sequence study, approximately 45% of the genes of proteins have not assigned any function. The group of hypothetical genes of proteins without any assigned functions is termed as hypothetical proteins[6]. Conserved hypothetical proteins are great number of genes in sequence of organisms genome. The hypothetical proteins are not described functionally[7].

In functional annotation of hypothetical proteins, Interproscan seems very helpful. Its purpose is to combine different proteins signature recognition methods into one resource with look up of corresponding interPro and GO annotation. The inter Pro database incorporates Prosite, Prints, Pfam, ProDom, SMART, TIGRFAMs, PIR superfamily, SUPERFAMILY, Gene3D and PANTHER databases[8]. It is imperative that the framework built for managing the InterProScan analyses is highly configurable, so that the heterogeneous nature of the different applications can be represented appropriately, and the software can run in a multitude of computing environments[9].

Naturally occurring homologous proteins usually have similar stable tertiary structures[10]. The three dimensional structures of all proteins can be predicted by using bioinformatics methods such as ab-initio method, threading and homology method.Ab-intio can predict structure of proteins on the basis of physio-chemical principle. Ab-intio

1. Abbottabad International Medical College, Abbottabad.
2. Hazara University, Mansehra.

Correspondence: Dr. Anila Farid
Abbottabad International Medical College, Abbottabad.
Email: anila.farid@gmail.com

Received: December 11th 2018 Accepted: January 21th 2022

is unreliable and unrealistic method for future use[11].

Threading which is also called as fold recognition is the method which is used for sequence with sequence identity = 30. Template provides a base for the prediction of structure of the proteins by Homology modeling.of proteins having sequence identity of about > 30%. Recently with the arrival of structural genomics, the importance of homology modeling has been extremely increased. Homology modeling is reliable method for model building. The best homology models were selected according to Global Model Quality Estimation (GMQE) and QMEAN statistical parameters. GMQE is a quality estimation which combines properties from the target-template alignment. The quality estimate ranges between 0 and 1 with higher values for better models[12].

Homology modeling is a representation of the similarity of environmental residues at topologically corresponding positions in the reference proteins. In the absence of experimental data, model building on the basis of a known 3D structure of a homologous protein is at present the only reliable method to obtain the structural information[13].

Homology modeling can be completed in four steps[14].

1. Template recognition of known protein from PDB databank.
2. Sequence alignment of the query with that of the template.
3. Model building on the basis of this alignment.
4. Assessment and refinement of the newly built model.

Homology modeling has become a widespread technique for the construction of GPCR models intended to study the structurefunction relationships of the receptors and aid the discovery and development of ligands capable of modulating their activity[15].

Main purpose of the study is to identify the function and structure of hypothetical proteins of Salmonella typhi CT18 by using bioinformatic programs.

MATERIALS AND METHODS

Complete protein sequence of *Salmonella typhi*CT18was downloaded from the databases of comprehensive microbial resource (CMR) (http; cmr.jcvi.org).The comprehensive microbial resource is a free website used to display information on all of the publicly available, complete prokaryotic genome.

Salmonella *typhi* contains 4934 proteins. Amongst these 780 were hypothetical proteins. We separated hypothetical proteins from total genome and saved them in the word document format.

The hypothetical proteins sequences were analyzed for function annotation for presence of conserved enzymatic domain by interproscan.

Interproscan is bioinformatics webtool which provides functional analysis of different proteins by classifying them into different families and predicting domains and important sites. InterProScan tool, which combines different protein signature recognition methods from InterPro consortium member databases into one resource [16].

After functional annotation of proteins, the three structures of proteins were modeled.

Homology modeling of predicted functional proteins such as protein Transcriptional regulator (STY 4289),Crisper associated proteins (STY 3065), Methyl transferases (STY 3264) ,Transcriptional regulator (STY2107), FeS cluster protein (STY 1754) was carried out using modeler.

Template plays a vital role in model building.Template was searched by BLAST against PDB (protein data bank). The best template was selected which showed maximum identity, low E-value (less than 1), high querry coverage (=70%) with equivalent target protein.
Target protein Model was built on the basis of target - template alignment file. Alignment of target and template was carried out by BLAST.

After the target template protein alignment file construction, the three dimensional homology model of protein was constructed with the help of MODELLER 9.10 program. MODELLER 9.10 automatically builds the models.

The newly built models were visualized by Discovery Studio Viewer (DS viewer).

During execution of modeler, ten models were constructed.Each model was evaluated tocheckthe reliability of the models. The evaluation was carried out by Procheck and Prosa-web.

Procheck helps to demonstrate if the structural features of the models are reliable or not.

The stability and energy of the target with that of the template wascalculatedbyPROSA-web.

Energy and Z-score files were obtained as a result of Prosa.

RESULTS

Table -1:Total proteins and their percentage

S. No	Proteins	Numbers	Percentage
1	Total proteins	4934	100%
2	Hypothetical proteins	600	12.16 %
3	Functionally detected	362	60.3%
4	Uncharacterized proteins	120	20%
5	No hits were found for remaining proteins	118	19.6%

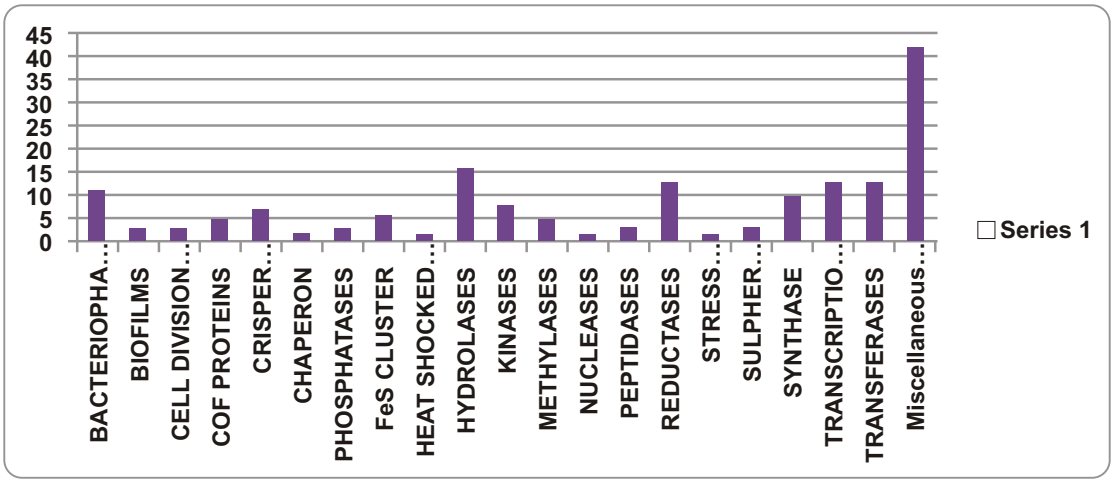

Fig 1:
Graph representing different functional groups and number of hypothetical proteins in each group.

Figure -2:Three dimensional model of transcriptional regulator (STY 2107)

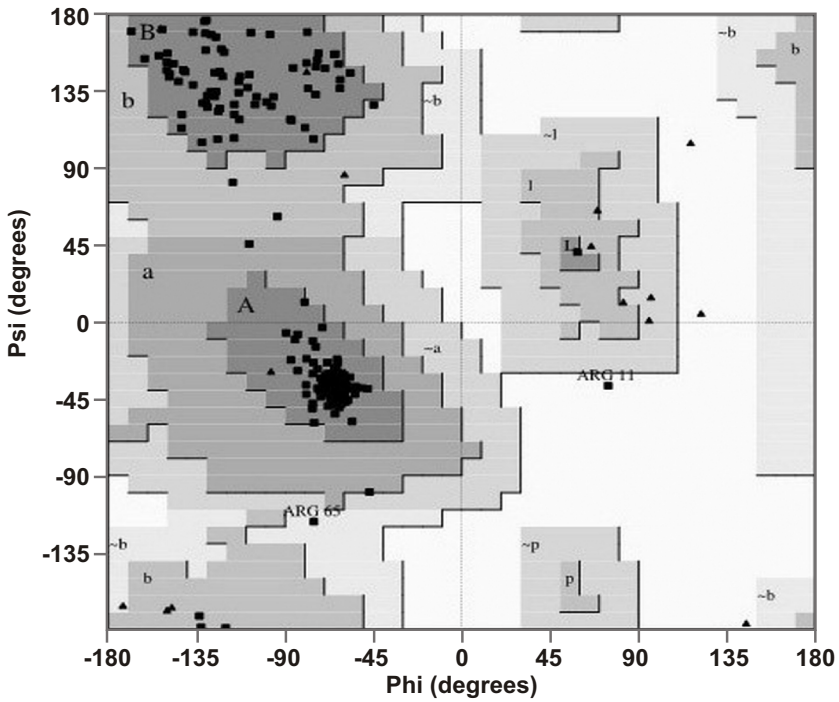

Figure- 3: Ramachandran plot of transcriptional regulator (STY2107) model using PROCHECK

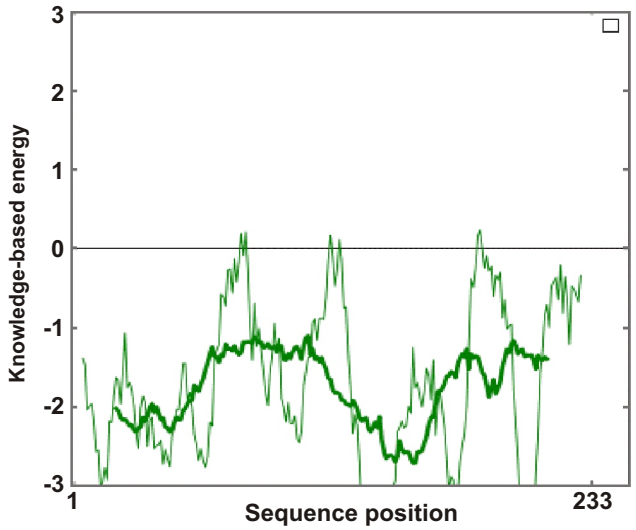

Fig -4: Energy peaks of transcriptional regulator STY 2107obtained through Prosa-web

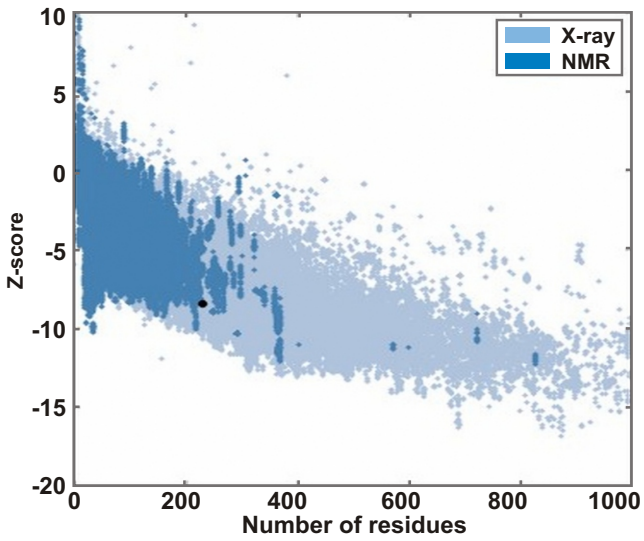

Fig -5: Z-score scheme of transcriptional regulator STY2107 obtained through Prosa-web

DISCUSSION

The function and structure of hypothetical proteins can be accessible by conventional biological methods but these methods require a lot of time and are costly. Therefore, homology modeling is becoming important for obtaining three dimensional structure of protein because genome project produces sequences at much higher rate than we can solve three dimensional structure by NMR or X-Rays crystallography. Ion channels and carriers are difficult to express and purify in amounts for X-ray crystallography and nuclear magnetic resonance (NMR) studies. The homology modeling approach is a valuable approach for obtaining structural information about carriers and ion channels[16,17].

During present study we predicted the function and structure of hypothetical proteins of Salmonella typhi strain CT18 by using bioinformatics tools. *Salmonella typhi* CT18 is the most virulent among serotype of *Salmonella typhi*. Therefore, Salmonella typhi strain CT18 was preferred during present study. The entire protein sequence of *Salmonella typhi* strain such as Salmonella typhi strain CT 18 is available at

Comprehensive Microbial Resource (CMR). The sequences of all proteins were downloaded from CMR. Total of 4934 proteins were found in *Salmonella typhi* CT18. Out of these, 600 were hypothetical proteins which were 12.16% of the total protein[8]. The functional prediction of hypothetical proteins of Salmonella typhi CT18 was carried out by interproscan and structure was predicted by modeller.

Function of hypothetical proteins was predicted by bioinformatic tool called Interproscan. Among 600 hypothetical proteins, 362 proteins (60.3% of total hypothetical proteins) were functionally detected and 120 (20%) hypothetical proteins were proteins of unknown function as shown in table 3.3. No hits were also found for 118 proteins (19.6% of total hypothetical proteins)[19].

The functionally predicted proteins were categorized into particular groups on the basis of specific conserved enzymatic domain. These groups are kinases, hydrolases, peptidases, ligases, redutases, synthetases, transferases, ligases, carboxylase, phosphatases transcriptional regulators etc. The proteins with similar functions were placed in same group.

During present study, homology modeling of predicted transcriptional regulator (STY2107), methyl transferases (STY 3264) , crisper associated protein (STY 3065) , transcriptional regulator (STY4289) and FeS cluster proteins (STY 1754) were carried out. These proteins were selected for homology modeling as representative group.

Transcriptional regulators are regulatory proteins which activate transcription of DNA by binding near the promoter of DNA. Sites of DNA sequence where regulatory proteins bind are called enhancer sequences. The expression of eukaryotic genes is controlled primarily at the level of initiation of transcription. In prokaryotes such as Salmonella typhi, transcription regulator makes the cell to quickly adopt to the different changes occurring in external environment. Repressor bind to the regions called operators which are located downstream the promoter region and activators bind to the upstream portion of promoter. A combination of repressor and activator in Salmonella typhi helps to determine whether a gene is transcribed.

The three dimensional structure building of hypothetical protein STY2107 was constructed by homology modeling. The first and most important step in homology modeling is the target and template comparison. Template (crystal structure of E.coliYebc ,Pdb id 1KON) for target (transcriptional regulator) was retrieved by BLAST. As the name implies, BLAST performs "local" alignments. Most proteins are modular in nature, with one or more functional domains occurring within a protein[20]. The same domains may also occur in proteins from different species. The BLAST algorithm is tuned to find these domains or shorter stretches of sequence similarity[18].

It was selected because it has greater similarity with the target sequences .The target and template has 98% identity and query coverage of 100%. Identities comprises same amino acids in both sequences whereas positive residues mean amino acids which are biochemically similar to each other. Alignment was done by BLAST .Modeller 9.10 was used for construction of model. Ten models were constructed. These models were evaluated by Procheck. Ten plots were generated byProcheckfor each of the model in which plot 1 is Ramachandran plot.Programmed Coral Draw was used for viewing these plots. Ramachandran plot is divided into four regions. Favored region,Allowed region, generously allowed region and disallowed region. Out of 10 models ,one best model was selected. Selected model contains the greatest number of residues in favored region while all other 9 models constitute less number of residues in favored region. It shows that among 233 residues, 201 were present in favoredregion ,7residues were in allowed regions, one was in generously allowed region and one was in disallowed region constituting percentage 95.7%,3.3%,0.5% and 0.5% respectively[11].

Energy graph of target by Prosa showed the overall quality of protein structure. The energy of hypothetical protein transcriptional regulator (2107) was estimated using PROSA. Z-score scheme of transcriptional regulator (STY 2107) was obtained through Prosa-web[19]. It showed highly negative value (-8.5) which indicated that model was highly stable.

CONCLUSIONS

By knowing the structure, drug can be manufactured to get rid of the disease. Since drugs interact with receptors that consist mainly of

proteins, protein 3D structure determination, and thus homology modeling is important in drug discovery. Human organic cation transporters (hOCTs) belong to solute carriers (SLC) 22 family of membrane proteins that play a central role in transportation of chemotherapeutic drugs for several clinical and pathological conditions, including cancer and diabetes.

REFERENCES

1. Pawar S, Ashraf M,Mehata K,Lahiri C.Computational identification of indispensable virulence proteins of Salmonella Typhi CT18.Curr.Top.Salmonella Salmonellosis.2017 April 5.
2. Yousef AE, Carlstrom C. Food microbiology: a laboratory manual. John Wiley & Sons; 2003 May 5.
3. Ong SY, Pratap CB, Wan X,Hou S,Rahman AY,Saito JA, et al. Complete genome sequence of Salmonella enteric subsp.enterica serovar Typhi P-stx-12.Journal of bacteriology.2012 April;194(8):211-5.
4. Deng W,Liou SR, Plunkett G, Mayhew GF,Rose DJ,Burland V, et al. Comparative genomics of Salmonella enteric serovar Typhi strains Ty2 and CT18.Journal of bacteriology.2003 Apr1;185(7):2330-7.
5. Janson G,Zhang C,Prado MG,Paiardini A,PyMod 2.0:improvement in protein sequence-structure analysis and homology modeling within PyMOL.Bioinformatics.2017 Feb 1;33(3):444-6.
6. Hoskeri JH, Krishna V, Amruthavalli C. Functional annotation of conserved hypothetical proteins in Rickettsia massiliae MTU5. Journal of Computer Science & Systems Biology. 2010;3(2):50-1.
7. Galperin MY,Koonin EV.Conserved hypothetical proteins:prioritization of targets for experimental study. Nucleic acids research. 2004 Jan1;32(18):5452-63.
8. Zdobnov EM, Apweiler R. InterProScanan integration platform for the signature-recognition methods in InterPro. Bioinformatics. 2001 Sep 1;17(9):847-8.
9. Jones P,Binns D,Chang HY,Fraser M,Li W,Mc Anulla C, et al. InterProScan 5:genome- scale protein function classification. Bioinformatics.2014 May 1;30(9):1236-40.
10. Kaczanowski S, Zielenkiewicz P. Why similar protein sequences encode similar three-dimensional structures?. Theoretical Chemistry Accounts. 2010 Mar 1;125(3-6):643-50.
11. Venclovas C, Zemla A, Fidelis K, Moult J. Assessment of progress over the CASP experiments. Proteins: Structure, Function, and Bioinformatics. 2003;53(S6):585-95.
12. Ekins S,Liebler J,Nerves BJ,Lewis WG,Coffee M,Bienstock R, et al. Illustrating and homology modeling the proteins of the Zika virus. F1000Research.2016;5.
13. Vyas VK, Ukawala RD, Ghate M, Chintha C. Homology modeling a fast tool for drug discovery: current perspectives. Indian journal of pharmaceutical sciences. 2012 Jan;74(1):1.
14. Meier A,Soding J.Automatic prediction of protein 3D. structures by probabilistic multi-template homology modeling.PLoS Comput Biol.2015 Oct 23;(10):e1004343.
15. Costanzi S.Homology modeling of class ag protein-coupled receptors.In Homology Modeling 2011(pp.259-279).Humana Press.
16. Qiu J, Zang S,Ma Y,Owusu L,Zhou L ,Jiang,T,Xin Y.Homology modeling and identification of amino acids involved in the catalytic process of Mycobacterium tuberculosis serine acetyltransferase.Molecular Medicine Reports.2017 Mar 1;15(3):1343-7.
17. Ravna AW,Sylte I. Homology modeling of transporter proteins(carriers and ion channels).In Homology Modeling 2011(pp.281-299).Humana Press.
18. Madden T.The BLAST sequence analysis tool.In The NCBI Handbook.2nd edition2013 March 15.National Center for Biotechnology Information(US).
19. Muhammed MT, Aki-Yalcin E.Homology modeling in drug discovery:Overwiew,current applications,and future perspectives.Chemical biology and drug design.2019Jan;93(1):12-20.
20. Dakal TC, Kumar R, Romator D.Structural modeling of human organic cation transporters.Computational Biology and Chemistry.2017 Jun 1;68:153-6.

DATA SHARING STATEMENT: The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

CONFLICT OF INTEREST: Authors declared no conflict of interest.
GRANTED SUPPORT AND FINANCIAL DISCLOSURE: Nil

AUTHOR'S CONTRIBUTION
The following authors full fill authorship criteria as per ICMJE guidelines;

Farid A:	Idea conception, drafting the work, final approval, agreed to be accountable for all the work.
Sadiq S:	Design of the work, data acquisition, critical rivision, final approval, agreed to be accountable for all the work.
Haider B:	Data analysis, drafting of the work, final approval, agreed to be accountable for all the work.
Arshad A:	Data interpretation, critical rivision, final approval, agreed to be accountable for all the work.