An in silico Approach for Structural and Functional Annotation of Salmonella enterica serovar typhimurium Hypothetical Protein R_27

Arif Khan¹, Hossain Ahmed², Nasreen Jahan³, Syed Raju Ali⁴, Al Amin⁴, Mohammad Neaz Morshed¹*

¹Department of Science and Humanities
Military Institute of Science and Technology (MIST)
Mirpur Cantonment, Dhaka 1216, Bangladesh
Emails: arifkhanbge35@gmail.com, mmneaz@hotmail.com

²Department of Biotechnology and Genetic Engineering
University of Development Alternative
Dhanmondi, Dhaka 1209, Bangladesh
Email: ahmedshohag.bd@gmail.com

³Department of Biotechnology and Genetic Engineering
Mawlana Bhashani Science and Technology University
Santosh, Tangail 1902, Bangladesh
Email: naza.bge@gmail.com

⁴Department of Biotechnology
Bangladesh Agricultural University
Mymensingh 2202, Bangladesh
Emails: m15121307@bau.edu.bd, josnasnan@gmail.com

*Corresponding author

Received: October 11, 2015 Accepted: February 15, 2016
Published: March 31, 2016

Abstract: Typhoid fever is a major cause of illness in most developing countries, including Bangladesh. In quest of new potential drug against Typhoid fever, the current study was designed to elucidate structural and functional details of S. typhi hypothetical protein (HP) R_27. HP R_27 has the primary amino acid sequences available only. The structural annotation was determined by ProtParam, SOPMA, and CELLO. The three-dimensional (3D) structure of HP R_27 predicted through homology modeling by using Phyre2. The 3D structure then refined and verified by ModRefiner, PROCHECK, ERRAT, QMEAN. The functional annotation was also performed by InterPrоScan, SMART, Pfam, NCBI-CDD and found Phospholipase D-like and DNA repair activity. Multiple sequence alignment also supported the existence of PLD-like domain and DNA repair protein domain in the selected hypothetical protein sequences. Finally, the cavity of drug binding was also identified to assist further molecular docking study and potent inhibitor identification. This in silico approach can be further utilized in molecular drug design for other clinically significant pathogens.

Keywords: Typhoid fever, Hypothetical protein, Homology modeling, PLD-like activity.

Introduction
Salmonella enterica serovar typhimurium (S. typhi) is an enteric gram-negative bacteria agent of typhoid fever and other related clinical diseases that are distributed worldwide mainly in regions with limited sanitary conditions [28]. Typhoid fever is a serious invasive bacterial
disease of humans with an annual global burden of approximately 16 million cases, leading to 600,000 fatalities [31].

Typhoid fever caused an anticipated 21.7 million cases and 217,000 deaths in 2000. Adolescent’s children and infants in south-central and Southeast Asia experience the most burden of illness. About 400 incident cases are accounted and the fever is projected to occur in about 6,000 people per year in the United States [7, 27, 38]. Invasive strains of non-typhoidal salmonella, for instance S. typhi ST313 have recently been labeled as rising diseases in Africa. Associations among malaria, HIV and malnutrition, have thrown in to spread of this fever, explicit the necessity of using costly antimicrobial drugs in the poorest health services in the world [10].

Typhoid fever is take epidemic proportions in Bangladesh as it is a round the year problem. The specific reason for such kind of occurrence are having tainted water, unhygienic food managing practice and flawed sewerage system. A high frequency of multidrug resistant typhoid fever has been reported. So, new and cost effective drug development is a basic need for developing country like Bangladesh [26, 29, 33].

In more recent time, genome sequences of different organisms are available on various databases by utilizing next generation sequencing technology to gather information to a greater extent. As a result, increasing amounts of information about hypothetical proteins deposited in sequence databases rather than experimentally isolated data in Protein Data Bank (PDB) [3, 4]. Hypothetical proteins are generally forecasted to be expressed from an open reading frame (ORF). These proteins have no experimental evidence regarding their functions [20]. At present, it is assumed that 50% proteins of a genome are hypothetical proteins [35]. This encourages in silico study of a hypothetical protein utilizing experimental data [30].

In the current study, the S. typhi hypothetical protein (HP) R_27 was selected as the primary amino acid sequences is available but there have no structural and functional details. S. typhi HP R_27 was subjected to extensive in silico study to determine its molecular and structural properties. Further, we tried to predicted good quality model of HP R_27 using protein homology modeling techniques and successive computer aided active site prediction for the development of an effective drug against S. typhi.

Materials and methods

Sequence retrieval
The UniProt Knowledge Base (UniProtKB) database for S. typhi was primarily explored to find out hypothetical proteins with possible interest in research and application. The amino acid sequence of the HP R_27 was retrieved from the Uniprot database (http://www.uniprot.org/). UniProtKB is used to collect functional information on any kind of proteins, with accurate and rich annotation [24]. The primary accession ID of the selected protein is Q9L5J1 and it contains 259 amino acids. The amino acid sequence was then collected as a FASTA format for further proteomic analysis.

Structural annotation
A tool of ExPASy server, ProtParam (http://web.expasy.org/protparam/), was used for the analysis of the physicochemical properties from the retrieved protein sequence. This tool can predict physicochemical properties for instance the molecular weight, theoretical pI, aliphatic index, amino acid composition, grand average of hydropathicity (GRAVY), instability index, and extinction coefficients.
Self-optimized prediction method with alignment (SOPMA) was used for secondary structure prediction [12]. Protein’s secondary structural properties are including α helix, β helix, 3_10 helix, Beta bridge, Extended strand, Bend region, Beta turns, Ambiguous states, Random coil and Other states. In addition, the PSIPRED ((http://bioinf.cs.ucl.ac.uk/psipred/) [5] server was also exploited to authenticate the results found from SOPMA.

Subcellular localization of R_27 was predicted by CELLO v.2.5 [39]. Results were also cross-checked with subcellular localization predictions obtained from PSORTb version 3.0.2 [40], TBPred [34], Signal Peptide (Signal P 4.1) [9], Secretory Protein (Secretome P) [1], HMMTOP [21], TMHMM [36], SOSUI [15].

Homology modeling
Phyre2 (Protein Homology/Analogy Recognition Engine) is a protein fold identification server [17]. It was used to predict the three-dimensional (3D) homology model of HP R_27. The input data of this protein was in FASTA format. After homology modeling, it is necessary to refine the structure for better accuracy.

ModRefiner (http://zhanglab.ccmb.med.umich.edu/ModRefiner/) is an algorithm for high-resolution protein structure refinement, which used to refine the predicted protein model. Amino acid sequences were given in the FASTA format. Model refinement was done for couple of times to get the best structure [36]. The predicted model was evaluated to verify the stereo chemical quality with PROCHECK by Ramachandran plot which was done through “Protein structure and model assessment tools” [21]. The best model was selected from PROCHECK depending on overall G-factor, number of amino acids in different regions. ERRAT and QMEAN were used for further analysis of the selected protein structure [2, 6].

Functional annotation
Hypothetical protein R_27 was analyzed due to find the presence of conserved domains based on sequence similarity search with close orthologous family members. Four bioinformatics tools and databases including InterProScan [40], Simple modular architecture research tool (SMART) [22], Proteins Families Database (Pfam) [11], and NCBI Conserved Domains Database (NCBI-CDD) [24] were used for functional annotation.

Active site identification
Active site is defined as a definite region of protein which is responsible for its functional activity. This consists of several amino acid residues. The active site of the selected hypothetical protein predicted by Computed Atlas of Surface Topography of proteins (CASTp) [8].

Results and discussion
UniProtKB delivers an authoritative resource for protein sequences and functional information. Sequences of HP R_27 of S. typhi were obtained from UniPortKB.

The blastp result against non-redundant database showed homology with DNA repair protein and Phospholipase D (PLD)-like protein (Table 1). Pfam server predicted the PLD-like domain at 14-135 amino acid residues with an e-value of 2.6e-11. The PLD-like domain domain was also found in InterProScan server and NCBI-CDD at 14-116 amino acid residues.
Table 1. Analogous proteins obtained from nonredundant database

<table>
<thead>
<tr>
<th>Entry name</th>
<th>Organism</th>
<th>Protein name</th>
<th>Score</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi</td>
<td>742850038i</td>
<td><em>Escherichia coli</em></td>
<td>DNA repair protein</td>
<td>343</td>
</tr>
<tr>
<td>gi</td>
<td>633894482i</td>
<td><em>Klebsiella pneumoniae</em></td>
<td>DNA repair ATPase</td>
<td>326</td>
</tr>
<tr>
<td>gi</td>
<td>490126160i</td>
<td><em>Escherichia coli</em></td>
<td>PLD-like domain protein</td>
<td>325</td>
</tr>
<tr>
<td>gi</td>
<td>695806886i</td>
<td><em>Klebsiella oxytoca</em></td>
<td>DNA repair protein</td>
<td>325</td>
</tr>
<tr>
<td>gi</td>
<td>490302223i</td>
<td><em>Klebsiella pneumoniae</em></td>
<td>PLD-like domain protein</td>
<td>324</td>
</tr>
<tr>
<td>gi</td>
<td>608787356i</td>
<td><em>Raoultella planticola</em></td>
<td>DNA repair ATPase</td>
<td>317</td>
</tr>
<tr>
<td>gi</td>
<td>779902985i</td>
<td><em>Vibrio campbellii</em></td>
<td>DNA repair protein</td>
<td>224</td>
</tr>
<tr>
<td>gi</td>
<td>659054291i</td>
<td><em>Vibrio parahaemolyticus</em></td>
<td>DNA repair protein</td>
<td>223</td>
</tr>
<tr>
<td>gi</td>
<td>701174448i</td>
<td><em>Vibrio harveyi</em></td>
<td>DNA repair protein</td>
<td>222</td>
</tr>
</tbody>
</table>

The ExPASy’s ProtParam server was used to analyze the theoretical physiochemical characteristics from the plain amino acid sequence of the hypothetical protein R_27. The protein was predicted to be consisting of 259 amino acids, with a molecular weight of 29617.6 Daltons and pI of 5.92 indicating a negatively charged protein. The instability index of the protein was computed to be 23.17, classified this protein as stable. The negative GRAVY index of -0.531 is indicative of a hydrophilic and soluble protein. The most abundant amino acid residue was found to be Leucine (34), followed by Lysine, Asparagine and Glutamate (24 each). The sequence had 33 positively charged residues and 36 negatively charged residues. The molecular formula of the protein was found as C1315H2108N352O413S5 (Table 2).

Table 2. Different physio-chemical properties of HP R_27

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular weight</td>
<td>29617.6</td>
</tr>
<tr>
<td>Extinction coefficients</td>
<td>29005</td>
</tr>
<tr>
<td>Ext. coefficient</td>
<td>28880</td>
</tr>
<tr>
<td>Theoretical pl</td>
<td>5.92</td>
</tr>
<tr>
<td>Total number of negatively charged residues (Asp + Glu)</td>
<td>36.00</td>
</tr>
<tr>
<td>Total number of positively charged residues (Arg + Lys)</td>
<td>33.00</td>
</tr>
<tr>
<td>Instability index</td>
<td>23.17</td>
</tr>
<tr>
<td>Grand average of hydropathicity (GRAVY)</td>
<td>-0.53</td>
</tr>
<tr>
<td>Aliphatic index</td>
<td>95.68</td>
</tr>
</tbody>
</table>

The secondary structure of HP R_27 was predicted by SOPMA with standard parameters which are presented in a tabulated form in Table 3 respectively. It contains 35.14% α helix, 21.62% extended strand, 9.65% β turn and 33.59% random coil. The graphical secondary structure presentation of HP R_27 obtained from PSIPRED server is shown in Fig. 1.

Predicting subcellular localization of hypothetical proteins can give information about their cellular functions. This information could be utilized in understanding disease mechanism and developing drugs [41]. It was analyzed by CELLO and authenticated by PSORTb v3.2.0 and
Predict Protein server. The subcellular localization of the query protein was predicted to be a cytoplasmic protein.

Fig. 1 Predicted secondary structure of HP R_27 generated by PSIPRED server
Table 3. Calculated secondary structure elements of by SOPMA

<table>
<thead>
<tr>
<th>Secondary structure</th>
<th>No. of residue</th>
<th>Percentage, (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>α helix (Hh)</td>
<td>91</td>
<td>35.14</td>
</tr>
<tr>
<td>3_{10} helix (Gg)</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>π helix (Ii)</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>Beta bridge (Bb)</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>Extended strand (Ee)</td>
<td>56</td>
<td>21.62</td>
</tr>
<tr>
<td>Beta turn (Bt)</td>
<td>25</td>
<td>9.65</td>
</tr>
<tr>
<td>Bend region (Ss)</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>Random coil (Cc)</td>
<td>87</td>
<td>33.59</td>
</tr>
<tr>
<td>Ambiguous states (?)</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>Other states</td>
<td>0</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Homology modeling of the selected protein was done by using Phyre2 in order to obtain 3D structure. 3D structure of proteins gives important insights about the molecular basis of function and thereby permits an effective design of experiments [13]. That is why; the high resolution 3D structure of a protein is the main key in the understanding and manipulation of biochemical and cellular functions of proteins, [18]. Modrefiner refine again the Phyre2 generated model. Refined model is depicted in Fig. 2.

![Fig. 2 Refined model of HP R_27 generated by Modrefiner](image)

ModRefiner derived refined model of HP R_27 was analyzed. The percent of residues in different regions remain same in the final model. The predicted structure then validated through Ramachandran plot analysis. The results of this analysis are depicted in Fig. 3 and Table 4. PROCHECK, another tool was used to measure the accuracy of protein models. Ramachandran plot statistics of HP R_27 revealed that 90.04% of amino acid residues were found in the most favored regions. Thus, the selected protein model was very good as all of the amino acid residues were within the limits of Ramachandran plot. Verification was also done by ERRAT and QMEAN server. The respective values Z-scores of Cbeta interaction energy, torsion angle energy, solvation energy, secondary structure, and solvent accessibility in case of HP R_27 are -2.06, -2.86, -2.2, -2.26 and -3.62. The overall QMEAN score for HP R_27 is -4.71. QMEAN generated results confers HP R_27 as a qualified model for drug target scopes.
Table 4. Ramachandran plot analysis of HP R_27

<table>
<thead>
<tr>
<th>Ramachandran plot statistics</th>
<th>Residue</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residues in the most favored regions [A, B, L]</td>
<td>217</td>
<td>90.04</td>
</tr>
<tr>
<td>Residues in the additional allowed regions [a, b, l, p]</td>
<td>22</td>
<td>9.13</td>
</tr>
<tr>
<td>Residues in the generously allowed regions [a, b, l, p]</td>
<td>2</td>
<td>0.83</td>
</tr>
<tr>
<td>Residues in the disallowed regions [xx]</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>Number of non-glycine and non-proline residues</td>
<td>241</td>
<td>100.0</td>
</tr>
<tr>
<td>Number of end residues (excl. Gly and Pro)</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>Number of glycine residues</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>Number of proline residues</td>
<td>6</td>
<td>-</td>
</tr>
<tr>
<td>Total number of residues</td>
<td>259</td>
<td>-</td>
</tr>
</tbody>
</table>

Fig. 3 Ramachandran plot analysis of HP R_27 through Swiss model workshop

The ligand binding site of the hypothetical protein R_27 was determined through CASTp server. It has been found that 57 amino acids are involved in potent active site. The best active site was found in areas with 1382.7 and a volume of 1754.3 amino acids. Some of them are also found in active site of other proteins [14, 16, 31]. The active site of hypothetical protein R_27 depicted in Fig. 4.
Conclusion
We have analyzed HP R_27 from *Salmonella enterica* serovar *typhimurium* (*S. typhi*) through *in silico* approach. The physicochemical parameters prediction, sub-cellular localization and functional annotation are useful in distinguishing the hypothetical protein with DNA repair mechanism and PLD-like activity. The results might assist in developing drugs against *S. typhi*. This *in silico* approach can be further utilized in drug design to identify putative drug targets for other clinically significant pathogens.

Acknowledgements
We cordially thank Abdullah Al-Emran, Assistant Professor of the Department of Biotechnology and Genetic Engineering, Mawlana Bhashani Science and Technology University and Md. Anayet Hasan, Department of Genetic Engineering and Biotechnology, University of Chittagong, for their suggestions and inspiration during research proceedings.

References


Md. Arif Khan, B.Sc. (Hons.)
E-mail: arifkhanbge35@gmail.com
Md. Arif Khan is working as an M.Phil. (Chemistry) student at Chemoinformatics Laboratory of Department of Science and Humanities, Military Institute of Science and Technology. He completed his B.Sc. (Hons.) from the Department of Biotechnology and Genetic Engineering, Mawlana Bhashani Science and Technology University, Bangladesh. His areas of research are bioinformatics, chemoinformatics, medicinal chemistry, cancer biology, immunology, computational biology and computational drug design.

Hossain Ahmed
E-mail: ahmedshohag.bd@gmail.com
Hossain Ahmed is doing M.Sc. in Biotechnology and Genetic Engineering at the Department of Biotechnology and Genetic Engineering, University of Development Alternative, Bangladesh. Currently he is working in pharmacology, bioinformatics and medicinal plant.

Nasreen Jahan, M.Sc.
E-mail: naza.bge@gmail.com
Nasreen Jahan is working as a Lecturer in the Department of Biotechnology and Genetic Engineering, Mawlana Bhashani Science and Technology University, Bangladesh. He completed B.Sc. (Hons.) and M.Sc. in Biotechnology and Genetic Engineering from Mawlana Bhashani Science and Technology University, Bangladesh. Her areas of research are chemical biology, molecular genetics, bioinformatics, and immunology.
Syed Raju Ali, B.Sc.
E-mail: m15121307@bau.edu.bd

Syed Raju Ali is a M.Sc. student of Biotechnology Department, Bangladesh Agricultural University, Mymensingh, Bangladesh. He completed his B.Sc. in Biotechnology and Genetic Engineering from Mawlana Bhashani Science and Technology University. His scientific interests are in molecular biology, molecular biotechnology, human genetics, bioinformatics and biologics.

Md. Al Amin, M.Sc.
E-mail: josnasnan@gmail.com

Md. Al Amin completed his M.Sc. from Biotechnology Department, Bangladesh Agricultural University, Mymensingh, Bangladesh. He completed his B.Sc. in Biotechnology and Genetic Engineering from Mawlana Bhashani Science and Technology University. His scientific interests are in molecular genetics, cell and developmental biology, cellular biology of cancer and bioinformatics.

Assoc. Prof. Major Mohammad Neaz Morshed, Ph.D.
E-mail: mmneaz@hotmail.com

Mohammad Neaz Morshed is working as an Associate Professor in Department of Science and Humanities, Military Institute of Science and Technology. He has completed his Ph.D. in Medicinal and Pharmaceutical Chemistry from Korea Institute of Science and Technology, Seoul, South Korea. He completed his B.Sc. in Chemistry from Dhaka University and M.Sc. in Organic Chemistry from Cheju National University, South Korea. His research interests are chemoinformatics, computational chemistry, computer-aided drug design, virtual screenings, molecular modelling, QSAR/QSPR, pharmacophore modeling, chemical databases, molecular docking.