Identification of potential drug targets by subtractive genome analysis of Escherichia coli O157:H7: an in silico approach

Shakhinur Islam Mondal1,6,*
Sabina Ferdous1,6,*
Nurnabi Azad Jewel1
Arzuba Akter1,6
Zaber Mahmud1
Md Muzahidul Islam1
Tanzila Afrin3
Nurul Karim4,5

1Genetic Engineering and Biotechnology Department, Shahjalal University of Science and Technology, Sylhet, Bangladesh; 2Biochemistry and Molecular Biology Department, Shahjalal University of Science and Technology, Sylhet, Bangladesh; 3Department of Pharmacy, East West University, Aftabnagar, Bangladesh; 4Biochemistry and Molecular Biology Department, Jahangirnagar University, Savar, Bangladesh; 5Division of Parasitology, 6Division of Microbiology, Department of Infectious Diseases, Faculty of Medicine, University of Miyazaki, Miyazaki, Japan

*These authors contributed equally to this work

Abstract: Bacterial enteric infections resulting in diarrhea, dysentery, or enteric fever constitute a huge public health problem, with more than a billion episodes of disease annually in developing and developed countries. In this study, the deadly agent of hemorrhagic diarrhea and hemolytic uremic syndrome, Escherichia coli O157:H7 was investigated with extensive computational approaches aimed at identifying novel and broad-spectrum antibiotic targets. A systematic in silico workflow consisting of comparative genomics, metabolic pathways analysis, and additional drug prioritizing parameters was used to identify novel drug targets that were essential for the pathogen’s survival but absent in its human host. Comparative genomic analysis of Kyoto Encyclopedia of Genes and Genomes annotated metabolic pathways identified 350 putative target proteins in E. coli O157:H7 which showed no similarity to human proteins. Further bioinformatic approaches including prediction of subcellular localization, calculation of molecular weight, and web-based investigation of 3D structural characteristics greatly aided in filtering the potential drug targets from 350 to 120. Ultimately, 44 non-homologous essential proteins of E. coli O157:H7 were prioritized and proved to have the eligibility to become novel broad-spectrum antibiotic targets and DNA polymerase III alpha (dnaE) was the top-ranked among these targets. Moreover, druggability of each of the identified drug targets was evaluated by the DrugBank database. In addition, 3D structure of the dnaE was modeled and explored further for in silico docking with ligands having potential druggability. Finally, we confirmed that the compounds N-coelesterazine and N-(1,4-dihydro-5H-tetrazol-5-ylidene)-9-oxo-9H-xanthene-2-sulfonamide were the most suitable ligands of dnaE and hence proposed as the potential inhibitors of this target protein. The results of this study could facilitate the discovery and release of new and effective drugs against E. coli O157:H7 and other deadly human bacterial pathogens.

Keywords: E. coli O157:H7, KEGG metabolic pathways, novel and broad-spectrum antibiotic targets, DNA polymerase III alpha, homology modeling

Introduction

Enteropathogenic Escherichia coli and enterohemorrhagic E. coli (EHEC) infections in humans are a major source of morbidity and mortality in both developing and developed countries.1 Among various pathogenic E. coli strains that cause intestinal or extra-intestinal diseases in humans, the most devastating are shiga toxins producing EHEC strains, because they cause not only diarrhea and hemorrhagic colitis but also life-threatening hemorrhagic uremic syndrome and encephalopathy.2 Over 100 serotypes of shiga toxin-producing E. coli have been associated with human infections and the most common serotype is E. coli O157:H7. Several deadly outbreaks of E. coli O157:H7 were reported in Canada, United States, Great Britain, and Japan.3–5
However, the massive outbreak in Sakai city, Japan, in 1996 is of great concern as a number of deaths from the infection were reported.6

The genome of EHEC O157:H7 Sakai strain was sequenced in 2001.7 The sequence analysis revealed that this strain contains 18 prophages (Sp1 to Sp18), six phage-like elements (SpLE1 to SpLE6), and two plasmids (pO157 and pOSAK1). To know the mechanism underlying pathogenicity of this bacterium, a substantial number of virulence-related genes or functions associated with various stages of infection have been identified.7 However, lack of details of functional annotations often limit the possibility to use them as targets for designing new drugs against this pathogen.

The treatment of E. coli O157:H7 mostly relies on conventional antibiotic therapy although some studies have highlighted that there is no evidence that this improves the course of disease and antibiotic treatment of patients with E. coli O157:H7 infection increases the risk of hemolytic–uremic syndrome.8,9 Moreover, an increase of antibiotic resistance has been reported in E. coli O157:H7 over the last 30 years which is also alarming.10–14 The accumulated results strongly suggest that there is an urgent and continuing need to find new drug and vaccine candidates to tackle this deadly pathogen.

Drug target identification is the first step in the drug discovery process.15 However, traditional drug discovery methods are time-consuming, expensive, and often yield few drug targets. In contrast, advances in complete genome sequencing, bioinformatics, and cheminformatics represent an attractive alternative approach to identify drug targets worthy of experimental follow-up. Because of the availability of both pathogen and host–genome sequences, it has become easier to identify drug targets at the genomic level for any given pathogen.16,17 In recent years, computational methods have been used widely for the identification of potential drug and vaccine targets in different pathogenic microorganisms.18–21 Subtractive and comparative genomics approach combined with metabolic pathway analysis was found to be an efficient way to identify the protein-set essential for the pathogen’s survival but absent in the host.22 Subtraction of the host genome from essential genes of pathogens helps in searching for non-human homologous targets which ensures no interaction of drugs with human targets. On the other hand, comparative genomics method emphasizes the selection of conserved proteins amongst several species as most favorable targets.23–26 The use of advanced bioinformatics tools with integrated genomics, proteomics, and metabolomics may ensure the discovery of potential drug targets for most of the infectious diseases. Once the target(s) have been identified, the in silico virtual screening of different chemical databases could provide unprecedented opportunity to select and design the best possible inhibitor(s).27

In this study, we took an in-depth in silico approach to identify novel therapeutic targets in E. coli O157:H7 Sakai strain by combining analysis of metabolomics and genomics data. Instead of analysis of whole genome, we particularly considered the key essential or survival proteins of the pathogen which are non-homologous to the host. We elucidated a good number of novel targets in E. coli O157:H7 to design effective drugs against broad-spectrum pathogenic bacteria. Moreover, we provided a modeled 3D structure of DNA polymerase III alpha (dnaE) which was selected as the best possible target for inhibition and designing potential drugs. To the best of our knowledge this was the first in silico identification of drug targets in E. coli O157:H7.

Materials and methods
Pathway analysis and protein retrieval
Figure 1 showed the strategies for identification of suitable drug targets and prediction of inhibitor used. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database28,29 was searched for metabolic pathways for both human genomes and E. coli O157:H7 Sakai strain. The identification numbers of all pathways from both organisms were listed. A manual comparison was made, and pathways that did not appear in the human genomes but were present in the pathogen, according to the KEGG database annotations, were selected as unique to E. coli O157:H7, while the remaining pathways were listed as common. Amino acid sequences of proteins from common and unique pathways were obtained from UniprotKB.30

Identification of essential proteins and non-homologous proteins in humans
Database of Essential Genes (DEG)31 was used to identify the essential proteins involved in host-pathogen pathways. The DEG 6.8 database was retrieved from http://www.tubic.tju.edu.cn/deg/. The BioEdit Sequence Alignment Editor (version 7.1.3) was used for Protein Basic Local Alignment Search Tool (BLASTP) search to screen for and eliminate the probable essential proteins of the organism setting e-value cut off 10−4, sequence identity >35%, bit score >100 and others as default.

All human protein sequences were retrieved from Refseq database ftp site (ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/protein/) and essential proteins were subjected to BLASTP search against the human proteins with BioEdit. Only the non-hit proteins at e-value cut off 10−10, were
selected as non-homologous proteins to avoid any functional similarity with host proteome.

**Subcellular localization prediction and targets’ prioritization**

Subcellular localization prediction of the essential non-human proteins was done by PSORTb version 3.0.2, which predicts three types of localization such as: cytoplasmic, membrane, and extracellular proteins for Gram-negative bacteria. The potential drug targets were evaluated by several molecular and structural criteria for prioritizing suitable drug targets. Drug targets’ prioritization involved calculation of molecular weight (MW) using computational tools and drug targets associated literature available at Swiss-Prot database. Protein Data Bank and ModBase databases were searched for identifying experimentally and computationally solved 3D structures respectively. The selected protein was searched for any structural identity with
the 3D ligand binding site of any human protein structure on the web server SMAP-WS at a cut off value of 30% sequence identity.

Moreover, druggability is another important prioritization criterion for therapeutic targets; that is defined as the likelihood of being able to modulate the activity of the therapeutic target protein with a small-molecule drug. The druggability of identified drug targets was measured by mining DrugBank contents. BLASTP with default parameters was performed against the list of targets of compounds found within DrugBank to align the potential drug targets from E. coli O157:H7. Alignments with e-values less significant than $10^{-25}$ were removed as described previously as selection criteria for filtering BLAST results in identifying drug targets of bacterial genomes.

**Identification of novel targets and searching for common proteins**

To identify novel targets among the potential targets, databases DrugBank, SuperTarget, and Therapeutic Target Database, were searched for similarity with the cytoplasmic proteins. Parameters were set as e-value $<10^{-3}$, sequence identity $>35\%$, and bit score $>100$. The non-hit proteins at the threshold value were selected as novel drug targets. To search for the common proteins amongst pathogenic bacteria, all protein sequences of 73 different strains of pathogenic bacteria were retrieved from PATRIC database. The novel targets were subjected to BLASTP against these proteomes at e-value cut off $10^{-3}$, sequence identity $>35\%$, bit score $>100$ with BioEdit software. The proteins that were found to be common in at least 40 pathogenic strain proteomes were listed as broad-spectrum targets. Different bacterial species were used as references.

**Homology modeling**

As no exact protein data bank (PDB) structure was available for dnaE in PDB, it was subjected to BLAST search against PDB structures using 0.001 e-value cut off. The template for homology modeling was chosen considering X-ray diffraction resolution and highest sequence similarity. Homology modeling was done on ESyPred3D server.

**Structure validation and active site prediction**

The modeled structure was assessed through SWISSMODEL structure assessment tool and ANOLEA (atomic non-local environment assessment) assessing the packing quality of the models. PROCHECK suite of programs checked the stereochemical quality of protein structures. Energy minimization was carried out by GROMOS96 with default parameters implemented in Swiss PDB Viewer (version 4.0.4). Active site of the modeled structure was determined by CASTp server.

**Virtual screening, drug likeliness, and toxicity analysis**

For visual analysis and comparison of the active site interaction with the ligands, 24 ligand molecules of E. coli K-12 dnaE subunit were extracted from BindingDB database and docked with the subject receptor. Virtual screening was done with a total 6,460 molecules, 5,040 experimental and 1,447 approved molecules deposited in DrugBank, based on selected active sites into the dnaE. Virtual screening was performed on Linux (Ubuntu 10.04) based cluster with 32 core systems. The top 100 molecules were selected based on lowest binding energy after being docked several times. The selected molecules were analyzed by Lipinski’s rule of five. The ligand interaction analysis and visualization was done with the help of Pymol and Discovery Studio (Accelrys, San Diego, CA, USA). Absorption, distribution, metabolism, excretion, and toxicity (ADMET) prediction was carried out with PreADMET server. PreADMET predicts mutagenicity and carcinogenicity of a compound and helps to avoid toxic compound. Oral bioavailability was predicted with FAF-Drugs2 program of Mobyle@RPBS server.

**Results and discussion**

**Identification of pathogen-specific pathways**

Here we report the first computational comparative and subtractive genomics analysis of different metabolic pathways from E. coli O157:H7, for the identification of potential drug targets. A systematic workflow was defined involving several bioinformatics tools, databases, and drug target prioritization parameters (Figure 1), with the goal of obtaining information about proteins that were involved in various metabolic pathways of E. coli O157:H7, but absent in its host, therefore avoiding any potential side effects. When we searched the KEGG database for pathogen metabolic pathways, a total of 105 different pathways appeared. In order to identify drug targets involved in pathogen-specific metabolic pathways, comparative analysis of the metabolic pathways of the host and pathogen was performed. Detailed pathways analyses revealed that a total of 35 pathways were present only in the pathogen and termed as pathogen-specific pathways, and the remaining 70 pathogen pathways were defined as common host-pathogen pathways as listed in Table 1.
### Table 1 Host–pathogen common and pathogen-specific pathways from KEGG database

<table>
<thead>
<tr>
<th>Pathway IDs</th>
<th>Pathway names</th>
<th>Pathway IDs</th>
<th>Pathway names</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Host–pathogen common pathway</strong></td>
<td></td>
<td><strong>Pathway IDs</strong></td>
<td><strong>Pathway names</strong></td>
</tr>
<tr>
<td>ecs00010</td>
<td>Glycolysis/Gluconeogenesis</td>
<td>ecs00860</td>
<td>Porphyrin and chlorophyll metabolism</td>
</tr>
<tr>
<td>ecs00020</td>
<td>Citrate cycle (TCA cycle)</td>
<td>ecs00900</td>
<td>Terpenoid backbone biosynthesis</td>
</tr>
<tr>
<td>ecs00030</td>
<td>Pentose phosphate pathway</td>
<td>ecs00920</td>
<td>Sulfur metabolism</td>
</tr>
<tr>
<td>ecs00040</td>
<td>Pentose and glucuronate interconversions</td>
<td>ecs00970</td>
<td>Aminoacyl-tRNA biosynthesis</td>
</tr>
<tr>
<td>ecs00051</td>
<td>Fructose and mannose metabolism</td>
<td>ecs01040</td>
<td>Biosynthesis of unsaturated fatty acids</td>
</tr>
<tr>
<td>ecs00052</td>
<td>Galactose metabolism</td>
<td>ecs02010</td>
<td>ABC transporters</td>
</tr>
<tr>
<td>ecs00053</td>
<td>Ascorbate and aldarate metabolism</td>
<td>ecs03010</td>
<td>Ribosome</td>
</tr>
<tr>
<td>ecs00061</td>
<td>Fatty acid biosynthesis</td>
<td>ecs03018</td>
<td>RNA degradation</td>
</tr>
<tr>
<td>ecs00071</td>
<td>Fatty acid metabolism</td>
<td>ecs03020</td>
<td>RNA polymerase</td>
</tr>
<tr>
<td>ecs00080</td>
<td>Ubiquinone and other terpenoid-quinone biosynthesis</td>
<td>ecs03030</td>
<td>DNA replication</td>
</tr>
<tr>
<td>ecs00090</td>
<td>Oxidative phosphorylation</td>
<td>ecs03060</td>
<td>Protein export</td>
</tr>
<tr>
<td>ecs00100</td>
<td>Purine metabolism</td>
<td>ecs03410</td>
<td>Base excision repair</td>
</tr>
<tr>
<td>ecs00110</td>
<td>Pyrimidine metabolism</td>
<td>ecs03420</td>
<td>Nucleotide excision repair</td>
</tr>
<tr>
<td>ecs00120</td>
<td>Alanine, aspartate and glutamate metabolism</td>
<td>ecs03430</td>
<td>Mismatch repair</td>
</tr>
<tr>
<td>ecs00130</td>
<td>Glycine, serine and threonine metabolism</td>
<td>ecs03440</td>
<td>Homologous recombination</td>
</tr>
<tr>
<td>ecs00140</td>
<td>Valine, leucine and isoleucine degradation</td>
<td>ecs03450</td>
<td>Sulfur relay system</td>
</tr>
<tr>
<td>ecs00150</td>
<td>Valine, leucine and isoleucine biosynthesis</td>
<td>ecs03460</td>
<td>Glycerolipid metabolism</td>
</tr>
<tr>
<td>ecs00160</td>
<td>Lysine biosynthesis</td>
<td>ecs03470</td>
<td>Inositol phosphate metabolism</td>
</tr>
<tr>
<td>ecs00170</td>
<td>Lysine degradation</td>
<td>ecs03480</td>
<td>Glycerophospholipid metabolism</td>
</tr>
<tr>
<td>ecs00180</td>
<td>Arginine and proline metabolism</td>
<td>ecs03490</td>
<td>Arachidonic acid metabolism</td>
</tr>
<tr>
<td>ecs00190</td>
<td>Histidine metabolism</td>
<td>ecs03500</td>
<td>alpha-Linolenic acid metabolism</td>
</tr>
<tr>
<td>ecs00200</td>
<td>Tyrosine metabolism</td>
<td>ecs03510</td>
<td>Sphingolipid metabolism</td>
</tr>
<tr>
<td>ecs00210</td>
<td>Phenylalanine metabolism</td>
<td>ecs03520</td>
<td>Pyruvate metabolism</td>
</tr>
<tr>
<td>ecs00220</td>
<td>Tryptophan metabolism</td>
<td>ecs03530</td>
<td>Glyoxylate and dicarboxylate metabolism</td>
</tr>
<tr>
<td>ecs00230</td>
<td>Phenylalanine, tyrosine and tryptophan biosynthesis</td>
<td>ecs03540</td>
<td>Propanoate metabolism</td>
</tr>
<tr>
<td>ecs00240</td>
<td>beta-Alanine metabolism</td>
<td>ecs03550</td>
<td>Butanoate metabolism</td>
</tr>
<tr>
<td>ecs00250</td>
<td>Taurine and hypotaurine metabolism</td>
<td>ecs03560</td>
<td>One carbon pool by folate</td>
</tr>
<tr>
<td>ecs00260</td>
<td>Selenocompound metabolism</td>
<td>ecs03570</td>
<td>Thiamine metabolism</td>
</tr>
<tr>
<td>ecs00270</td>
<td>Cyanoamino acid metabolism</td>
<td>ecs03580</td>
<td>Riboflavin metabolism</td>
</tr>
<tr>
<td>ecs00280</td>
<td>d-Glutamine and d-glutamate metabolism</td>
<td>ecs03590</td>
<td>Vitamin B6 metabolism</td>
</tr>
<tr>
<td>ecs00290</td>
<td>d-Alanine metabolism</td>
<td>ecs03600</td>
<td>Nicotinate and nicotinamide metabolism</td>
</tr>
<tr>
<td>ecs00300</td>
<td>Lysine, serine and threonine metabolism</td>
<td>ecs03610</td>
<td>Pantothenate and CoA biosynthesis</td>
</tr>
<tr>
<td>ecs00310</td>
<td>Valine, leucine and isoleucine degradation</td>
<td>ecs03620</td>
<td>Biotin metabolism</td>
</tr>
<tr>
<td>ecs00320</td>
<td>Valine, leucine and isoleucine biosynthesis</td>
<td>ecs03630</td>
<td>Lipoic acid metabolism</td>
</tr>
<tr>
<td>ecs00330</td>
<td>Lysine biosynthesis</td>
<td>ecs03640</td>
<td>Lipoic acid metabolism</td>
</tr>
<tr>
<td>ecs00340</td>
<td>Lysine degradation</td>
<td>ecs03650</td>
<td>Folate biosynthesis</td>
</tr>
<tr>
<td><strong>Pathogen-specific pathway</strong></td>
<td></td>
<td><strong>Pathway IDs</strong></td>
<td><strong>Pathway names</strong></td>
</tr>
<tr>
<td>ecs00281</td>
<td>Geraniol degradation</td>
<td>ecs06270</td>
<td>Aminobenzoate degradation</td>
</tr>
<tr>
<td>ecs00361</td>
<td>Chlorocyclohexane and chlorobenzene degradation</td>
<td>ecs06330</td>
<td>Nitrotoluene degradation</td>
</tr>
<tr>
<td>ecs00362</td>
<td>Benzoate degradation</td>
<td>ecs06420</td>
<td>Ethylbenzene degradation</td>
</tr>
<tr>
<td>ecs00363</td>
<td>Biphenol degradation</td>
<td>ecs06600</td>
<td>CS-Branchied dibasic acid metabolism</td>
</tr>
<tr>
<td>ecs00364</td>
<td>Fluorobenzoate degradation</td>
<td>ecs06800</td>
<td>Methane metabolism</td>
</tr>
<tr>
<td>ecs00401</td>
<td>Novobiocin biosynthesis</td>
<td>ecs09030</td>
<td>Limonene and pinene degradation</td>
</tr>
<tr>
<td>ecs00440</td>
<td>Phosphonate and phosphinate metabolism</td>
<td>ecs09100</td>
<td>Nitrogen metabolism</td>
</tr>
<tr>
<td>ecs00473</td>
<td>D-Alanine metabolism</td>
<td>ecs09300</td>
<td>Caprolactam degradation</td>
</tr>
<tr>
<td>ecs00521</td>
<td>Streptomycin biosynthesis</td>
<td>ecs01053</td>
<td>Biosynthesis of siderophore group</td>
</tr>
<tr>
<td>ecs00523</td>
<td>Polyketide sugar unit biosynthesis</td>
<td>ecs01110</td>
<td>non-ribosomal peptides</td>
</tr>
<tr>
<td>ecs00540</td>
<td>Lipopolysaccharide biosynthesis</td>
<td>ecs01120</td>
<td>Biosynthesis of secondary metabolites</td>
</tr>
<tr>
<td>ecs00550</td>
<td>Peptidoglycan biosynthesis</td>
<td>ecs02020</td>
<td>Microbial metabolism in diverse environments</td>
</tr>
<tr>
<td>ecs00621</td>
<td>Dioxin degradation</td>
<td>ecs02030</td>
<td>Two-component system</td>
</tr>
<tr>
<td>ecs00622</td>
<td>Toluene degradation</td>
<td>ecs02040</td>
<td>Bacterial chemotaxis</td>
</tr>
<tr>
<td>ecs00623</td>
<td>Polychlorinated aromatic hydrocarbon degradation</td>
<td>ecs02060</td>
<td>Phosphotransferase system</td>
</tr>
<tr>
<td>ecs00624</td>
<td>Polychlorinated aromatic hydrocarbon degradation</td>
<td>ecs03070</td>
<td>Bacterial secretion system</td>
</tr>
<tr>
<td>ecs00625</td>
<td>Chloroalane and chloroalkene degradation</td>
<td>ecs05130</td>
<td>Pathogenic Escherichia coli infection</td>
</tr>
<tr>
<td>ecs00626</td>
<td>Naphthalene degradation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Abbreviation:** KEGG, Kyoto Encyclopedia of Genes and Genomes.
Identification of non-homologous essential proteins

To be an effective drug target, the protein should be crucial for the survival of the pathogen in the host body but non-homologous to human proteins, and this criterion is a prerequisite for avoiding cross binding of drugs with human proteins, and drug side effect probability. Unique pathways are those that are specific to the pathogen but absent in its host. Proteins in these pathways can also be considered as unique to the pathogen and might serve as potential drug and vaccine targets. Moreover, several unique proteins are known to be present in common pathways as identified during our analysis of E. coli O157:H7 (data not shown) and in several previous studies on other bacteria.

Additionally, we also identified that a single unique protein can also take part in multiple pathways. Proteins that are involved in more than one pathway could be more effective drug targets when, in addition, they are non-homologous proteins. Nevertheless, being unique or non-human and involved in metabolic pathways are not the sole criteria for selecting favorable drug targets. It is possible that a bacterial protein showing no similarity to host proteins might be involved in multiple metabolic pathways, but its disruption might be of no therapeutic benefit. The reasons may include presence of paralogs, isoenzymes, and most importantly, being non-essential for the pathogen’s survival. Again, not all essential proteins are non-homologous in nature. Therefore, pathogen proteins that fulfill the criteria of being unique and essential at the same time represent more attractive drug targets.

Based on the criteria discussed above, we identified 780 probable essential proteins from host–pathogen common pathways, and 234 from pathogen unique pathways (Supplementary materials). These proteins showed good similarity with the experimentally proven essential proteins recorded in DEG database. BLASTP search against human proteome narrowed down the target proteome to only 220 and 130 proteins from common pathways and unique pathways respectively resulting in 350 proteins which were essential for the pathogen’s survival and non-homologous to the host (Supplementary materials).

Subcellular localization and prediction of drug target prioritization

Localization of the proteins in the cell is an important factor for identification of suitable and effective drug targets. Membrane localized proteins are difficult to purify and assay and therefore, cytoplasmic proteins are more favorable as drug targets. Other major factors are: accessibility value of a target protein; preferably low MW (<100 kDa); whether a potential drug is a transmembrane protein; and availability of 3D structural information. Based on these essential features, the identified non-homologous essential proteins of E. coli O157:H7 were further characterized. Most of the proteins had MW less than 100 kDa indicating the possibility to experimentally study these proteins for drug development.

From the common pathways, 152 proteins were found to be cytoplasmic, 54 proteins to be membrane localized, and the other 14 proteins to be of unknown localization. From pathogen unique pathways, 70, 52, and eight proteins were found to be cytoplasmic, membrane localized, and of unknown localization respectively (Supplementary materials). Based on these results, 152 and 70 cytoplasmic proteins from common pathways and unique pathways respectively were considered for further analysis to identify suitable drug targets (Figure 2).

The DrugBank is a unique bioinformatics and cheminformatics resource which combines detailed drug data with comprehensive information about drug targets. By utilizing the DrugBank database, druggability of non-homologous essential proteins of E. coli O157:H7 was measured by sequence similarity to the targets of small-molecule drugs and BLASTP search was performed to align the non-homologous essential proteins to the list of drug-targeted proteins from DrugBank. A total of 129 proteins of E. coli O157:H7 showed high similarities with the binding partners of US Food and Drug Administration (FDA)-approved drugs, experimental small-molecule compounds, or nutraceutical compounds supporting the potential of comparative genomics in drug discovery. Among these, 80 proteins and 49 proteins were from common and unique pathways respectively (Supplementary materials).

By this comparison with drug-targeted proteins additionally a list of approved drug and drug-like compounds was identified that bind to proteins with similar sequences to those of E. coli O157:H7. It is reasonable that careful filtering
of this set could reveal a number of potential compounds that were primed for optimization and derivatization using traditional medicinal chemistry although protein sequence similarity does not guarantee identical structures or binding pockets.\textsuperscript{38} We searched the presence of 3D structures of the non-host essential proteins of \textit{E. coli} O157:H7. Such information could greatly facilitate a structure-based drug design, including homology modeling, docking, virtual screening or pharmacophore-based screening.\textsuperscript{54} The PDB and ModBase were used as sources for the 3D structural information. Out of 350 non-host essential proteins, ten proteins were identified as having experimentally determined 3D structures in PDB and 312 were found to have 3D models in ModBase (Supplementary materials).

**Novel targets' identification**
Proteins which showed significant similarity with the databases were discarded and the remaining protein sequences were taken as novel targets. Forty-four proteins from pathogen unique pathways and 76 proteins from common pathways totaling 120 proteins were defined as novel proteins (Supplementary materials). Metabolic pathway analysis indicated that these 120 proteins were involved in 12 biological processes unique to pathogens and 49 biological processes that were common in both host and pathogen. Moreover, all of these 61 biological processes were classified into 12 classes: amino acid metabolism, carbohydrate metabolism, energy metabolism, glycan biosynthesis and metabolism, lipid metabolism, metabolism of cofactors and vitamins, metabolism of other amino acids, nucleotide metabolism, genetic information processing, environmental information processing, cellular processes, and others (Supplementary materials). Figure 3 showed the percentage distribution of novel drug targets involved in different biological process.

**Novel targets in pathogens' unique pathways**
Our study revealed that 44 proteins were uniquely involved in pathogen-specific 12 unique pathways and these were lipopolysaccharide (LPS) biosynthesis, peptidoglycan biosynthesis, methane metabolism, C5-branched dibasic acid metabolism, nitrogen metabolism, phosphonate and phosphinate metabolism, bacterial secretion system, phosphotransferase system (PTS), flagellar assembly, two-component system, biosynthesis of siderophore group non-ribosomal peptides, and bacterial chemotaxis.

Three enzymes of LPS core biosynthesis and three enzymes of lipid A pathway were found uniquely present in LPS biosynthesis pathway (KEGG Pathway: map00540). In LPS of the Enterobacteriaceae, the core oligosaccharide is responsible for many of the biological properties of the antigenic O-polysaccharide.\textsuperscript{55} On the other hand, enzymes of lipid A pathway required for bacterial growth could be excellent targets for the development of new antibiotics.\textsuperscript{56} Moreover, murF; EC 6.3.2.10 was the only enzyme present in peptidoglycan biosynthesis (KEGG Pathway: map00550). In Gram-positive bacteria, the cell wall composed of peptidoglycan macromolecules and many surface proteins of Gram-positive bacteria is thought to be important for survival within an infected host.\textsuperscript{57}
In methane metabolism (KEGG Pathway: map00680), AckA; EC 2.7.2.1 was found to be uniquely present. Methanotrophs involved in the global methane cycle consume methane as their sole source of carbon and energy for growing, whereas methanogens can obtain energy for growth by converting a limited number of substrates to methane. The protein ilvH, acetolactate synthase III small subunit, is a unique protein present in C5-branched dibasic acid metabolism (KEGG Pathway: map00660) which provides alternative sources of carbon and energy.

There was a unique presence of four proteins in nitrogen metabolism (KEGG Pathway: map00910). In the nitrogen cycle, different reductive or oxidative reactions are utilized by prokaryotes for energy conservation. In case of oxygen-deprived growth conditions for E. coli, anaerobic respiration nitrate is the preferred electron acceptor. The putative resistance protein was identified as unique protein present in phosphonate and phosphinate metabolism (KEGG Pathway: map00440). Natural products containing carbon-phosphorous bonds, so-called C-P compounds have been found in many organisms, but only protists and bacteria, mostly actinobacteria, have biosynthetic capacity. Moreover, the secB; protein-export protein secB and etpN; type 4 prepilin-like proteins leader peptide-processing enzyme, that are the parts of bacterial secretion system (KEGG Pathway: map03070), were found as unique. Many proteins implicated in efflux of different toxins and drugs; virulence and biogenesis of different organelles (pili and flagella) are secreted. The secB is involved in efficient export of proteins across the cytoplasmic membrane in E. coli. In PTS (KEGG Pathway: map02060), n9 proteins were found to be uniquely present. It has been known that PTS is involved in transportation of more than 20 carbohydrates in bacteria and plays a major role in phosphorylation and uptake of carbohydrates and controlling their metabolism. Flagellar FljJ protein which is part of flagellar assembly (KEGG Pathway: map02040) as well as a putative general chaperone and cytoplasmic protein was identified as unique in the pathogen. The bacterial flagellum extending from the cytoplasm to the cell exterior serves as both a motor organelle and a protein export/assembly apparatus. A total of 19 proteins were identified as unique in the pathogen-specific pathway in a two-component system (KEGG Pathway: map02020). Bacterial two-component system is required for adaptation to external stimuli and can affect changes in cellular physiology. A single protein was identified to be involved in biosynthesis of siderophore group non-ribosomal polypeptide pathway. Polyketide synthases and non-ribosomal polypeptide synthetases are known to be responsible for the biosynthesis of several siderophores such as enterobactin in E. coli. Two proteins were detected in case of bacterial chemotaxis. In chemotaxis, bacteria sense chemical gradients in the environment and move toward favorable conditions. The pathway is arguably best characterized in the case of E. coli.

**Novel targets in common host-pathogen pathways**

We also identified 76 proteins in 49 host-pathogen common metabolic pathways as novel targets. The pathways were grouped as metabolism of amino acid, carbohydrate, energy, lipid, other amino acids, nucleotide, cofactors and vitamins, and genetic information processing, environmental information processing, cellular processes, and others (Supplementary materials).

**Identification of broad-spectrum targets**

Common proteins among several species would be well broad-spectrum antibiotic targets. We used 73 species as reference and proteins which were common in at least 40 different species were listed as broad-spectrum targets (Supplementary materials). Forty-four proteins were identified as broad-spectrum targets (Supplementary materials). In addition, broad-spectrum proteins involved in multiple pathways would be better targets as their inhibition of activity will hamper more than one system in the pathogen. dnaE subunit and AckA were involved in a maximum number of pathways. AckA catalyzes the reversible reaction of formation of acetyl phosphate from acetate and ATP. Whereas dnaE participates in some critical pathways of the pathogen like purine metabolism, pyrimidine metabolism, DNA replication, mismatch repair, and homologous recombination. There is no resolved X-ray crystallography structure for both dnaE (E. coli O157:H7 Sakai strand) and AckA (E. coli O157:H7 Sakai strand) as we intended to do homology modeling. However, dnaE subunit (Uniprot ID: Q8X8X5) was preferred over AckA (Uniprot ID: P0A6A5) based on suitability of homology modeling and docking studies. During the BLAST search against PDB structures (threshold e-value <0.001), dnaE subunit showed 98% sequence identity with E. coli replicative dnaE subunit (PDB ID: 2HNN), whereas 94% for AckA with Salmonella enterica subspecies enterica serovar Typhimurium AckA (PDB ID: 3SK3 and 3SLC) (data not shown). Furthermore, an inter-domain motion during ligand binding as well as a ligand binding pocket located at the...
The dimeric interface of form-II AckA has been reported\textsuperscript{24} that is very hard to address through in silico docking software. In case of dnaE subunit, literature searching helped us to identify the pocket for ligand binding as well as important and conserved residues within the pocket that are important for catalytic activity.\textsuperscript{75}

Moreover, we conducted BLAST searching for dnaE (Uniprot ID: Q8X8X5) with e-value cut off of 0.001 against UniprotKB. We selected the organisms whose proteins showed at least 75\% sequence identity with dnaE. The organisms were searched in PATRIC database to check their host and pathogenicity. Only the human hosts were considered and predicted to be pathogenic if involved in disease(s) according to PATRIC database.\textsuperscript{42} We found that all the organisms are pathogenic except Lelliottia, Kluyvera, Hafnia, Ewingella, Cedecea, and Yokenella. However, literature searching helped us to conclude that Kluyvera, Yokenella, Ewingella, Hafnia, and Cedecea are also involved in diseases.\textsuperscript{76–81} We did not find sufficient information about Lelliottia. Thus it is clearly demonstrated that \textit{E. coli} O157:H7 dnaE does not share any sequence similarity with non-pathological bacterial dnaE. Therefore, due to the above advantages, dnaE was selected for homology modeling and subsequent structure-based drug designing.

**Homology modeling**

DNA polymerase III holo-enzyme has ten different peptides arranged in an asymmetric dimer and contains a 3'-5' exo-nuclease activity. The alpha subunit is at the core enzyme and mainly functions in the polymerase activity. The best hit of similarity search identified the crystal structure of the catalytic alpha subunit of \textit{E. coli} replicative dnaE DNA polymerase III (PDB ID: 2HNH).\textsuperscript{82} It showed 99.98\% sequence similarity with our target sequence and was used as the template for homology modeling. The modeled structure was shown in Figure 4.

**Structure validation and energy minimization**

Advancing toward the way of in silico drug design with protein models depends largely on the quality of the models. Inspection of the Psi/Phi Ramachandran plot analysis showed that the model built by ESyPred3D has residues in most favored regions 92.4\%, residues in additional allowed regions 7.1\%, and residues in generously allowed regions 0.3\%. A good quality model would be expected to have over 90\% in the most favored regions (Figure 5A).

This structure was also verified with ANOLEA. The y-axis of the plot represents the energy for each amino acid of the protein chain and it showed that maximum residues are in favorable energy environment (Figure 5B). As a result the structure modeled was considered as a good quality model for further analysis. To obtain a better refined model energy minimization was done using Swiss PDB viewer. It minimizes energy using GROMOS96. The force-field energies of the overall structure before and after minimization were $-13,816.546$ KJ/mol and $-36,224.391$ KJ/mol respectively.

**Active site analysis**

Active site is the region on the surface of an enzyme to which a specific substrate (ligand) or set of substrates (ligands) binds. The properties of the active site are determined by the sequences of amino acids and the 3D arrangement of the polypeptide chains of the enzyme. Identification of the active site of \textit{E. coli} O157:H7 Sakai strain dnaE was done by CASTp server. This server calculates the surface area and volume of the pocket of the given structure. In addition, it also shows the active site residues. The active site residues with a volume of 17,116 were shown in Figure 6. Literature searching helped us to identify the important residues within the pocket. The structure of catalytic alpha subunit of \textit{E. coli} is known to have three conserved residues of aspartate (Asp 402, Asp 404, and Asp 556 in our model, ensured by aligning with the reference structure, data not shown) and four conserved residues.
Advances and Applications in Bioinformatics and Chemistry 2015:8

Notes: (A) Result of PROCHECK verification program, showing number and percentages of residues in most favored regions (red); additional allowed regions (yellow); generously allowed regions (creamy white); and in disallowed regions (white). Based on an analysis of 118 structures of resolution of at least 2.0 angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favored regions. (B) Result of the 3D structure verification tool ANOLEA. This figure shows residues in favorable energy environment (green) and residues in unfavorable energy (red).

Abbreviation: ANOLEA, atomic non-local environment assessment.

Virtual screening, drug likeliness, and toxicity analysis

Twenty-four ligand molecules of *E. coli* K-12 obtained from BindingDB database were docked on the receptor. Molecules showing lowest binding energy were listed in Table 2. The 251D molecule is a potent inhibitor of the bacterial replicative strand DNA polymerase III alpha.
Table 2: Lowest docking energies and important residues of the binding site observed to be interactive with the ligands from BindingDB database

<table>
<thead>
<tr>
<th>No</th>
<th>Compounds from BindingDB</th>
<th>Important amino acid residues involved</th>
<th>Docking energy (Kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ZINC 5117079</td>
<td>SER365, PHE392, ARG391, ARG397, MET400, ASP402, ASP404, ARG711</td>
<td>-8.8</td>
</tr>
<tr>
<td>2</td>
<td>CID 9809878</td>
<td>SER365, ARG391, ARG711, PHE392, ASP402, ASP404, GLY364, MET400, ARG397, SER545, GLY56, VAL52, LYS30, LYS53, ALA57, TYR549</td>
<td>-8.7</td>
</tr>
<tr>
<td>3</td>
<td>ZINC 28356629</td>
<td>SER365, PHE392, ARG391, ARG397, MET400, ASP402, ASP404, ARG711, PHE757, ASN758, HIS761</td>
<td>-8.5</td>
</tr>
</tbody>
</table>

![A](image1.png)

2-[(3,4-dichlorophenyl)methylamino]-3,7-dihydropurin-6-one (ZINC: 5117079)

![B](image2.png)

N-Coeleneterazine (Drug Bank ID: DB04118)

![C](image3.png)

251D (PubChem ID: CID 9809878)

![D](image4.png)

N-(1,4-Dihydro-5H-tetrazol-5-ylidene)-9-oxo-9H-xanthene-2-sulfonamide (Drug Bank ID: DB04698)

Figure 7: Three-dimensional representation.

Notes: Three-dimensional representation of the interactive residues on the binding site of the protein when it interacts with (A) active inhibitors (ligands) respectively with CID 9809878, ZINC 5117079 and ZINC 28356629 and (B) top binding affinity molecules DB04118 and DB04698. The color indicator on the left side shows the types of interaction of particular residues.

Information on these molecules’ binding interaction was taken as reference to predict convenient inhibitors of dnaE of *E. coli* O157:H7 Sakai strand from the molecules of DrugBank (Figure 7A). The top 100 hits from DrugBank showing lower binding energies after being docked five times were filtered for Lipinski’s rule of five that reduced the compounds to 59. To avoid off target binding, the compounds having human targets were excluded. The top two hits showed stable or nearly stable binding energy and good clustering performances from the docking results (Table 3) and structures are shown in Figure 8. Among the final high affinity binding molecules, DB04118 (N-coeleneterazine) and DB04698 (N-(1,4-dihydro-5H-tetrazol-5-ylidene)-9-oxo-9H-xanthene-2-sulfonamide) were found to interact with important residues (Figure 7B) required for DNA
binding and catalysis. According to DrugBank database, both molecules are experimental drugs. DB04118 has no specific target yet. However, 3-dehydroquinate dehydratase of *Helicobacter pylori* (American Type Culture Collection strain 700392/26695), that catalyzes a trans-dehydration via an enolate intermediate, is the target for DB04698. These molecules could be proposed as potential inhibitors of the *E. coli* O157:H7 Sakai strand dnaE. For the identification of potential drug candidates human intestinal absorption is imperative.86 By using PreADMET we found human intestinal absorption was 95.36% and 88.10% for DB04118 and DB04698 respectively, indicating well-absorbed compounds (70%–100% according to PreADMET) that is desirable for drug candidates. As only the unbound drug is necessary for diffusion or transport across the cell membranes and target-drug interaction, we predicted percent drug bound in plasma protein. Plasma protein binding prediction results showed 99.28% and 100.00% plasma protein binding for DB04118 and DB04698 respectively, indicating strongly bound chemicals which are not desirable. Caco-2 cell model serves as a reliable in vitro model for the prediction of oral drug absorption. PreADMET predicted 17.45 (nm/second) and 0.36 (nm/second) Caco-2 cell permeability for DB04118 and DB04698 respectively that are considered middle and low permeability respectively. Both drugs exhibit good oral bioavailability and mostly negative result in Ames test as predicted by FAF-Drugs2 and PreADMET toxicity prediction (data not shown) and no evidence of carcinogenic activity for only DB04698 as predicted by PreADMET rodent carcinogenicity prediction. We also analyzed other ADMET properties like blood–brain barrier penetration, skin permeability, MDCK cell, and P-gp inhibition, and in all of these cases results are positive for our proposed compounds (data not shown).

**Conclusion**

The overall picture emerging from this study was the identification of broad-spectrum antibiotic target as well as prediction of potential inhibitors of DNA III alpha of deadly pathogen *E. coli* O157:H7 using extensive in silico tools. Bacterial replicative dnaE is a member of C family of polymerases87,88 that are unique in terms of sequence. Bacterial replicative DNA polymerase III share no sequence similarity and is strikingly different from canonical DNA polymerases including those of eukaryotic replicative polymerases.89 It has been reported that DNA polymerases are responsible for pathogen survival and drug resistance so they have been considered to be a drug target in a broad group of Gram-positive pathogens such as *Staphylococcus*, *Streptococcus*, *Enterococcus*, and *Mycoplasma*.90 It has also been reported that DNA polymerase III inactivation would impede survival of *Mycobacterium tuberculosis* within the host.91

Since bacterial evolution aims to acquire resistance against single or multiple antibiotics to ensure their survival in the environment, the development, not only of new conventional antibiotics but also of novel compounds and alternative strategies for the battle against bacterial infections, is becoming a topical and widely recognized need. In this study, a number of criteria such as essentiality of proteins, dissimilarity with host, conservation among pathogens,
availability in drug databases, virtual screening etc, were utilized to explore potential drug targets as well as to predict a drug that can block dnaE of deadly E.coli O157:H7 strain. This in silico strategy can be used for screening novel and alternative targets in a way to design and develop new drugs against other emerging human pathogens.

Acknowledgments
We acknowledge Professor Dr. Muhammad Zafar Iqbal for his cooperation and support throughout the research. We would like to thank the Department of Genetic Engineering and Biotechnology, and the Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet, Bangladesh for providing us research facilities.

Disclosure
The authors report no conflicts of interest in this work.

References