Identification and functional annotation of hypothetical proteins of uropathogenic *Escherichia coli* strain CFT073 towards designing antimicrobial drug targets

Harpreet Kaur\(^a\), Vikram Singh\(^b\), Manmohit Kalia\(^c\), Balvinder Mohan\(^a\) and Neelam Taneja\(^a\)

\(^a\)Department of Medical Microbiology, Postgraduate Institute of Medical Education and Research, Chandigarh, India; \(^b\)Center of Computational Biology and Bioinformatics, Central University of Himachal Pradesh, Dharamshala, India; \(^c\)Department of Biology, State University of New York, Binghampton, NY, USA

Communicated by Ramaswamy H. Sarma

**ABSTRACT**

Urinary tract infections are a serious health concern worldwide, especially in developing countries. *Escherichia coli* strain CFT073 is a highly virulent pathogenic bacterial strain. CFT073 proteome contains 4897 proteins, out of which 992 have been classified as hypothetical proteins. Identification and characterization of hypothetical proteins can aid in the selection of targets for drug design. In this study, we studied the hypothetical proteins from the UPEC strain CFT073 using various computational tools. By NCBI-CDD, 376 protein sequences showed conserved domains. Based on the functional motifs in their primary sequences, we classified these 376 hypothetical proteins into 7 functional categories. Further KEGG database was used to find the roles of these hypothetical proteins in several pathways. Protein interaction network analysis of hypothetical proteins identified 53 proteins as highly interacting metabolic proteins. Virulence factor analysis of the proteins identified 8 proteins as virulent. We conducted a non-homology search for the identified proteins of UPEC in the available human proteome. We observed that 35 proteins are non-homologous to humans and hence could be selected for drug designing targets. Qualitative characterization of the selected 35 non-homologous hypothetical proteins including essentiality analysis and evaluation of druggability by similarity search against drug bank database was performed. Out of these 35 proteins, three-dimensional structures of six proteins (NP_752562.1, NP_756345.1, NP_754893.1, NP_756600.2, NP_755264.1 and NP_752994.1) could be successfully modelled. These new annotations can help to better understand disease mechanisms at the molecular level, as well as provide new targets for drug development against the UPEC strain CFT073.

**Introduction**

Urinary tract infections (UTIs) are one of the most common infections responsible for high morbidity and economic costs both in the community as well as in the healthcare settings (Nicolle, 2005). Uncomplicated UTIs usually occur in healthy non-pregnant women, while complicated UTIs (cUTIs) can affect people of both sexes and across different ages and are often linked to structural or functional urinary tract abnormalities (Flores-Mireles et al., 2015). Uropathogenic *Escherichia coli* (UPEC) are implicated in the causation of >90% of uncomplicated and about 50% of complicated UTIs (Terlizzi et al., 2017). UPEC is becoming highly drug-resistant and poses a therapeutic challenge in low to middle-income countries where antibiotic resistance is rampant (Wang et al., 2018). Currently, the emergence of carbapenem-resistant UPEC such as New Delhi Metallo-beta-lactamase-1 (NDM-1) has gained notoriety and become a grave challenge as carbapenems are the last-resort antibiotics (Khan et al., 2017). UPEC strains have a variety of virulence and fitness factors that help them colonize the mammalian urinary tract successfully (Johnson, 1991; Kaper et al., 2004). It is important to understand the functional integrity of the UPEC genome for understanding the mechanism of pathogenesis and bacterial colonization in different stages of infection. Availability of the whole genome sequence of pathogenic microorganisms in the National Centre for Biotechnology Information (NCBI) database has revealed that one-third of the proteins are hypothetical. Hypothetical proteins possibly may play an important role in the disease progression as well as the survival of pathogen (Desler et al., 2009; Kumar et al., 2014). Using sequence and structure-based methods, the precise role of hypothetical proteins from several pathogenic organisms including *Shigella flexneri* (Sen & Verma, 2020), *Klebsiella pneumoniae* (Pranavathiyani et al., 2020) and *Shigella dysenteriae* (Rabbi et al., 2021) has already been published. Although several drug targets have been identified against UPEC by various means, newer drug targets identified from the characterization of hypothetical proteins might be useful for designing newer antimicrobials (Frisinger et al., 2021).
UPEC strain CFT073, a highly virulent UPEC strain belonging to serotype O6:K2:H1 was isolated from the blood of a woman suffering from acute pyelonephritis. It was sequenced in 2002 (Gen bank: AE014075.1) and its genome consists of a single chromosome of size 5,231,428 bp without any plasmids (D. M. Green et al., 1990; Welch et al., 2002). It is 5,90,209 bp longer than the well-studied K-12 MG1655 nonpathogenic strain. In this strain of UPEC, 5179 genes express 4897 proteins, out of which 992 proteins are still uncharacterized and come under the ‘hypothetical’ category (the protein encoded by a known open reading frame but has yet to be identified as a protein product) (Desler et al., 2009).

In this study, we used a variety of computational algorithms to infer the possible functions of hypothetical proteins from UPEC strain CFT073. We conducted an in silico analysis of these hypothetical proteins at the molecular and structural level to identify potential therapeutic targets for drug development and design.

Materials and methods

Various tools were used in the functional annotation of hypothetical proteins as shown in the flow chart (Figure 1).

Sequence retrieval

The full genome of UPEC strain CFT073 was retrieved from NCBI (http://www.ncbi.nlm.nih.gov/) with accession No. NC_004431.1. The sequences of hypothetical proteins were mined and their Uniprot accession-ids were retrieved using the Uniprot database. Proteins with specific accession-ids were considered for further analysis.

Domain identification and sequence characterization

Conserved motifs and domains in protein sequences were identified for their functional assignment using various methods. NCBI-CDD (National Centre for Biotechnology Information- Conserved Domain database) (Derbyshire et al., 2015), Interproscan (Jones et al., 2014), and Smart tools (Letunic et al., 2015) were used for the identification of functional domains of all hypothetical proteins. InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites (Quevillon et al., 2005). HAMAP (High-quality Automated and Manual Annotation of Proteins) was used to identify and annotate many protein families (Pedruzzi et al., 2015). ScanProsite tool was used to match prosite motifs of these sequences (de Castro et al., 2006; Sigrist et al., 2009). Protein families were identified by using Pfam and the superfamilies of proteins were predicted through the structural classification of proteins (SCOP) superfamily (Gough et al., 2001; Punta et al., 2012). Molecular weight and theoretical isoelectric point (pI) of protein sequences were determined by using the Compute pI/Mw tool. The EMBOSs Pep Stat method was used to measure the aliphatic and aromatic properties of protein sequences with the average number of polar and non-polar amino acids, as well as the basic and acidic nature of protein sequences (Rice et al., 2000). The GRAVY CALCULATOR was used to measure the grand average of hydropathy (GRAVY) value of protein sequences (http://www.gravy-calculator.de).

Proteins localization prediction and annotation

After obtaining the functional sequence signatures of selected hypothetical proteins, attempts were made to locate these proteins. For the prediction of particular secretome and transmembrane helices, JVirGel v.2.0 was used. It is a software for simulation and analysis of proteomics data.

It creates a virtual two-dimensional (2D) protein gel based on the migration behavior of proteins which is dependent on their theoretical molecular weights and calculated isoelectric points (Hiller et al., 2006). The membrane proteins’ topology was predicted by using TOPCONS (Tsirigos et al., 2015). TOPCONS is a web server for consensus prediction of membrane protein topology. The web interface allows for constraining parts of the sequence to a known inside/outside location. TMHMM v.2.0 (Krogh et al., 2001), HMMTOP (Tusnady & Simon, 2001), SignalP 5.0 server (Emanuelsson et al., 2007) were used to identify cellular locations of these proteins, while CELLO v.2.5 (C. Yu et al., 2004), PSORTb v.3.0 (N. Y. Yu et al., 2010), and SubLoc v.1.0 (Hua & Sun, 2001) were used to identify sub-cellular locations like cytoplasm, inner membrane and outer membrane. TMHMM is a tool used to predict the presence of transmembrane helices in proteins. The results will indicate the segments of the protein that lie inside, outside or within the membrane (Krogh et al., 2001). HMMTOP predicts membrane topology of integral membrane proteins using hidden Markov model (Tusnady & Simon, 2001). SignalP server predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram-negative prokaryotes, and eukaryotes (Emanuelsson et al., 2007). CELLO2GO (C.-S. Yu et al., 2014) was used to find the functional processes of these hypothetical proteins and their gene ontology after thoroughly exploring the sequence-based parameters. The cellular components, biological processes, and molecular functions in which these hypothetical proteins could be involved were all covered by CELLO2GO.

Pathway analysis of proteins

These hypothetical proteins were analyzed using the Kyoto Encyclopedia of Genes and Genome (KEGG) database, which was used to predict pathways regulated by these proteins (Kanehisa et al., 2014).

Interactional network analysis of hypothetical proteins

We performed protein interaction network analysis to identify the hypothetical proteins with higher protein-protein interactions by using STRING database version 10 (Szklarczyk et al., 2015). STRING v.10 (Search Tool for the Retrieval of
Interacting Genes/Proteins (http://string.embl.de/) is a database of documented and predicted protein interactions. Different prediction channels like co-evolution, fusion, gene co-occurrence, and experimental methods in the STRING database were used for the identification of metabolic interaction networks. The interaction confidence score for the hypothetical proteins was calculated using the formula described below so that the potential metabolic interactions will not be missed from the analysis (Kushwaha & Shakya, 2010).

Confidence score of the target

\[
\frac{\text{Number of interactions of target protein by used methods}}{\text{Total number of the method used}}
\]

Virulence factor analysis of hypothetical proteins

To aid in the understanding of pathogenesis mechanisms and the search for new therapeutic targets, we used VICMpred (Saha & Raghava, 2006) and VirulentPred (Garg & Gupta, 2008) to identify hypothetical proteins that may be responsible for virulence.

Nonhomology analysis by BLASTp human genome

As humans are the hosts of E. coli, hypothetical proteins present in the pathways, proteins with higher protein-proteins interactions, and proteins involved in virulence were compared with the proteome of Homo sapiens (human; taxid 9606) available in the NCBI database using the BLASTp (Basic Local Alignment Search Tool for proteins) (Altschul et al., 1997). The proteins having BLAST hits with expectation value (E-value) less than 0.001 were considered to be homologous to the human proteome and the remaining proteins were believed to be specific to UPEC.

Essentiality analysis

Essentiality analysis for the selected non-homologous hypothetical proteins was performed using BLASTp against a database of essential genes (Luo et al., 2014).

Druggability analysis

The druggability of the non-homologous hypothetical proteins was assessed by examining their efficacy in binding to
the probable drug candidates. In the present study, a homology search was performed for each of the non-homologous hypothetical proteins against the Drug Bank 5.0 target collection database (Law et al., 2014). Drug Bank is a database that contains 229 FDA-approved biotech (peptide/protein) drugs, 8250 drug entries, including 2016 FDA-approved small drugs and more than 6000 experimental drugs. If the predicted hypothetical protein matches any target with similar biological functions from the list of the non-homologous hypothetical proteins from the Drug Bank database, it complements its druggable property, and if there is no match, then that target is regarded as a ‘Novel target’ (Crowther et al., 2010).

Structural modeling of functionally important hypothetical proteins

For structural analysis of the proteins, we used SWISS PDB Viewer (http://swissmodel.expasy.org/interactive). Since choosing the best-fit model is critical for avoiding structure prediction errors, this structure modeling process included six steps: template search, template choice, model building, model quality evaluation, ligand modeling, and oligomeric state conversion. The sequence similarity between the target and approximated template sequences influenced the model quality. The SAVES software was used to check the developed model for stereochemical efficiency, residue parameters, non-bonded associations, model stability, and atom macromolecular volume (Colovos & Yeates, 1993). PROCHECK was used to analyze the Ramachandran plot to distinguish residues in the most favored regions (Laskowski et al., 1996).

Results and discussion

Selected sequences and their analysis

Using UPEC strain CFT073 from the NCBI Genome database, a total of 570 proteins were found as hypothetical proteins containing specific Uniprot accessions-ids. The CDD database at NCBI is a resource for protein domains with specific hits having pre-calculated position-specific scoring matrices to investigate conserved motifs and domains. It identified 376 proteins to have conserved domains (Supplementary File 1). The SMART tool, which contains data of fully sequenced genomes, identified functional domains in 374 protein sequences. The Interproscan tool, which uses the InterPro protein signature database to classify specific domains in different sequences, predicted 159 functional domains. Additionally, HAMAP which is a proteome database of microbial genomes and ScanProsite that matches proteins against PROSITE selection returned 45 and 144 functional regions respectively. The Pfam database, which contains a large number of protein families, each represented by multiple sequence alignments and Hidden Markov Models (HMM) was used to find families of conserved domains for specific functions. It identified 229 proteins to have specific families. Further SCOP-super family also based on HMM identified the superfamily of 352 proteins. Using Compute pI/Mw to calculate physicochemical characteristics, 173 proteins were found to be basic (\(pI \geq 7\)), whereas 203 proteins were found to be acidic (\(pI < 7\)) (Supplementary File 2). The grand average of hydropathy (GRAVY) was calculated by dividing the total sum of hydropathy values for all amino acids by the total length of the protein and results (Supplementary File 3). Results of EMBOSS Pep Stat tool which determined aliphatic and aromatic properties, basic/acidic nature, and an average number of polar and non-polar amino acids contained in these proteins (Supplementary File 4).

Cellular and subcellular localization of hypothetical proteins

Cellular and subcellular location is crucial in determining the potential functional role of the proteins. This also determines their interaction with other proteins and their potential to be chosen as a significant target for drug designing (Mohan & Venugopal, 2012). A total of 51 proteins having transmembrane helices and 28 proteins with secretome were predicted by JVirGel while TOPCONS predicted 94 proteins with transmembrane helices. The TMHMM algorithm, which is based on a Hidden Markov Model that differentiates between soluble and membrane proteins with adequate specificity and sensitivity, was used to predict transmembrane proteins. TMHMM identified 75 hypothetical proteins to have transmembrane helices whereas HMMTOP identified 123 proteins having segments present in the transmembrane. SignalP which predicts the signal peptide cleavage site’s location based on Artificial Neural Network (ANN) found 76 sequences with transmembrane residues. By combining results from all the above tools, we concluded that 83 out of 376 proteins had residues in the cell membrane.

For sub-cellular location other tools like SubLoc v1.0, Cello and PSORTb were used. SubLoc v1.0 predicts the localization of proteins at the subcellular level in gram-negative bacteria. By using SubLoc v1.0, 269 sequences were found to be in the cytoplasm, 35 proteins were extracellular and 72 were predicted to be present in the periplasm. Cello uses a multi-class SVM method to predict the localization of protein domains based on the physicochemical properties of proteins. The algorithm used in Cello is separately trained for the class of gram-negative bacteria and therefore provides more reliable results. Cello predicted localization of 376 proteins into 449 domains: 211 domains in the cytoplasm, 72 periplasmic, 55 innermembronic, 23 extracellular, and 16 outermembronic. PSORTb is another method for predicting the position of proteins in gram-positive, gram-negative, and archaeal sequences. In our dataset, it returned 103 proteins as cytoplasmic, 64 cytoplasmic-membranes, 3 as outer-membrane, 7 as extracellular, and 7 as periplasmic proteins. By comparing the results from all the three above tools, and taking a consensus, we concluded that 216 proteins were cytoplasmic, 8 were extracellular, 27 were periplasmic and no location could be predicted for 125 proteins (Supplementary File 5). Furthermore, functional annotation predicted that these proteins may have 27 molecular functions (Figure 2). The largest cluster of proteins was predicted to have a role in ion binding followed by hydrolase and oxidoreductase activity, DNA and protein binding, and isomerase activity. At the biological level, 40 processes were
identified (Figure 3), where most of the proteins were observed to be involved mainly in the biosynthetic processes, response to stress, various metabolic processes, nitrogen compound metabolic process, carbohydrates metabolic process, and transport process. Based on cellular components, 10 functional categories were identified (Figure 4).

Classification of hypothetical proteins based on conserved sequences and protein families

It was observed that 376 hypothetical proteins belonged to various families like enzymes, transporters, outer membrane proteins, proteins involved in pathogenesis, stress response, binding, and various metabolic processes (Supplementary File 1). Seventy hypothetical proteins were identified to be involved in binding activities like nucleic acid-binding, ion binding, protein binding, cytoskeleton protein binding, etc. In nucleic acid binding, several proteins were predicted to have roles during DNA replication and repair after DNA damage. These included ribosomal proteins, transcription factors, and post-translational modifiers of mRNA involved in stress conditions (Hantke, 2001). The outer membrane (OM) proteins protect the gram-negative bacteria in a harsh environment. We found 7 outer membrane proteins exclusive to E. coli functioning like porins. OM porins are transmembrane pore-forming proteins with a β-barrel structure, which form a water-filled open channel and allow the passive transport of hydrophilic compounds (Choi & Lee, 2019). The porins are the most abundant proteins of the OM in gram-negative bacteria, are of various types, and seem to play an important role in maintaining the envelope integrity of the gram-negative bacteria. Because porins mediate the passive diffusion of antibiotics across the OM, they are closely associated with antibiotic resistance in gram-negative bacteria. For example, β-lactams and fluoroquinolones are known to penetrate the OM through the non-specific porin OmpF (Delcour, 2009). The proteins NP_757261.1, NP_752506.1, and NP_754615.1 were identified as OmpF which acts as a receptor for T-even like phages. The OmpA is a non-specific porin that allows the passive transport of many small chemicals. It is also a peptidoglycan-associated protein with a flexible periplasmic domain that is involved in the non-covalent interaction with peptidoglycan (Choi & Lee, 2019). The protein NP_755264.1 was predicted as a mechanosensitive ion channel that allows efflux of solvent and solutes in the cytoplasm hence making its role significant in the survival of pathogens (Pao et al., 1998). The proteins NP_755375.1 and NP_755277.1 contain major OM facilitator superfamily domains and are representative of a class of membrane transporters that are involved in the transportation of sugars, amino acids, drugs, various metabolites, and varieties of ions (Pao et al., 1998). The protein NP_753820.1 belonged to the porin superfamily. It forms trimeric channels that allow export of a variety of substrates in Gram negative bacteria.

A total of 65 hypothetical proteins were predicted to behave as enzymes like transferases, GTPases, ATP synthases, ligases, kinases, peptidases, nucleases, helicases, and many others. Enzymes regulate the internal and external environment for the survival and virulence of bacteria by providing essential nutrients for their growth (Beacham, 1979; Humann & Lenz, 2009). Hence, all of these enzymes are important for bacterial survival and are potential targets for drug designing. Forty-six proteins were found to be involved in metabolic processes like biosynthetic process, catabolic process, lipid metabolic process, DNA metabolic process, secondary metabolic process, carbohydrate metabolic process, and cellular amino acid metabolic process. The study of bacterial metabolism focuses on the chemical diversity of substrate oxidations and dissimilation reactions (reactions by which substrate molecules are broken down), which normally function in bacteria to generate energy. Biological oxidation of organic compounds by bacteria results in the synthesis of ATP as the chemical energy source. This process also permits the generation of simpler organic compounds (precursor molecules) needed by the bacteria cell for biosynthetic or assimilatory reactions (Amato et al., 2014). Proteins involved in metabolic processes may be important for bacterial pathogenesis and can be treated as possible drug targets. Transporter proteins are involved in the transportation of nutrients, that are helpful in various metabolic processes, and hence the survival of the organism. These proteins also facilitate the transfer of virulence factors and are directly involved in infection (E. R. Green & Mecsas, 2016). Nine proteins were found to be transporters that influx/efflux the essential ions and nutrients. In the transporter system, SNAREs (soluble N-ethylmaleimide-sensitive factor attachment receptor) have an essential role in membrane fusion to stabilize the host intracellular mechanism (Ahn, 1998). Membrane fusion is a fundamental biological process for organelle formation, nutrient uptake, and the secretion of molecules. It is central to all aspects of immune function, including the secretion of immune mediators and the ingestion and destruction of pathogens (Wesolowski & Paument, 2008). Two proteins NP_752687.1 and NP_757257.1 were predicted to be ABC transporters. The protein NP_752687.1 was a V-type ATP synthase (subunit C) which may be involved in ATP synthesis and hence may be involved in providing energy for various metabolic processes of UPEC (Rappas et al., 2005). Similarly, protein NP_757257.1 was identified as an anion permease ArsB/NhaD which may translocate sodium, arsenate, antimonite, sulfate, and organic anions across biological membranes in all kingdoms of life.

Eight proteins were found to be involved in pathogenesis like a response to stress and signaling pathways. Cellular processes such as enzymatic activities, membrane fusion and transportation of ions, and many other processes depend upon signaling mechanisms. Signaling pathway proteins are of great importance and can be exploited for the novel therapeutic agent’s formulation against the disease caused by UPEC.

Pathway analysis of functionally characterized proteins

The regulatory function of hypothetical proteins in various pathways was predicted by using the KEGG database. Pathway analysis provided us additional features for the identification of key roles of hypothetical proteins. We
identified three key metabolic pathways i.e. purine, glycerophospholipid, and carbon metabolism in the UPEC strain CFT073 where these hypothetical proteins might be playing crucial roles. One protein each (NP_752562.1, NP_7537336.1, and NP_756345.1) was predicted to be involved in purine, glycerophospholipid metabolism, and carbon metabolism respectively. Adenine and guanine are purines present in nucleotides and nucleic acids and their final oxidation product in humans is uric acid. Purine binds to sugar residues through their 9-nitrogen position on the nucleoside (Samant et al., 2008) and a nucleotide is formed when the sugar residue is phosphorylated. Carbon metabolism plays an important role in the growth of bacteria. Glycerophospholipid and glycerolipid pathways contain major components of the bacterial cell wall, such as phosphatidylglycerol (PG), phosphatidylethanolamine (PE), and cardiolipin (CDL), as well as the precursor for fatty acid biosynthesis, 1, 2Diacyl-sn-Glycerol (Leithner et al., 2018).

**Protein-protein interactions analysis**

Proteins with highly interacting functional associations were analyzed by the STRING database and 53 proteins were found as highly interacting proteins with a confidence level greater than 1 (Supplementary File 6). Uncovering protein-protein interaction information helps in the identification of drug targets (Kaur et al., 2021). Studies have shown that proteins with a more significant number of interactions (hubs) include families of enzymes, transcription factors. For a more accurate understanding of their importance in the cell, one has to identify various interactions and determine the resultant effect of the interactions. Biological networks are now considered as a starting point of many studies for understanding and curing many human diseases (Kushwaha & Shakya, 2010).

**Virulence factor analysis of hypothetical proteins**

Gram-negative pathogens frequently evolve to improve their virulence in the host environment by modifying features such as motility, cell adhesion, and the ability to deal with the host’s immune response. Based on the consensus sequence review of VICMpred and VirulentPred, 8 hypothetical proteins were classified as virulence factors (Supplementary File 7). Targeting virulence factors has long been thought to be a better therapeutic intervention against bacterial pathogenesis and they can be used as an alternative therapy to antibiotics or as potentiators of the host immune response (Fleitas Martinez et al., 2019). The recent announcement of preclinical proof of concept for anti-virulence molecules like parthenolide could enable the anti-virulence concept to become a reality as a new antibacterial strategy (Cegelski et al., 2008; Kalia et al., 2018).

**Figure 2.** Bar chart depicting the molecular functions of hypothetical proteins.
We conducted a non-homology search for the identified proteins of UPEC present in pathways (3 proteins), proteins with higher protein-protein interactions (53 proteins), and virulence proteins (8 proteins) in the available human proteome for predicting drug development targets. We excluded human homologs as drug targets to avoid potential adverse reactions and cytotoxicity in humans (Sarkar et al., 2012). Using BLASTp against human proteome (human; taxid: 9606) available at NCBI database, 29 proteins were found to be homologous with the human host (E-value < 0.001) (Kaur et al., 2021). The remaining 35 proteins (2 pathways proteins, 6 virulent proteins, and 27 proteins with higher protein-protein interactions) present exclusively in uropathogenic E. coli were identified as a potential target for drug discovery (Supplementary Table 1).

**Essentiality analysis of selected non-homologous hypothetical proteins**

All the non-homologous hypothetical proteins were found as essential for the survival of pathogens by similarity search in DEG, a database of essential genes (Supplementary Table 2).

**Druggability analysis of selected non-homologous hypothetical proteins**

Potential non-homologous hypothetical proteins as drug targets identified in the present study were further analyzed for their druggability by similarity search against targets present in the Drug Bank database. A manual search for the presence of the thirty-five non-homologous hypothetical proteins in the DrugBank revealed that five have similarities with the Drug Bank database (E-value 0.00001) and therefore were designated druggable candidates. The remaining 30 non-homologous hypothetical proteins have no homologs in the Drug Bank database (Supplementary Table 3), and therefore, they were regarded as novel drug targets.

**Structural modeling of functionally important hypothetical proteins**

Identification and characterization of hypothetical proteins at the structural level can be helpful for the selection of potential targets for drug designing (Teh et al., 2014). Out of 35 non-homologous proteins, tertiary structures of only six proteins NP_752562.1, NP_756345.1, NP_754893.1, NP_756600.2, NP_755264.1 and 752994.1 could be successfully modeled by
using SWISS PDB viewer. It was done by selecting suitable templates. After models were built, these were validated for no error rate in the predicted structure. The SAVES software was used to check the developed model for residue parameters, stereochemical consistency, model compatibility, non-bonded interactions, model compatibility, and atom macromolecular volume. PROCHECK was used for checking the number of residues in the most favored regions of the Ramachandran plot.

The protein NP_752562.1 [EC: 3.5.3.26] belongs to cupin like superfamily. It has an important role in purine metabolism (KEGG ID: K14977). Purines act as metabolic signals, control cell growth, provide energy, are part of essential coenzymes, contribute to sugar transport and donate phosphate groups in phosphorylation reactions (Fumagalli et al., 2017). Most bacteria can produce nucleotides de novo, while others, including some lactic acid bacteria, require the addition of either purines or pyrimidines to the growth medium (Kilstorp et al., 2005). These auxotrophic bacteria utilize salvage pathways for the conversion of the required nucleobases or nucleosides to nucleotides. Many bacteria can utilize nucleotides as sources of purines or pyrimidines, but these have to be dephosphorylated by extracellular nucleotidases before entering the cell (Xi et al., 2000). For structural analysis, its selected template was 1rc6.1.A.1, and the predicted structure (Supplementary Figure 1). During validation by Procheck, 71.9% of residues were found to be in the most preferred region.

The protein NP_756345.1 belongs to the superfamily of triosephosphate isomerase (TIM) phosphate binding. It has an important role in carbon metabolism involved in the growth of bacteria. Maintaining proper intracellular carbon levels is crucial in cell physiology to maximize nutrient utilization and cell growth (Kawai et al., 2019; Tang et al., 2011). For structural analysis, its selected template was 1rvg.1.B, and the predicted structure (Supplementary Figure 2). During validation by Procheck, 89.3% of residues were found to be in the most preferred region.

The protein NP_754893.1 (virulent protein) was identified as a zinc metalloprotease belonging to the peptidase M48 superfamily. Zinc metallopeptidases of bacterial pathogens are widely distributed virulence factors and represent promising pharmacological targets (Vemula et al., 2016). Metalloproteases, in which zinc metal ion is essential for catalytic activity, are produced by various human pathogenic bacteria (Miyoshi & Shinoda, 1997). Vemula et al. (2016) predicted the immunomodulatory function of a secreted M. tuberculosis protein, Zinc metalloprotease-1 (Zmp1) that alters phagosome maturation and is considered essential for intracellular survival of M. tuberculosis (Vemula et al., 2016). A wide variety of pathological actions of bacterial metalloproteases have been documented, which highlights its importance as a unique target against UPEC for the discovery of inhibitors. For structural analysis, its selected template was 3c37.1.A and the predicted structure (Supplementary Figure 3). During validation by Procheck, 90.2% of residues were found to be in the most preferred region.

The protein NP_756600.2 was identified as a protein with the highest protein-protein interactions with a confidence score of 6.142. For structural analysis, its selected template was 3e8p.1.c, and the predicted structure (Supplementary Figure 4). During validation by Procheck, 88.4% of residues were found to be in the most preferred region. This protein belongs to the superfamily of hotdog superfamily. The hotdog fold was first discovered in the structure of E. coli FabA (beta-hydroxy-decanoyl-acyl carrier protein (ACP)-dehydratase) and then in Pseudomonas 4HBT (4-hydroxybenzoyl-CoA thioesterase). Using sequence analysis, a large superfamily of hotdog domains has been identified. It includes numerous archaeal, eukaryotic, and prokaryotic proteins involved in several roles ranging from catalytic actions, metabolic roles such as thioester hydrolysis in fatty acid metabolism to degradation of phenylacetic acid and the environmental pollutant 4-chlorobenzoate. This superfamily also includes FapR, a non-catalytic bacterial homolog that is involved in the transcriptional regulation of fatty acid biosynthesis (Dillon & Bateman, 2004).

Another protein (NP_755264.1) identified as a protein with high protein-protein interactions (confidence score of 4.142) belongs to the OmpA_C-like superfamily. OmpA is a key virulence factor that facilitates the maturation of
intracellular bacterial communities (IBC) and chronic bacterial persistence in UPEC (Nicholson et al., 2009; Selvaraj et al., 2007). For structural analysis, its selected template was 2n48.1A, and the predicted structure (Supplementary Figure 5). During validation by Procheck, 91.1% of residues were found to be in the most preferred region.

The protein NP_752994.1 was also identified as a protein with the highest protein-protein interactions with a confidence score of 4. This protein is a Metallo-beta-lactamase protein fold contained in the class B beta-lactamases. These proteins include thioesterases, members of the glyoxalase II family, that catalyze the hydrolysis of S-D-lactoyl-glutathione to form glutathione and D-lactic acid and a competence protein that is essential for natural transformation and could be a transporter involved in DNA uptake. Except for the competence protein, these proteins bind two zinc ions per molecule as a cofactor. Metallo-beta-lactamases are important enzymes because they are involved in the breakdown of antibiotics by antibiotic-resistant bacteria. (Carfi et al., 1995).

For structural analysis, its selected template was 2xf4.1.A and the predicted structure (Supplementary Figure 6). During validation by Procheck, 92.4% of residues were found to be in the most preferred region.

Summary and future perspectives

Important information about a pathogen may be missed if the data annotated as ‘hypothetical proteins’ is ignored. Despite the massive amount of genomic data available, more than a third of genes have no known function. Understanding an organism’s entire cell cycle is critical, especially in terms of hypothetical proteins that could be essential players in metabolic pathways, having higher protein-protein interactions, have pathogenic qualities, and therefore could be used as therapeutic targets.

Using a variety of computational tools, the current study focused on the functional annotation of hypothetical proteins from the UPEC strain CFT073 which is a highly virulent strain causing UTIs. To the best of our understanding, this is the first comprehensive study against a highly virulent uropathogenic Escherichia coli strain CFT073, though functional annotation of hypothetical proteins has been done for several strains of organisms including E. coli, Helicobacter pylori and Staphylococcus aureus etc. We characterized 376 hypothetical proteins and classified their physicochemical properties as well as known domains and families. Many of the hypothetical proteins have molecular functions including DNA binding and isomerase activity, as well as are involved in biological processes like cell adhesion, stress response, biosynthesis, metabolism, and catabolism. Identification and characterization of hypothetical proteins, as well as structural level analysis, can aid in the selection of targets for drug design. The current study identified 35 hypothetical proteins that are non-homologous to the human host and can be targeted as a drug target for the discovery of newer antimicrobial compounds. Further, all selected thirty-five non-homologous hypothetical proteins were found in the DEG database and hence were considered to be essential for the survival of UPEC. The druggability was assessed by a similarity search against the DrugBank database and revealed that five non-homologous hypothetical proteins have similarities in the Drug Bank database and remaining 30 have no homologs were absent in the DrugBank and therefore regarded as novel drug targets. Out of these 35 non-homologous hypothetical proteins, three-dimensional structures of six proteins [(NP_752562.1 (involved in purine metabolism), NP_756345.1 (involved in carbon metabolism), NP_754893.1 (virulence protein, peptidase M48 superfamily), NP_756600.2 (superfamily of hotdogps), NP_755264.1 (OmpA-C like superfamily) and NP_752994 (Metallo-beta-lactamase superfamily)] could be successfully modeled. Our work aids in the rapid detection of hypothetical protein’s unknown function, which could be used as therapeutic targets. Further research for new inhibitors can be performed against these hypothetical proteins.

Acknowledgements

Authors are highly thankful to the Department of Science and Technology (DST) and Council for Scientific and Industrial Research (CSIR) for providing financial support (DST/INSPIRE/Fellowship/2015/IF150340 and 09/141 (0205)/2019-EMR-1 respectively).

Disclosure statement

The authors declare that no conflict of interest exists among them.

Funding

The author(s) reported there is no funding associated with the work featured in this article.

References

Ahn, R. E. J. (1998). Conserved structural features of the synaptic fusion complex: SNARE proteins reclassified as Q- and R-SNAREs. *Proceedings of the National Academy of Sciences of the United States of America*, 95(26), 15781–15786.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D., Arciszewska, L., Drake, D., Craig, N., Chauvatcharin, N., Atchartpongkul, S., Utamapongchai, S., … Craig, N. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 3389–3402. https://doi.org/10.1111/j.1574-6968.2009.01707.x

Amato, S. M., Fazen, C. H., Henry, T. C., Mok, W. W. K., Orman, M. A., Sandvik, E. L., Volzing, K. G., & Brynildsen, M. P. (2014). The role of metabolism in bacterial persistence. *Frontiers in Microbiology*, 5(MAR), 70–79. https://doi.org/10.3389/fmicb.2014.00070

Beacham, I. R. (1979). Periplasmic enzymes in gram-negative bacteria. *International Journal of Biochemistry*, 10(11), 877–883. https://doi.org/10.1016/0020-711X(79)90117-4

Cegelski, L., Marshall, G. R., Eldridge, G. R., & Hultgren, S. J. (2008). The biology and future prospects of antivirulence therapies. *Nature Reviews. Microbiology*, 6(1), 17–27. https://doi.org/10.1038/nrmicro2244

Choi, U., & Lee, C. R. (2019). Distinct Roles of Outer Membrane Porins in Antibiotic Resistance and Membrane Integrity in Escherichia coli. *Frontiers in Microbiology*, 10(APR), 953–959. https://doi.org/10.3389/fmicb.2019.00953
Colovos, C., & Yeates, T. O. (1993). Verification of protein structures: Patterns of nonbonded atomic interactions. *Protein Science: A Publication of the Protein Society*, 2(9), 1511–1519. https://doi.org/10.1002/pro.5560020916

Crowther, G. J., Shamugam, D., Carmona, S. J., Doyle, M. A., Hertz-Fowler, C., Berminan, M., Nwaka, S., Ralph, S. A., Roos, D. S., Van Voorhis, W. C., & Agüero, F. (2010). Identification of attractive drug targets in neglected-disease pathogens using an in silico approach. *PLoS Neglected Tropical Diseases*, 4(8), e804. https://doi.org/10.1371/journal.pntd.0000804

de Castro, E., Sigrist, C. J. a., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P. S., Gasteiger, E., Bairoch, A., & Hulo, N. (2006). ScanProsite: Detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Research*, 34(Web Server), W362–365. https://doi.org/10.1093/nar/gkl124

Delcour, A. H. (2009). Outer membrane permeability and antibiotic resistance. *Biochim Biophys Acta*, 1794(5), 808–816. https://doi.org/10.1016/j.bbabap.2008.11.005

Derbyshire, M. K., Gonzales, N. R., Lu, S., He, J., Marchler, G. H., Wang, Z., & Marchler-Bauer, A. (2015). Improving the consistency of domain annotation within the Conserved Domain Database. *Database*, 2015(7), 1–8. https://doi.org/10.1093/database/bav012

Desler, C., Suravajhala, P., Sanderhoff, M., Rasmussen, M., & Rasmussen, L. J. (2009). In Silico screening for functional candidates amongst hypothetical proteins. *BMC Bioinformatics*, 10(1), 289. https://doi.org/10.1186/1471-2105-10-289

Dillon, S. C., & Bateman, A. (2004). The Hotdog fold: Wrapping up a superfamily of thioesterases and dehydrogenases. *BMC Bioinformatics*, 5, 109. https://doi.org/10.1186/1471-2105-5-109

Emanuelsson, O., Brunak, S., von Heijne, G., & Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols*, 2(4), 953–971. https://doi.org/10.1038/nprot.2007.131

Fleitas Martínez, O., Cardoso, M. H., Ribeiro, S. M., & Franco, O. L. (2019). Recent advances in anti-virulence therapeutic strategies with a focus on dismantling bacterial membrane microdomains, toxin neutralization, quorum-sensing interference and biofilm inhibition. *Frontiers in Cellular and Infection Microbiology*, 9(APR), 74. https://doi.org/10.3389/fcimb.2019.00074

Flores-Mireles, A. L., Walker, J. N., Caparon, M., & Hultgren, S. J. (2015). Urinary tract infections: Epidemiology, mechanisms of infection and treatment options. *Nature Reviews Microbiology*, 13(5), 269–284. https://doi.org/10.1038/nrmicro3432

Frisinger, F. S., Jana, B., Donadio, S., & Guardabassi, L. (2021). In Silico screening for functional candidates amongst hypothetical proteins. *BMC Bioinformatics*, 10(1), 289. https://doi.org/10.1186/1471-2105-10-289

Dillon, S. C., & Bateman, A. (2004). The Hotdog fold: Wrapping up a superfamily of thioesterases and dehydrogenases. *BMC Bioinformatics*, 5, 109. https://doi.org/10.1186/1471-2105-5-109

Kalia, M., Yadav, V. K., Singh, P. K., Sharma, D., Narvi, S. S., & Agarwal, V. (2018). Exploring the impact of parthenolide as anti-quorum sensing and anti-biofilm agent against Pseudomonas aeruginosa. *Life Sciences*, 199, 96–103. https://doi.org/10.1016/j.lfs.2018.03.013

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2014). Data, information, knowledge and principle: Back to metabolomics in KEGG. *Nucleic Acids Research*, 42(D1), D119–205. https://doi.org/10.1093/nar/gkt1076

Kaper, J. B., Nataro, J. P., & Mobley, H. L. (2004). Pathogenic Escherichia coli. *Nature Reviews Microbiology*, 2(12), 123–140. https://doi.org/10.1038/nrmicro818

Kaur, H., Kalia, M., & Taneja, N. (2021). Identification of novel non-homologous drug targets against Acinetobacter baumannii using subtractive genomics and comparative metabolic pathway analysis. *Microbial Pathogenesis*, 152(July), 104608. https://doi.org/10.1016/j.mpath.2020.104608

Kawai, Y., Mercier, R., Mickiewicz, K., Serafini, A., Sório de Carvalho, L. P., & Errington, J. (2019). Crucial role for central carbon metabolism in the bacterial L-form switch and killing by β-lactam antibiotics. *Nature Microbiology*, 4(10), 1716–1726. https://doi.org/10.1038/s41564-019-0497-3

Khan, A. U., Maryam, L., & Zarrilli, R. (2017). Structure, Genetics and Worldwide Spread of New Delhi Metallo-β-lactamase (NDM): A threat to public health. *BMC Microbiology*, 17(1), 101–112. https://doi.org/10.1186/s12866-017-1012-8

Kilstrup, M., Hammer, K., Jensen, P. S., & Martinussen, J. (2005). Nucleotide metabolism and its control in lactic acid bacteria. *FEMS Microbiology Reviews*, 29(3), 555–590. https://doi.org/10.1016/j.femsre.2004.06.006

Krogh, a., Larsson, B., von Heijne, G., & Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology*, 305(3), 567–580. https://doi.org/10.1006/jmbi.2000.4315

Kumar, K., Prakash, A., Tasleem, M., Islam, A., Ahmad, F., & Hassan, M. I. (2014). Functional annotation of putative hypothetical proteins from Candida dubliniensis. *Gene*, 543(1), 93–100. https://doi.org/10.1016/j.gene.2014.03.060

Kushwaha, S. K., & Shakhya, M. (2010). Protein interaction network analysis-Approach for potential drug target identification in Mycobacterium tuberculosis. *Journal of Theoretical Biology*, 262(2), 284–294. https://doi.org/10.1016/j.jtbi.2009.09.029

Laskowski, R. a., Rullmann, J. a., MacArthur, M. W., Kaptein, R., & Thornton, J. M. (1996). AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *Journal of Biomolecular NMR*, 8(4), 477–486. https://doi.org/10.1007/BF00228148

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., Macleavieski, A., Arndt, D., Wilson, M., Neveu, V., Tang, A., Gabriel, G., Ly, C., Adamjée, S., Dame, Z. T., Han, B., Zhou, Y., & Wishart, D. S. (2014). DrugBank 4.0: Shedding new light on drug metabolism.
salvage. *Journal of Bacteriology*, 182(19), 5332–5341. https://doi.org/10.1128/JB.182.19.5332-5341.2000

Yu, C., Lin, C., & Hwang, J.-K. (2004). Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Science*, 13(5), 1402–1406. https://doi.org/10.1110/ps.03479604

Yu, C.-S., Cheng, C.-W., Su, W.-C., Chang, K.-C., Huang, S.-W., Hwang, J.-K., & Lu, C.-H. (2014). CELLO2GO: A web server for protein subCELular LOcalization Prediction with Functional Gene Ontology annotation. *PloS One*, 9(6), e99368. https://doi.org/10.1371/journal.pone.0099368

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., Dao, P., Cenk Sahinalp, S., Ester, M., Foster, L. J., & Brinkman, F. S. L. (2010). PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokar-yotes. *Bioinformatics (Oxford, England)*, 26(13), 1608–1615. https://doi.org/10.1093/bioinformatics/btq249