Ramesh et al. Beni-Suef University Journal of Basic and Applied Sciences (2021) 10:37
https://doi.org/10.1186/s43088-021-00126-7

Beni-Suef University Journal of Basic and Applied Sciences

RESEARCH Open Access

# Comparative analysis of Rosetta stone events in *Klebsiella pneumoniae* and *Streptococcus pneumoniae* for drug target identification

Poornima Ramesh[1], Jayashree Honnebailu Nagendrappa[1] and Santosh Kumar Hulikal Shivashankara[1,2*]

## Abstract

**Background:** Drug target identification is a fast-growing field of research in many human diseases. Many strategies have been devised in the post-genomic era to identify new drug targets for infectious diseases. Analysis of protein sequences from different organisms often reveals cases of exon/ORF shuffling in a genome. This results in the fusion of proteins/domains, either in the same genome or that of some other organism, and is termed Rosetta stone sequences. They help link disparate proteins together describing local and global relationships among proteomes. The functional role of proteins is determined mainly by domain-domain interactions and leading to the corresponding signaling mechanism. Putative proteins can be identified as drug targets by re-annotating their functional role through domain-based strategies.

**Results:** This study has utilized a bioinformatics approach to identify the putative proteins that are ideal drug targets for pneumonia infection by re-annotating the proteins through position-specific iterations. The putative proteome of two pneumonia-causing pathogens was analyzed to identify protein domain abundance and versatility among them. Common domains found in both pathogens were identified, and putative proteins containing these domains were re-annotated. Among many druggable protein targets, the re-annotation of EJJ83173 (which contains the GFO_IDH_MocA domain) showed that its probable function is glucose-fructose oxidoreduction. This protein was found to have sufficient interactor proteins and homolog in both pathogens but no homolog in the host (human), indicating it as an ideal drug target. 3D modeling of the protein showed promising model parameters. The model was utilized for virtual screening which revealed several ligands with inhibitory activity. These ligands included molecules documented in traditional Chinese medicine and currently marketed drugs.

**Conclusions:** This novel strategy of drug target identification through domain-based putative protein re-annotation presents a prospect to validate the proposed drug target to confer its utility as a typical protein targeting both pneumonia-causing species studied herewith.

**Keywords:** Rosetta stones, Protein modeling, Drug target discovery, Virtual screening, Pneumonia, Protein domains

* Correspondence: sk.genesan@gmail.com
[1]Department of PG Studies & Research in Biotechnology, Kuvempu University, Jnana Sahyadri, Shankaraghatta, Shimoga District, Karnataka 577451, India
[2]Department of Biotechnology and Bioinformatics, Biosciences Complex, Kuvempu University, Shankaraghatta, Karnataka State 577451, India

## 1 Background

A protein domain is a well-defined region within a protein that performs a specific function. Thus, the duplication of a protein domain may enhance the function of the protein. The fact that numerous proteins contain duplicated domains indicates that multifarious present-day proteins have evolved from simple proteins mainly through domain duplication. Recombination as a reason for domain duplication and domain shuffling is imaginably the most important forces driving protein evolution culminating in the complex proteome. The gene duplications and domain coding-exon duplications have resulted in an increased abundance of domains in the proteome, while domain shuffling increases versatility which is the number of discrete contexts in which a domain can occur [38].

Two polypeptides in one organism are likely to interact if their homologs express as a single polypeptide is called Rosetta stone protein. Such events help link different proteins together, leading to functional interactions between linked proteins, which may be the reason for local and global relationships within the proteome. These relationships help us to understand the role of proteins within the context of their associations and facilitate the assignment of functions to uncharacterized proteins based on their linkages with proteins of known function. Every genome projects aim to annotate the proteins coded by the genome under investigation. However, when a genome sequencing project is completed and released into public domains, researchers take a second look at the original annotation of proteins to curate them using various annotation methods. This is referred to as "re-annotation" [2]. The re-annotation of putative proteins has been attempted previously and resulted in identifying novel drug targets for many infectious diseases [26].

Pneumonia is caused by many classes of microbes, and each microbial manifestation of the disease is attributed to some specific protein interaction. *Streptococcus species* and *Klebsiella species* are known to cause the highest proportion of pneumonia. Pneumonia is an inflammatory condition of the lung primarily affecting the alveoli. It affects approximately 450 million people globally (7% of the population) and results in about 4 million deaths per year [20, 27]. In this study, we have considered two such species whose pathogenesis pattern may vary, but their protein repertoire resemblance can throw light on finding a specific drug target usable in both species. We have considered KPNIH11 and SPD39 species for our study. *Streptococcus pneumoniae* is isolated in nearly 50% of cases of community-acquired pneumonia (CAP) [1, 30]. *Klebsiella pneumoniae* accounts for hospital-acquired pneumonia infections [12].

Putative proteins are a conceptually translated sequence of amino acids from open reading frames (ORFs) with no known protein/peptide evidence. Only the putative protein data was considered in this study since putative proteins may potentially have an expression in natural biological systems. Therefore, they may serve as novel potential drug targets. Domains are well-known functional modules of proteins, making them ideal candidates to study protein-specific functions rather than targeting the entire protein.

KPNIH11 and SPD39, respectively, had 18.64% (1006 out of 5397 proteins) and 5.9% (256 out of 4366 proteins) putative proteins in their total proteomes. The structure and function of putative proteins are often poorly understood due to no known evidence of translational expression. The possible reason for many putative proteins in *Klebsiella pneumoniae* as compared to *Streptococcus pneumoniae* might be because the latter is highly studied due to its well-known pathogenesis. Putative proteins are re-annotated to determine the possible function of the protein.

Drug target identification for an infectious disease like pneumonia has been attempted by many previous studies using either comparative genomics, metabolic network modeling and simulation, multi-omics approach, or subtractive genomics approach and has resulted in identifying a significant number of potential drug targets with considerable success. The present study focuses on the comparison of the Rosetta stone events followed by the domain-based re-annotation of unannotated putative protein population in the two most common pneumonia-causing pathogens culminating in potential drug target identification.

## 2 Methods

### 2.1 Data collection and protein domain repertoire analysis

The protein sequences of *Klebsiella pneumoniae* subsp. *pneumoniae KPNIH11* (KPNIH11) and *Streptococcus pneumoniae strain D39* (SPD39) were retrieved from the NCBI protein database (https://www.ncbi.nlm.nih.gov/protein/). As mentioned in the introduction section about the importance of re-annotation of putative protein datasets, the study was concentrated on putative proteins, and hence out of the total proteins, only putative proteins were selected for the study. Domain repertoire was cataloged using the National Centre for Biotechnology Information (NCBI) Conserved Domain Database (CDD) Batch search, and domain architecture was cross-verified and ascertained using the SMART database [19]. Domain list was uploaded into Venny graphical tool version 2.1.0 (https:// bioinfogp.cnb.csic.es/tools/ venny/index.html). It gave the number and names of the shared domains between the two species of

bacteria. Putative proteins containing the shared domains were separated and re-annotated using Position-Specific Iterated (PSI-BLAST), and annotated proteins were searched for homology against the human genome using Domain Enhanced Lookup Time Accelerated BLAST (DELTA-BLAST) according to Telkar et al. [36].

### 2.2 Druggability prediction, protein modeling, and virtual screening

Proteins from KPNIH11 and SPD39 that did not show any homology to human proteins were evaluated for druggability scores using the EMBL-EBI DrugEBilitytool [24]. Druggable proteins were modeled using the SWIS S-MODEL database [17]. Ramachandran plot was used to check model quality using the PROCHECK server [18]. The models thus generated were then energy minimized according to Pawan et al. [14] before using them for further analysis. The structure of inositol 2-dehydrogenase orthologs from both the genomes was superimposed at the HOMSTRAD database, according to Khazanov et al. [16], and it was observed that there exists very least amount of deviation and high degree structural alignment between them (Fig. 5). The superimposed structure was visualized using the UCSC CHIMERA software [25].

The active pocket of modeled proteins was determined using the CASTp server [37], which gives a list of amino acids lining the possible active pocket. For virtual screening, the supercomputer facility at TACC server was used, which scans for ZINC database [15] and TCM database [5] as default settings for the given protein structure and provide possible drugs interacting with the modeled protein active site [9]. These reported molecules were assessed for ADMET properties using the DATAWARRIOR software [29].

## 3 Result

### 3.1 Domain composition analysis

We found 22 domains common in the two pathogens (Table 1). During evolution, different organisms tend to gain or lose their genome and proteome homology by exchanging DNA sequences through processes like duplication and recombination. These can arise from adaptations in response to environmental changes or the immune response of the host. As a result of their rapid doubling time and large population sizes, bacteria can evolve rapidly. The domains shared between KPNIH11 and SPD39 proteomes are the evidence that these organisms have had domain sharing and shuffling process during their evolution. One hundred twenty-three and 34 proteins in KPNIH11 and SPD39, respectively, contained these common domains. Supplementary Tables 1and 2 show the accession number, name of the common domain, and their tethering pattern in putative proteins of the two species. The results clearly show that some domains such as GFO_IDH_MocA, DeoRC, HATPase_c, rve, and Transketolase_C tend to tether with the same domain each time they form a protein attributing to a specific function. For example, GFO_IDH_MocA domains tether with GFO_IDH_MocA_C domain whose function is attributed to utilizing nicotinamide adenine dinucleotide phosphate (NADP) or nicotinamide adenine dinucleotide (NAD) for glucose-fructose redox reactions [34]. Using such tethering data, each domain can be assigned with versatility and abundance scores. Versatility is the number of different tethering combinations a domain can form, and abundance is the number of times a domain is identified in a set of proteins. In the current study, we have considered all the putative proteins which contain the common domains in two organisms. The domain versatility and abundance of these common domains in the putative proteins are shown in Table 1.

**Table 1** List of shared domains with their versatility-abundance scores

| Common domains (NCBI CDD) | SPD39 | | KPNIH11 | |
|---|---|---|---|---|
| | Abundance | Versatility | Abundance | Versatility |
| AAA | 1 | 0 | 5 | 5 |
| ApbE | 1 | 0 | 1 | 0 |
| DeoRC | 1 | 0 | 8 | 1 |
| DEXDc | 1 | 1 | 1 | 1 |
| EamA | 3 | 0 | 9 | 0 |
| FocA | 1 | 0 | 1 | 0 |
| FtsX | 1 | 0 | 1 | 0 |
| GFO_IDH_MocA | 1 | 1 | 11 | 1 |
| GFO_IDH_MocA_C | 1 | 1 | 8 | 1 |
| HAMP | 2 | 3 | 4 | 6 |
| HATPase_c | 2 | 3 | 2 | 3 |
| HELICc | 2 | 1 | 1 | 1 |
| His_kinase | 2 | 3 | 1 | 2 |
| HsdM_N | 1 | 1 | 1 | 1 |
| HTH | 11 | 5 | 22 | 8 |
| LplA | 1 | 1 | 1 | 0 |
| MFS | 1 | 0 | 55 | 0 |
| PerM | 1 | 0 | 2 | 0 |
| Rve | 1 | 1 | 11 | 1 |
| TenA | 1 | 0 | 1 | 0 |
| Transketolase_C | 1 | 1 | 2 | 1 |
| ulaA | 1 | 0 | 1 | 0 |

The table shows a list of 22 domains found in putative proteins of both KPNI H11 and SPD39. Each domain has a versatility to tether with other domains and to occur multiple times in different proteins. Abundance = the number of domain occurrences in a protein set that has one or more common domains. Versatility = the number of different tethering domains

Furthermore, when proteins containing shared domains were re-annotated using PSI-BLAST, 27 and 11 proteins, respectively, in KPNIH11 and SPD39 were annotated (Supplementary tables 3 and 4). Position-Specific Iterated Basic Local Alignment Search Tool (PSI-BLAST) uses protein-protein BLAST to derive a position-specific scoring matrix (PSSM) to identify protein similarity. This matrix is used to search the database for new matches. Thus, PSI-BLAST provides a means of detecting distant relationships between proteins. Hence, domain-based re-annotation resulted in annotating a set of proteins in the putative protein data set.

### 3.2 Identification of druggable proteins

DELTA-BLAST was used for finding the homology between bacterial proteins and the human proteome. The results showed that all annotated 27 proteins in KPNI H11 had no homology with the human proteome. However, two proteins (ABJ54307 and ABJ55037) of SPD39 had homology with the human proteins, making nine proteins available for further analysis. For a protein to be used as a drug target, it has to be checked for druggability. This was carried out using the EMBL-EBI DrugEBility database. All short-listed proteins were found to be druggable in this procedure. The result confirmed that all annotated 27 proteins in KPNIH11 and nine proteins in SPD39 could be used as drug targets.

### 3.3 Comparative modeling

All the 36 druggable proteins were subjected to 3D modeling using the SWISS-MODEL server [13], in which only two proteins in each organism could be modeled resulting in four structures available for further analysis. Out of the models generated, EJJ83173 and EJJ80284 were found to be complete protein models and other proteins were truncated protein models (Fig. 1). CASTp server was used for determining the active pockets of the protein models. Binding sites and active sites of proteins are often associated with structural pockets and cavities (Fig. 2). This analysis provided the identification and measurements of accessible surface pockets and inaccessible interior cavities for proteins.

### 3.4 Protein interaction analysis for the protein EJJ83173

The protein-protein interaction of modeled protein was studied using the STRING database [33] (Fig. 3). Putative protein EJJ83173 showed interaction with eight different proteins in a biclique architecture where all proteins interact. From the interaction data, it is evident that targeting this protein may potentially affect the inositol metabolism in microbes under study. It is to be noted here that this annotation is based on the domain composition of the protein; hence, targeting the protein is like eventually targeting the GFO_IDH_MocA (oxidoreductase family, NAD-binding Rossman fold) domain, which is common in both organisms.

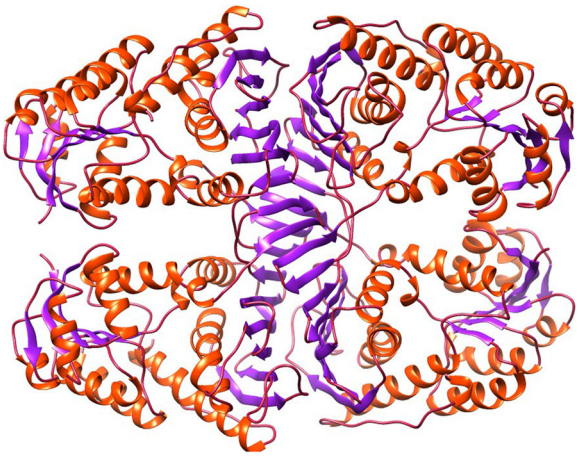

**Fig. 1** Protein model of inositol 2-dehydrogenase (EJJ83173). The model was a homotetramer protein generated using the SWISS-MODEL server

### 3.5 Model evaluation of EJJ83173

The 3D model generated by the SWISS-MODEL server showed −2.89 QMEAN value (a scoring function to estimate global and local model quality; higher numbers indicate higher reliability of the residues), 57.14% sequence identity, and 0.47 sequence similarity. The modeling was done using PDB ID: 3NT5 as a template (Fig. 1). Ramachandran plot constructed for EJJ83173 protein (Fig. 4) using PROCHECK in PDBsum database gave the following result: 1094 residues (92.3%) are found in the most favored region, 84 residues (7.1%) are found in the additionally allowed region, and only eight residues (0.7%) are found in the disallowed region which is negligible. Over 90% of residues in the favored region assures a good quality protein model.

### 3.6 Structure superimposition study

The structures of inositol 2-dehydrogenase orthologs from both the genomes were superimposed at the HOMSTRAD database, according to Khazanov et al. [16], and it was found that there exists very least amount of deviation and high degree structural alignment between them (Fig. 5). The superimposed structure was rendered in the UCSC Chimera visualizer [25]. Figure 5 depicts the inositol 2-dehydrogenase from KPNIH11 (red color) and SPD39 (green color). The orthologues have RMSD on Ca atoms = 2.837 Å units.

### 3.7 Virtual screening

The 3D model for EJJ83173 was subjected to pocket prediction (section 3.3), and it identified a probable active pocket, which was subsequently used for molecular

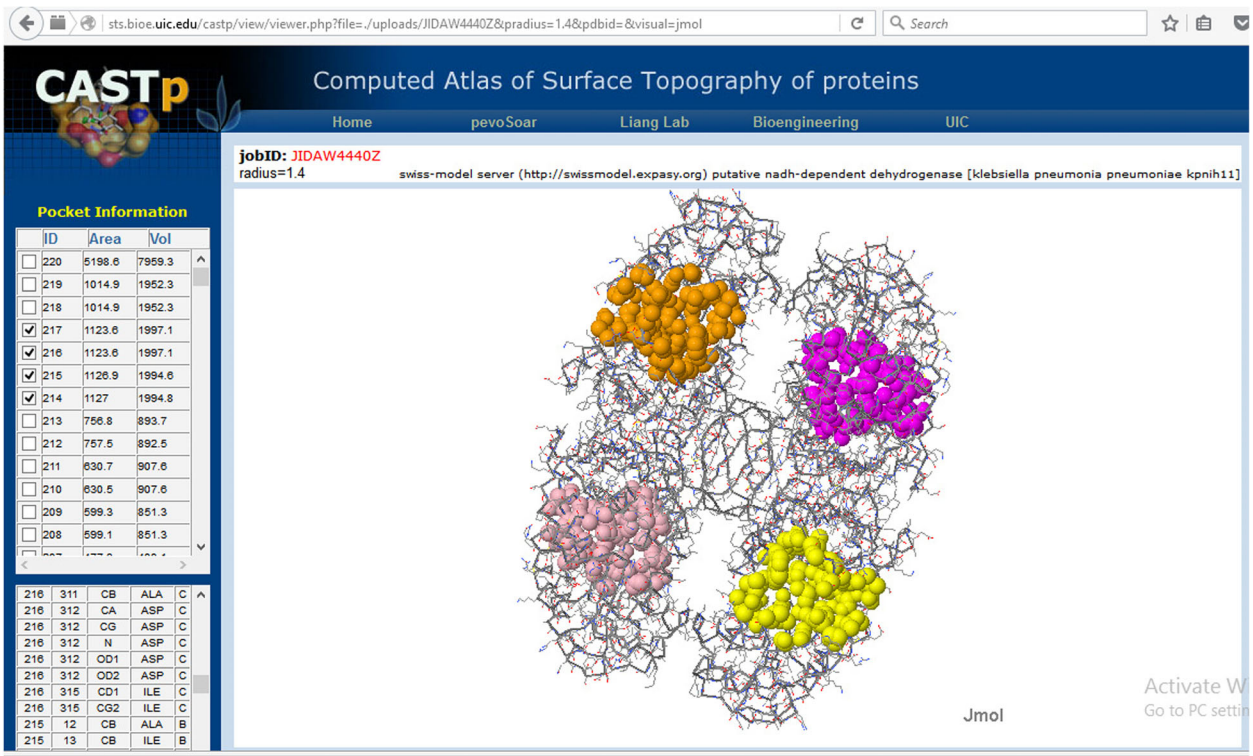


**Fig. 2** Active sites of protein EJJ83173.This protein is a homotetramer whose active sites were determined using the CASTp server. Four active sites are highlighted on 4 monomers, which show approximately the same area and volume variables

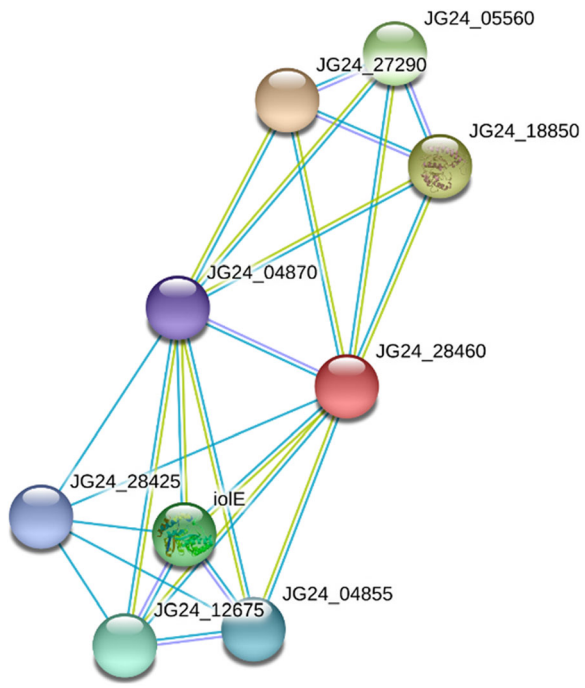


**Fig. 3** Protein interaction network of A. EJJ83173 (JG24_28460). The interactions were constructed using the STRING database. Most proteins in the interaction are involved in inositol metabolism, which supports the re-annotation of EJJ83173 as inositol 2-dehydrogenase

docking studies to find potential inhibitors. For virtual screening, grid box (Fig. 6) was set (with size x-10, y-18, z-12 and grid center x-11.739, y-30.608, z-18.306) using AutodockTools 1.5.6 software and was docked against natural ligands at drugdiscovery@TACC database.

Virtual screening was carried out against two ligand databases (ZINC database of commercially available drugs and TCM database of Traditional Chinese Medicine) to check for specific interaction with protein EJJ83173. These two databases were set as default ligand libraries in the server. Results obtained for the ZINC database and TCM database showed the binding energy, drug likeliness, tumorigenic property, mutagenic property, reproductive effect, and irritation properties of the ligands (Supplementary tables 5 and 6). The highest negative binding energy was shown by ligand with TCM ID: 45055 (−9.8) in the TCM database and by ligand with ZINC ID: ZINC68563949 (−9.5) in the ZINC database (Fig. 7). The top 15 results are tabulated in both TCM and ZINC database results, which can be used to inhibit the inositol dehydrogenase enzyme.

## 4 Discussion

In the light of many reports, which argue that the complexity of an organism is loosely linked to the number of genes, the complexity of protein and their architecture is

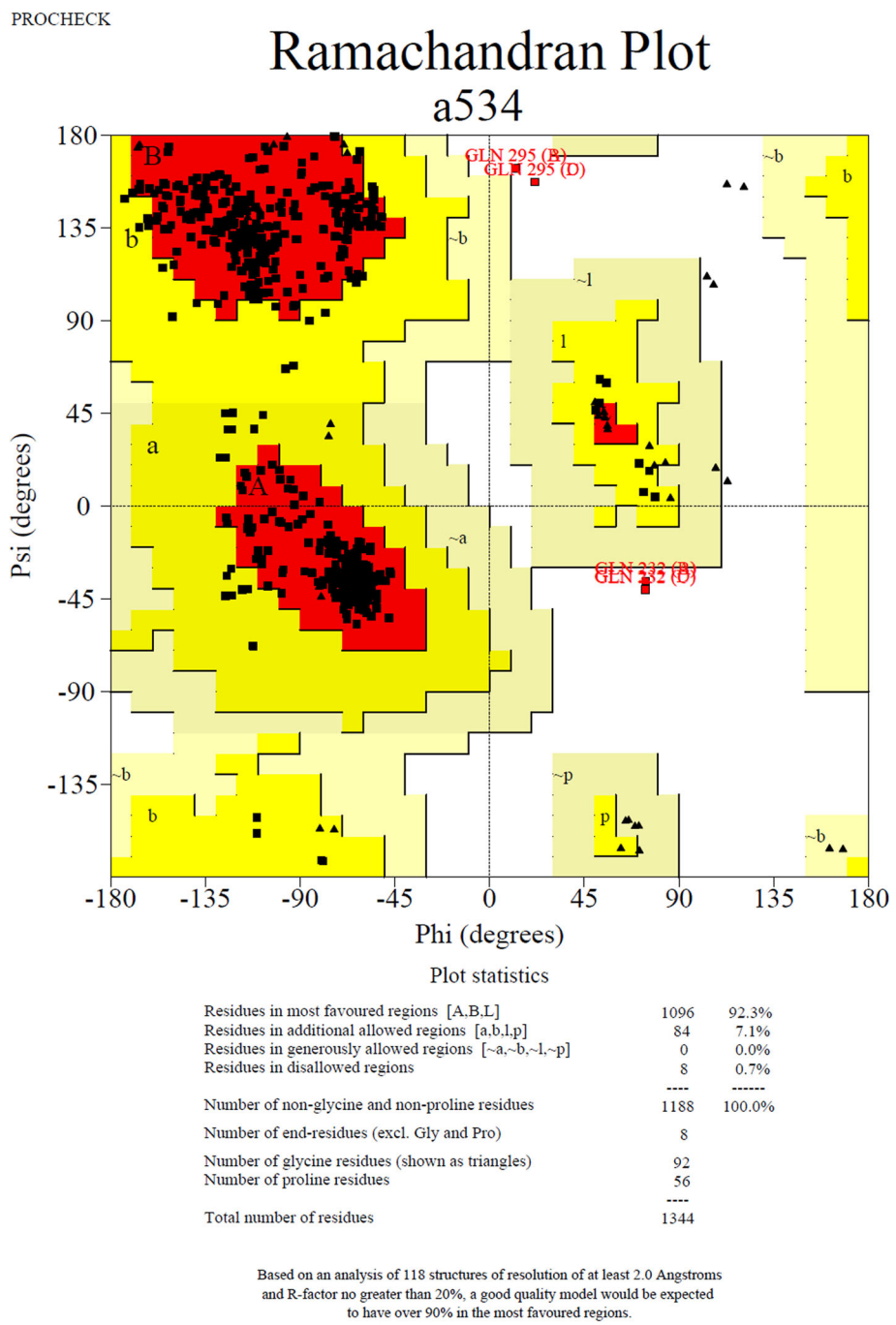


**Fig. 4** Ramachandran plot for EJJ83173 protein. The plot was constructed using PROCHECK in the PDBsum database to ascertain the protein stability and model quality. One thousand ninety-four residues (92.3%) are found in the most favored region, proving a good model quality

centered on some of the observations that flies have fewer genes than nematodes and humans have fewer than rice. Even the simplest bacterial genome is the product of extensive gene duplication and recombination [7]. Hence, the increase in protein repertoire is due to (i) duplication of domain coding sequences; (ii) divergence and modification of duplicated genes through mutations, deletions, and insertions; and (iii) gene recombination. These mechanisms are believed to be the origin of the diverse proteome [21].

Several protein-protein interactions facilitate through autonomously folding modular domains. Proteome-wide efforts to model protein-protein interaction or "interactome" networks have largely ignored this modular

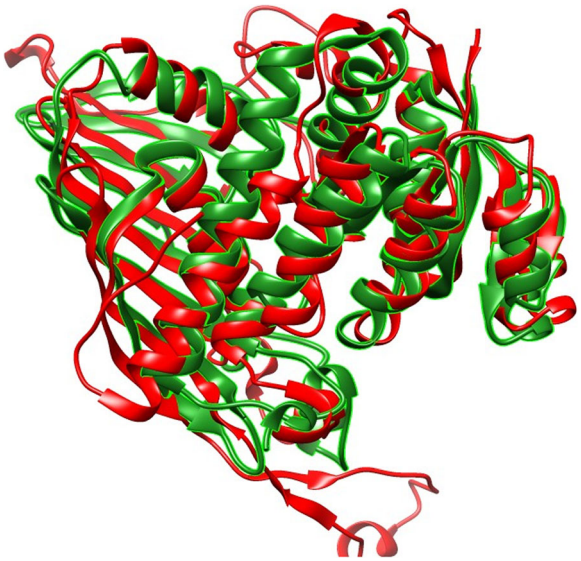

**Fig. 5** Structure superimposition of the inositol dehydrogenase. The figure represents superimposed structures of inositol 2-dehydrogenase using the HOMSTRAD database and visualized using UCSC chimera. The orthologues have RMSD on Ca atoms = 2.837 Å units

organization of proteins [3]. The protein-protein interactions are, in turn, domain-domain interactions. Hence, the complexity of domain architecture may increase the complexity of protein interaction also. This study was taken up to identify the domains and domain-based re-annotation of the putative protein datasets of the two pneumonia-causing bacteria *K. pneumoniae* and *S. pneumoniae*. Since the putative proteins are annotated based on the similarity but not experimentally validated, the re-annotation may help find the previously unreported novel drug targets [2, 4, 31].

The putative protein sequence datasets from both organisms under study were scanned for their domain composition using the CDD database. Twenty-two domains were found to be commonly present between the two putative protein datasets. The study was further explicitly focused on the putative protein dataset of only those proteins containing common domains. It helps discover the ortholog, which may have structural homology, which in turn is advantageous if they are druggable proteins where, hypothetically, one ligand may inhibit both the orthologues [8].

The protein domain repertoire shows different levels of abundance and versatility for each of the proteomes. It appears that *K. pneumoniae* has more versatility than *S. pneumoniae*. MFS domain is abundant in *K. pneumonia*,

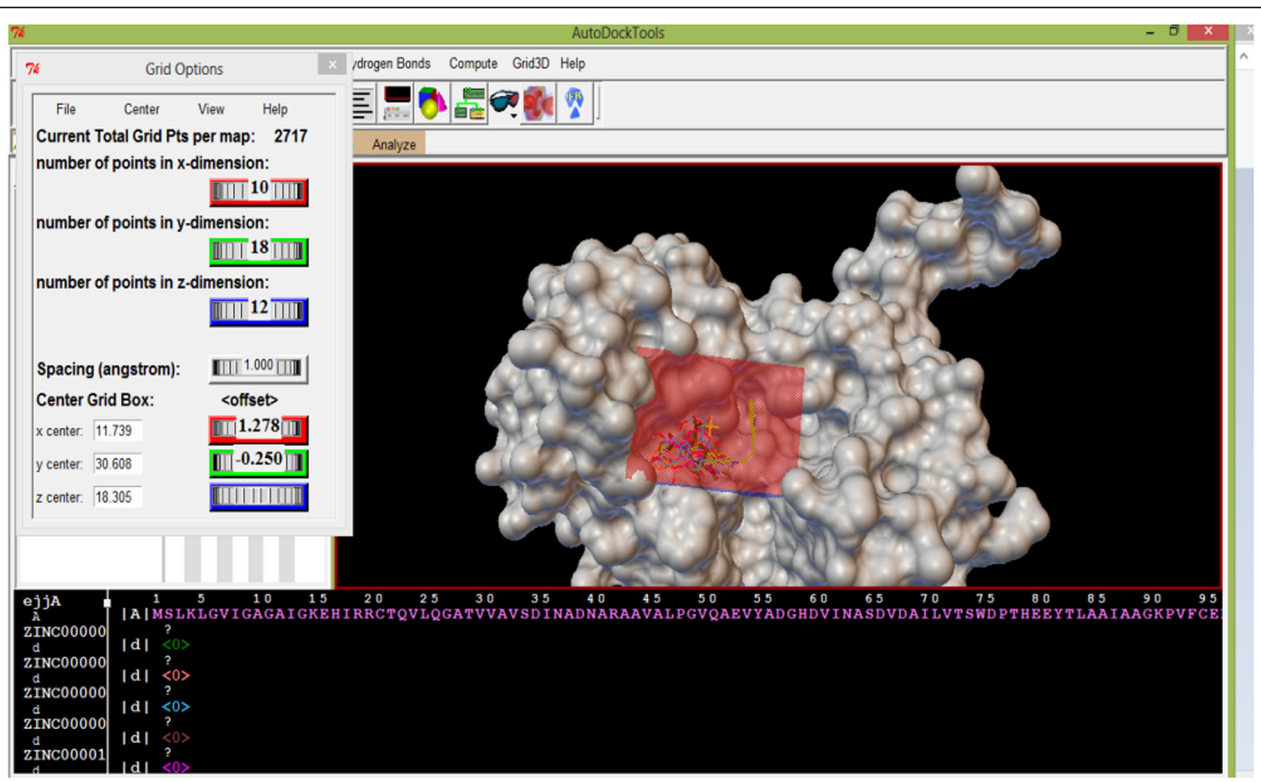


**Fig. 6** Docking grid box for the protein model of EJJ83173. The grid box of dimensions x-10, y-18, z-12 and grid center x-11.739, y-30.608, z-18.306 was set using AutodockTools 1.5.6 software. The colored projections in the center represent the active pocket

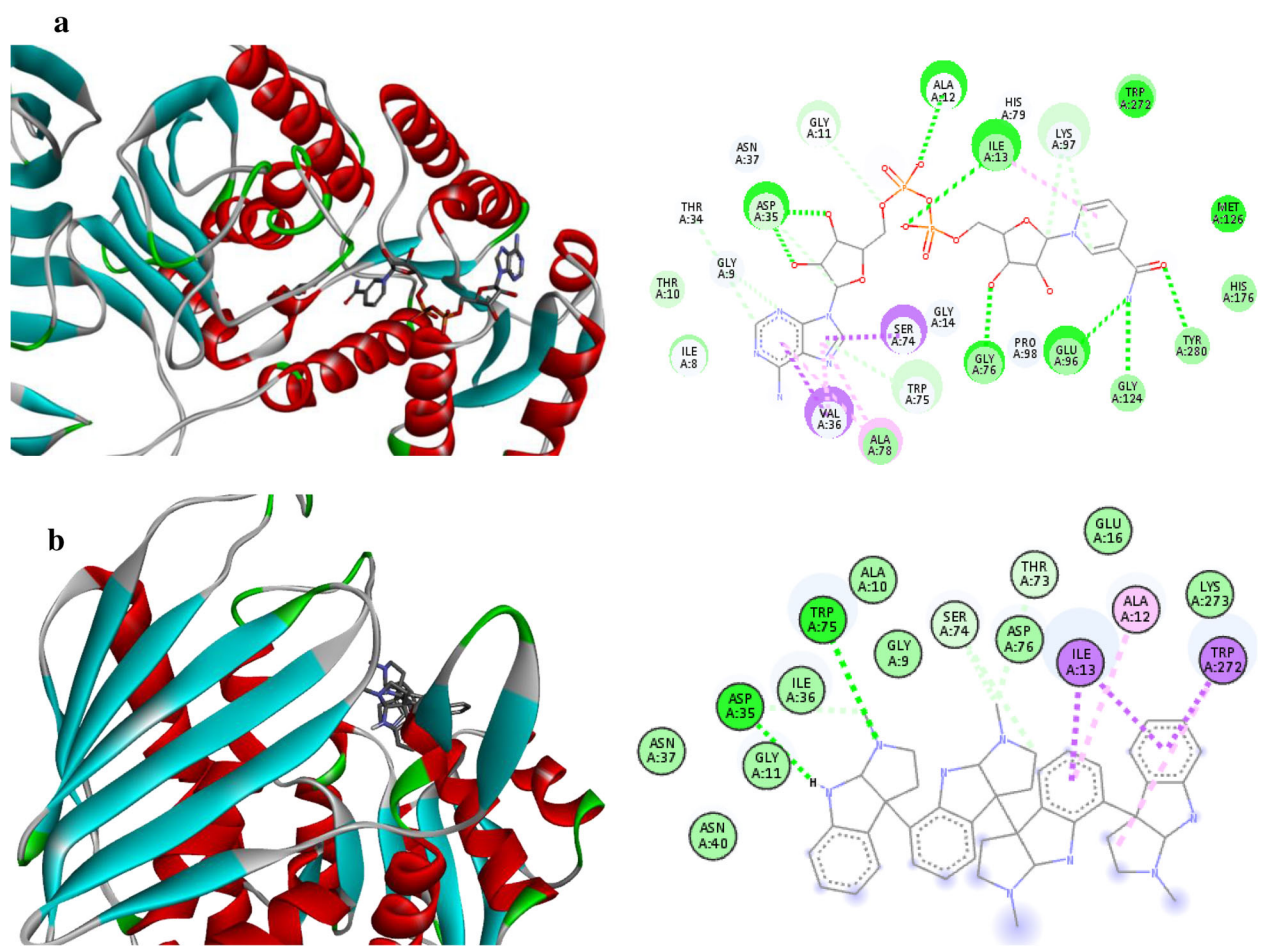


**Fig. 7** Interaction of ligands with EJJ83173 protein. **a** ZINC68563949, a commercially available drug from the ZINC database binds to EJJ83173 with a binding energy of −9.5. **b** TCM ligand 45055, a traditional Chinese medicine, binds to EJJ83173 protein with a binding energy of −9.8. Higher negative binding energy accounts for better protein-ligand interaction and inhibition

whereas the HTH domain is abundant in *S. pneumoniae* putative proteome. However, these domains have shown a different pattern of domain association in both proteomes. Interestingly, the MFS domain does not tether with any other domain in both the proteomes. Such domains can be called reserved domains. And domains like HTH domains can be called social domains since they tether with different types of domains giving it higher versatility and hence may perform a wide range of functions. However, some domain combinations like the GFO_IDH_MocA domain tether with the GFO_IDH_MocA_C domain whose function is utilizing NADP or NAD for glucose-fructose redox reactions [34]. Reserved domains contribute to the abundance but not to versatility, and hence, the biochemical versatility of the proteome may diminish. Our study was able to re-annotate the protein by using PSI-BLAST, subsequently verifying by reciprocal BLAST against the related organism. Among the re-annotated protein lists, *S. pneumoniae* comprised many functionally different proteins than *K. pneumoniae*. In the present work, the protein domain repertoire analysis leading to the identification of Rosetta stone events helped in identifying the more abundant domains and versatile protein domains. Only the proteins with these selected domains were further used for the identification of potential drug targets through DELTA-BLAST and DrugEbility analysis.

A potential drug target is a protein that does not have homology with the host genome when it comes to infectious diseases [23]. Genomes have always been investigated for non-homologous proteins while searching for potential drug targets [10]. DELTA-BLAST searches a database of pre-constructed position-specific scoring matrices before searching a protein sequence database, to yield better homology detection. For its position-specific score matrix (PSSMs), DELTA-BLAST employs a subset of NCBI's CDD. It implies that DELTA-BLAST performance is directly dependent on CDD that contains information regarding conserved domains. Hence, using DELTA-BLAST is more appropriate to determine the domain-based homology search. Therefore, the results

showed that 27 proteins in KPNIH11 and nine proteins in SPD39 had no homology with the human proteome. In principle, all these proteins could be used as drug targets because they are specific to the microbes. Theoretically, any drug administered against these proteins should not interact with human proteins to alter human physiology.

The main motto of understanding common domain fusions in two different pathogens is to finally identify an ideal druggable target, common to both organisms. In this respect, the EMBL-EBI DrugEBllity tool was used to calculate the possibility of using proteins as druggable targets. This database predicts the druggability of any given protein by comparing it to existing protein models by performing a BLAST search [39]. This result suggested that 27 proteins in KPNI H11 and nine proteins in SPD39 can be used as drug targets based on the druggability and tractability score. DrugEBIlity server has been used in many previous studies and proven a reliable calculation method for evaluating the druggability of the protein [6, 11].

For any drug design and development, the presence of a target protein 3D structure is essential. If unavailable, at least the protein should be available for the 3D modeling with a proper template. Most druggable proteins did not find templates when sequence search was performed against the PDB database except for two proteins, EJJ83173 (inositol 2-dehydrogenase with GFO_IDH_MocA domain) and EJJ80284 (antitoxin with HTH_XRE domain). Furthermore, the protein also needs to be the hub in the protein interaction network with several interacting proteins. The more the degree, the more crucial the protein will be since knocking down the protein will knock down the entire protein interaction network. The putative protein EJJ83173 has more advantage over EJJ80284 because of three reasons. Firstly, EJJ83173 was found to have more protein interacting with it compared to EJJ80284. Secondly, the biological relevance of EJJ83173 for the survival of an organism is more because it is a part of carbohydrate catabolism where EJJ80284 is predicted as an antitoxin molecule for which no significant biological relevance was found in terms of essentiality to the survival of the organism. And finally, no suitable active pockets were predicted for EJJ80284.

The putative protein EJJ83173 which is identified as inositol 2-dehydrogenase as described in the manuscript has been shown as an attractive drug target by the previous reports which are quoted in the manuscript. Furthermore, this enzyme is reported to be an integral part of myo-inositol catabolism which is the sole source of carbon for many bacteria including *Legionella pneumophila*, *Bacillus subtilis*, *Lactobacillus casei*, *Salmonella enterica*, and *Sinorhizobium meliloti* in previous studies [22]. Therefore, it was thought that this protein may serve as a good drug target since it may disrupt the carbon utilization process by the bacteria under study. EJJ83173 can serve as a promising drug target because it has a confirmed orthologue in *S. pneumoniae*, it is predicted as a druggable protein, it has no homologous domain/protein in human proteome as indicated by DELTA_BLAST, and it is having more degree of interacting proteins. The model generated for EJJ83173 using SWISS-MODEL has more than 92.3% residues in the allowed region; hence, it is considered a good model. Furthermore, the structure superimposition of orthologues of inositol 2-dehydrogenase shows only 2.8Å root mean squared deviation (RMSD), suggesting the topological and geometrical similarity between the orthologues. Hence, it can be hypothesized that a ligand that binds to EJJ83173 can also bind to its counterpart in *S. pneumoniae*.

Many servers are available for ligand screening. Among them, the drugdiscovery@TACC server is the most robust one. It is a web resource that provides controlled access to molecular docking software running on the Lonestar 5 supercomputer at TACC. The database has collaborated with other databases containing sets of natural and synthetic ligands. In our study, we used two such datasets, namely, the ZINC database and the TCM database. ZINC database is the curated collection of commercially available chemical compounds created for keeping virtual screening as the primary objective. TCM database stands for Traditional Chinese Medicine database, which contains traditionally used medicines and their three-dimensional structure data ready for virtual screening. Previously, this server has been used for virtual screening and discovery of novel inhibitors [35]. The results show the availability of a good number of inhibitors for the protein. The top ligand from the ZINC database showed a good number of physical interactions along with an affinity of −9.5μM, also with no predicted side effects. The ligands which show very close affinity and no predicted toxicity can be taken further for the development process.

It is reported that the Gfo_Idh_MocA protein family contains many different proteins, which almost exclusively consist of NAD(P)-dependent oxidoreductases that have a diverse set of substrates, typically pyranoses. The members of this protein family have a two-domain structure consisting of an N-terminal nucleotide-binding domain and a C-terminal α/β-domain. The C-terminal domain contributes to the substrate binding and catalysis and contains a βα-motif with a central α-helix carrying common essential amino acid residue. The β-sheet of the α/β-domain contributes to the oligomerization in most of these proteins [34]. Domain-based annotation of

EJJ83173 (putative NADH-dependent dehydrogenase of KPNIH11) has not been reported yet. Our study has re-annotated this putative protein as inositol 2-dehydrogenase protein. Combining this information, it is evident that protein EJJ83173 can serve as an attractive drug target. Inosine monophosphate dehydrogenase was targeted in previous studies in the case of tuberculosis [32]. This protein is an attractive drug target [28]. Therefore, the study explains the potency of EJJ83173 protein as a probable drug target. The list of ligands obtained from virtual screening may be further used for clinical testing, which targets both KPNIH11 and SPD39. This study provides a rich source for further experiments to elucidate the role of putative protein EJJ83173 as a drug target for pneumonia infection.

## 5 Conclusion

This study has focused on analyzing the domain repertoire of two pneumonia-causing pathogens to identify Rosetta stone events. Putative proteins of these pathogens were selected in particular to analyze any missing links in protein domain shuffling. Analysis of domain tethering and domain sharing showed that the two species shared many domains in common. The re-annotation of the putative protein by utilizing position-specific iterations has provided several druggable protein candidates. However, an attempt to 3D model these re-annotated putative proteins resulted in only one protein (EJJ83173) as the most likely drug target with good model parameters. Re-annotation of protein EJJ83173 (which contains the GFO_IDH_MocA domain) showed that its probable function is glucose-fructose oxidoreduction. This protein also has sufficient interactor proteins and homolog in both pathogens but no homolog in the host (human), indicating it as an idol drug target. Through virtual screening, several traditional medicines and existing marketed drugs were found to effectively interact with the protein. This study provides a model for drug target identification through domain-based protein re-annotation. However, protein/peptide evidence for this protein target has to be identified to validate these findings and to analyze the usability of this protein as a reliable drug target.

### Abbreviations

BLAST: Basic Local Alignment Search Tool; CAP: Community-acquired pneumonia; CDD: Conserved Domain Database; DELTA-BLAST: Domain Enhanced Lookup Time Accelerated BLAST; KPNIH11: *Klebsiella pneumoniae* subsp. *pneumoniae KPNIH11*; NAD: Nicotinamide adenine dinucleotide; NADP: Nicotinamide adenine dinucleotide phosphate; NCBI: National Centre for Biotechnology Information; ORFs: Open reading frames; PSI-BLAST: Position-Specific Iterated BLAST; RMSD: Root mean squared deviation; SPD39: *Streptococcus pneumoniae strain D39*

## 6 Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s43088-021-00126-7.

**Additional file 1: Supplementary table 1.** Proteins containing common domains and their domain repertoir. These proteins of SPD39 contain one or more of the common domains between the two microbes. Their tethering patterns with other domains account for the domain versatility, while the number of occurence accounts for abundance. **Supplementary table 2.** Proteins containing common domains and their domain repertoir. These proteins of KPNIH11 contain one or more of the common domains between the two microbes. Their tethering patterns with other domains account for the domain versatility, while the number of occurence accounts for abundance. **Supplementary table 3.** Protein annotation using PSI-BAST. These proteins of SPD39 have been annotated using PSI-BLAST and checked for homology with human proteome using DELTA-BLAST. All the proteins except ABJ54307 and ABJ55037 were found non-homologous to human proteomemaking them ideal drug targets. **Supplementary table 4.** Protein annotation using PSI-BLAST. These proteins of KPNIH11 have been anntated using PSI-BLAST and checked for homology with human proteome using DELTA-BLAST. All the proteins were found non-homologous to human proteomemaking them ideal drug targets. **Supplementary table 5.** Virtual screening results of TCM databse ligands. This list of ligands showed affinity towars EJJ83173 protein during docking. These are Traditional Chineses Medicine that can be targeted to interfere protein function of EJJ83173. **Supplementary table 6.** Virtual screening results of ZINC databse ligands. This list of ligands showed affinity towars EJJ83173 protein during docking. These are commercially available drugs that can be targeted to interfere protein function of EJJ83173.

### Acknowledgements

Not applicable.

### Authors' contributions

PR was involved in data curation, analysis, writing, reviewing, and editing of the manuscript. JHN was involved in data curation and analysis; SKHS conceptualized, designed the methodology, and prepared the original draft manuscript. All the authors have read and approved the final manuscript.

### Authors' information

Poornima Ramesh and Jayashree Honnebailu are post-graduate students at the Department of Biotechnology and Bioinformatics, Kuvempu University. Santosh Kumar H S is Assistant Professor at the Department of Biotechnology and Bioinformatics, Kuvempu University. He is interested in understanding the Rosetta stone events in proteins and exploiting them for drug target identification.

### Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

### Availability of data and materials

All data generated or analyzed during this study are included in this published article [and its supplementary information files].

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

Received: 7 September 2020 Accepted: 26 May 2021
Published online: 10 June 2021

## References


1. Anevlavis S, Bouros D (2010) Community acquired bacterial pneumonia. Expert Opin Pharmacother 11(3):361–374. https://doi.org/10.1517/14656560903508770
2. Bocs S, Danchin A, Medigue C (2002) Re-annotation of genome microbial coding-sequences: finding new genes and inaccurately annotated genes. BMC Bioinformatics 3:5. https://doi.org/10.1186/1471-2105-3-5
3. Boxem M, Maliga Z, Klitgord N, Li N, Lemmens I, Mana M, de Lichtervelde L, Mul JD, van de Peut D, Devos M, Simonis N, Yildirim MA, Cokol M, Kao HL, de Smet AS, Wang H, Schlaitz AL, Hao T, Milstein S, Fan C, Tipsword M, Drew K, Galli M, Rhrissorrakrai K, Drechsel D, Koller D, Roth FP, Iakoucheva LM, Dunker AK, Bonneau R, Gunsalus KC, Hill DE, Piano F, Tavernier J, van den Heuvel S, Hyman AA, Vidal M (2008) A protein domain-based interactome network for C. elegans early embryogenesis. Cell 134(3):534–545. https://doi.org/10.1016/j.cell.2008.07.009
4. Camus J-C, Pryor MJ, Médigue C, Cole ST (2002) Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. Microbiology (Reading) 148(Pt 10):2967–2973. https://doi.org/10.1099/00221287-148-10-2967
5. Chen CY-C (2011) TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico. Plos one 6(1):e15939. https://doi.org/10.1371/journal.pone.0015939
6. Chen Y-A et al (2019) Assessing drug target suitability using TargetMine. F1000Res 8:233. https://doi.org/10.12688/f1000research.18214.2
7. Chothia C, Gough J, Vogel C, Teichmann SA (2003) Evolution of the protein repertoire. Science (New York) 300(5626):1701–1703. https://doi.org/10.1126/science.1085371
8. Corbi-Verge C, Kim PM (2016) Motif mediated protein-protein interactions as drug targets. Cell Commun Signal 14:8. https://doi.org/10.1186/s12964-016-0131-4
9. Cross JB, Thompson DC, Rai BK, Baber JC, Fan KY, Hu Y, Humblet C (2009) Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. J Chem Inform Model 49(6):1455–1474. https://doi.org/10.1021/ci900056c
10. Damte D, Suh JW, Lee SJ, Yohannes SB, Hossain MA, Park SC (2013) Putative drug and vaccine target protein identification using comparative genomic analysis of KEGG annotated metabolic pathways of Mycoplasma hyopneumoniae. Genomics 102(1):47–56. https://doi.org/10.1016/j.ygeno.2013.04.011
11. Giuliani S et al (2018) Computationally-guided drug repurposing enables the discovery of kinase targets and inhibitors as new schistosomicidal agents. Plos Comput Biol 14(10):e1006515. https://doi.org/10.1371/journal.pcbi.1006515
12. Gorrie CL, Mirčeta M, Wick RR, Edwards DJ, Thomson NR, Strugnell RA, Pratt NF, Garlick JS, Watson KM, Pilcher DV, McGloughlin SA, Spelman DW, Jenney AWJ, Holt KE (2017) Gastrointestinal carriage is a major reservoir of Klebsiella pneumoniae infection in intensive care patients. Clin Infect Dis 65(2):208–215. https://doi.org/10.1093/cid/cix270
13. Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18(15):2714–2723. https://doi.org/10.1002/elps.1150181505
14. Hanumanthappa M (2016) Conformational flexibility and dynamic properties in allosteric regulation of Mycobacterium tuberculosis pyruvate kinase. MOJ Proteomics Bioinformatics 4(3):269–283. https://doi.org/10.15406/mojpb.2016.04.00128
15. Irwin JJ, Shoichet BK (2005) ZINC--a free database of commercially available compounds for virtual screening. J Chem Inform Model 45(1):177–182. https://doi.org/10.1021/ci049714+
16. Khazanov NA, Damm-Ganamet KL, Quang DX, Carlson HA (2012) Overcoming sequence misalignments with weighted structural superposition. Proteins 80(11):2523–2535. https://doi.org/10.1002/prot.24134
17. Kiefer F, Arnold K, Kunzli M, Bordoli L, Schwede T (2009) The SWISS-MODEL Repository and associated resources. Nucleic Acids Res 37(Database issue):D387–D392. https://doi.org/10.1093/nar/gkn750
18. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) {\it PROCHECK}: a program to check the stereochemical quality of protein structures. J Appl Crystallography 26(2):283–291. https://doi.org/10.1107/S0021889892009944
19. Letunic I, Bork P (2018) 20 years of the SMART protein domain annotation resource. Nucleic Acids Res 46(D1):D493–D496. https://doi.org/10.1093/nar/gkx922
20. Lodha R, Kabra SK, Pandey RM (2013) Antibiotics for community-acquired pneumonia in children. Cochrane Database Syst Rev 2013(6):CD004874. https://doi.org/10.1002/14651858.CD004874.pub4
21. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290(5494):1151–1155. https://doi.org/10.1126/science.290.5494.1151
22. Manske C, Schell U, Hilbi H (2016) Metabolism of myo-inositol by Legionella pneumophila promotes infection of amoebae and macrophages. Appl Environ Microbiol 82(16):5000–5014. https://doi.org/10.1128/AEM.01018-16
23. Melak T, Gakkhar S (2014) Potential non homologous protein targets of mycobacterium tuberculosis H37Rv identified from protein-protein interaction network. J Theor Biol 361:152–158. https://doi.org/10.1016/j.jtbi.2014.07.031
24. Owens J (2007) Determining druggability. Nat Rev Drug Discov 6(3):187–187. https://doi.org/10.1038/nrd2275
25. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem 25(13):1605–1612. https://doi.org/10.1002/jcc.20084
26. Resende T et al (2013) Re-annotation of the genome sequence of Helicobacter pylori 26695. J Integrative Bioinformatics 10(3):233. https://doi.org/10.2390/biecoll-jib-2013-233
27. Ruuskanen O, Lahti E, Jennings LC, Murdoch DR (2011) Viral pneumonia. Lancet 377(9773):1264–1275. https://doi.org/10.1016/S0140-6736(10)61459-6
28. Saiardi A, Azevedo C, Desfougères Y, Portela-Torres P, Wilson MSC (2018) Microbial inositol polyphosphate metabolic pathway as drug development target. Adv Biol Regul 67:74–83. https://doi.org/10.1016/j.jbior.2017.09.007
29. Sander T, Freyss J, von Korff M, Rufener C (2015) DataWarrior: an open-source program for chemistry aware data visualization and analysis. J Chem Inform Modeling 55(2):460–473. https://doi.org/10.1021/ci500588j
30. Sharma S, Maycher B, Eschun G (2007) Radiological imaging in pneumonia: recent innovations. Curr Opin Pulm Med 13(3):159–169. https://doi.org/10.1097/MCP.0b013e3280f3bff4
31. Siezen RJ, Francke C, Renckens B, Boekhorst J, Wels M, Kleerebezem M, van Hijum SAFT (2012) Complete resequencing and re-annotation of the Lactobacillus plantarum WCFS1 genome. J Bacteriol 194(1):195–196. https://doi.org/10.1128/JB.06275-11
32. Singh V, Donini S, Pacitto A, Sala C, Hartkoorn RC, Dhar N, Keri G, Ascher DB, Mondésert G, Vocat A, Lupien A, Sommer R, Vermet H, Lagrange S, Buechler J, Warner DF, McKinney JD, Pato J, Cole ST, Blundell TL, Rizzi M, Mizrahi V (2017) The inosine monophosphate dehydrogenase, GuaB2, is a vulnerable new bactericidal drug target for tuberculosis. ACS Infect Dis 3(1):5–17. https://doi.org/10.1021/acsinfecdis.6b00102
33. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 43(Database issue):D447–D452. https://doi.org/10.1093/nar/gku1003
34. Taberman H, Parkkinen T, Rouvinen J (2016) Structural and functional features of the NAD(P) dependent Gfo/Idh/MocA protein family oxidoreductases. Protein Sci 25(4):778–786. https://doi.org/10.1002/pro.2877
35. Tan Z, Chen L, Zhang S (2016) Comprehensive modeling and discovery of mebendazole as a novel TRAF2- and NCK-interacting kinase inhibitor. Scie Rep 6(1):33534. https://doi.org/10.1038/srep33534
36. Telkar S, Kumar H, Mahmood R (2014) Synteny approach of drug target prediction among unique hypothetical proteins of *Streptococcus gordonii* causing infective endocarditis. Sci Technol Arts Res J 2(4):34. https://doi.org/10.4314/star.v2i4.7
37. Tian W, Chen C, Lei X, Zhao J, Liang J (2018) CASTp 3.0: computed atlas of surface topography of proteins. Nucleic Acids Res 46(W1):W363–W367. https://doi.org/10.1093/nar/gky473
38. Vogel C, Teichmann SA, Pereira-Leal J (2005) The relationship between domain duplication and recombination. J Mol Biol 346(1):355–365. https://doi.org/10.1016/j.jmb.2004.11.050
39. Wang S, Wei W, Cai X (2015) Genome-wide analysis of excretory/secretory proteins in Echinococcus multilocularis: insights into functional characteristics of the tapeworm secretome. Parasit Vectors 8:666. https://doi.org/10.1186/s13071-015-1282-7


## Publisher's Note