Introduction
Introduction

1.1. Motivation

Infectious tropical diseases are the major cause of mortality and disability in developing countries. Despite high investment in pharmaceutical research, the number of new drugs developed for such diseases has declined in recent years (Khanna, 2012). Development of a new drug is financially advantageous for a company, but also carries a lot of risk. In the current market scenario, the encouragement for companies to identify new therapeutic targets and develop drugs for neglected infectious diseases is low (Buckup, 2008), because profits in drugs for diseases that affects a small group of peoples, particularly the poor, are less than an equivalent investment in drugs for the developed world (Webber et al., 2001). Therefore, non-profit academic research bodies need to take up potential target identification and drug development for infectious diseases.

Determination of a minimal set of genes or their corresponding proteins, which are essential for the survival of a particular pathogen, enables the identification of potential drug targets. However high-throughput experimental assays for essential gene identification of pathogenic microorganisms based on gene knock-out method are not feasible for all organisms since such knock-out strains need to be constructed for each gene, a formidable task.

Advancements in a variety of post-genomic techniques have resulted in generation of a large amount of data about metabolic pathways, protein-protein interaction networks and cellular localization of proteins. Cellular proteins interact with each other and form biological networks to accomplish various cellular functions. Hence understanding the behavior of and interactions among various cellular components can contribute to identification of new drug targets. Thus, for potential target identification and drug development, computational methods that support the prediction of essential proteins are needed.
1.1.1. Hypothesis of the work

It has been shown that homology mapping of a particular pathogen’s proteome with the experimentally-verified essential proteins in another model organism can help identify the essential proteins of the former with a reasonable accuracy. Based on these two computational tools namely FindTarget (Chetouani et al., 2001) and TiDT (Singh et al., 2006) were developed for the identification of putative essential genes in pathogens. However these tools lack the provision for identification and removal of paralogous sequences of pathogen, removal of very short sequences of the pathogen prior to homology mapping and do not have the choice to update the host and essential protein database. Also, homology mapping using BLAST requires the use of an arbitrary cut-off score to differentiate between putative essential and non-essential proteins.

Flux Balance Analysis (FBA) is a mathematical modeling approach widely used for identification of essential genes or proteins (Feist et al., 2007). However, the FBA-based method requires clear information regarding nutrition availability and biomass production under specific environmental conditions (Feist et al., 2010).

In addition, different centrality measures obtained from protein-protein interaction network, such as degree, closeness, eccentricity, radiality, betweenness and clustering coefficient etc, can be used for prediction of essential proteins (Silander et al., 2009).

However, none of the above methods by itself provides a good estimate of gene or protein essentiality. Therefore there is a need to develop a tool that integrates more than one approach, e.g. sequence homology approach as well as the protein-protein interaction network topology approach, for effective prediction of drug targets in a pathogen.

1.1.2. Significance

The proposed tool will be able to identify potential drug targets against a pathogen not only for the use in humans, but also in other organisms. With this tool, the users will have the liberty to use the formatted host proteome and the essential protein database of their interest, and to update the existing databases when new updates are available.
1.1.3. Objective

The main objectives of this thesis are

(i.) To design a pipeline software for exogenous drug target identification.

(ii.) To develop a classifier model for effective classification of essential proteins in microbial proteome using support vector machine (SVM) algorithm.

(iii.) Combine sequence homology and protein-protein interaction network analysis approaches for effective identification of putative drug targets in nine common human pathogenic bacteria.

(iv.) Structure-based virtual screening of one identified potential drug target penicillin binding protein 4 (PBP4) of *Listeria monocytogenes* strain EGDe for potential lead identification.

1.2. Development of pipeline tool for putative drug target identification

The target protein must be essential for the growth, replication and viability of the pathogen, but should not have any homolog in the human host, in order to address cytotoxicity issues. Intersection of two datasets, namely (i) a pathogen’s subtractive proteome dataset with the host proteome, and (ii) the pathogen’s minimal essential protein dataset, should represent a set of proteins which are essential for the pathogen but are not needed by the host, and therefore whose manipulation may reasonably be expected to interfere with the pathogen’s survival without adversely affecting the host. These proteins could thus act as potential targets for drugs acting against the particular pathogen.

Proteome subtraction approach is based on sequence homology mapping using BLASTP search of a pathogen’s proteome sequence against the human proteome to identify the pathogen’s proteins that are non-homologous to human proteome. The identified non-human homolog protein sequence of the pathogen is then subjected to BLASTP search against pathogen essential proteins from database of essential gene (DEG). Different authors have suggested the use of different e-value thresholds ranging
from $10^{-2}$ to $10^{-10}$ for identification of bacterial proteins non-homologous to human, and for identification of pathogen specific essential proteins.

However, essential protein identification based on proteome subtraction approach alone has some limitations, including the need to use an arbitrary cut-off score to differentiate between putative essential and non-essential proteins, and the inability of computer techniques to reliably distinguish between related and unrelated proteins when pair-wise sequence identity is low (twilight zone), e.g. below $\sim 25\%$ (Rost, 1999).

To address this problem we have designed a pipeline tool for bacterial drug target identification by integrating in silico proteome subtraction approach and protein-protein interaction network topology based approach.

Protein-protein interaction network analysis of pathogen’s subtractive proteome dataset with the host proteome can be done to identify pathogen’s proteins that are functionally related tend to clusters into hubs. These hub proteins are essential not only for the stability and integrity of protein-protein interaction network, but also for the growth and survival of the pathogen (Jeong et al., 2001), hence can be considered as putative drug targets.

1.3. **Bacterial pathogen considered for putative drug target identification**
Using this tool, we have identified pathogen specific essential proteins of nine bacterial pathogens, and the identified putative targets of these pathogens were documented in the form of a database, which has been made available at www.stbmi.ac.in/BDTI.

1.3.1. **Campylobacter jejuni**
*Campylobacter jejuni* infection is the prime cause of bacterial enteritis globally and its frequency has been increasing with time. Majority of the Campylobacter infections result in acute diarrhea, but only immune-compromised patients with enteritis require antibiotic treatment. Fluoroquinolones such as Ciprofloxacin and macrolides such as Erythromycin are the most commonly used antibacterial agents for the treatment of campylobacteriosis. Resistance to macrolides, fluoroquinolones and other most widely used antibiotics gives rise to a challenge in campylobacteriosis control worldwide (Ruiz-Palacios, 2007).
In recent years, there is a significant increase in fluoroquinolone resistance in Campylobacter in India (Ghosh et al., 2013). In northern India, the resistance to macrolides has increased from 6.1% (Jain et al., 2005) to 22.2% (Ghosh, et al., 2013) during the period 2005 to 2013.

1.3.2. *Clostridium difficile*

*Clostridium difficile* is an enteric pathogen which is the prime cause of antibiotic – associated diarrhea (AAD), which results from alteration due to reduction of natural intestinal flora following antibiotic treatment. Patients suffering from *Clostridium difficile* associated diarrhea (CDAD) show symptoms from mild diarrhea to pseudomembranous colitis, a severe form of intestinal infection (Kilicarslan et al., 2014). CDAD accounts for 20% to 30% of antibiotic-associated diarrhea with mortality up to 25% in elderly patients (Vaishnavi, 2011). Apart from mortality, CDAD imposes a huge economic burden on global health service (Verma et al., 2011).

Metronidazole and oral vancomycin have been the most common treatment for *Clostridium difficile* infections. In recent years, there has been a substantial rise in the rate of *Clostridium difficile* infections as well as the emergence of virulent and antibiotic resistant *Clostridium difficile* strains (Shah et al., 2010). Alternative therapeutics used for treatment of *Clostridium difficile* infections using gut-derived bacteriocins, probiotics and phage (Rea et al., 2013) are being tried but have not yet been very successful. Other treatment methods based on antibodies (Humphreys et al., 2014) or vaccines (Leuzzi et al., 2014) are currently under development. So there is a need for development of new drugs for the treatment and prevention of *Clostridium difficile* infections.

1.3.3. *Francisella tularensis*

*Francisella tularensis*, a gram-negative facultative intracellular bacterium, causes tularemia, a life-threatening zoonotic disease. *Francisella tularensis* subspecies tularensis (type A) is a highly infectious organism which is classified as a ‘CDC category A’ bioterrorism agent (He et al., 2009). Use of antibiotics streptomycin and gentamicin is restricted to severe tularemia cases because of their side effects. Treatment of tularemia is confined to a few antimicrobial agents, such as fluoroquinolone and tetracycline.
Introduction

(Hepburn et al., 2008). In recent years, evolution towards high level fluoroquinolone resistance in *Francisella* species has been reported (Sutera et al., 2014). Live attenuated vaccines to combat tularemia are under development (Rockx-Brouwer et al., 2012, De Pascalis et al., 2012). No other vaccine is currently available for human use. In view of the potential for use as a bioterrorism agent, it would be useful to have newer agents that can deal with this infection.

1.3.4. *Haemophilus ducreyi*

Chancroid or soft chancre (ulcus molle), is a sexually transmitted, genital ulcerative disease, caused by facultative anaerobic gram-negative coccobacillus *Haemophilus ducreyi*. This disease is prevalent in developing countries of Asia, Africa, Latin America and Caribbean among the people of low socioeconomic group (Lewis et al., 2006). The significance of the disease was recently enhanced with the finding that it has the capacity to boost transmission of human immunodeficiency virus (HIV) infection.

1.3.5. *Leptospira interrogans*

Leptospirosis, a zoonotic infection of global significance caused by *Leptospira interrogans* was identified in India since 1931 (Sambasiva et al., 2003). The disease is characterized by jaundice, acute renal failure, and bleeding and is widespread in southern, central, eastern and western India, where substantial rain fall, animal husbandry practices, and agrarian society, increase the risk of this infection (Sethi et al., 2010). Leptospirosis Burden Epidemiology Reference Group (LERG) at the WHO has estimated the number of human cases of severe leptospirosis to be over 500,000 cases per year globally (Abela-Ridder et al., 2010). In recent years, emergence of severe pulmonary hemorrhagic leptospirosis, with high mortality rate (Truong et al., 2011), has drawn global attention to this pathogen.

1.3.6. *Listeria monocytogenes*

*Listeria monocytogenes* is a gram-positive, intracellular, pathogenic bacterium, which can resist low temperature, high salt concentrations and acidic pH values (Milillo et al., 2012). In healthy humans, it causes a mild febrile gastroenteritis. However, infection during pregnancy may lead to premature birth or miscarriage, as well as meningitis in
newborn infants (Ramaswamy et al., 2007). Also, in immune-compromised adults being treated for cancer, leukemia or AIDS, Listeriosis is associated with septicemia and meningitis (Swaminathan et al., 2007). The majority of sporadic cases and outbreaks of listeriosis are caused by *L. monocytogenes* strains belonging to serovars 1/2a, 1/2b and 4b (Orsi et al., 2011). Though *L. monocytogene* is sensitive to a broad range of antibiotics, resistance to several antibiotics has been reported (Nwachukwu et al., 2010, Walsh et al., 2001). In India, multidrug-resistant strains of *L. monocytogene* have been reported in river water, milk and human clinical specimens (Sharma et al., 2012, Soni et al., 2013), and may pose serious threats to human and animal health. Hence, there is a need to identify of novel therapeutic drug targets and to develop new drugs against *L. monocytogenes*.

1.3.7. *Neisseria gonorrhoeae*

*Neisseria gonorrhoeae* causes gonorrhea is one among the most common sexually transmitted diseases around the globe. High disease burden evidenced by recent increase of reported cases of gonococcal infection is a public health concern worldwide (Workowski et al., 2008). Antigenic variation capability of *Neisseria gonorrhoeae* makes it difficult to design an efficient vaccine against this bacterium (Zhu et al., 2011). In the absence of a vaccine, antimicrobial therapy is used for the treatment and control of the disease. Excess use of antibiotics and genetic mutations within the organism has resulted in the development of antibiotic resistance in *Neisseria gonorrhoeae* to multiple classes of antibiotics, namely, penicillins, tetracyclines and quinolones (Sood et al., 2013). Emergence of *N. gonorrhoeae* strains resistance to ceftriaxone has been reported from India (Bala et al., 2007).

1.3.8. *Neisseria meningitides*

Meningococcal disease is a life-threatening illness, which particularly affects children and adolescents, and has a high mortality rate (Manchanda et al., 2006). Of the 13 serogroups of *Neisseria meningitidis* identified, six serogroups (A, B, C, W135, X and Y) are responsible for majority of the infections worldwide (Manchanda, et al., 2006). Serogroups A, C and W-135 are found to be predominant in Asia; in particular serogroup
A in India (Sinclair et al., 2010). Polysaccharide and polysaccharide-protein conjugate vaccines against serogroups A, C, Y and W-135 have been licensed and are available globally (Khatami et al., 2010).

The first case of Group B meningococcal meningitis infection in India was identified in 1994 (Suri et al., 1994) and the second in 2012 (Aggarwal et al., 2013). Cross reactivity between the ‘B capsular polysaccharide’ of serogroup B and human immunoglobulin M, hinders the development of an effective vaccine against serogroup B (Granoff et al., 1995). This inability to develop vaccines against serogroup B calls for the identification of potential therapeutic targets of Neisseria meningitides serogroup B.

1.3.9. *Staphylococcus aureus*

*Staphylococcus aureus* is a gram-positive cooccal bacterium, accountable for both community-acquired as well as hospital-associated infections. In India, over the last decade, methicillin-resistant *S. aureus* strains (MSRA) has become endemic with prevalence of such strains reaching 20% in the eastern part and 50% in the western part of the country (Ray et al., 2013). Emergence of community acquired MRSA has been increasingly reported as an important pathogen from India (D'Souza et al., 2010). Rising prevalence of MSRA has led to an increased use of the antibiotic vancomycin, which has ultimately resulted in emergence of *S. aureus* strains that are less susceptible to vancomycin (Ramana et al., 2012). Emergence of multidrug-resistance *S. aureus* strains have evoked the need to search for alternative potential therapeutic targets for the treatment of this infection.

Since essential proteins can be expected to share certain features distinct from those of non-essential proteins, it should theoretically be possible to develop a machine learning based classifier to predict whether a protein belongs to one of these two categories. This approach can significantly reduce the time required for identification and characterization of microbial essential proteins.
2. Support vector machine based approach for classification of bacterial essential proteins

In order to overcome this issue of twilight zone, we implemented Support Vector Machine algorithm (Ben-Hur et al., 2008) to develop a classifier model for in silico classification of prokaryotic essential proteins from non-essential proteins based on the physico-chemical properties of their amino acid sequences. This classifier was designed based on aminoacid physico-chemical descriptor vectors calculated from amino acid sequences using PROFEAT (Rao et al., 2011) and PseAAC (Shen et al., 2008) web servers. The ten descriptor vectors considered were: amino acid composition; pseudo aminoacid composition; amphiphilic pseudo amino acid composition; di-peptide composition; normalized Moreau–Broto autocorrelation; Moran autocorrelation; Geary autocorrelation; composition, transition and distribution; quasi sequence order and total aminoacid properties.

3. Characterization of putative drug targets

We combined the homology mapping based proteome subtraction and protein-protein interaction network topology approaches to identify essential proteins in the nine selected pathogens. The essential proteins so identified were then subjected to druggability analysis, sub-cellular localization prediction, metabolic pathway analysis and broad spectrum analysis to find out potential drug targets against these organisms. Based on the above analysis, we selected an essential protein of Listeria monocytogenes EGDe (Penicillin binding protein 4 [PBP4]; PDB: 3ZGA) that fulfilled several criteria for a promising drug target, since it (i) was a broad-spectrum target, (ii) was membrane-associated, and (iii) was involved in a pathogen-specific pathway and subjected it to virtual screening for identification of promising lead compounds.

4. Virtual screening

Listeria monocytogenes EGDe (Penicillin binding protein 4 [PBP4]; PDB: 3ZGA) was used for a virtual screening approach that included sequential ligand-based and structure-based screening steps, followed by absorption, distribution, metabolism, excretion and toxicity (ADMET) profiling. Ligand set which had passed the tests for ADMET properties were subjected to reverse docking (Kharkar et al., 2014) against a set of 113
known bacterial targets to identify compounds that may be active against multiple pathogen proteins.

5. Thesis outline

The problem addressed in this thesis mainly focuses on putative drug target identification, the methods used in work belongs to fairly independent subject areas, namely software development, machine learning and virtual screening. As the methods used in different chapters are dissimilar and independent of each other, these are discussed in their respective chapters.

The thesis is divided into four chapters. Chapter 1 describes the development of a tool for exogenous drug target identification. Chapter 2 describes the attempt to construct a classifier model for classification of prokaryotic essential proteins. Chapter 3 describes the identification and characterization of putative drug targets of nine common bacterial pathogens of human using sequence similarity and protein interaction network topology based approaches. Chapter 4 describes the ligand and structure-based virtual screening of one identified potential drug target penicillin binding protein 4 (PBP4) of Listeria monocytogenes strain EGDe for promising lead identification followed by reverse docking to detect multi target leads.