Discussion
5.1 Synteny based comparative genomics

The area of genomics is one of the most significant developments in the modern science of biomedical research over the past decade. Ever since the first bacterial genome was sequenced in 1995, prokaryotic genome sequencing has been progressing very rapidly. All complete genome projects taken up till date are 12923, 2709 bacterial genomes have been finished completely and made publicly available and many are ongoing (http://www.genomesonline.org) on these aspects. Earlier when the first two bacterial genomes were sequenced, it was proposed that comparison of the two genomes may indicate the minimal gene set required for bacterial life (Mushegian and Koonin, 1996). Sequence data on a large scale has already begun metamorphosing the view of modern biological framework and processes at the taxon, population, molecular and biochemical levels.

Although, the Human Genome Project was the quintessential “big science” project, data derived from it are having the effect of standardizing and globalizing science. Sequence data are now freely available everywhere and require only reasonably priced consumer-grade computing tools to process. Research Laboratories throughout the world can now add this genomic module to their biological investigations with petite additional investment in terms of equipments.

As all the species considered in this study were pathogenic to humans or mammals, the genome sequences provided insights into bacterial biology and pathogenicity. On the other hand, as the bacterial genomes were amongst the smallest and closely related, the comparison of the genomes was crucial. However, it was clear that survival in different environmental niches has made them to add additional genes, for example those associated with pathogenicity with specific hosts. As the genome sequences became available and more
genomic variations were identified, it seemed impossible that "one size fits all" in the entire bacterial domain. Even with the availability of hundreds of bacterial genomes, it is lot of scope for understanding of the bacterial genomics. Firstly, continuously expanding genomic data have provided a foundation for various applications, indeed most significantly for comparative genomics. The analysis of a single genome gives biological insights into any given organism and the comparative genome analysis of multiple genomes provides substantially more insight on the physiology and evolution of bacteria. It expands better functional annotation for the predicted coding sequences, and identifies new genes (Fraser et al., 2000).

Most of the genome sequence projects have been targeted on pathogenic bacteria and comparative genomics provided a new route for the discovery of bacterial virulence factors. Predominantly, in vitro culture seems to be impossible for some of the obligate intracellular bacterial pathogens like C.pneumoniae, and genomics has provided the effective method for the identification of virulence genes which can be considered as drug targets (Raskin et al., 2006).

Comparative genomics sheds light on many genomic variations within closely related bacteria. Although, the chromosomal organisation among more closely related bacteria is more conserved, genomic diversity exists and the extent of it was revealed to be far greater than expected even within a single bacterial species. For example, in C.pneumoniae, the largest genomes possess over ~1 Mb more DNA than the smallest genomes. This variation in genome size reflects the genomic composition, which differs vastly within the same or related species.

The complete genome of C.pneumonia AR39, J138, TW183 and CWL029 strain revealed that 99.5% of the genes are homologous with strain-specific genes (greater than 50 bp) and the related species considered in this study are
comparatively heterogeneous. So the identification of genomic differences can provide insights into biological function, as well as about the evolution of bacterial pathogens.

Particularly, gene gain or loss, and gene context can provide clues to the identity and function of virulence factors. Additionally, genomic comparisons with closely related species can accelerate functional annotation of novel genes and other features such as gene fusions and pseudogenes.

Increasingly, sequences of closely related genomes are available, meaning that genome differences are minimal, and comparative genomics allows the identification of single nucleotide polymorphisms (SNPs). Technologies for comparing bacteria at genomic, transcriptomic and proteomic levels in recent years have been extensively reviewed recently (Binnewies et al., 2006). Here, two commonly used methods are summarized as identification of orthologous and non-orthologous genes from synteny based comparative genomics as a modern approach in the field of comparative genomics.

Synteny refers to conservation of gene order. The word synteny was originally applied to describe the occurrence of gene orthologues on equivalent chromosomes in two different eukaryotic genomes. In recent years the synteny based comparison has been adapted by many of the genomicists and is now widely accepted with the context of referring to multigene regions, where, the DNA sequence and gene order are conserved between genomes. The detailed genetic maps of C.pneumonia indicated that genes are not necessarily occur at the same relative position in all bacterial genomes but it was known that certain gene clusters were syntenic (orthologous).

Genome sequencing has allowed a more detailed assessment, and it was soon accomplished that although there seems to be a positive selection for clustering of physically interacting proteins, there is no absolute requirement for
concurrence of any genes in a bacterial genome and synteny is lost at a much faster rate than sequence similarity.

Nonetheless, synteny remains a useful indicator in assessment of genome evolution and prediction of gene function. Gene clusters that appear resistant to dispersal include the ribosomal protein operon, the *nuo* operon (NADH dehydrogenase), and the *dcm* cluster.

5.1.1 Classification based on synteny blocks.

The chlamydial genomes of close related organisms showed extensive synteny and revealed a striking feature of bacterial genome evolution. Using dot plots to represent similarity between aligned genomes, it has been shown that large-scale symmetrical inversions centered on the origin of replication are common in bacteria.

In such plots and synteny alignment tools like SynteView revealed that collinear genomes are represented as an unbroken diagonal and runs of conserved regions observed throughout the whole *C.pneumoniae* genomes. A single symmetrical inversion around the origin results in a counter diagonal with multiple inversions appearing as a broken X pattern.

Using these methods, synteny can be easily detected in between phylogenetically distant genomes such as *S.coelicolor* and *C.pneumonia*. The prevalence of these inversions over other types of rearrangements is intriguing and has provoked several plausible explanations.

An unanswered question is whether such rearrangements are more likely to be occur per se, or they voluntarily fixed by selection with other rearrangements. One factor may be that the inversion does not disturb the orientation or distance of a gene relation to the origin (Bentley & Parkhill, 2004).

Under certain circumstances, multiple replication events may be in operation, distance from the origin would have a gene-dosage effect. Furthermore, such mutual inversions does not disturb the equality of replicochrome
lengths and the differential mutational pressures on the leading and lagging strands may have an unfavorable effect on genes whose strand is switched by other types of inversion (Szczepanik et al., 2001).

In connection with replication, it seems likely that the unwound and unpackaged DNA within replication forks serving as hotspots for mutual recombination. *Chlamydia* and *Chlamydophila* genomes had a slower rate of inversions relative to phylogenetic distance, possibly due to absence of proteins from their replication machinery that are involved in recombination. By performing the comparative genome analysis, gradual reduction of synteny due to mutual inversions and other gene rearrangements was observed. It implies that, conserved gene order may be useful as a orthology measure for studying the relationship between genomes. But the other factors need to be taken into account such as; the potentially calamitous effect on synteny due to insertion element expansion may lead to non-orthologous regions.

The organization of bacterial genomes changes during evolution as a result of gene losses acquisitions by duplication or horizontal transfers and transpositions (Challacombe et al., 2011). Genes can be located based on their random arrangement along the genomes. Synteny as discussed above revealed the conservation of their relative positions in genomes of diverse species reflects the fundamental constraints on the natural evolution. The present approaches also based on synteny infer pairs of co-localized genes from multiple chlamydial genomes.

The organization of synteny facilitates is useful to study their evolutionary history (Junier and Rivoire, 2013). In genomes, we thus identify synteny units or "syntons", which are clusters of proximal genes that encompass and extend operons. The size distribution of these syntons divides them into large syntons, which correspond to fundamental macro-molecular complexes of bacteria and smaller ones, which display a remarkable exponential distribution of sizes.
The graph plots clearly depicts that the syntons based on the synteny block size distribution reveals the number of syntons conserved ranging from two blocks to thirty seven blocks size along with the number of blocks varying from three to eighty five as an average among the C.pneumoniae strains which is higher than comparable to the C.pneumoniae and other Chlamydothila genomes as shown in the result.

The comparative analysis showed that the block size and numbers distribution moves in depreciation rate as the organisms considered for the study become distantly related comparatively. Based on the block size distribution graphs was plotted. We can clearly rank the relationship among the genomes, where the C.pneumonia, C.abortus, C.felis, C.caviae, C.muridiarium and C.trachomatis genomes revealed that the block size varies approximately from two to thirty seven and the number of these distributed blocks varies from four to eighty, four to eighty, three to eighty, two to eighty five and two to eighty seven respectively. The graphs obtained were directly proportional to the arrangement of synteny blocks as shown in the SyntevieView result.

These distributions are "universal" in two respects: it holds for vastly different genomes, and for functionally distinct genes. Similar statistical laws have been reported previously in studies of bacterial genomes, and generally attributed to purifying selection or neutral processes. The position of genes along genomes is crucial as it affects their function and evolution. In bacterial genomes, functional constraints are thus responsible for the concentration of highly expressed genes near the origin of replication, and the clustering of co-functional genes into operons of co-regulated genes are observed (Rocha, 2004). Likewise, evolutionary constraints on gene order are evidenced in bacterial genome by the highly variable rates of recombination at different chromosomal regions, and by the inclination of co-localized genes to be co-displaced through horizontal gene transfer (Junier and Rivoire, 2013). Therefore, the activities like
gene loss, gain or duplication and rearrangement during evolution occur, the comparison of evolutionary related species shows a remarkable stability of genomic organization.

The conservation of genomic organization also raises three questions: (i) How to infer, from a comparison of multiple genomes and the pairs of genes with significant conservation of proximity? (ii) How to describe, beyond pairwise relationships and the organization of conserved properties of co-localization? (iii) How to explain, from an evolutionary perspective, the origin of this organization?

Based upon several previous genomic studies, these questions in our study are tackled through a multi-genome comparative analysis, using *C. pneumonia* as reference species as input over the remaining chlamydial genomes which are completely sequenced and that are presently available.

Our study has been successful in answering (a) yield, at each given phylogenetic level, a list of pairs of genes that remain homologous across multiple chlamydial genomes considered in this study. Later, it addresses (b) by projecting these pairs on individual genomes to define synteny units, or syntons, as clusters of mutually proximal genes.

From the comparative analysis we have found that these syntons relates to functional features of chlamydial genomes encompassing operons, and that the distribution of their size partition them in two classes: large syntons, associated with fundamental functions of bacterial cells, and smaller ones, which in many cases do not correspond to previously defined genomic units. These smaller syntons, however, follow a remarkable statistical property, with their sizes being exponentially distributed. This data collectively is helpful in characterization of candidate genes present in these locations as large syntons and small syntons to be drug targets which can be considered for a successful drug discovery pipeline.
Chapter 5

Discussion

The input data is a set of chlamydial genomes whose genes are partitioned into N orthologous classes. Specifically, we consider here a set of 1128 synten blocks of *C. pneumonia* AR39 genome as reference species compared with *C. psitassi*, *C. trachomatis*, *C. muridarum Nigg*, *C. caviae GPIC*, *C. abortus*, *C. felis* and *C. pneumonia* TW183, CWL029 and J138 were the species considered for SBCG analysis, as the genes found in the orthologous regions of *C. pneumonia* are 70%-99% conserved across the evolutionarily related chlamydial species.

These are annotated in terms of clusters of orthologous genes (COGs). COGs are defined from gene sequences only in the form of rectangular synten blocks in SyntView, with no reference to gene positions, based on the principle that any group of at least four genes from distant genomes that are more similar to each other than to any other genes from the same genome should belong to the same COG.

Available complete genome sequences of *Chlamydophila* are phylogenetically related and unevenly sampled. By treating them equivalently amounts to biasing the statistics towards the most represented species and clades. To correct this bias, we underweight each genome in proportion to the number of other genomes to which it is similar, thus naturally defining an effective number of genomes. This approach indicates that the evolutionary distance below which two genomes are considered to be similar.

Comparative analysis of orthologous genomic regions is an effective method for the identification of segments which are likely to mediate a sequence-specific biological function. A synteny based *in silico* analysis using 1128 chlamydial (*C. pneumonia* AR39) synten blocks was performed to classify the *C. pneumonia* AR39 genome which resulted in 460 synten blocks as orthologous and 354 non-orthologous synten blocks.
5.2 Characterization of orthologous and non-orthologous syntenic blocks

By considering the data set of identified orthologous syntenic blocks alone which was classified in the earlier step was considered for critical proteome analyses of conserved regions classified. The next step was to find non-host proteins from the *C.pneumonia* genome as it is an important strategy to predict non-host proteins in the knowledge based drug target discovery pipeline. The term "non-host" refers here to those bacterial proteins that showed no homology with human proteins.

It has been suggested that such proteins can serve as better drug targets, in terms of avoiding the likely side effects and cross-reactivity caused by antibiotics. Prediction of non-host proteins from bacterial genomes remains therefore a critical step in computational drug discovery.

For the identification of such proteins in the *C.pneumonia* genome, amino acid sequences of protein-coding genes from orthologous dataset were obtained from the NCBI database, and compared with the human proteome using NCBI BLASTP. A total of 414 proteins from the *C.pneumonia* genome and were classified as non-host proteins. Initially, this information regarding non-host proteins can be used and has been previously used in many earlier studies for the prediction of drug targets.

Though, it minimizes the time required for drug testing and development, the inclusion of gene essentiality information and drug prioritization parameters offers great advantage in the careful selection of candidates for drug discovery pipelines (Butt *et al.*, 2012).

The computational identification of essential genes of *C.pneumonia* via the homology search method is an important strategy in identification of potential therapeutic targets based on knowledge base discovery. Essential genes
predictions from the individual features of essential genes and from algorithms that combine many of these features have shown significant sensitivity and accuracy when they are applied to experimentally verified essential genes (Gustafson et al., 2006). A detailed evaluation of these computational algorithms, in parallel with experimental verification, has yet to be done. Also, in many instances, prediction classifiers trained on one species dataset do not produce the same prediction accuracy when applied to other bacterial genomes.

The predictive potentials of homology searching are based on the genes conserved and it is general that essential genes are evident in a recent study of essential gene identification in Yersinia pseudotuberculosis (Butt et al., 2012). Here most of the computationally predicted essential genes via DEG-based homology search were also experimentally validated as essential. Furthermore, these essential genes were identified as essential and conserved across more than nine bacterial genomes that were present in the DEG at the time of study (Duffield et al., 2010).

In general, essential gene prediction via homology searching is based on the notion that a gene is likely to be essential if a homolog is the member of another bacterium as an experimentally validated essential gene (Butt et al., 2012). It can be expected, therefore, that as more information regarding essential bacterial genes becomes available from experimental studies, prediction results will increase in accuracy, including in genomes for which experimental approaches have not yet been conducted or are difficult to perform. Therefore, by taking advantage of essential genes information from 31 bacteria's experimentally we verified the essential genes. Here we are reporting 58 protein-coding genes from the C.pneumonia genome as essential from orthologous data set.
Furthermore, the conserved nature of these orthologous proteins across these chlamydial species may allow them to be vaccine components for all the chlamydial organisms considered in this study. This list of eleven drug targets contains an array of different proteins which are crucial and essential for various crucial biochemical and metabolic functional processes in the pathogen.

Eleven proteins were short-listed as drug targets from synteny based comparative genomics strategy. These proteins were classified based on functional class into four proteins as transmembranes, four proteins as signal peptides and three as lipoproteins which are considered as favorites for the researchers involved in drug target discovery.

Where transmembranes are important class of proteins involved in providing a channel through which molecules and ions can pass into the cell. It is a protein which spans the entire length of the cell membrane and is embedded between the phospholipids. Transmembrane proteins mediate the communication between cells and the interacting chemical messengers. These proteins are involved in activation of many important biological processes and played a central role in basically all physiological processes. Presently transmembrane proteins constitute around 60% of the approved drug targets and inhibition of these proteins will absolutely control the pathogenicity of C. pneumonia to cause COPD.

Signal peptides are an imperative class of proteins involving in the significant factors determining interaction with the protein transport system (secretory pathway), hence directs the protein to the destination to which that protein is need to be delivered. Different classes of signal peptide are present which are specific to different cellular placement. Signal peptide motifs are more frequent in drug targets.
Many studies suggested that lipoproteins are functionally diverse class of secreted bacterial proteins universally present in bacteria characterized by an N-terminal lipid moiety. Typically 1% to 3% of bacterial genomes encode lipoproteins. The lipid moiety serves to anchor these proteins to the cell surface. Lipoproteins are synthesized as pre-prolipoproteins and mature by post-translational modifications (Kovacs-Simon et al., 2011).

Inhibition of genes involved in lipoprotein synthesis in Gram-negative bacteria attenuates a variety of Gram-negative pathogens, including *Chlamydia pneumoniae*. The attenuated phenotype of these mutants indicates a significant role of lipoproteins and lipoprotein production in bacterial virulence (Kovacs-Simon et al., 2011). Targeting these proteins will unquestionably facilitate the inhibition of the pathogenicity of *C. pneumoniae*.

The two prioritized drug targets predicted using the present strategy are Ribonucleotide-diphosphate reductase subunit alpha (NP_445410.1) and DNA-directed RNA polymerase subunit beta (NP_445236.1) involved in the key transcriptional process in *C. pneumoniae*.

Sequence search against DrugBank was promising in identifying FDA approved drugs presently available in the market which have the potential to inhibit the above mentioned prioritized drug targets, the drugs predicted were Gemcitabine, Hydroxyurea, Clofarabine and Rifampin, Rifaximin, Rifabutin obtained by performing BLAST search against DrugBank. Our study revealed that <1% of the proteins are found to contribute as the drug targets in the *C. pneumonia* genome using the present strategy.

Further by considering the data set of classified 354 non-orthologous synteny blocks in a genome-wide comparative study, the concealed relationship between genes and genomes of different Chlamydial species were unveiled using Synteny based comparative genomics approach, based on synteny block
similarity and genomic comparison, the hypothetical proteins were filtered to avoid the noise and 70 protein coding sequences were retrieved for the further analysis.

A critical proteome analysis of non-conserved regions led to the characterization of 60 proteins as non-host and 18 proteins as essential. The identified 18 proteins were ranked as drug targets, the predicted 18 drug targets comprising 1% of the total number of protein coding sequences in *C.pneumonia* AR39 genome. A detailed discussion has been provided about the identification of non-host and essential proteins in earlier step for the identification of drug targets from the classified orthologous synteny blocks dataset.

Further the non-host essential proteins were considered for the prioritization step which led to the discovery of 1 protein target ranking it as potential drug target. Proteome analysis of the *C.pneumoniae* predicted targets are involved in chlamydial developmental cycle and these *in silico* predicted targets are expressed during replication process.

Bacterial protein toxin prediction step is crucial as they are the most potent poisons known and may show activity at very high dilutions where toxins are the major determinant of virulence, usually virulent strains of bacterium produce a range of toxins which is not observed in non-virulent strains the protein toxins are soluble proteins secreted by living bacteria which are essential proteins for their survival, attack and defensive mechanism. It has been successful in identifying these toxins in *C.pneumonia*, which can be classified into exotoxins and endotoxins where the latter is specific to gram negative bacteria's like *Chlamydia*.

Subcellular localization prediction step is vital as it is a key functional attribute of a protein (Yu *et al.*, 2004). The cellular functions are often localized in specific compartments; predicting the subcellular localization of unknown
proteins may perhaps be used to obtain information about their functionalities and is helpful to select proteins for further analysis.

Additionally, studying the subcellular localization of proteins is also helpful in understanding disease mechanisms and developing novel drugs. All bacterial proteins are synthesized in the cytoplasm, and most remain there to carry out their unique functions. Other proteins, however, contain export signals that direct them to other cellular locations.

In Gram-positive bacteria, these include the cytoplasmic membrane, cell wall and extracellular region. In Gram-negative bacteria, they include the cytoplasmic membrane, the periplasm, the outer membrane and the extracellular space. In most cases the whole protein is located in a single compartment; however, proteins can also cover multiple localization sites (Wang et al., 2005).

Among the 18 putative uncharacterized essential proteins predicted toxins were analyzed using the BTXPred tool which predicts the bacterial toxins based on the protein sequence and it classifies them into exotoxins and endotoxins, 5 proteins were identified as bacterial exotoxins proteins and 2 as bacterial endotoxins proteins which are involved in toxigenesis with the help of lipopolysaccharide components attached to produce cytotoxicity and play an important role in pathogenesis, so we report these bacterial toxins as potential candidate targets.

Unavailability of specific drugs was found when searched in the DrugBank database based on sequence search (Pharmasearch), so there is an urgent need to develop specific inhibitors of these targets as these are the primary cause for the pathogenesis of the C.pneumoniae and related organisms which are involved in pathogenicity.
The present study gives an insight about the putative drug targets which leads to novel drug discovery by applying the modern computational approaches like chemoinformatics; structure based drug design, molecular docking and virtual screening.

To perform functional analysis of proteins listed out from earlier analysis, proteins were classified into families and domains to predict the important sites involved in vital functionalities. We extrapolated the listed proteins using InterPro, which helps in analyzing the given protein’s function and classifies based on their family and domains in which they are present with additional information on important sites of protein and integrates the protein signature’s information from the member databases interlinked.

Successful classification of the short-listed proteins was performed; 1 RNA binding protein, 6 DNA binding proteins, 4 proteins involved in catalytic activity, 3 ATP binding proteins, 1 chaperonin protein, 1 protein involved in oxidoreductase activity, 1 protein involved in hydrolase activity, 1 protein involved in proteolysis. All the proteins listed above are involved in the transcriptional process during the replication.

Four proteins were ranked as potential drug targets based on the virulence, functionality and the availability of the drugs, namely signal recognition particle protein FtsY and cysteiny1-tRNA synthetase which are endotoxins involved in pathogenesis, serine protease and Clp Protease are class of proteins which are ubiquitous and serves as virulence factor in causing the disease.

Signal recognition particle protein FtsY (SRP) is a multimeric protein, which along with its conjugate receptor (SR), is involved in targeting secretory proteins to the rough endoplasmic reticulum (RER) membrane in eukaryotes, or
to the plasma membrane in prokaryotes. SRP recognizes the signal sequence of the nascent polypeptide on the ribosome.

Cysteiny1-tRNA synthetase is an alpha monomer and belongs to class Ia. It is highly specific despite not possessing the amino acid editing activity characteristic of many other tRNA ligases. It catalyzes the attachment of an amino acid to its cognate transfer RNA molecule in a highly specific two-step reaction. These proteins differ widely in size and oligomeric state, and have limited sequence homology.

Serine, proteases are proteins that have serine, an amino acid, bonded at the active site. Their main function is involved in processes such as inflammation, blood clotting, and the immune system in both prokaryotes and eukaryotes. The serine protease is the enzyme that catalyzes the hydrolysis of ester or amide. Serine proteases are sequence specific, these are involved in signalling pathways, enzyme activation and degradative functions in different cellular or extracellular compartments.

Clp Protease is a member of the HSP (heat-shock protein) 100 family. It functions as an ATP-dependent molecular chaperone and is the regulatory subunit of the ClpXP protease. ClpXP is involved in DNA damage repair, protein folding, stationary-phase gene expression, and ssrA-mediated protein quality control. To date more than 50 proteins includes transcription factors, metabolic enzymes, and proteins involved in the starvation and oxidative stress responses have been identified as substrates. These predicted proteins are also attractive chlamydial vaccine components due to their predicted promiscuity for being essential genes for the survival of the organism.
5.3 Unique Metabolic pathway identification and interactome analysis

Metabolic network-based pathway analysis has a comparatively short history. It was started by Bruce Clark in 1980 that relied mainly on kinetic information and used convex analysis techniques (Rockafellar, 1970), to study stability in chemical reaction networks.

An actual mathematical formalism for analyzing metabolic pathways was then introduced by Seressiotis and Bailey (Seressiotis and Bailey, 1988) who presented an artificial intelligence based algorithm to generate a set of pathways able to transform a given substrate to a given product by using a database containing enzyme and metabolite information. This algorithm was further advanced by Mavrovouniotis et al. (Mavrovouniotis et al., 1990), to deal with multiple reactants and products. The generated solutions could then be compared to the pathways found in nature.

The completion of the human genome project has revolutionized the field of drug-discovery against threatening human respiratory pathogens. Prediction of non-homologous proteins as drug targets is first preference for host–pathogen diseases. Most of the potential drug targets can be identified among large number of non-homologous protein through protein interaction network analysis.

The information about specific pathways in several pathway databases such as the Kyoto Encyclopaedia of Genes and Genomes (KEGG), MetaCyc and BioCyc that have become available in the past few years varies enormously in the level of detail. For example, the glycolysis pathway is usually curated in greater detail than an organism-specific pathway such as that for mycolic acid synthesis. In practice, therefore, the pathway databases provide initial scaffolds for constructing more elaborate models by analysis of biochemical literature. A good
model is an indispensable pre-requisite for performing reliable predictions on the behavior of a system.

Drug target must possess the non-homologous property which is associated with maximum number of functional proteins (Anishetty et al., 2005). Various existing experimental methods and computational approaches are genomics, proteomics, metabolomics, gene disruption, interactomics, protein functional cascading, bioinformatics etc (Sakharkar et al., 2004).

Metabolic pathway network involves the integrated interconversion of thousands of metabolites through enzyme-catalyzed biochemical reactions; it can be reduced to a composite network of interconnected molecules and reactions that facilitate to generate mass, energy, and information exchange.

Whereas, reductionist approaches remain critically important to develop an understanding of the individual functions of genes and proteins inevitably most cellular components exert their function through a network of metabolic and genetic interactions (Tyson and Novak 2010).

The latest, genome-driven in silico modeling and analysis of metabolic networks have been recognized as a promising approach for characterizing complex cellular systems and for deriving novel strategies for biomedical and biotechnological applications in the context of systems biology and biotechnology (Boran and Iyengar, 2010).

It is a reduced representation of cellular metabolism at a highly conceptual level, where each node represents a metabolite that is connected to others by links, each of which represents a metabolic reaction catalyzed by a gene-encoded metabolic enzyme (Barabasi and Oltvai, 2004).

With the availability of full-genome sequences since the mid-1990s metabolic network analysis provides a crucial framework for understanding the
cellular metabolism in both physiological and pathophysiologica states, which will eventually facilitate comparative metabolic network analysis-based drug discovery. In this thesis, our objective is to employ a metabolic network analysis approach to identify novel therapeutic targets (a pathophysiological state) and the metabolism of the bacterial pathogen, C.pneumonia (a physiological state), in order to understand ultimately target cell metabolism for drug discovery.

Additionally, the comparative metabolic pathway analysis approach has also revealed to be capable of predicting the drug targets against pathogen metabolism when completely reconstructed metabolic networks are available.

Databases storing a vast amount of "ready-to-use" metabolic pathways are currently available on the internet. Many of these pathways are simply derived from high-throughput experiments, like microarray experiments.

5.3.1 Comparative metabolic pathway analysis

Computational comparative metabolic network analysis of C.pneumonia AR39 was performed with a goal to identify potential therapeutic targets. A methodical workflow was designed that involved various bioinformatics tools, databases, and drug target prioritization parameters. Primarily the data source of the C.pneumonia AR39 metabolic networks and its human host was downloaded from the KEGG database. The KEGG presently contains information about 65 metabolic pathways in C.pneumonia AR39 and 110 in H.sapiens.

Names and total numbers of proteins present in each pathway were calculated, and comparative analysis was performed manually based on two-list for the identification of pathways specific to C.pneumonia, and pathways common to C.pneumonia AR39 and H. sapiens. Ten different metabolic pathways
were identified as unique to *C. pneumonia*, and fifty five pathways were shared (see Table 7).

Metabolic pathways of the host *H. sapiens* and the pathogen *C. pneumonia* AR39 have been compared by using KEGG database. Pathways which do not appear in the *Homo sapiens* but present in the *C. pneumonia* AR39 have been designated as unique pathways. All the proteins of pathways were subjected to a BLASTp search specifically using the option in BLAST against the *H.sapiens* database (Altschul et al., 1997).

The next step was to find non-host proteins from the *C. pneumonia* genome. As mentioned above the term "non-host" refers here to those bacterial proteins that show no homology or less homology with *H.sapiens* proteins. It has been suggested that such proteins can serve as better drug targets, in terms of avoiding the likely side effects and cross-reactivity caused by antibiotics.

Identification of non-host proteins from bacterial genomes remains therefore a critical step in computational drug discovery. For the identification of such proteins in the *C. pneumonia* genome, amino acid sequences of protein-coding genes from common and unique metabolic pathways were obtained from the KEGG and NCBI databases, and compared with the human proteome using NCBI BLASTP. A total of 199 proteins from the *C. pneumonia* genome showed "no homology" against the human proteome.

Our results signify that the comparative metabolic network analysis approach will facilitate the understanding of cellular metabolism by identifying potential constraints and predicting as well as ultimately targeting the metabolism of the organisms whose complete metabolic networks are available through the seamless integration of virtual screening with experimental validation.
A detailed discussion on unique pathways identified is as follows:

**Lipopolysaccharide biosynthesis**

Lipopolysaccharide (LPS) that constitutes the outer surface of the outer membrane of most Gram-negative bacteria is referred to as an endotoxin. It is comprised of a hydrophilic polysaccharide and a hydrophobic component referred to as lipid A. Lipid A is responsible for the major bioactivity of endotoxin, and is recognized by immune cells as a pathogen-associated molecule.

Most enzymes and genes coding for proteins are responsible for the biosynthesis and export of lipopolysaccharide in *Chlamydia pneumoniae* have been identified, and they are shared by most Gram-negative bacteria based on genetic information.

The complete structure of lipopolysaccharide differs from one bacterium to another, consistent with the recent discovery of additional enzymes and gene products that can modify the basic structure of lipopolysaccharide in some bacteria, especially pathogens like *Chlamydia pneumoniae*.

A detailed knowledge of LPS biosynthesis is of the utmost importance in understanding the function of the outer membrane of Gram-negative bacteria. The regulation of LPS biosynthesis affects many more compartments of the bacterial cell than the outer membrane and thus contributes to the understanding of the physiology of Gram-negative bacteria in general, on the basis of which only mechanisms of virulence and antibiotic resistance can be studied to find new targets for antibacterial treatment.

The study of LPS biosynthesis is also an excellent example to demonstrate the limitations of "genomics" and "proteomics", since secondary gene products
can be studied only by the combined tools of molecular genetics, enzymology and analytical structural biochemistry (Gronow and Brade, 2001).

LPS comprises three parts namely i) polysaccharide (O) side chains (O-antigen); ii) core oligosaccharide; iii) lipid A. Lipid A contains unusual fatty acids (e.g. hydroxy-myristic acid) and is inserted into the outer membrane while the rest of the LPS projects from the surface. LPS forms the amphipathic interface between Gram-negative bacteria and their environment and contributes protection against antibiotics and the complement system.

**Peptidoglycan biosynthesis**

Peptidoglycan is a macromolecule made of long aminosugar strands cross-linked by short peptides. It forms the cell wall in bacteria surrounding the cytoplasmic membrane. The glycan strands are typically comprised of repeating N-acetylglucosamine (GlcNAc) and N-acetylmuramic acid (MurNAc) disaccharides. Each MurNAc is linked to a peptide of three to five amino acid residues.

Disaccharide subunits are first assembled on the cytoplasmic side of the bacterial membrane on a polyisoprenoid anchor (lipid I and II). Polymerization of disaccharide subunits by transglycosylases and cross-linking of glycan strands by transpeptidases occur on the other side of the membrane. Bacterial cell wall biosynthesis inhibitors form a major class of antibiotics.

Peptidoglycan biosynthesis of bacterial cell wall is a complex process that involves number of enzyme reactions that take place in the cytoplasm (synthesis of the nucleotide precursors) and on the inner side (synthesis of lipid-linked intermediates) and outer side (polymerization reactions) of the cytoplasmic membrane. It is important to understand the cytoplasmic steps of peptidoglycan biosynthesis, which can be divided into four sets of reactions that lead to the
Chapter 5

Discussion

syntheses of (1) UDP-N-acetylglucosamine from fructose 6-phosphate, (2) UDP-N-acetylMuramic acid from UDP-N-acetylglucosamine, (3) UDP-N-acetylMuramyl-pentapeptide from UDP-N-acetylMuramic acid and (4) D-glutamic acid and dipeptide D-alanyl-D-alanine.

Furthermore, special attention is given to (a) the chemical and enzymatic synthesis of the nucleotide precursor substrates that are not commercially available and (b) the search for specific inhibitors that could act as antibacterial compounds (Barreteau et al., 2008).

Recent data concerning the different enzymes involved (see Table 7) are presented in earlier studies. It has significantly contributed to the understanding of the biosynthesis of undecaprenyl phosphate, the carrier lipid required for the anchoring of the peptidoglycan hydrophilic units in the membrane, and to the characterization of the MraY and MurG enzymes which catalyze the successive transfers of the N-acetylMuramoyl-peptide and N-acetylglucosamine moieties onto the carrier lipid, respectively. Enzyme inhibitors and antibacterial compounds interfering with these essential metabolic steps and interesting targets are presented in the pathway (Bouhss et al., 2008).

Polycyclic aromatic hydrocarbon degradation

PAHs are aromatic hydrocarbons with two or more fused benzene rings with natural as well as anthropogenic sources. They are widely distributed environmental contaminants that have detrimental biological effects, toxicity, mutagenecity and carcinogenicity. Due to their ubiquitous occurrence, recalcitrance, bioaccumulation potential and carcinogenic activity, the PAHs have gathered significant environmental concern. Even though PAH may go through adsorption, volatilization, photolysis, and chemical degradation, microbial degradation is the main degradation process (Haritash and Kaushik, 2009).
PAH degradation depends on the environmental conditions, number and type of the microorganisms, nature and chemical structure of the chemical compound being degraded. They are biodegraded/biotransformed into less complex metabolites, and through mineralization into inorganic minerals, H(2)O, CO(2) (aerobic) or CH(4) (anaerobic) and rate of biodegradation depends on pH, temperature, oxygen, microbial population, degree of acclimation, accessibility of nutrients, chemical structure of the compound, cellular transport properties, and chemical partitioning in growth medium.

A number of bacterial species are known to degrade PAHs and most of them are isolated from contaminated soil or sediments. *Pseudomonas aeruginosa, Pseudomonas fluorescens, Mycobacterium spp.*, *Haemophilus spp.*, *Rhodococcus spp.*, *Paenibacillus* spp are some of the commonly studied PAH-degrading bacteria.

Lignolytic fungi too have the property of PAH degradation. *Phanerochaete chrysosporium, Bjerkandera adusta, and Pleurotus ostreatus* are the common PAH-degrading fungi. Enzymes involved in the degradation of PAHs are oxygenase, dehydrogenase and lignolytic enzymes. Fungal lignolytic enzymes are lignin peroxidase, laccase, and manganese peroxidase. They are extracellular and catalyze radical formation by oxidation to destabilize bonds in a molecule.

The biodegradation of PAHs has been observed under both aerobic and anaerobic conditions and the rate can be enhanced by physical/chemical pretreatment of contaminated soil. Addition of biosurfactant-producing bacteria and light oils can increase the bioavailability of PAHs and metabolic potential of the bacterial community (Haritash and Kaushik, 2009).

**Biosynthesis of secondary metabolites**

This class of pathway is responsible for the biosynthesis of secondary metabolites, which are organic compounds that are not directly involved in
growth, development and reproduction of an organism. Unlike primary metabolites, which are essential for life, the absence of secondary metabolites has less dramatic effect on an organism. Examples for secondary metabolites includes plant chemical defenses against herbivory and antibiotics. The phylogenetic distribution of secondary metabolites is often limited to a narrow set of taxa.

This pathway involves Antibiotic Biosynthesis, Fatty Acid Derivatives Biosynthesis, Nitrogen-Containing Secondary Compounds Biosynthesis, Phenylpropanoid Derivatives Biosynthesis, Phytoalexins Biosynthesis, Polyketides Biosynthesis, Sugar Derivatives Biosynthesis, Terpenoids Biosynthesis, Terpenophenolics Biosynthesis, Xanthones Biosynthesis (Lancini and Lorenzetti, 1993). The enzymes involved in this pathway are listed in the Table 7.

Microbial metabolism in diverse environments

The microorganisms effectively utilize the specific metabolic pathways for the biodegradation of various compounds occurs under both aerobic and anaerobic conditions. Depending upon the environmental conditions, different types of bacteria, fungi, and enzymes are involved in the degradation process of these compounds and also by using different pathways to biotransform a substrate at a certain transformation rate (Kaiser et al., 1996).

a) Sulfur metabolism: Sulfur is a ubiquitous element of the Earth crust where it is mostly present as sulfate salts. It is an essential component of life. In *C. pneumonia* oxido-reduction and assimilation of sulfur takes place in the presence of oxygen, sulfur metabolism is particularly energy costly. As sulfate, it must first permeate the cell, usually against the intracellular electric potential which is usually strongly negative (-70 mV), then change from a highly oxidized state to a reduced state (Grein et al., 2013). This
requires a significant consumption of energy, as well as the maintenance of a very low oxido-reduction potential, a process that seems difficult to achieve with the simultaneous presence of oxygen molecules. CP0946 (3'2'), 5'-bisphosphate nucleotidase) and CP0204 (oxido-reductase) are the enzymes responsible for sulphur metabolism in C. pneumonia.

b) Lysine degradation: The 2-oxoglutarate dehydrogenase complex catalyzes the overall conversion of 2-oxoglutarate to succinyl-CoA and CO₂. In Gram-negative bacteria it is made up of a single type of subunit and multiple copies of the E1 component, along with multiple copies of the E3 component, are assembled around an E2 core of icosahedral symmetry. CP0378 (sucA 2-oxoglutarate dehydrogenase subunit E1) and CP0225 (branched-chain alpha-keto acid dehydrogenase subunit E2) is a multienzyme complex that plays an important role in branched-chain amino acid catabolism. It occupies a key position in intermediary metabolism and provides metabolic precursors for secondary metabolism in bacteria (Li et al., 2003).

c) Tyrosine metabolism: Tyrosine or 4-hydroxyphenylalanine, is one among the 20 amino acids that are used by cells to synthesize proteins. Tyrosine is an aromatic amino acid important in the synthesis of thyroid hormones, catecholamines, and melanin. It occurs in proteins that are part of signal transduction processes. It functions as a receiver of phosphate groups that are transferred by way of protein kinases. Tyrosine degradation is catalyzed by a series of five enzymatic reactions that yield acetoacetate, which is ketogenic, and the Krebs cycle intermediate fumarate, which is glucogenic (Suda and Takeda, 1950). In C. pneumoniae CP0005 (aromatic amino acid aminotransferase) play a catabolic role in proteolysis, where tyrosine is catalyzed into pyruvate and CP0981 (RNA methyltransferase) is involved in catalysis of N-methyl tyramine to Hordenine.
d) **Pyruvate metabolism:** The pyruvate metabolic pathway is the sum of biochemical reactions involving pyruvate, the intersection of pathways facilitates for glucose and energy homeostasis. At the end of glycolysis, carried out by pyruvate kinases, converts phosphoenolpyruvate (PEP) to ATP and pyruvate which can be converted to acetyl-CoA or anaerobically to lactate involved in anaerobic glycolysis (Zupke and Stephanopoulos, 1994; Liu, 2003). The enzymes like CP1018 (phosphoenolpyruvate carboxykinase); CP0677 (pyruvate kinase); CP0453 (pyruvate dehydrogenase subunit E1 beta); CP1037 (dihydrolipoamide dehydrogenase); CP1037 (dihydrolipoamide dehydrogenase); CP0452 (branched-chain alpha-keto acid dehydrogenase subunit E2); CP0824 (malate dehydrogenase); CP0340 (acetyl-CoA carboxylase carboxyltransferase subunit alpha) were identified as unique to *C.pneumoniae*.

e) **Nicotinate and nicotinamide metabolism:** Nicotinate is also known as niacin or vitamin B3. Nicotinamide is the amide derivative of nicotinic acid. Nicotinate and nicotinamide metabolism are essential for the bacterial pathogens as the precursors for generation of coenzymes, NAD⁺ and NADP⁺, which are essential for redox reactions and electron transfer reactions (Ahn et al., 2014). They therefore exist in oxidised (NAD(P)⁺) and reduced (NAD(P)H) forms. These coenzymes are crucial for many metabolic pathways including glycolysis, TCA cycle, pentose phosphate cycle, fatty acid biosynthesis and metabolism pathways and many others. The enzyme (CP0499, surE) stationary phase survival protein is the member of Nicotinate and nicotinamide metabolism pathway in *C.pneumoniae*.
Two component system

Two-component systems (TCSs) are common signal transduction pathways found abundantly in most phyla except animals. The basic TCS pathway involves two multi-domain proteins. The first is a histidine protein kinase (HPK) whose autokinase activity is dependent upon an environmental stimulus. The second is a response regulator (RR), onto which a phosphoryl group is transferred from the phosphorylated HPK, and which mediate phosphorylation-dependent effects within the cell. TCS proteins can be readily identified from genomic sequences, however, approaches for identifying and classifying TCS proteins (see Table 7) are non-trivial because TCSs are multi-domain, multi-gene pathways, which exhibit considerable heterogeneity in their gene and domain organization.

Two-component systems (TCSs) constitute a signal transduction mechanism, mainly found in prokaryotes, which is relatively simple as they are basically formed by two proteins: a histidine kinase (HK) and a response regulator (RR). These two proteins are able to transmit all kinds of signals productively and elegantly by the use of phosphorylations in order to ensure cell survival.

For this purpose, the HK possesses many qualities; first it is able to sense a signal, second it binds ATP, and third it autophosphorylates on a catalytic His residue. Subsequently, the RR comes into play to promote the final response, but first it needs to be phosphorylated on a catalytic Asp residue via a phosphoryl group transfer from the phosphorylated His of the HK.

The phosphorylated RR (RR~P) can thus elicit many diverse responses, which generally involve binding to DNA to activate specific genes. Finally, the system is shut down by the dephosphorylation of the
RR→P, a mechanism which can be performed by the RR itself, or assisted by the HK.

Undoubtedly, understanding the molecular basis of specificity in the interaction between HK and RR couples as well as the complete catalytic mechanism between these two molecules in the signaling process is especially important from many viewpoints, including the medical science and recent advances in the structural and biochemical field have contributed decidedly.

**Flagellar assembly**

The molecular mechanism of bacterial flagellar assembly has been an ongoing study that spans three decades. Early studies revealed that regulation of flagellar gene transcription was coupled to the assembly process. Latest advances in the understanding of the regulation of flagellar assembly had shown that translational and post-translational regulation too plays a significant role in flagellar assembly (Aldridge and Hughes, 2002).

The flagellum provides a model system for understanding how gene regulation functions to ensure the efficient assembly of a complex structure and fundamental mechanisms common to all type III secretion systems.

Certain classes of pathogenic bacteria (*C.pneumonia*) secrete virulence proteins in a Sec-independent manner, by a mechanism known as type III secretion. The central body of the export machinery specific for virulence proteins is known as a needle complex, which has a related structural organization to flagella. The two structures share several proteins (see Table 7) with highly homologous amino acid sequences. Even where the
sequence identity is low among flagellar proteins from various species, the physico-chemical properties of each protein remain homologous. Therefore, by comparing the physico-chemical properties of unidentified proteins, it is possible to find homologs among type III secretion systems (Aizawa, 2001).

The bacterial flagellum is a complex multicomponent structure which serves as the propulsive organelle for many species of bacteria. Rotation of the helical flagellar filament, driven by a proton-powered motor embedded in the cell wall, enables the flagellum to function as a screw propeller.

It seems likely that roughly all of the genes required for flagellar structure and function have been identified. Stable analysis of the portions of the genome enclosing these genes may disclose the existence of a few more genes. Transcription of the flagellar genes is under the control of the products of a single operon, and so these genes constitute a regulon. Other controls, both transcriptional and post-transcriptional, have been identified. Many of these genes have been sequenced, and the information obtained will aid the drug discovery process and helps to clarify the various regulatory mechanisms of the flagellar regulon (Jones & Aizawa, 1990).

The flagellum is composed of several substructures. The long helical filament is connected via the flexible hook to the complex basal body which is located in the cell wall. The filament is composed of many copies of a single protein, and can adopt a number of distinct helical forms. Structural analyses of the filament are adding to our understanding of this dynamic polymer. The component proteins of the hook and filament have all been identified.
Phosphotransferase system

The Phosphotransferase system (PTS) is a sugar transport system that couples the transport of a sugar to its phosphorylation. It is found only in bacteria, where it catalyzes the transport and phosphorylation of numerous monosaccharides, disaccharides, amino sugars, polyols, and other sugar derivatives (Deutscher et al., 2006). To carry out its catalytic function in sugar transport and phosphorylation, the PTS uses PEP as an energy source and phosphoryl donor.

The phosphoryl group of PEP is usually transferred via four distinct proteins (domains) to the transported sugar bound to the respective membrane component(s) (EIIC and EIID) of the PTS. The organization of the PTS as a four-step phosphoryl transfer system, in which all P derivatives exhibit similar energy (phosphorylation occurs at histidyl or cysteyl residues), is surprising, as a single protein (or domain) coupling energy transfer and sugar phosphorylation would be sufficient for PTS function.

A possible explanation for the complexity of the PTS was provided by the discovery that the PTS also carries out numerous regulatory functions. Depending on their phosphorylation state, the four proteins (domains) forming the PTS phosphorylation cascade (EI, HPr, EIIA, and EIIB) can phosphorylate or interact with numerous non-PTS proteins and thereby regulate their activity.

Additionally, in certain bacteria, one of the PTS components (HPr) is phosphorylated by ATP at a seryl residue, which increases the complexity of PTS-mediated regulation. In this study, we have specifically listed enzymes involved in protein phosphorylation-related regulatory functions.
of the PTS. As we shall see, the PTS regulation network not only controls carbohydrate uptake and metabolism but also interferes with the utilization of nitrogen and phosphorus and the virulence of certain pathogens (Deutscher et al., 2006).

Further the phosphotransferase system is built around a collection of specific transporters that import resources into the cell. Typical bacterial cells make many similar transporters that all stand ready to import whatever sugars are available. A complex regulatory network decides which transporters are expressed and used at any given time.

The phosphotransferase system is also particularly energy-efficient when compared to many of the other transport systems in the cell. Many transporters use ATP to power the import of nutrients, but PTS transporters also add a phosphate group to them at the same time.

Instead of ATP, a molecule of phosphoenolpyruvate (one of the intermediates in glycolysis) is used to power the PTS reaction and provide the phosphate. PTS member proteins can be ranked as potential drug targets as they are involved in crucial cellular processes to obtain energy, by inhibiting the activity of these enzymes leads to the inhibition of energy production in the pathogen.

**Bacterial secretion system**

Gram-negative bacteria possess many protein secretion-membrane insertion systems that apparently evolved independently. Bacteria secrete a wide range of proteins whose functions include biogenesis of organelles, such as pilli and flagella, nutrient acquisition, virulence, and efflux of drugs and other toxins. Six distinct secretion systems have been shown to mediate protein export through the inner and outer membranes of Gram-negative bacteria (Kostakioti et al., 2005).
These pathways are highly conserved throughout the Gram-negative bacterial species. In Gram-positive bacteria, secreted proteins are commonly translocated across the single membrane by the Sec pathway or the two-arginine (Tat) pathway.

Protein secretion plays a central role in modulating the interactions of bacteria with their environments. This is particularly the case when symbiotic bacteria (whether pathogenic, commensal or mutualistic) are interacting with larger host organisms. In the case of Gram-negative bacteria, secretion requires translocation across the outer as well as the inner membrane, and a diversity of molecular machines has been elaborated for this purpose.

A number of secreted proteins are destined to enter the host cell (effectors and toxins), and thus several secretion systems include apparatus to translocate proteins across the plasma membrane of the host also. Many studies have attempted to describe biological processes and cellular components that play important roles in the interactions of microbes with plant and animal hosts, including the processes of bacterial secretion (Tseng et al., 2009).

In Gram-negative bacteria, where secretion involves translocation across inner and outer membranes, there are now known six general classes of protein secretion systems namely 1) Type I protein secretion system (T1SS) 2) Type II secretion system (T2SS) 3) Type III secretion system (T3SS) 4) Type IV secretion system (T4SS) 5) Type V secretion system (T5SS), 6) Type VI secretion system (T6SS) each of which shows considerable diversity (Tseng et al., 2009).
Methane metabolism

Methane is metabolized principally by methanotrophs and methanogens in the global carbon cycle. Methanotrophs consume methane as the only source of carbon, while methanogens produce methane as a metabolic byproduct. Methylotrophs, which are microorganisms that can obtain energy for growth by oxidizing one-carbon compounds, such as methanol and methane, are situated between methanotrophs and methanogens (Amanda, 2010).

Methanogens can obtain energy for growth by converting a limited number of substrates to methane under anaerobic conditions. Three types of methanogenic pathways are known: CO2 to methane, methanol to methane, and acetate to methane.

Methanogens use 2-mercaptoethanesulfonate (CoM; coenzyme M) as the terminal methyl carrier in methanogenesis and have four enzymes for CoM biosynthesis. Coenzyme B-Coenzyme M heterodisulfide reductase (Hdr), requiring for the final reaction steps of methanogenic pathway, is divided into two types: cytoplasmic HdrABC in most methanogens and membrane-bound HdrED in Methanosarcina species.

In methanotrophs and methyltrophs methane is oxidized to form formaldehyde, which is at the diverging point for further oxidation to CO2 for energy source and assimilation for biosynthesis. There are three pathways that convert formaldehyde to C2 or C3 compounds: serine pathway, ribulose monophosphate pathway, and xylulose monophosphate pathway (Chistoserdova et al., 2005).
The first two pathways are found in prokaryotes and the third is found in yeast. As a special case of methylo trophs, various amines can be used as carbon sources in trimethylamine metabolism.

Computational based studies suggest most potential drug targets can be identified among large number of non-homologous protein through the analysis of protein interaction network. It gives deep insight into the structural biology of many biochemical pathways. Through high-resolution structures of protein-interactions structural mechanisms of diseases can be identified and engineer proteins towards specific functions. In this study, the entire interactome has been explored for identification of potential drug targets to design drugs that disrupt pathogenesis.

It is more apparent that the differences in species are due to more interactions between the component proteins, rather than the components themselves. Therefore, several efforts have been made to identify these protein interactions, in an attempt to understand biological systems better. Up to the late 1990's, protein function analyses primarily focused on single proteins (LaCount et al., 2005). But the majority of proteins interact with other proteins for proper function; they should be studied in the context of their interacting partners to fully understand their function. Proteins control all biological systems in a cellular process, and most of the proteins perform their functions independently, the vast numbers of proteins interact with others for proper biological activity.

Proteins are the key catalysts that facilitate most biological processes in a cell, including gene expression, cell growth, proliferation, nutrient uptake, morphology, motility, intercellular communication and apoptosis. Understanding how proteins interact with each other is vital in
identifying biological networks and proteins functionality within the cell. They form the basis for several signal transduction pathways in the cell, in addition to various transcriptional regulatory networks. During protein-protein interaction analysis we can find different types of protein interactions.

Protein interactions are fundamentally characterized as stable or transient, and both types of interactions can be either strong or weak. Stable interactions are those associated with proteins that are purified as multi-subunit complexes, and the subunits of these complexes can be identical or different. Hemoglobin and core RNA polymerase are examples of multi-subunit interactions that form stable complexes.

Transient interactions are expected to control the majority of cellular processes. As the name implies, transient interactions are temporary in nature and typically require a set of conditions that promote the interaction, such as phosphorylation, conformational changes or localization to discrete areas of the cell. Transient interactions can be strong or weak, and fast or slow. While in contact with their binding partners, transiently interacting proteins are involved in a wide range of cellular processes, including protein modification, transport, folding, signaling, and cell cycling (Ozbabacan et al., 2011).

It is indeed a need to understand protein structure and function which has been a critical driving force for biological research in the recent years. With the advent of high-throughput algorithms/tools to identify protein-protein interactions, more information and knowledge on protein function can been obtained, along with the development of several methods to predict and study the interactions between virulence proteins (Eisenberg et al., 2000).
Based on the proteins catalyzing each reaction, the protein-protein interactions are then computed. Further, based on the knowledge of the proteins that participate in various pathways and the pathways that are involved in cellular processes, influences have been derived.

The result of two or more proteins that interact with a specific functional objective can be demonstrated in several different ways. 1) Alter the kinetic properties of enzymes, which may be the result of subtle changes in substrate binding or allosteric effects. 2) Allow for substrate channeling by moving a substrate between domains or subunits, resulting ultimately in an intended end product. 3) Create a new binding site, typically for small effector molecules. 4) Inactivate or destroy a protein. 5) Change the specificity of a protein for its substrate through the interaction with different binding partners; e.g., demonstrate a new function that neither protein can exhibit alone. 6) Serve a regulatory role in either an upstream or a downstream event (Golemis, 2002).

The identified unique, non-host essential protein's interaction network for *C.pneumonia* AR39 was derived from the STRING database. Based on the methods provided in STRING was effective to predict the possible protein-protein interactions were incorporated (Snel et al., 2000).

Twenty two protein targets were considered for the protein interaction network analysis. The results obtained from the STRING's confidence view suggested that among the proteins analyzed, strong associations were between, nrdA and nrdB, waaA-murE and murF, cydB and cydA as shown in Fig: 4.14.

The interaction between hemA and hemL implies that (HemA) glutamyl-tRNA reductase; Catalyzes the NADPH-dependent reduction of
glutamyl-tRNA(Glu) to glutamate 1-semialdehyde (GSA) and glutamate-1-semialdehyde aminotransferase (hemL) interacts with binding score 0.216. The second step of the pathway is catalyzed by glutamate-1-semialdehyde aminomutase which results in the formation of 5-aminolevulinic acid; functions in porphyrin (tetrapyrroles) biosynthesis.

nrDA - ribonucleotide-diphosphate reductase subunit alpha(nrdA) and nrDB - ribonucleotide-diphosphate reductase subunit beta interaction suggests that the protein complex catalyzes the biosynthesis of deoxyribonucleotides from the corresponding ribonucleotides.

The strong interaction between waaA - 3-deoxy-D-manno-octulosonic-acid transferase, murE - UDP-N-acetylglucosamylalaninyl-D-glutamate-2, 6-diaminopimelate ligase and murF- UDP-N-acetylglucosamylalaninyl-D-glutamyl-2, 6-diaminopimelate-D-alanyl-D-alanyl ligase suggest that it catalyzes the addition of meso-diaminopimelic acid to the nucleotide precursor UDP-N-acetylglucosamyl-L-alanyl-D-glutamate (UMAG) in the biosynthesis of bacterial cell-wall Peptidoglycan and involved in cell wall formation. It catalyzes the final step in the synthesis of UDP-N-acetylglucosamyl-pentapeptide, the precursor of murein.

cydA - cytochrome D ubiquinol oxidase, subunit I and cydB - cytochrome D ubiquinol oxidase, subunit II interaction results in catalyzation of quinol dependent, Na⁺ independent oxygen uptake. Cytochrome bd protein may play an important role in micro-aerobic nitrogen fixation and oxidation-reduction process in the respiratory pathogen like C.pneumonia.

The subsequent interactome analyzed is vital in identifying possible pathways for the emergence of pathogenicity and drug
resistance. A set of proteins involved in both intrinsic and extrinsic pathogenic mechanisms were considered and shortest paths from different drug targets to these identified proteins were computed and scored.

Top-scoring paths were identified, which contained proteins that were up-regulated upon exposure to antichlamydial drugs. Different short listed targets appear to have different propensities for, giving rise to a very important direction to explore in drug discovery.

The study leads to the identification of possible proteins for virulence and drug resistance, providing novel insights to significantly impact the drug discovery process.

5.4 Modeling and docking analysis of Ribonucleotide-diphosphate reductase subunit beta (RNR)

Three-dimensional protein structures are wealth of information for the functional annotation of protein molecules which is difficult to detect in a linear sequence. Though structures are best determined by experimental methods such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, experimental methods cannot always be preferable due to many limitations. In such situations, prediction of the protein 3D structure by theoretical methods using computational resource can result in a reliable model.

Similarity over complete sequence or large sequence fragments facilitates prediction and modeling of total structural domains. The statistics derived from distributions of structural features of known protein structures emphasizes the possibility to predict such features in proteins with unknown protein structures. The accuracy of models of protein structures is adequate for many experimental purposes such as
analysis of point mutation effects, enzymatic reactions, interactions of protein complexes, and active sites for drug design. The *in silico* modeling approach has been useful in saving the time and cost needed for an experimental structure determination process.

Protein 3D structures can be modeled either *ab initio* from protein sequence alone or by comparative methods that depends on a database of known protein structures. *Ab initio* methods are essentially based on the laws of physics, while comparative methods, including comparative modeling, are based primarily on statistical knowledge. However there have been significant improvements in the *ab initio* methods (Madhusudhan et al., 2005). Comparative modeling gives the most accurate results if a known protein structure that is sufficiently similar to the modeled sequence is available.

By comparative modeling, prediction of the protein 3D structure must fulfill the two following criterions. First, the target sequence to be modeled must have a significant similarity to the compared template sequence of known structure. Second, it should be possible to compute a precise alignment between the target sequence and the template structure. The complete structure prediction process involves; a) fold assignment, b) target-template alignment, c) model building, and d) model evaluation.

**RNR Modeling**

In this study, a 3D RNR protein structure was generated through homology modeling. By performing sequence alignment between ribonucleotide-diphosphate reductase subunit beta (*C pneumonia* AR39) with the ribonucleotide-diphosphate reductase subunit beta (*Chlamydia trachomatis*) was executed. It had significant sequence identity of 87%, the alignment, thus obtained is shown in Fig 4.15.
Chapter 5

Note that the extent of the alignment used in subsequent homology modeling did not extend beyond the corresponding termini in 1SYY X-ray structure. The first 21 residues and the last 28 residues of nrdB are missing from the homology model, which thus corresponds to the coil region, which neither code for functional units like alpha helix or a beta strand.

Later the sequence alignment was extended which resulted in identification of other templates 4D8G, 4D8F and 1ANI along with 1SYY which also showed significant sequence identity. The Homology modeling approach was implemented to generate 3D structure of protein using Modeler 9v12 with a graphical user interface Easy Modeler. The three-dimensional structure of *C. pneumonia* AR39 ribonucleotide-diphosphate reductase subunit beta (nrdB) was modeled based on crystallographic coordinates of four bacterial 1SYY, 2ANI, 4D8F and 4D8G proteins.

**Model validation**

The protein model of ribonucleotide-diphosphate reductase subunit beta from *Chlamydia pneumoniae* was inspected for validation using many different criterias. The RMSD analysis of the developed model was evaluated by means of deviation from its templates using SwissPDB viewer (Fig. 4.18). The Ca RMSD and the backbone RMSD deviations for the model and the template crystal structure are 0.36 Å and 0.42 Å respectively.

The stereo chemical quality of the predicted model was evaluated using the Ramachandran plot of phi / psi distribution in the model is developed using RAMPAGE and SwissPDBviewer for checking non-GLY residues at the disallowed regions. Standard bond lengths and bond angles of the model were determined using Verify3D. The analysis revealed RMS Z-scores for bond lengths and bond angles to be 1.40 and 0.32 respectively.
The results showed that ~98% residues of nrdB are in favored region and 2% in allowed region, with 100% of accuracy to that of amino acid non-substituted protein structures. The values obtained suggest high model quality of the protein.

**Protein-ligand docking**

Computational prediction of protein-ligand complexes using docking tools, like AUTODOCK, DOCK, FLEEx, and GOLD (Norgan et al., 2011; Ewing et al., 2001; Kramer et al., 1999; Verdonk et al., 2003), where they use the combination of empirical scoring functions are used to predict ligand interactions in binding sites and binding affinities of ligands to proteins.

Whereas, the binding geometries depend on the docking methods, binding energy estimates rely heavily on the potential functions used to calculate them. Knowledge based potentials follow rules based on statistical analysis of binding affinities and geometries of experimentally determined protein-ligand complexes.

These rules are converted to "pseudo-potentials" which are then applied to score computer generated ligand orientations (Gohlke et al., 2000; Muegge and Martin, 1999). A major concern with such methods is that they might only select for those orientations that have been observed in the crystal structures used to derive the potential.

From our earlier druggability search the identified FDA approved drug molecule Cladribine ligand was docked into the assumed catalytic site binding pocket resulting in relatively high docking scores.
Docking studies have resulted in docking score as high as -5.51 for compound Cladribine. These possible off-target effects give further direction to the study of new RNR inhibitors to discover their role in other oxidation-reduction process. Taken together, the computational tools proved to be useful in many aspects of small molecule discovery processes including exploring the possible mechanism of a bioactive compound, guiding future optimization, or finding possible off-targets of small molecules.