TARGET IDENTIFICATION AND DRUG DESIGN FOR HUMAN PATHOGEN *CHLAMYDOPHILA PNEUMONIAE -IN SILICO* ANALYSIS

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Technology

in

Biochemical Engineering & Biotechnology

By

E. HARIKISHAN REDDY

(20600010)

Department of Chemical Engineering

National Institute of Technology

Rourkela-769008, Orissa, India

2008

TARGET IDENTIFICATION AND DRUG DESIGN FOR HUMAN PATHOGEN *CHLAMYDOPHILA PNEUMONIAE -IN SILICO* ANALYSIS

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Technology

in

Biochemical Engineering & Biotechnology

By

E. HARIKISHAN REDDY

Under the Guidance of

Prof. Gyana R. Satpathy

Dept. of Biotechnology and Medical Engineering

Department of Chemical Engineering

National Institute of Technology

Rourkela-769008, Orissa, India

2008

National Institute of Technology

Rourkela

CERTIFICATE

This is to certify that the thesis entitled, "**Target identification and drug design for human pathogen *Chlamydophila pneumoniae -in silico* analysis**" submitted by Sri E Harikishan Reddy in partial fulfillment of the requirements for the award of Master of Technology in Chemical Engineering with specialization in "**Biochemical Engineering & Biotechnology**" at the National Institute of Technology, Rourkela (Deemed University) is an authentic work carried out by him under my supervision and guidance.

To the best of my knowledge, the matter embodied in the thesis has not been submitted to any other University / Institute for the award of any Degree or Diploma.

Date :

Prof. Gyana R. Satpathy

Dept. of Biotechnology & Medical Engineering,

National Institute of Technology,

Rourkela – 769008.

ACKNOWLEDGEMENT

I express my sincere gratitude and appreciation to many people who helped keep me on track toward the completion of my thesis. Firstly, I owe the biggest thanks to my supervisor and HOD of Biotechnology and Medical Engineering, **Prof. Gyana Ranjan Satpathy**, whose advice, patience, and care boosted my morale. I also thank my guide for providing us with best facilities in the department and his timely suggestions.

I am very much thankful to HOD of Chemical Engineering dept., **Prof. K.C.Biswal**, for his valuable suggestions. I also thank **Prof. B.N.Misra**, ACS-Bioinformatics, Biotech Park, Lucknow for his motivation and guidance during the summer training program.

My special thanks to **Mr.Shadrack Jabes B, Mr.Sripad Chandan Patnaik, Mr.Koteswra Reddy Gujjula** for helping me and providing me good company in the lab. I also thank all my friends, without whose support my life might have been miserable here.

I wish to express my gratitude to my parents and brothers, whose love and encouragement have supported me throughout my education.

E. Harikishan Reddy

Roll No: 20600010

M.Tech(Biochemical Engg. & Biotechnology)

Dept. Chemical Engineering

N.I.T, Rourkela – 769008

Email: hariehkr@gmail.com

CONTENTS

Abstract

Whole genome sequence of the human pathogen *Chlamydophila pneumoniae* and four other strains of same species were analyzed to identify drug targets. Total number 4388 protein coding genes were studied from four strains; in which 3948 genes were having more than 100 amino acids in their coding sequence were selected; we found 147 genes were identified as non-human homologs and conserved proteins among four strains. These non-human homologs genes and their encoding protein were categorized on the basis of the pathways involved in the basic survival mechanisms of the bacterium. Further, MSA of these genes showed eight different types of proteins as a novel drug target to design a drug. The modeled Holliday junction DNA helicase RuvB protein has more appropriate active sites among all other target proteins. Though all chosen drugs bind to Holliday junction DNA helicase RuvB protein, the binding site on the target protein with the minimum binding energy was selected. By using the active site prediction tools, under the optimized conditions we designed a set of antibiotics. Docking was done with the Autodock 4.0 with the different conformations of each ligand. This is the better drug that binds to the active site of target protein and inhibits their activities, which will effects one of the most essential pathways involved in DNA replication, recombination, modification and repair. Therefore, this *in silico* analysis provides rapid and potential approach for identification of drug target and designing of drug.

Keywords: *Chlamydophila pneumoniae*, homology modeling, drug targets, docking, drug design, Holliday junction DNA helicase Ruv-B, MSA (Multiple Sequence Alignment).

LIST OF TABLES

List of Figure

ABBREVIATIONS

BLAST: Basic Local Alignment Search Tool

C.P_AR39: *C. pneumoniae* AR39

C.P_CWL029: *C. pneumoniae* CWL029

C.P_J138: *C. pneumoniae* J138

C.P_TW183: *C. pneumoniae* TW183

C. pneumonia: *Chlamydophila pneumoniae*

CASTp: Computed Atlas of Surface Topography of proteins

DEG: Database Essential Genes

DNA: Deoxyribonucleic acid

EB: Elementary body.

E-value: expectation value

KEGG: Kyoto Encyclopedia of Genes and Genomes

MSA: Multiple sequence alignment.

NCBI: National Center for Biotechnology Information

PASS: Putative Active Sites with Spheres

PDB: Protein Data Bank

Chapter 1

INTRODUCTION

Introduction:

The growing number of microbial genome sequencing projects has generated a large number of sequences. To date, sequence information from approximately 400 complete genomes has been deposited into various public domains, completion of the human genome project has revolutionised the field of drug-discovery against threatening human pathogens. These data pose a major challenge in the post-genomic era, i. e. to fully exploit this treasure trove for the identification and characterization of virulent factors in these pathogens, and to identify novel putative targets for therapeutic intervention [1].

The strategies for drug design and development are progressively shifting from the genetic approach to the genomic approach [2]. Novel drug targets are required in order to design new defense against antibiotic sensitive pathogens. Comparative genomics and bioinformatics provide new opportunities for finding optimal targets among previously unexplored cellular functions based on an understanding of their related biological processes in bacterial pathogens and their hosts.

The genome information is also useful in the identification, validation, selection of the potential candidates and screening based on "essentiality" and "selectivity" criteria of the microbial systems [3]. The target must be essential for the growth, replication, viability or survival of the microorganism, i. e. encoded by genes critical for pathogenic life-stages. The microbial target for treatment should not have any well-conserved homolog in the host, in order to address cytotoxicity issues. This can help to avoid expensive dead-ends later drug discovery process. Genes that are conserved in different genomes often turn out to be essential. A gene is deemed to be essential if the cell cannot tolerate its inactivation by mutation, and its status is confirmed using conditional lethal mutants. A good candidate is a gene essential for bacteria survival, yet cannot be found in the mammalian host [4]. Inactivation of essential genes results in the lethal phenotype in the bacteria [5] and these drugs should function as a ‘magic bullet’ against bacteria. This would help to avoid costly dead-ends when a lead target is identified and investigated in great detail to find all its inhibitors are invariably toxic for humans[6].

The possibilities of selecting targets through genomics-related methodologies are increasing. An interesting approach designated "differential genome display" has been proposed for the prediction of potential drug targets[7,8]. This approach relies on the fact

that genomes of parasitic microorganisms are generally much smaller and encode fewer proteins than the genomes of free-living organisms. The genes that are present in the genome of a parasitic bacterium, but absent in the genome of a closely related free-living bacterium, are therefore likely to be important for pathogenicity and may be considered candidate drug targets. A complementary approach to target identification by bioinformatics was reported in a concordance analysis of microbial genomes. A simple and efficient computational tool was developed that can determine concordances of putative gene products showing sets of proteins conserved across one set of user-specified genomes, but are not present in another set of user-specified genomes. The functions encoded by essential genes are considered to constitute the foundation of life of the organism, and are therefore likely to be common to all cells. Identification and characterization of essential genes for the establishment and/or maintenance of infection may be the basis to elaborate novel and effective antimicrobials against bacteria, especially if these genes are conserved in various bacterial pathogens, suggested searching for drug targets among previously characterized proteins that are specific and essential for a particular pathogen. Recently,compiled a list of all currently available essential genes into the Database of Essential Genes (DEG)[9], which includes the essential genes identified in the genomes of *Mycoplasma genitalium, Haemophilus influenzae, Vibrio cholerae, Staphylococcus aureus, Escherichia coli, Helicobacter pylori, Bacillus subtilis, Mycobacterium tuberculosis H37Rv, Streptococcus pneumoniae, Salmonella typhimurium and Saccharomyces cerevisiae.* Concurrently, the recent availability of the human genome sequence represents a major step in drug discovery. These resources provide a basis for addressing the "complexities and conundrums" in drug discovery by computational methods. The application of a subtractive genomics approach for the identification of essential genes that may be considered as candidates for antibacterial drug discovery, using the completely sequenced bacteria.

Subtractive genomics has been successfully used by authors to locate novel drug targets in *Pseudomonas aeruginosa* [3]. The work has been effectively complemented with the compilation of the Database of Essential Genes (DEG) for a number of pathogenic microorganims. Concurrently, the recent availability of the human genes can eliminate potential drug targets that have close human homologs.

Whole genome sequence of the human pathogen *Chlamydophila pneumoniae* and four other strains of same species were analyzed to identify drug targets. Furthermore, we

were successfully identified a number of promising drug targets, among these targets we are taken DNA helicase RuvB protein for new antibiotic development. DNA helicase RuvB protein involved in the repair of DNA, DNA recombination and in an SOS response. Unknown structure protein can be predicated computationally by using different algorithm. Homology modeling builds on the observation that the three-dimensional structure of proteins is better conserved during evolution than its sequence [10].

Protein performs its function through interaction with other molecules such as substrate, ligand, DNA and other domains of proteins. The three-dimensional structure of protein provides the necessary shape and physicochemical texture to facilitate these interactions. Structural information of protein surface regions enables detailed studies of the relationship of protein structure and function. Specifically, characterization of protein surface regions helps to analyze enzyme mechanism, to determine binding specificity and to plan mutation studies. It can also help to identify the biological roles of newly solved protein structures with an unknown function.

The identification and visualization of protein cavities is the starting point for many structure-based drug design (SBDD) applications. Sites of activity in proteins usually lie in cavities, where the binding of a substrate typically serves as a mechanism for triggering some event, such as a chemical modification or conformational change. Consequently, binding sites are often targeted in attempts to interrupt molecular processes via therapeutics. Although binding site locations are often furnished by x-ray data or fold recognition, tools that automatically predict these locations have become quite popular in SBDD, especially as front-ends to molecular docking or when alternate binding sites are sought [11, 12]. The size and shape of protein cavities dictates the three-dimensional geometry of ligands that can strongly bind there; i.e. they must fit like a hand in glove. Thus, a minimal requirement for drug activity is that the molecule sterically fit the region of buried volume inscribing the active site cavity, with some allowance for induced fit. The determination and visualization of these volumes is critical in drug design, particularly since manual intervention is still fruitfully employed in most design scenarios. An ordinary stick representation of a protein, unfortunately, provides little insight regarding the location, shape, or size of its buried volumes. While surface representations [13, 14] are a step in the right direction, they still fall short in that they require the user to infer buried volumes from often-occluded void space. Consequently, methods for direct display of regions of buried volume in proteins have become prevalent in recent years

[15]. Moreover, as molecular docking and virtual screening become more predictive and prevalent, the possibility of interfacing such tools with functional genomics via threading or homology modeling becomes increasingly tempting. A versatile tool like PASS, CASTp that can rapidly predict binding sites should, therefore, find a niche as a front-end to such automated screening efforts.

Protein-ligand docking methods aim to predict the binding energy of the protein-ligand complex given the atomic coordinates. Recent improvements in search algorithms and energy functions, computational docking methods have become a valuable tool to probe the interaction between protein and its inhibitors. The interaction energy between the protein and its ligand is calculated by a simplified, often grid-based force field [16]. Generally various docking methods followed by various energy scoring functions. Basic components may include steric and electrostatic energies, sometimes supplemented by other terms accounting for hydrogen bonding and salvation effects. Gibbs free energy of binding is ΔG then related to the binding constant by $\Delta G_0 = -RT \ln K_i$ At best, ΔG is determined by statistical thermodynamics resulting in a master equation that considers all contributing effects.

This can be written out conceptually by the following equation.

$$\Delta G_{Binding} = \Delta G_{Motion} + \Delta G_{Interaction} + \Delta G_{Solvent} + \Delta G_{Cofiguration}$$

The accurate prediction of enzyme-substrate interaction energies is one of the major challenges in computational biology.

Active sites of a protein are key factor for the flexible docking. Autodock4.0 [17]. is an automated docking tool that was designed to predict how small molecules bind to receptor of known 3D structure and it also optionally enables to model Binding parameters of ligand with number of distinct conformational clusters and to find all possible minimum binding energy.

Chapter 2

LITERATURE REVIEW

2.1 Chlamydia Pneumoniae:

Chlamydia pneumoniae is a common obligate intracellular bacterium that causes upper and lower respiratory infections worldwide [18]. In addition to acute infections, several chronic inflammatory diseases have been presumptively associated with *C. pneumoniae* infection. Increasing evidence implicates that a persistent lung infection caused by *C. pneumoniae* may contribute to the initiation, exacerbation and promotion of asthma symptoms. A causal association between *C. pneumoniae* infection and asthma is biologically plausible based on the observations that asthma is a chronic inflammatory disease of the airways, and that Chlamydia are known to produce chronic inflammatory damage in target organs. Whether *C. pneumoniae* lung infections activate the same immunopathologic mechanisms that have been demonstrated for other chlamydial diseases has not been explored systematically. Chlamydophila pneumoniae also infects and causes disease in Koalas, emerald tree boa (Corallus caninus), iguanas, chameleons, frogs and turtles.

C pneumoniae. causes infection approximately 50% of young adults and 75% of elderly persons have serological evidence of previous infection. The pathogen is estimated to cause 3-10% of community-acquired pneumonia cases among adults. The estimated number of cases of *C pneumoniae* pneumonia is 300,000 cases per year[19].

C. pneumoniae infection has also been linked with atherosclerosis — another chronic inflammatory disease. Since then, a large number of seroepidemiological studies have confirmed these findings. The actual presence of *C. pneumoniae* in atherosclerotic lesions has also been demonstrated in a number of studies and by various methods. Moreover, the presence of *C. pneumoniae*-specific T lymphocytes in atherosclerotic tissue specimens suggests that *C. pneumoniae* participates in the maintenance of the inflammatory response in the tissue and may thus be involved in the progression of the disease. In experimental animals, *C. pneumoniae* infection has been found to induce inflammatory changes and calcified lesions containing *Chlamydia* and to accelerate the development of atherosclerosis.

Table1: The complete genome information about C. pneumoniae strains from NCBI.

Tax name	**C.P_AR39**	**C.P_CWL029**	**C.P_J138**	**C.P_TW183**
Accession	NC_002179.2	NC_000922.1	NC_002491.1	NC_005043.1
GI	58021288	15617929	15835535	33241335
Tax id	115711	115713	138677	182082
Genome_ID	154	140	160	311
DNA length	1229853	1230230	1226565	1225935
Genetic Code	11	11	11	11
Publications	10684935	10192388	10871362	10452345
Protein count	1112	1052	1069	1113
CDS count	1112	1052	1069	1113
RNA count	41	43	41	41
Gene count	1167	1122	1110	1155
Others	0	2	0	2
Total	2320	2219	2220	2311

2.1.1 Life cycle of Chlamydophila pneumoniae

Chlamydophila pneumoniae is a small bacterium (0.5 micrometres) that undergoes several transformations during its life cycle. It exists as an elementary body (EB) in between hosts. The EB is not biologically active but is resistant to environmental stresses and can survive outside of a host. The EB travels from an infected person to the lungs of a non-infected person in small droplets and is responsible for infection. Once in the lungs, the EB is taken up by cells in a pouch called an endosome by a process called phagocytosis. However, the EB is not destroyed by fusion with lysosomes as is typical for phagocytosed material. Instead, it transforms into a reticulate body and begins to replicate

within the endosome. The reticulate bodies must utilize some of the host's cellular machinery to complete its replication. The reticulate bodies then convert back to elementary bodies and are released back into the lung, often after causing the death of the host cell. The EBs are thereafter able to infect new cells, either in the same organism or in a new host. Thus, the life cycle of Chlamydophila pneumoniae is divided between the elementary body which is able to infect new hosts but cannot replicate and the reticulate body which replicates but is not able to cause new infection.

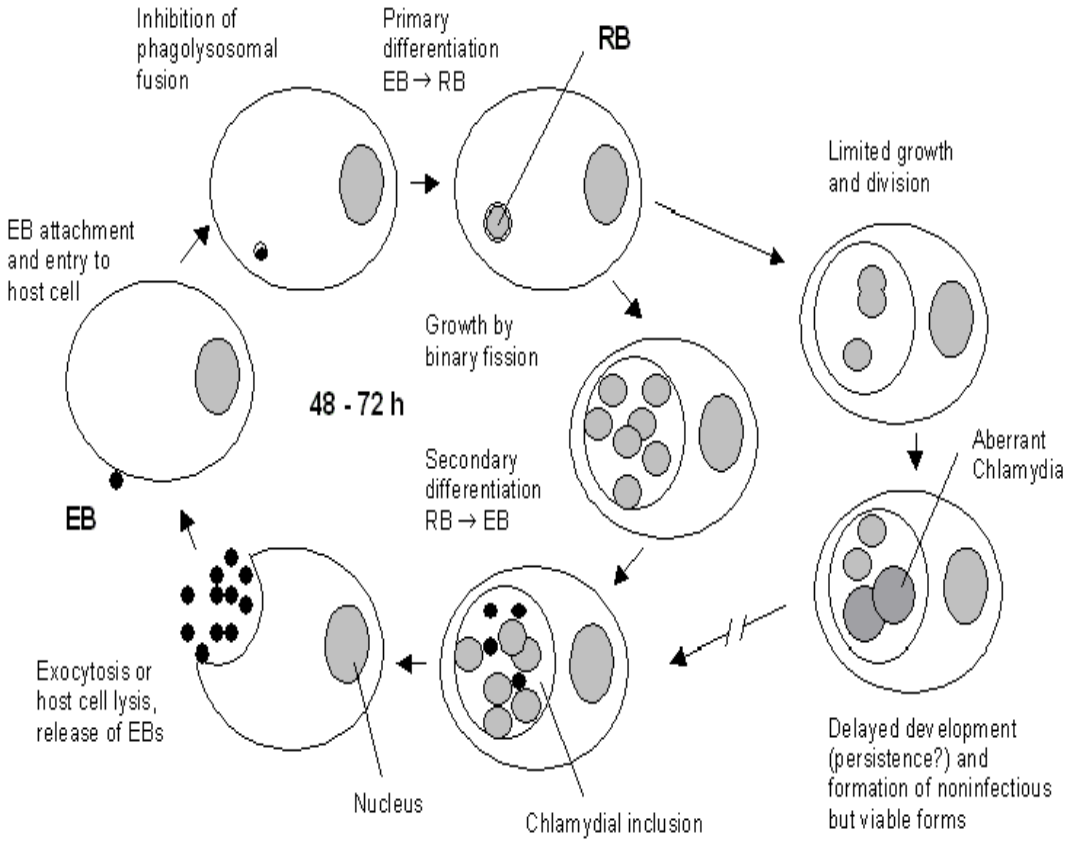


Fig.1: Developmental life cycle of *Chlamydia*

2.1.2 Pneumonia caused by Chlamydophila pneumoniae

Chlamydophila pneumoniae is a common cause of pneumonia around the world. Chlamydophila pneumoniae is typically acquired by otherwise healthy people and is a form of community-acquired pneumonia. Because treatment and diagnosis are different from historically recognized causes such as Streptococcus pneumoniae, pneumonia caused by Chlamydophila pneumoniae is categorized as an "atypical pneumonia."

2.1.3 Symptoms and diagnosis

Symptoms of infection with Chlamydophila pneumoniae are indistinguishable from other causes of pneumonia. These include cough, fever, and difficulties breathing. Chlamydophila pneumoniae more often causes pharyngitis, laryngitis, and sinusitis than other causes of pneumonia; however, because many other causes of pneumonia results in these symptoms, differentiation is not possible. Likewise, a physical examination by a health provider does not typically provide information which allows for a definite diagnosis.

Diagnosis of Chlamydophila pneumoniae may be confounded by prior infections with this microorganism. Examination of sputum or the secretions of the respiratory tract may reveal signs of the bacteria. Otherwise, examination of the blood may reveal antibodies against the bacteria. If there has been a prior infection, this may have resulting in pre-existing antibodies. Therefore, interpretation may require a period of six weeks in order to reanalyze the antibodies and to determine whether the infection was new or old. Examination of the blood may also show proteins (antigens) from Chlamydophila pneumoniae, either through direct fluorescent antibody testing, enzyme-linked immunosorbent assay (ELISA), or polymerase chain reaction (PCR).

Chest x-rays of lungs infected with Chlamydophila pneumoniae often show a small patch of increased shadow (opacity). However, many different patterns are common and there is no appearance which allows for a specific diagnosis.

2.1.4 Treatment and prognosis

Typically, treatment for pneumonia is begun before the causative microorganism is identified. This empiric therapy includes an antibiotic active against the atypical bacteria, including Chlamydophila pneumoniae. The most common type of antibiotic used is a macrolide such as azithromycin or clarithromycin. If testing reveals that Chlamydophila pneumoniae is the causative agent, therapy may be switched to doxycycline, which is slightly more effective against the bacteria. Sometimes a quinolone antibiotic such as levofloxacin may be started empirically. This group is not as effective against Chlamydophila pneumoniae. Treatment is typically continued for ten to fourteen days for known infections.

Prognosis of pneumonia caused by Chlamydophila pneumoniae is excellent. Hospitalization is uncommon, complications are rare, and most people have no residual deficits. In fact, Chlamydophila pneumoniae is a common cause of walking pneumonia, so named because most people are able to continue to walk and participate in reduced activity during infection.

2.1.5 **Epidemiology and prevention**

Chlamydophila pneumoniae affects all age groups and is most common among the 60-79 year old age group. Reinfection is common after a short period of immunity. The incidence is one case out of one thousand per year and causes ten percent of community-acquired pneumonias treated without hospitalization.[citation needed] As of 2005, there are no vaccines or other ways to prevent infection other than good hygiene and healthy eating as well as active lifestyle some people with obesity face the same symptoms, a stress free life as well as active and conscious living are the best viral and physical prevention known.

2.2 Protein structure prediction and modeling

The ultimate goal of protein modeling is to predict a structure from its sequence with an accuracy that is comparable to the best results achieved experimentally. Most attempts to predict protein structure from basic physical principles alone try to reproduce the interatomic interactions in proteins, to define a computable energy associated with any conformation. Computationally, the problem of protein structure prediction then becomes a task of finding the global minimum of this conformational energy function. So far this approach has not succeeded, partly because of the inadequacy of the energy function and partly because the minimization algorithms tend to get trapped in local minima.

Other approaches to structure prediction are based on attempts to simplify the problem, to capture somehow the essentials. The alternative to a priori methods are approaches based on assembling clues to the structure of a target sequence by finding similarities to known structures. These empirical or 'knowledge-based' methods are becoming very powerful[20].

2.2.1 Protein structure prediction Methods

Methods for prediction of protein structure from amino acid sequence include:

1) Attempts to predict secondary structure without attempting to assemble these regions in three- dimensions. The results are lists of regions of the sequence predicted to form α-helices and regions predicted to form strands of β-sheet.

2) Homology modelling: prediction of the three-dimensional structure of a protein from the known structures of one or more related proteins. The results are a complete coordinate set for mainchain and sidechains, intended to be a high-quality model of the structure, comparable to at least a low-resolution experimental structure.

3) Fold recognition: given a library of known structures, determine which of them shares a folding pattern with a query protein of known sequence but unknown structure. If the folding pattern of the target protein does not occur in the library, such a method should recognize this. The results are a nomination of a known structure that has the same fold as the query protein, or a statement that no protein in the library has the same fold as the query protein.

4) Prediction of novel folds, either by a priori or knowledge-based methods. The results are a complete coordinate set for at least the mainchain and sometimes the sidechains also. The model is intended to have the correct folding pattern, but would not be expected to be comparable in quality to an experimental structure.

2.2.2 Homology modelling

Model-building by homology is a useful technique when one wants to predict the structure of a target protein of known sequence, when the target protein is related to at least one other protein of known sequence and structure. If the proteins are closely related, the known protein structures - called the parents - can serve as the basis for a model of the target. Although the quality of the model will depend on the degree of similarity of the sequences, it is possible to specify this quality before experimental testing. In consequence, knowing the quality of the model required for the intended application permits intelligent prediction of the probable success of the exercise.

2.2.3 Steps in Homology Modelling

1. Align the amino acid sequences of the target and the protein or proteins of known structure. It will generally be observed that insertions and deletions lie in the loop regions between helices and sheets.

2. determine mainchain segments to represent the regions containing insertions or deletions. Stitching these regions into the mainchain of the known protein creates a model for the complete mainchain of the target protein.

3. Replace the sidechains of residues that have been mutated. For residues that have not mutated, retain the sidechain conformation. Residues that have mutated tend to keep the same sidechain conformational angles, and could be modelled on this basis. However, computational methods are now available to search over possible combinations of sidechain conformations.

4. Examine the model - both by eye and by programs - to detect any serious collisions between atoms. Relieve these collisions, as far as possible, by manual manipulations.

5. Refine the model by limited energy-minimization. The role of this step is to fix up the exact geometrical relationships at places where regions of mainchain have been joined together, and to allow the sidechains to wriggle around a bit to place themselves in comfortable positions. The effect is really only cosmetic - energy refinement will not fix serious errors in such a model.

In a sense, this procedure produces 'what you get for free' in that it defines the model of the protein of unknown structure by making minimal changes to its known relative. Unfortunately, it is not easy to make substantial improvements. A rule of thumb is that if the two sequences have at least 40–50% identical amino acids in an optimal alignment of their sequences, the procedure described will produce a model of sufficient accuracy to be useful for many applications. If the sequences are more distantly related, neither the procedure described nor any other currently available method will produce a model, correct in detail, of the target protein from the structure of its relative.

In most families of proteins the structures contain relatively constant regions and more variable ones. The core of the structure of the family retains the folding topology, although it may be distorted, but the periphery can entirely re-fold. A single parent

structure will permit reasonable modelling of the conserved portion of the target protein, but will fail to produce a satisfactory model of the variable portion. Moreover, it will not be easy to predict which are the variable and constant regions. A more favourable situation occurs when several related proteins of known structure can serve as parents for modelling a target protein. These reveal the regions of constant and variable structure in the family. The observed distribution of structural variability among the parents dictates an appropriate distribution of constraints to be applied to the model.

2.3 Structure Analysis and active site prediction

The stereochemical validation of model structures of proteins is an important part of the comparative molecular modeling process. Firstly, the selection of high quality structures for inclusion in loop dictionaries is important for the simple reason that these coordinate sets will be used to build future models. Secondly, the structural evaluation of comparative modeling output must be used to identify possible problematic regions [21].

There is some measurements are good indicators of stereochemical quality; these include planarity; chirality; phi/psi preferences; chi angles; non-bonded contact distances; unsatisfied donors and acceptors.

Protein performs its function through interaction with other molecules such as substrate, ligand, DNA and other domains of proteins. The three-dimensional structure of protein provides the necessary shape and physicochemical texture to facilitate these interactions. Structural information of protein surface regions enables detailed studies of the relationship of protein structure and function. Specifically, characterization of protein surface regions helps to analyze enzyme mechanism, to determine binding specificity and to plan mutation studies. It can also help to identify the biological roles of newly solved protein structures with an unknown function

2.3.1 Ramachandran plot

The Sasisekharan-Ramakrishnan-Ramachandran plot describes allowed mainchain conformations. A Ramachandran plot is a way to visualize dihedral angles φ against ψ of amino acid residues in protein structure. It shows the possible conformations of φ and ψ angles for a polypeptide [22].

A fragment of the linear polypeptide chain common to all protein structures is shown in Fig.2 Rotation is permitted around the N-Cα and Cα-C single bonds of all

residues (with one exception: proline). The angles φ and ψ around these bonds, and the angle of rotation around the peptide bond, ω, define the conformation of a residue. The peptide bond itself tends to be planar, with two allowed states: trans, ω ≈ 180° (usually) and cis, ω ≈ 0° (rarely, and in most cases at a proline residue). The sequence of φ, ψ and ω angles of all residues in a protein defines the backbone conformation.

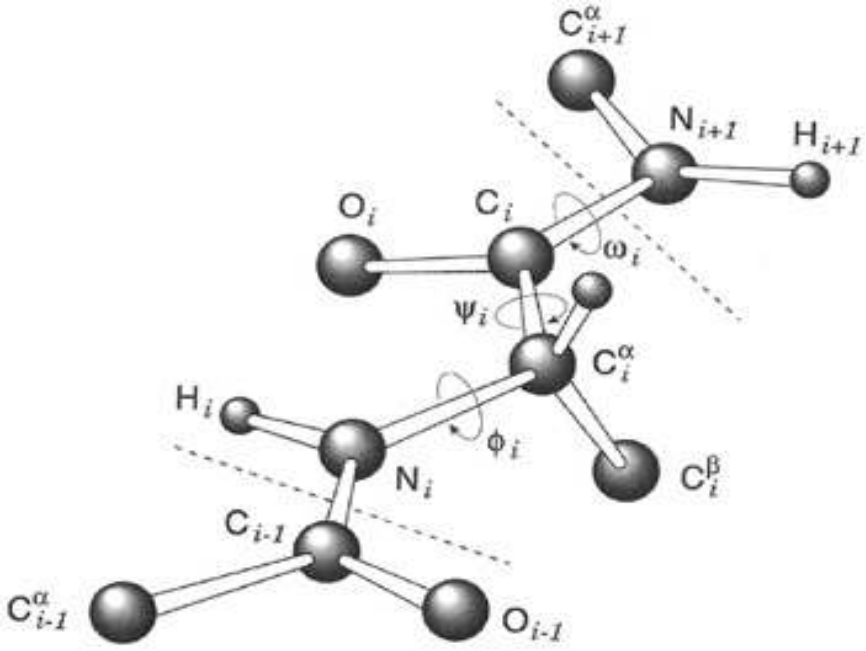


Fig.2**:** Definition of conformational angles of the polypeptide backbone.

The principle that two atoms cannot occupy the same space limits the values of conformational angles. The allowed ranges of φ and ψ, for ω = 180°, fall into defined regions in a graph called Sasisekharan- Ramakrishnan-Ramachandran plot - usually shortened to 'Ramachandran plot' (see Fig. 3). Solid lines in the figure delimit energetically-preferred regions of φ and ψ; broken lines in the figure delimit sterically-disallowed regions. The conformations of most amino acids fall into either the αR or β regions. Glycine has access to additional conformations. In particular it can form a left-handed helix: αL. Fig. 3 shows the typical distribution of residue conformations in a well-determined protein structure. Most residues fall in or near the allowed regions, although a few are forced by the folding into energetically less-favourable states.

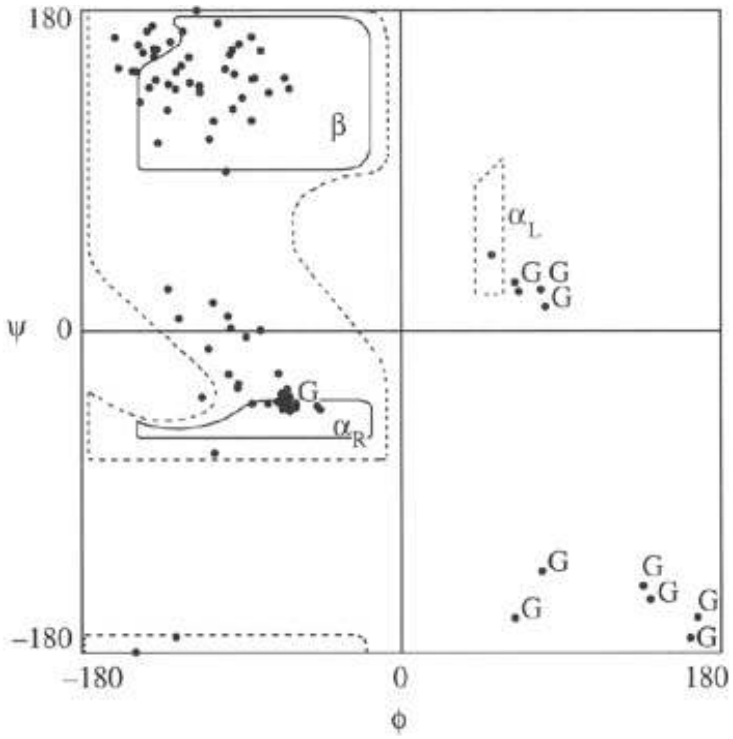


Fig.3**:** A Sasisekharan-Ramakrishnan-Ramachandran plot of acylphosphatase (PDB code 2ACY). Note the clustering of residues in the α and β regions, and that most of the exceptions occur in Glycine residues (labeled G).

The allowed regions generate standard conformations. A stretch of consecutive residues in the α conformation (typically 6–20 in native states of globular proteins) generates an α-helix. Repeating the β conformation generates an extended β-strand. Two or more β-strands can interact laterally to form β-sheets. Helices and sheets are 'standard' or 'prefabricated' structural pieces that form components of the conformations of most proteins. They are stabilized by relatively weak interactions, *hydrogen bonds,* between mainchain atoms. In some fibrous proteins all of the residues belong to one of these types of structure: wool contains α-helices; silk β-sheets.

Typical globular proteins contain several helix and/or sheet regions, connected by *turns.* Usually the ends of helix or strand regions appear on the surface of a domain of a protein structure. They are connected by turns, or loops: regions in which the chain alters direction to point back into the structure. Many but not all turns are short, surface-exposed regions that tend to contain charged or polar residues. Interactions involving sidechains must determine the mainchain conformation.

2.3.2 Sidechains

Sidechains offer the physicochemical versatility required to generate all the different folding patterns. The sidechains of the twenty amino acids vary in:

- **Size:** The smallest, glycine, consists of only a hydrogen atom; one of the largest, phenylalanine, contains a benzene ring.
- **Electric charge:** Some sidechains bear a net positive or negative charge at normal pH. Asp and Glu are negatively charged, Lys and Arg are positively charged. (Charged residues of opposite sign can form attractive pairwise interactions called salt bridges.)
- **Polarity:** Some sidechains are polar; they can form hydrogen bonds to other polar sidechains, or to the mainchain, or to water. Other sidechains are electrically neutral. Some of these contain chemical groups related to ordinary hydrocarbons such as methane or benzene. Because of the thermodynamically unfavourable interaction of hydrocarbons with water, these are called 'hydrophobic' residues, is an important contribution to protein stability.

Chapter 3

TOOLS FOR THE STUDY

3.1 Sequence alignments Tools

Sequence alignments provide a powerful way to compare novel sequences with previously characterized genes. Both functional and evolutionary information can be inferred from well designed queries and alignments. BLAST (Basic Local Alignment Search Tool), provides a method for rapid searching of nucleotide and protein databases. Since the BLAST algorithm detects local as well as global alignments, regions of similarity embedded in otherwise unrelated proteins can be detected. Both types of similarity may provide important clues to the function of uncharacterized proteins.

5 different versions of BLAST are

BLASTn –compares a nucleotide query sequence against a nucleotide sequence database.

BLASTp- compares a amino acid query seq. against a protein seq. database.

BLASTx- compares a 6 framed conceptual translation product of a nucleotide query seq. against the protein seq. database(ultimately protein is compared)

tBLASTn- compares a protein query seq. against a nucleotide seq. database. Dynamically translated in all the 6 reading frames.

tBLASTx- compares 6 framed(6 forms) translation of a nucleotide query seq. against 6 framed translation of a nucleotide seq. database.

3.2 On Line Homology Modelling Softwares

Swiss model (www.expasy.ch/swissmod/ SWISS-MODEL.html) is a fully automated protein structure Homology Modeling server. It has a first approach mode that helps performs Homology Modeling. The user has to enter his / her email id and input the protein sequence in Fasta format. It allows the user to choose the BLAST limit for template selection. It can search the pdb file from the pdb database with the user providing the name of the pdb file or the user can upload his / her own pdb file. The output file is a pdb file that is returned to the user's email address. The result can be forwarded by Swiss Model to PHD Secondary structure prediction at Columbia University and Fold Recognition Server (3D-pssm) of the ICRF. Swiss Model however does not accept the sequences for homology modelling when similarity is less than 25%[23].

Geno3D (http://geno3d-pbil.ibcp.fr) performs Comparative protein structure Modeling by spatial restraints (distances and dihedral) satisfaction. Geno3D is most frequently used for Homology or Comparative protein structure Modeling.Geno3d accepts input similar to Fasta format but only the one letter code has to be used. The result is obtained in the pdb format that can be viewed in any Molecular Modeling software.Geno3d offers many other features, it allows the user to select PDB entries as templates for Molecular Modeling after

a 3 step iterative PSI BLAST. It presents the output for each template, along with the secondary structure prediction, displays percent of agreement in secondary structure and repartition of information from template on query sequence. The output link is sent to the user's email address. It also notifies the user when it's server begins the Homology Modeling. It has an option where the user can decide how many models to generate. The main idea behind having more than one model generated is that the user may have a better flexibility and understanding. It also returns a superimposed pdb file which has the models superimposed on each other. This is one of the good points in Geno3d as it allows us to compare the various models generated in one window. All the results obtained can be downloaded as a archive.tar.Z that can be opened in WinZip in windows and in UNIX or Linux platforms. So the user does not have to save results in webpage effect or in a document file. It also displays the Ramachandran plot in the result[24].

CPHmodels Automated neural-network based protein modeling server (Http://www.cbs.dtu.dk/services/CPHmodels/). CPHmodels is a collection of databases and methods developed to predict protein structure. It performs prediction of protein structure using Comparative Modeling. It does not accept more than 900 amino acids in the input sequence. The sequences are kept confidential and are deleted after processing. This program did not give me appropriate results. The error it displayed was similar to the one displayed by Swiss Model[25].

3.3 Offline Homology Modelling Software:

MODELLER is used for homology or comparative modeling of protein three-dimensional structures. It is built in FORTRAN. It will runs on python script file commands. Modeller is most frequently used for homology or comparative protein structure modeling. Modeller helps determine the spatial restraints from the templates. It generates a number of 3D models of the sequence you submit satisfying the template restraints. MODELLER automatically calculate a full-atom model. MODELLER models protein 3D structure keeping in the constraints of spatial restraints. The restraints can be derived from a number of different sources. These include NMR experiments (NMR refinement),cross-linking experiments, fluorescence spectroscopy, rules of secondary structure packing (combinatorial modeling),image reconstruction in electron microscopy, homologous structures (comparative modeling),site-directed mutagenesis, residue-residue

and atom-atom potentials of mean force, etc. Modeller is not an automated homology modelling tool[26].

It is a very specific program. Any error in the format of the sequence alignment prevents the modeller from performing Homology Modeling. The program is very specific about the extension names of the file formats used for Homology Modeling. It is a very reliable program and it allows the user to specify what he wants in the end result. Modeller runs on platforms like Win XP, Linux, Sun Solaris and Macintosh.

DeepView - Swiss-PdbViewer is an application that provides a user friendly interface allowing to analyze several proteins at the same time. The proteins can be superimposed in order to deduce structural alignments and compare their active sites or any other relevant parts. Amino acid mutations, H-bonds, angles and distances between atoms are easy to obtain thanks to the intuitive graphic and menu interface. DeepView - Swiss-PdbViewer has been developped by Nicolas Guex (GlaxoSmithKline R&D). Swiss-PdbViewer is tightly linked to SWISS-MODEL, an automated homology modeling server developed within the Swiss Institute of Bioinformatics (SIB) at the Structural Bioinformatics Group at the Biozentrum in Basel[27].

3.4 Structure Analysis and Verification Server

PROCHEK Checks the stereochemical quality of a protein structure by analyzing residue-by-residue geometry and overall structure geometry. It is tell about: Covalent geometry, Planarity, Dihedral angles, Chirality, Non-bonded interactions[28].

WHAT_CHEK derived from a subset of protein verification tools from the WHATIF program; this does extensive checking of many sterochemical parameters of the residues in the model[29].

DOPE: The DOPE model score is designed for selecting the best structure from a collection of models built by MODELLER. DOPE uses the standard MODELLER energy function.

ERRAT is a protein structure verification algorithm that is especially well-suited for evaluating the progress of crystallographic model building and refinement. The program works by analyzing the statistics of non-bonded interactions between different atom types. A **single** output plot is produced that gives the value of the error function *vs.* position of a 9-residue sliding window. By comparision with statistics from highly refined structures,

the error values have been calibrated to give **confidence** limits. ERRAT will give an "overall quality factor" and if it is a high 90% range protein structure is good. This is extremely useful in making decisions about reliability[30].

VERIFY_3D determines the compatibility of an atomic model (3D) with its own amino acid sequence (1D) by assigned a structural class based on its location and environment (alpha, beta, loop, polar, nonpolar etc) and comparing the results to good structures. Then a database generated from vetted good structures is used to obtain a score for each of the 20 amino acids in this structural class. For each residue, the scores of a sliding 21-residue window (from -10 to +10) are added and plotted[31].

PROVE Calculates the volumes of atoms in macromolecules using an algorithm which treats the atoms like hard spheres and calculates a statistical Z-score deviation for the model from highly resolved (2.0 Å or better) and refined (R-factor of 0.2 or better) PDB-deposited structures[32].

3.5 Protein Active Site Prediction Tools

CASTp: Computed Atlas of Surface Topography of proteins (http://cast.engr.uic.edu.) provides an online resource for locating, delineating and measuring concave surface regions on three-dimensional structures of proteins. These include pockets located on protein surfaces and voids buried in the interior of proteins. The measurement includes the area and volume of pocket or void by solvent accessible surface model and by molecular surface model, all calculated analytically. CASTp can be used to study surface features and functional regions of proteins. CASTp includes a graphical user interface, flexible interactive visualization, as well as on the- fly calculation for user uploaded structures [33].

PASS: Putative Active Sites with Spheres is a simple computational tool that uses geometry to characterize regions of buried volume in proteins and to identify positions likely to represent binding sites based upon the size, shape, and burial extent of these volumes[34].. PASS'S utility as a predictive tool for binding site identification is tested by predicting known binding sites of proteins in the PDB using both complexed macromolecules and their corresponding apo-protein structures. The results indicate that PASS can serve as a front-end to fast docking. The main utility of PASS lies in the fact that it can analyze a moderate-size protein (~ 30 kD) in under twenty seconds, which

makes it suitable for interactive molecular modeling, protein database analysis, and aggressive virtual screening efforts.

As a modeling tool, PASS

(i) Rapidly identifies favorable regions of the protein surface,

(ii) Simplifies visualization of residues modulating binding in these regions, and

(iii) Provides a means of directly visualizing buried volume, which is often inferred indirectly from curvature in a surface representation.

PASS produces output in the form of standard PDB files, which are suitable for any modeling package, and provides script files to simplify visualization.

3.6 Docking Tools

Autodock 4.0

Autodock is used to perform computational molecular docking of small molecules to proteins, DNA, RNA and other important macromolecules, by treating the ligand and selected parts of the target as conformationally flexible. It uses a scoring function based on the AMBER force field, and estimates the free energy of binding of a ligand to its target. Novel hybrid global-local evolutionary algorithms are used to search the phase space of the ligand-macromolecule system.

The introduction of Autodock 4 comprises three major improvements:

1. The docking results are more accurate and reliable.
2. It can optionally model flexibility in the target macromolecule.
3. It enables Autodesk's use in evaluating protein-protein interactions.

Autodock 4 offers many new features and improvements over previous versions. The most significant is that it models flexible side chains in the protein. We can get both the 3D structure and the inhibition constants.

AutoDock4 scoring functions are van der Waals forces, Hydrogen Bonding, Electrostatics, Desolvation, Torsional.

Binding energy=Intermolecular energy +Torsional energy

$\Delta G_{bind} = \Delta G_{vdw} + \Delta G_{ele.} + \Delta G_{H\text{-}bond} + \Delta G \text{ desolv} + \Delta G_{tors}$

Here ΔG=change in free energy

Chapter 4

MATERIALS AND METHODS

MATERIALS AND METHODS:

4.1 Identification Novel Drug Targets:

Whole genome sequences were downloaded for *Chlamydophila pneumoniae* (*C. pneumoniae* AR39, *C. pneumoniae* J138, *C. pneumoniae* TW1839, *C. pneumoniae* CWl029) [35].from National Center for Biotechnology Information (NCBI) center [ftp://ftp.ncbi.nlm.nih.gov/genomes/bacteria/]. The strains having a circular genome with 1052-1112 predicted protein coding sequences. From the complete genome sequence data, the genes that code for proteins whose sequence were greater than 100 amino acids were selected out. These selected genes were subjected to BLASTX (parameter Matrix: BLOSUM62, Gap Penalties: Existence-11, Extension-1) against the DEG (**http://tubic.tju.edu.cu.deg**). A random expectation value (*E*-value) cut-off of 10 ^100 and a minimum bit-score cut-off of 100 was used to screen out genes that appeared to represent essential genes [36]. The screened essential genes of *Chlamydophila pneumoniae* were thus subjected to BLASTX against the human genome. The homologs were excluded and the list of non-homologs was compiled. The identified genes were than classified into different groups based on gene name and biological function, with the Swiss-Prot Protein Database (http://us.expasy.org/sprot), KEGG database[37]. The classified genes with same function were further analyzed to find homologs conserved genes with in all four *C. pneumoniae* strains.

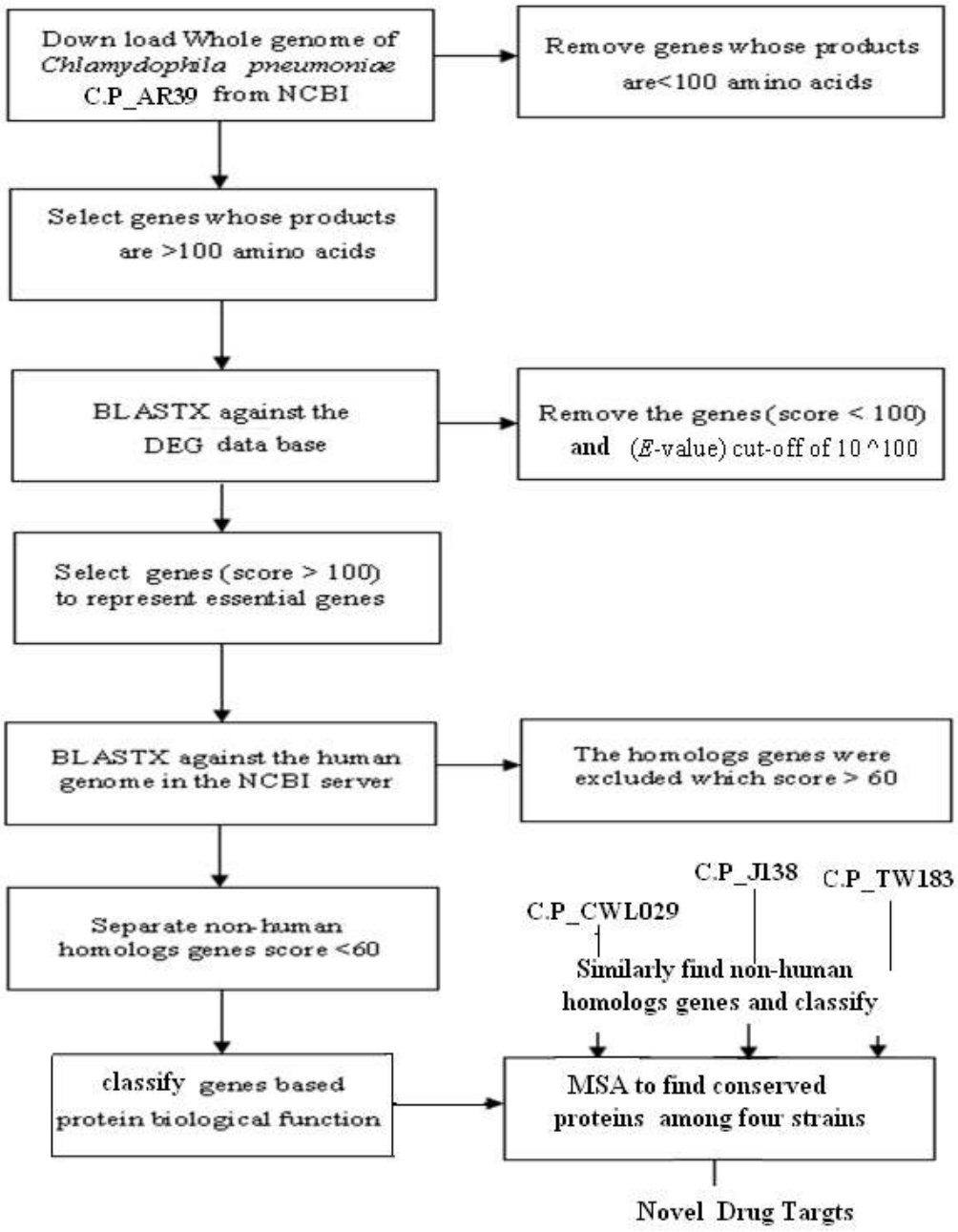


Fig.4: Flowchart for identification of novel drug targets.

4.2 Homology Modeling:

The homologs conserved protein coding sequence *i.e* DNA helicase Ruv-B was selected from *C. pneumoniae* strains for drug target [38, 39]. The three-dimensional structure of DNA helicase Ruv-B protein was modeled by considering the suitable well studied template proteins structure were identified by similarity search with the BLAST tool against the

protein databank. The homology modeling is done online software like Geno3D, Swiss model, CPHmodels by using different parameter. And offline homology modeling is done using deep view , priory the modeled protein was refined by the MODELER 9v2. The model was validated for the 3D-1D profile with VERIFY3D, and the stereochemical qualities were checked with PROCHECK, Errat, Prove and WHAT_IF (http://nihserver.mbi.ucla.edu/SAVS/). Finally, the structural properties of the target protein were validated by using the Ramachandran plot score. The different software models are compared with each other final best model is selected; it is used for further drug design process.

4.3 Active site identification:

Active sites of the target protein (DNA helicase Ruv-B) were predicted by using tools like PASS, CASTp which would be the key factor for the flexible docking. This provides resource for locating, delineating and measuring concave surface regions on three-dimensional structures of proteins. These include pockets located on protein surfaces and voids buried in the interior of proteins that are frequently associated with binding events. In addition, it measures the size of mouth openings of individual pockets, for better accessibility of binding sites to various ligands and substrates.

4.4 Ligand optimization and Docking:

Optimization of leads was done based on the Lipinski rule of five [40]. The ligand 3D structure is minimized with ACD labs chemsk 10.0(www.acdlabs.com)[41]. These optimized ligands are used to find its respective interactions with the target protein. The docking of the ligands with the target protein was done by using the Autodock 4.0. Prepare files like pbdqt for Ligands and Protein, map files for protein. Generate Grid box near to binding site of protein. Choose the Lamarckian genetic algorithm to search for the best conformers [42]. During the docking process, the docking parameters was set to, Maximum Number of GA runs 100, Population size of 150, Maximum number of evaluation 250000, Rate of Gene mutation 0.02 for each Compound. The parameters were set using the software Autodock Tools. The Calculations of Autogrid and Autodock were performed on Linux operating system having system properties (Intel(R) Pentium(R) D CPU 2.80GHz, 2.0 GB of RAM).

Chapter 5

RESULTS AND DISCUSSION

RESULTS AND DISCUSSION:

5.1 Identification of novel drug targets:

Whole genome sequence of the human pathogen *C. pneumoniae* and four other strains of same species were analyzed to identify drug targets. Total number 4388 protein coding genes were studied from four strains (*C. pneumoniae* AR39, *C. pneumoniae* J138, *C. pneumoniae* TW1839, *C. pneumoniae* CWl029) *via* an *in silico* genomic approach.; in which 3948 genes were having more than 100 amino acids in their coding sequence were selected; this was on the assumption that proteins less than 100 amino acids known to able to affect the catalytic activity of proteins and participate in protein complex formation which affect their enzyme activity [43].

Table 2: Computational results of *Chlamydophila pneumoniae.*

	Chlamydophila pneumoniae			
	C.P_AR39	C.P_CWL029	C.P_J138	C.P_TW183
Total number of protein coding genes	1112	1052	1069	1155
Genes where products are > 100 amino acids	961	970	977	1040
Genes where products are <100 amino acids	151	82	92	115
Essential genes having non-human homologs	31	47	35	34

5.2 Sequence analysis Sample results:

The query gene sequences BLASTX against DEG(database for essential genes) with blastX with parameter Matrix: BLOSUM62,Gap Penalties: Existence: 11, Extension: 1

DEG BLASTX output:

```
Query= ref|NC_002179.2|:201-1199
         (999 letters)
Database: deg.aa
           4509 sequences; 1,713,232 total letters
```

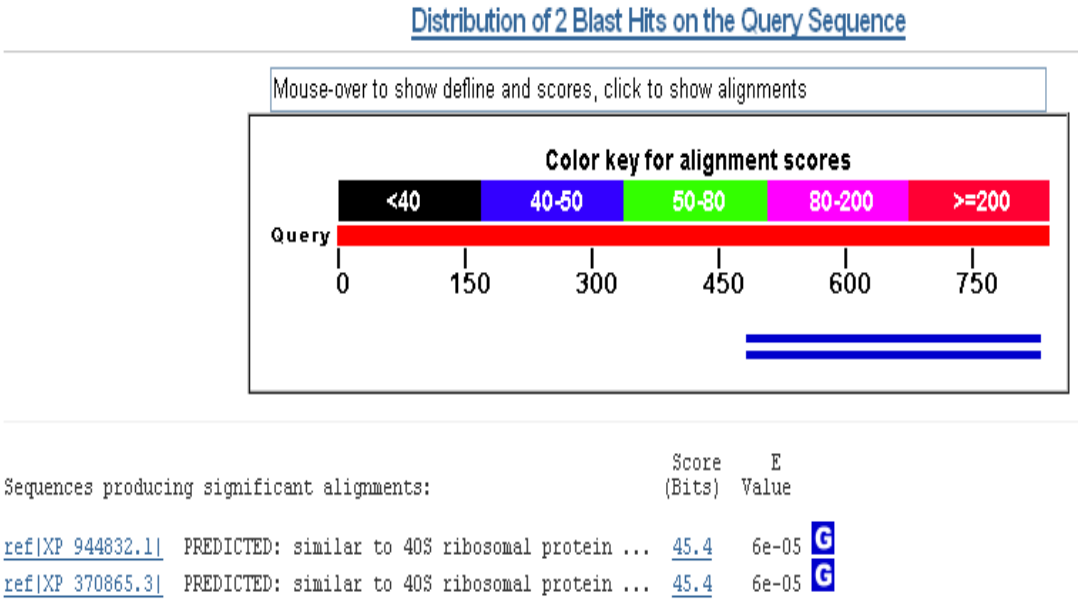
Sequences producing significant alignments:			Score (bits)	E Value
15927247_1	DEG10020269 Staphylococcus aureus,hemB,	...	235	5e-63
b0369_1	DEG10040070 Escherichia coli MG1655,hemB,	...	229	4e-61
15608158_1	DEG10090155 Mycobacterium tuberculosis H37Rv,gl	...	26	4.7
MG468_1	DEG10060378 Mycoplasma genitalium,MG468,	...	26	4.7

From above results , the score >100 identified as essential gene (our query sequence). These gene also blast with NMPDR ,the results are further blast against human protein sequences by using blast with default parameters. The results are shown below, from this the score <60, query sequence is taken as non-human homologs gene. This genes are known as a potential therapeutic targets. These gene are further classified in to different groups based on function.

NCBI BLASTX output:

Query= ref|NC_005043.1|:779091-779924
Length=834

Database: Homo sapiens RefSeq protein
13,039 sequences; 4,940,105 total letters



Sequences producing significant alignments:		Score (Bits)	E Value
ref|XP_944832.1|	PREDICTED: similar to 40S ribosomal protein ...	45.4	6e-05
ref|XP_370865.3|	PREDICTED: similar to 40S ribosomal protein ...	45.4	6e-05

Fig. 5: Graphical output of BLASTX results.

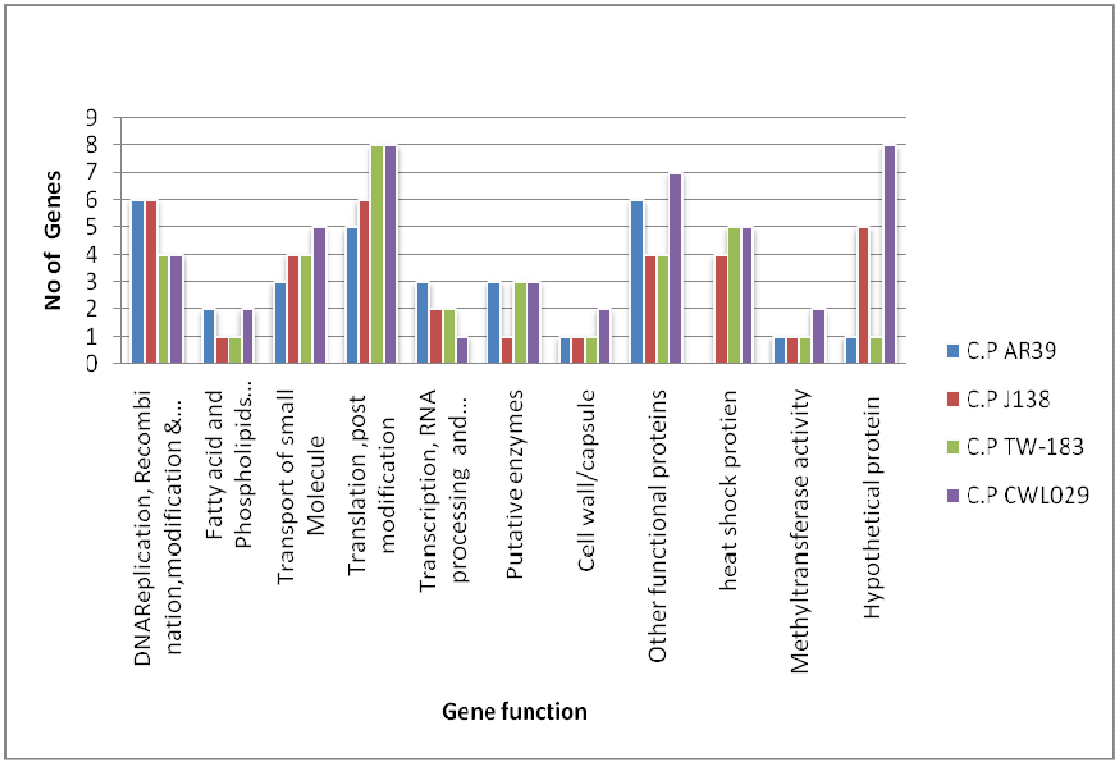


Fig.6: The graph showing non-human homologs essential genes encoding different proteins involved in a same biological function in comparison with four different strains.

147 genes were identified as non-human homologs and conserved proteins among four strains (Tab.2). These non-human homologs genes and their encoding protein were further categorized on the basis of the pathways involved in the basic survival mechanisms such as: genes belong to the DNA replication, recombination, modification and repair, translation and post translation modification, transport of small molecule, transcription, RNA processing and degradation. i. e. any disruption in the functioning of those genes will lead to bacterial death (Fig. 6). The pathway information for each target gene was obtained from the KEGG database. These essential genes were covering 3-4% of total genome of the organism. In MSA analysis we identified conserved regions among the protein sequences having same biological function (Tab.3). All such essential genes can be potential drug targets but including those genes whose products have sequence similarities with any human protein may lead to drug reactions with the host and, thus, to toxic effects. Therefore, homology modeling was done only with the DNA helicase RuvB protein encoding genes, which have no sequence similarities with the human genes.

Table 3: The predicted drug targets of Chlamydophila pneumoniae.

S.no	Protein name	Function of protein
1	excinuclease ABC subunit A	DNA Replication, Recombination, modification and Repair
2	Holliday junction DNA helicase RuvB	DNA Replication, Recombination, modification and Repair
3	30S ribosomal protein S10	Translation ,post modification
4	30S ribosomal protein S2	Translation ,post modification
5	GTP-binding protein EngA	GTP-dependent binding, GTPase of unknown physiological role.
6	hypothetical protein	hypothetical function
7	Acetyl glucosaminyl transferase	acetylglucosaminyl transferase
8	riboflavin-specific deaminase	Putative enzymes

5.3 Homology Modelling Results:

Three-dimensional structures will help in the identification of binding sites and may lead to the designing of new drugs. The 3D structure of DNA helicase RuvB protein of the *C. pneumoniae* was modeled with **Deep View; CPHmodels; Geno3D; Swiss model; MODELLER9v2** was used for fine building the model and global energy minimization.

Table4: homology modelling best results of different softwares of target protein.

S.no	protein	Procheck	Verify3D	Errat
1	Geno3Dmodel	75.7	87.18	95.03
2	Deep view model	93.9	92.53	84.12
3	Modeller model	94.2	92.90	80.80
4	CPHmodels	90.5	95.13	92.51
5	Swiss model	90.6	88.17	89.91

The above table shows the modeller showing better results than deepview, Swiss model. Modeller is the one of best homology modelling software. The details of Modeller results are explained below in details.

Results of modeller:

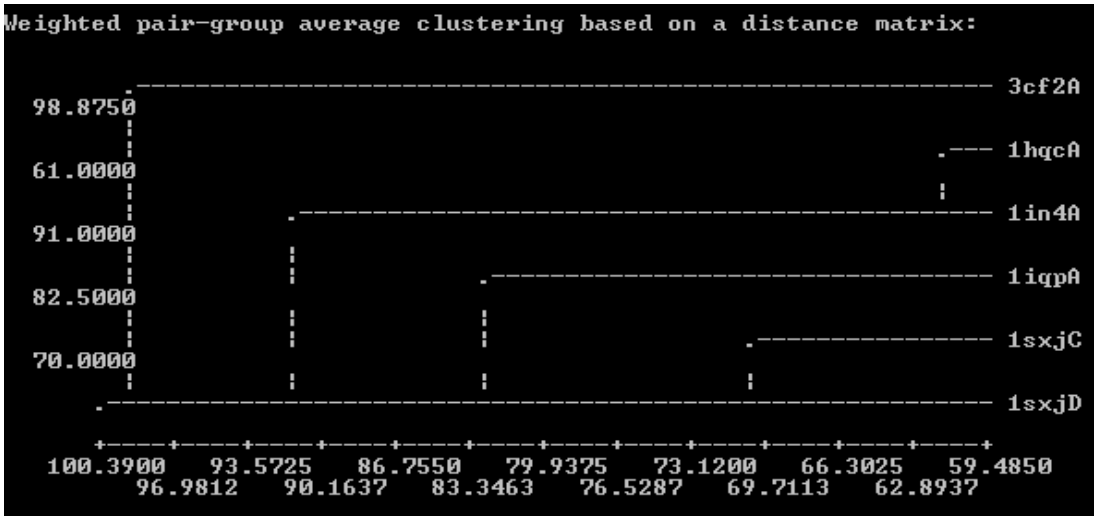


Fig. 7**:** clustering tree (dendrogram) from pairwise distance matrix .

The comparison above shows that 1hqc:A[44] and 1in4:A[45] are almost identical, both sequentially and structurally. However, 1in4:A has a better crystallographic resolution (3.6Å versus 1.6Å), eliminating 1hqc:A. A second group of structures (1iqp:A[46], 1sxj:c[47]) share some similarities. From this group, 1sxj has the poorest resolution leaving for consideration only 1iqp:A. 3cf2:A[48], is the most diverse structure of the whole set of possible templates. However, it is the one with the lowest sequence identity (26%) to the query sequence. We finally pick 1in4:A over 1iqp:A because of its better resolution versus 1.6Å , ts better crystallographic R-factor (23.4%) and higher overall sequence identity to the query sequence (53%).

Table 5: Detail known structure protein with target sequence.

Protein id	Identity	E-value	Resolution(A^o)	R-value
1e32:A	33	0.63E-02	2.90	0.224
1hqc:A	50	0	3.20	0.263
1in4:A	53	0	1.60	0.234
1iqp:A	26	0.52E-04	2.80	0.224
1sxj:C	50	0.90E-03	2.85	0.251
1sxj:D	49	0.20E-02	2.85	0.251
1ypw:A	26	0.40E-02	3.50	0.271

```
 _aln.pos          10         20         30         40         50         60
1in4A      ----------------QFLRPKSLDEFIGQENVKKKLSLALEAAKMRGEVLDHVLLAGPPGLGKTTLAH
RuvB       MTHQVAVLHQDKKFDVSLRPKGLEEFYGQHHLKERLDLFLCAALQRGEVPGHCLFFGPPGLGKTSLAH
 _consrvd                    **** * ** **   *  * * * **  ****  * *  ******** ***

 _aln.p    70         80         90        100        110        120        130
1in4A      IIASELQTNIHVTSGPVLVKQGDMAAILTSLERGDVLFIDEIHRLNKAVEELLYSAIEDFQI------
RuvB       IVAYTVGKGLVLASGPQLIKPSDLLGLLTSLQEGDVFFIDEIHRMGKVAEEYLYSAMEDFKVDITIDS
 _consrvd * *           *** * *  *    ****  *** *******  *  ** **** ***

 _aln.pos  140        150        160        170        180        190        200
1in4A      ---------DIQPFTLVGATTRSGLLSSPLRSRFGIILELDFYTVKELKEIIKRAASLMDVEIEDAAA
RuvB       GPGARSVRVDLAPFTLVGATTRSGMLSEPLRTRFAFSARLSYYSDQDLKEILVRSSHLLGIEADSSAL
 _consrvd           *  ************ ** *** **     *  *    ****  *   *   *    *

 _aln.pos    210        220        230        240        250        260        270
1in4A      EMIAKRSRGTPRIAIRLTKRVRDMLTVVKADRINTDIVLKTMEVLNIDDEGLDEFDRKILKTIIEIYR
RuvB       LEIAKRSRGTPRLANHLLRWVRDFAQIREGNCINGDVAEKALAMLLIDDWGLNEIDIKLLTTIIDYYQ
 _consrvd   ********** *  *   ***          ** *   *    * *** ** * * * * ***  *

 _aln.pos      280        290        300        310        320        330
1in4A      GGPVGLNALAASLGVEADTLSEVYEPYLLQAGFLARTPRGRIVTEKAYKHLKY---------EVP
RuvB       GGPVGIKTLSVAVGEDIKTLEDVYEPFLILKGFIKKTPRGRMVTQLAYDHLKRHAKNLLSLGEGQ
 _consrvd *****   *    *    **  **** *   **   ***** **  ** ***          *
```

Fig 8: The template(1in4:A) and target protein sequence alignment PAP alignment format.

The above five model are generated by modeller, the "best" model can be selected in several ways. The best model selected with the lowest value of the MODELLER objective function, the DOPE assessment score, with the highest percentage residue core region from Ramachandran plot and highest overall quality factor Errat, all of which are reporting the model one good structure of target protein. The molpdf and DOPE scores are not 'absolute' measures, in the sense that they can only be used to rank models calculated from the same alignment. Other scores are transferable.

Table 6: Summary of successfully produced models by single template model

S.no	Protein model	Procheck (Ramachandran plot: % core)	Verify3D (% of the residues had an averaged 3D-1D score > 0.2)	Errat (Overal quality factor)	Mol pdf score	Dope score	Final rank
1	Rnb1	94.2	86.09	85.67	1294.64	-36710.75	1
2	Rnb2	92.8	87.87	73.354	1316.65	-36555.46	3
3	Rnb3	93.8	83.73	78.89	1354.76	-36460.83	4
4	Rnb4	92.5	85.5	77.88	1458.02	-37035.18	5
5	Rnb5	94.2	87.87	78.638	1584.59	-36610.44	2

The above table analysis, confirms that model1 (Rnb1) is a reasonable model. However, the plotted DOPE score profile (below) shows regions of relatively high energy for the long active site loop between residues 194 and 205 and the long helices at the C-terminal end of the target sequence.

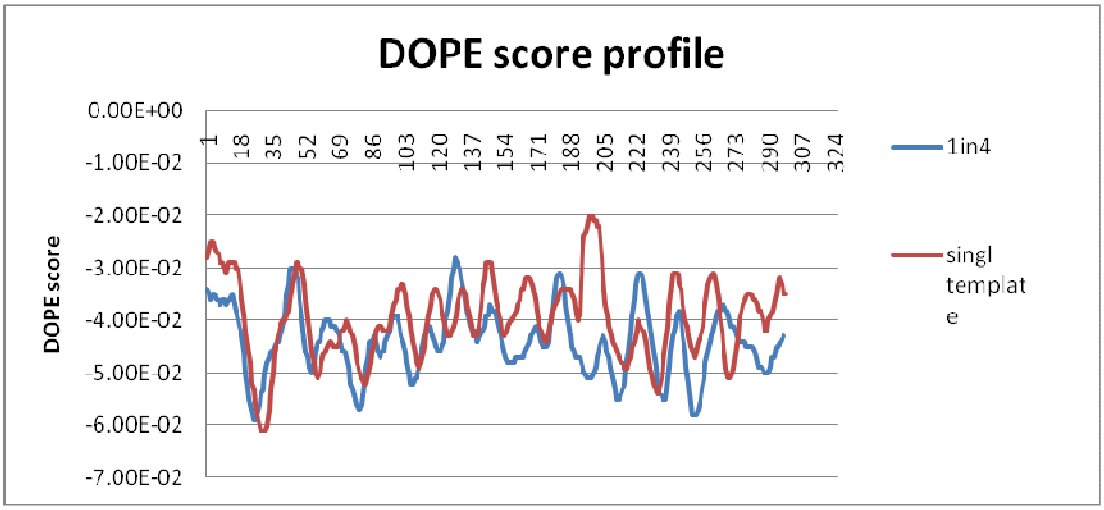


Fig 9: DOPE score profile for single template model1(Rnb1) and template 1in4.

The selected model is further refining with multiple templates and final Modeling loop using *ab-initio* methods. The structure of the 1in4 has been clustered in the DBAli database (http://salilab.org/DBAli/) within the family fm03090 of 2 members (1ixr:C,1in6:A). The multiple alignment generated by the with MODELLER.

Fm03090 family tree:

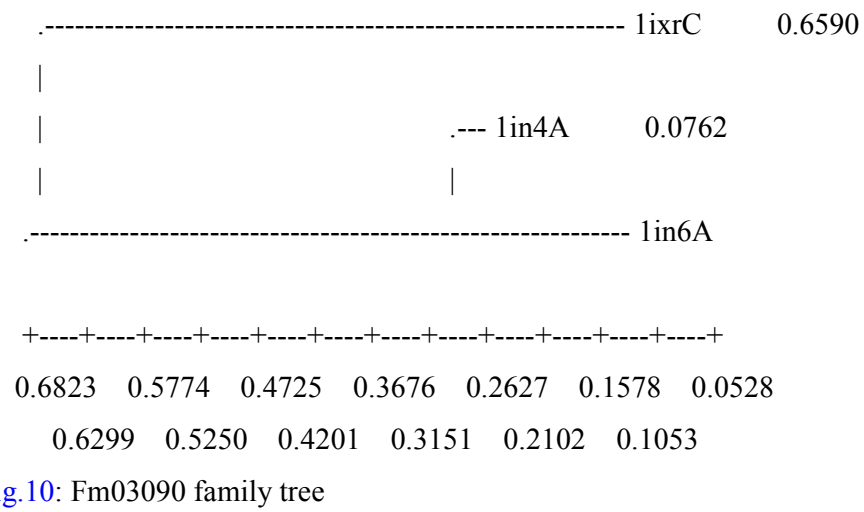


Fig.10: Fm03090 family tree

```
 _aln.pos         10        20        30        40        50        60
1ixrC     ---------------AL-RPKTLDEYIGQERLKQKLRVYLEAAKARKEPLEHLLLFGPPGLGKTTLAH
1in4A     ---------------QFLRPKSLDEFIGQENVKKKLSLALEAAKMRGEVLDHVLLAGPPGLGKTTLAH
1in6A     ---------------QFLRPKSLDEFIGQENVKKKLSLALEAAKMRGEVLDHVLLAGPPGLGRTTLAH
RuvB      MTHQVAVLHQDKKFDVSLRPKGLEEFYGQHHLKERLDLFLCAALQRGEVPGHCLFFGPPGLGKTSLAH
 _consrvd                   *** * *  **   *  *   * **  * *   * *  ****** * ***

 _aln.p   70        80        90       100       110       120       130
1ixrC     VIAHELGVNLRVTSGPAIEKPGDLAAILANSLEEGDILFIDEIHRLSRQAEEHLYPAMEDFVMDIVIG
1in4A     IIASELQTNIHVTSGPVLVKQGDMAAILTS-LERGDVLFIDEIHRLNKAVEELLYSAIEDFQI-----
1in6A     IIASELQTNIHVTSGPVLVKQGDMAAILTS-LERGDVLFIDEIHRLNKAVEELLYSAIEDFQIDI---
RuvB      IVAYTVGKGLVLASGPQLIKPSDLLGLLTS-LQEGDVFFIDEIHRMGKVAEEYLYSAMEDFKVDITID
 _consrvd   *          ***   *  *    *   *  **  *******     ** ** * ***

 _aln.pos  140       150       160       170       180       190       200
1ixrC     QGPAARTIRLELPRFTLIGATTRPGLITAPLLSRFGIVEHLEYYTPEELAQGVMRDARLLGVRITEEA
1in4A     ----------DIQPFTLVGATTRSGLLSSPLRSRFGIILELDFYTVKELKEIIKRAASLMDVEIEDAA
1in6A     ----------DIQPFTLVGATTRSGLLSSPLRSRFGIILELDFYTVKELKEIIKRAASLMDVEIEDAA
RuvB      SGPGARSVRVDLAPFTLVGATTRSGMLSEPLRTRFAFSARLSYYSDQDLKEILVRSSHLLGIEADSSA
 _consrvd               *** ***** *    **  **     *  *    *     *   *        *

 _aln.pos    210       220       230       240       250       260       270
1ixrC     ALEIGRRSRGTMRVAKRLFRRVRDFAQVAGEEVITRERALEALAALGLDELGLEKRDREILEVLILRF
1in4A     AEMIAKRSRGTPRIAIRLTKRVRDMLTVVKADRINTDIVLKTMEVLNIDDEGLDEFDRKILKTIIEIY
1in6A     AEMIAKRSRGTPRIAIRLTKRVRDMLTVVKADRINTDIVLKTMEVLNIDDEGLDEFDRKILKTIIEIY
RuvB      LLEIAKRSRGTPRLANHLLRWVRDFAQIREGNCINGDVAEKALAMLLIDDWGLNEIDIKLLTTIIDYY
 _consrvd    *  ***** * *  *   ***          *           *  *  **   *   *   *

 _aln.pos      280       290       300       310       320       330
1ixrC     GGGPVGLATLATALSEDPGTLEEVHEPYLIRQGLLKRTPRGRVATELARRHL--------------
1in4A     RGGPVGLNALAASLGVEADTLSEVYEPYLLQAGFLARTPRGRIVTEKAYKHLK---------YEVP
1in6A     RGGPVGLNALAASLGVEADTLSEVYEPYLLQAGFLARTPRGRIVTEKAYKHLK---------YEVP
RuvB      QGGPVGIKTLSVAVGEDIKTLEDVYEPFLILKGFIKKTPRGRMVTQLAYDHLKRHAKNLLSLGEGQ
 _consrvd  *****   *          **  * ** *   *    *****  *  *  **
```

Fig.11: The multiple structure alignment generated with MODELLER in PIR format.

Five model are generated with multiple structure alignment among these five models model1(Rnm1) showing high overall quality. The evaluation of the model indicates that the problematic loop (residues 194 to 205) has improved by using multiple structural templates. The global DOPE score for the models also improved from -36710.75 to -37227.23. MODELLER was able to use the variability in the loop region from the three templates to generate a more accurate conformation of the loop. However, the conformation of a loop in the region around the residue 46 at the C-terminal end of the sequence has higher DOPE score than for the model based on a single template.

Table 7: Results of modeling target protein with multiple templates.

S.no	Protein model	Procheck (Ramachandran plot: % core)	Verify3D (% of the residues had an averaged 3D-1D score > 0.2)	Errat (Overall quality factor)	Mol pdf score	Dope score	Final rank
1	Rnm1	94.2	92.31	80.625	9915.50	-37227.23	1
2	Rnm2	92.8	77.81	78.704	9922.36	-37316.92	4
3	Rnm3	93.5	87.28	80.435	9869.65	-37073.76	2
4	Rnm4	93.2	85.80	77.329	10459.26	-36820.12	5
5	Rnm5	92.8	85.21	77.064	9834.48	-37234.64	3

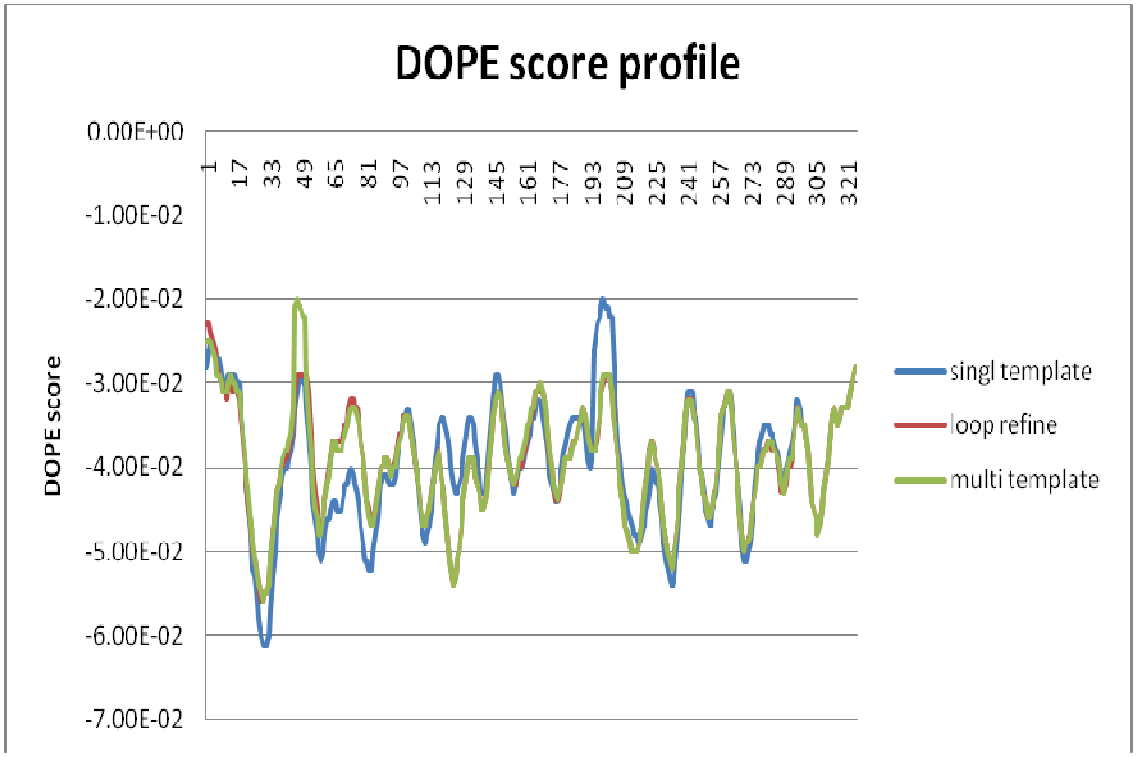


Fig.12: DOPE score profile for single template model1(Rnb1), multi template model1(Rnm1), loop refine model8(Rnl8).

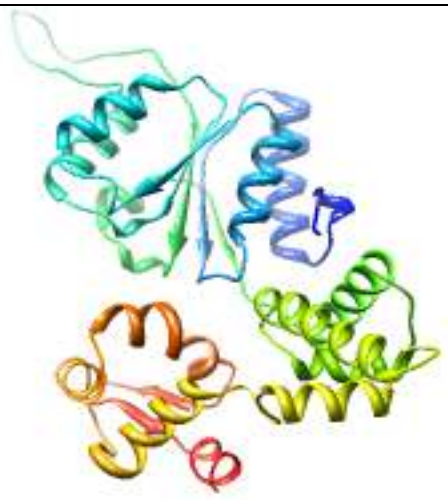
Loop refining

The loop between residues 44 and 51 is refining with modeller. The 10 different model are generated with MODELLER among these 8th model (Rnl8) showing good structural quality.

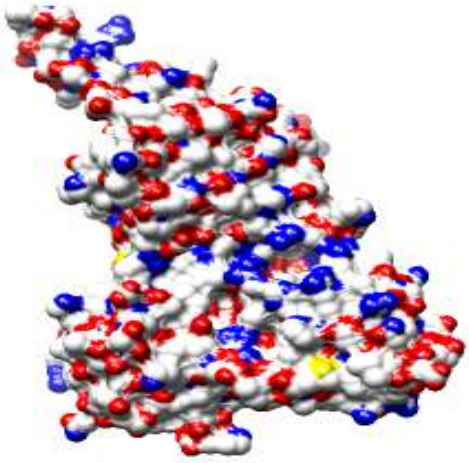
Table 8: Results of modeling target protein with loop refining.

S.no	Protein model	Procheck (Ramachandran plot: % core)	Verify3D (% of the residues had an averaged 3D-1D score > 0.2)	Errat (Overall quality factor)	Mol pdf score	Final rank
1	Rnl1	94.2	92.9	80.805	28.88	4
2	Rnl2	92.8	92.31	77.50	38.05	10
3	Rnl3	93.8	92.31	80.312	37.81	9
4	Rnl4	93.8	92.31	80.938	22.94	5
5	Rnl5	93.8	92.31	80.625	42.86	6
6	Rnl6	93.8	92.90	81.25	28.87	3
7	Rnl7	94.2	92.90	79.439	42.36	7
8	Rnl8	94.2	92.90	80.938	26.98	1
9	Rnl9	93.8	92.90	79.814	30.38	8
10	Rnl10	94.2	92.31	80.435	31.25	2

Final Total Energy of protein: -10633.516 KJ/mol. The final structure of protein is shown in fig.13. Homology modeling is only a viable technique because it produces models that can be used for further research. The structure of the target protein is structurally similar with the template if both the target and template sequences are similar. In general, above 40% sequence homology is required for generating useful models.



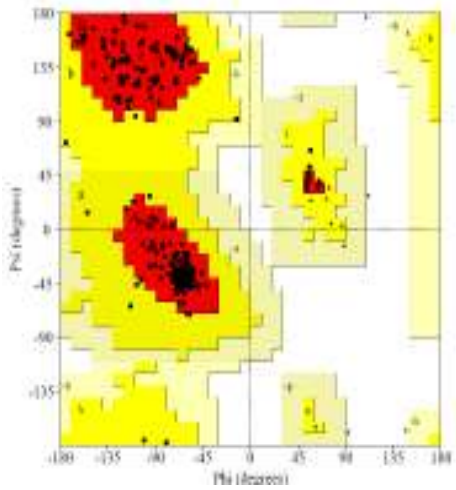
a) Modeled protein ribbon structure.



b) Modeled protein surface structure.

Fig.13: Predicted 3-D structure of Holliday junction DNA helicase RuvB protein

The modeled protein is validated with SAVES (Structure Analysis and Verification Server) it is located at NIH MBI Laboratory for Structural Genomics and Proteomics. The total energy values of the predicted 3-D model were calculated as 93.9% of Ramachandran plot (Fig. 14) value in 30 and 40 steepest descents and conjugate gradient, respectively.



Plot statistics

Residues in most favoured regions [A,B,L]	248	93.9%
Residues in additional allowed regions [a,b,l,p]	16	6.1%
Residues in generously allowed regions [~a,~b,~l,~p]	0	0.0%
Residues in disallowed regions	0	0.0%
Number of non-glycine and non-proline residues	264	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	28	
Number of proline residues	13	
Total number of residues	307	

M/c bond lengths: 99.3% within limits 0.7% highlighted
M/c bond angles: 98.2% within limits 1.8% highlighted
Planar groups: 87.4% within limits 12.6% highlighted

Fig.14: Ramachandran plot of Holliday junction DNA helicase RuvB protein from PROCHECK.

Errat analyzes the statistics of non-bonded interactions between different atom types and plots the value of the error function versus position of a residue. From Errat Overall

quality factor 84.122. Verify_3D determines the compatibility of an atomic model (3D) with its own amino acid sequence (1D) by assigned a structural class based on its location and environment (alpha, beta, loop, polar, nonpolar etc) and comparing the results to good structures. Verify_3D results shows 92.53% of the residues had an averaged 3D-1D score >0.2 and test passed. (See supplementary material 2 for the of the corresponding Structure validation results)

5.4 Active site prediction and docking study:

Active sites of the target protein were predicted by PASS, CASTp active site prediction tools. The feasible active sites predicted by the tools are as follows.

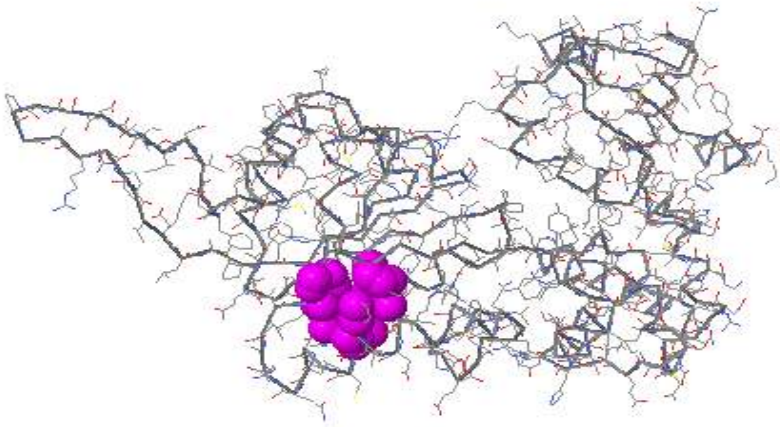


Fig.15: Visualization predicted active site binding pocket of target protein with void volume 162, area 208.9.

The feasible active sites:

20PRO 49VAL 51GLY 58PRO 60 GLY

65SER 70VAL 74VAl 101GLU 104VAL

151THR 157THR 177SER 179TYR 187ILE

219ASN 223ARG 294ASP 308LYS

Optimization of leads was done based on the Lipinski rule of five, in this poorly soluble compounds or compounds with poorer physical and chemical properties, as well as insoluble and non-permeable compounds would have been filtered out at earlier stages. The Molecular

weight known relationship between poor permeability and high molecular weight, number of hydrogen bond donors and acceptors – High numbers may impair permeability across membrane bilayer. the selected ligand are their properties are shown in Tab.3

Table 9: ligand properties are collected from NCBI Pubchem Compound database.

S.no	***Formula***	***ID***	***MW g/mol***	***A****	***TPSA***	***B****	***C****	***D****
1	$C_6H_{14}O_2$	452860	118.17	0.2	40.5	2	2	5
2	$C_5H_5N_5$	190	135.13	-0.3	80.5	2	5	0
3	$C_{10}H_{17}N$	2130	151.25	2.3		1	1	0
4	$C_{16}H_{19}N_3O_5S$	33613	365.40	0	133	4	6	4
5	$C_7H_7NO_3$	134085	153.13	-0.5	57.6	1	3	1
6	$C_{14}H_{18}N_4O_3$	5578	290.32	0.6	106	2	7	5
7	$C_{10}H_{11}N_3O_3S$	5329	253.28	0.7	98.2	2	6	3
8	$C_8H_5BrO_4S$	6475860	277.09	0.9	74.6	2	4	3
9	$C_{14}H_{18}O_4$	6475859	250.29	3	74.6	2	4	3
10	$C_{15}H_{13}NO_5$	5482292	287.27	1.9	99.8	3	6	5
11	$C_{18}H_{11}Cl_2NO$	5482291	392.18	4	108	4	6	6

A*- Octanol-water partition coefficient (XLogP)

TPSA - Topological polar surface area

MW - Molecular weight

B*- Hydrogen bond donors

C*- Hydrogen bond acceptors

D* - Number of rotatable bonds

These optimized ligands are used to find its respective interactions with the targeted protein by using the lamarkian genetic algorithm which gives the best 100 best possible interactions with the least binding energy.

The Docking was performed with all ligands to the target protein, below Table10 shows the different flexible conformation of ligands indicated by the run that binds to the target binding site with the respective binding energy. Among all the ligand compound, compound having id

5482291 was found to having lowest binding free energy -8.99 Kcal/mol and may considered potential lead for further investigation. It is binding to the predicted active site Ser177, Asn219 with lowest binding free energy , number of distinct conformational clusters found = 58, out of 100 runs, Using an rmsd-tolerance of 2.0 A and the sample result is shown below

Cluster Rank = 1 Run = 65

Number of conformations in this cluster = 9

RMSD from reference structure = 72.644 A

Estimated Free Energy of Binding = -8.99 kcal/mol [= (1)+(2)+(3)-(4)]

Estimated Inhibition Constant, Ki = 258.24 nM (nanomolar) [Temp= 298.15 K]

(1) Final Intermolecular Energy = -10.94 kcal/mol

vdW + Hbond + desolv Energy = -8.54 kcal/mol

Electrostatic Energy = -2.39 kcal/mol

(2) Final Total Internal Energy = -0.45 kcal/mol

(3) Torsional Free Energy = +2.20 kcal/mol

(4) Unbound System's Energy = -2.20 kcal/mol

Table 10: The binding free energy of DNA helicase RuvB protein with different compound and conformations.

S.no	*Pubchem Compound ID*	*Binding site*	*Chemical Formula*	*Binding energy(Kcal/mol)*	*Run*	*Rank*
1	5482291	SER177,ASN219	$C_{18}H_{11}Cl_2NO_5$	-8.99	65	1
2	5329	PRO58,THR157	$C_{10}H_{11}N_3O_3S$	-7.63	86	2
3	5482292	GLY60	$C_{15}H_{13}NO_5$	-7.28	38	3
4	6475859	ASP294	$C_{14}H_{18}O_4$	-7.24	86	4
5	6475860	GLY60,SER65	$C_8H_5BrO_4S$	-6.95	25	5
6	134085	ASN219,187ILE	$C_7H_7NO_3$	-6.01	4	6
7	33613	ASP294	$C_{16}H_{19}N_3O_5S$	-5.78	56	7
8	5578	ASN219,187ILE	$C_{14}H_{18}N_4O_3$	-5.28	62	8
9	190	GLY60	$C_5H_5N_5$	-4.37	59	9
10	2130	PRO20	$C_{10}H_{17}N$	-4.00	24	10
11	452860	THR157	$C_6H_{14}O_2$	-3.92	72	11

Chapter 6

CONCLUSION

Conclusion

The availability of full genome sequences and computer-aided software like modeler, Autodock help to identify probable antimicrobial drug targets to dock with protein targets, it has become a new trend in bioinformatics. *C. pneumoniae* is a multi-drug resistant bacterium and causes severe infection in humans. Active compound like 5482291 targeted to Holliday junction DNA helicase RuvB protein will be particularly useful in overcoming the detrimental consequence of *C. pneumoniae* infection. We present here a detailed *in silico* analysis of essential genes, Molecular Modeling of the target protein and followed by lead optimization that favours the docking. This paper present a detailed *in silico* thus the docking analysis proves the 5482291 is an active compound that binds to the targeted Holliday junction DNA helicase RuvB protein with the least binding energy.

On further study for the pharmacodynamic, pharmacokinetics, solubility and thermodynamics activity of these ligand receptor binding can inhibit the pathogenic activity of the organism.

REFERENCES

1. Miesel, L., Greene, J. and Black, T. A. (2003). Genetic strategies for antibacterial drug discovery. Nature Rev. Genet. 4, 442-456.

2. Galperin, M. Y. and Koonin, E. V. (1999). Searching for drug targets in microbial genomes. Curr. Opin. Biotechnol. 10, 571-578.

3. Sakharkar, K. R., Sakharkar, M. K. and Chow, V. T. K., (2004). A novel genomics approach for the identification of drug targets in pathogens, with special reference to Pseudomonas aeruginosa. In Silico Biol. 4, 0028.

4. Allsop, A. E. (1998). New antibiotic discovery, novel screens, novel targets and impact of microbial genomics. Curr. Opin. Microbiol. 1, 530-534.

5. Judson, N. and Mekalanos, J. J. (2000b). Transposon-based approaches to identify essential bacterial genes. Trends Microbiol. 8, 521-526.

6. Moir, D.T., Shaw ,K.J., Hare, R.S., Vovis, G.F.,(1999) Genomics and antimicrobial drug discovery. Antimicrob Agents Chemother. 43:439-446.

7. Huynen, M., Diaz-Lazcoz, Y. and Bork, P. (1997). Differential genome display. Trends Genet. 13, 389-390.

8. Huynen, M., Dandekar, T. and Bork, P. (1998). Differential genome analysis applied to the species-specific features of Helicobacter pylori. FEBS Lett. 426,1-5.

9. Zhang ,R., Ou, H.Y. and Zhang, C.T., (2004). DEG: a Database of Essential Genes. Nucleic Acids Research. 32, D271-D272.

10. Rost, B. (1999). Twilight zone of protein sequence alignments. Protein Eng. 12, 85-94.

11. Ringe, D., (1995). What makes a binding site a binding site?. Cur. Op. Struct. Biol. 5,825.

12. Ruppert, J., Welch, W. and Jain, A.N. (1997). Automatic identification and representation of protein binding sites for molecular docking. Prot. Sci. 6,524.

13. Connolly, M. L., (1983). Analytical molecular surface calculation. Journal of Appl. Crystallogr. 16,548.

14. Nicholls, A., Bharadwaj, R. and Honig. (1993). GRASP: Graphical representation and analysis of surface properties. Biophys. Journal. 64, A166.

15. Ho, C. M. W. and Marshall, G.R. (1990). De novo design of ligands. Journal Comput.-Aided Mol. Design, 4 337.

16. Luty, B.A., Wasserman, Z.R., Stouten, P.F.W. (1995). Molecular Mechanics/Grid Method for the Evaluation of Ligand-Receptor Interactions. J. Comp.Chem. 16, 454-464.

17. Huey, R., Morris, G. M., Olson, A. J. and Goodsell, D. S. (2007), A Semiempirical Free Energy Force Field with Charge-Based Desolvation, J. Computational Chemistry. 28,1145-1152.

18. Stephens, R. S, (1999). Chlamydophila: intracellular biology, pathogenesis,and immunity. American Society for Microbiology. Washington, D.C.

19. Andersen, P. (1998). Pathogenesis of lower respiratory tract infections due to Chlamydia, Mycoplasma, Legionella and viruses. Thorax. Apr. 53(4),302-7.

20. Lesk, M., (2005), Introduction to Bioinformatics, Second edition. Oxford University Press Inc., New York.

21. Morris, A. .L., MacArthur, M.W., Hutchinson, E.G. and Thornton, J. M. (1992). Stereochemical quality of protein structure coordinates. PROTEINS: Structure, Function, and Genetics. 12,345-364.

22. Ramachandran, G.N., Ramakrishnan, C. and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. In: J. Mol. Biol. 7, 95-99.

23. Schwede, T., Kopp, J., Guex, N., and Peitsch, M.C. (2003). SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res. 31, 3381-3385.

24. Combet, C., Jambon, M., Deleage, G. and Geourjon, C. (2002). Geno3D: Automatic comparative molecular modelling of protein. Bioinformatics. 18, 213-214.

25. Lund, O., Frimand ,K., Gorodkin, J., Bohr, H., Bohr, J., Hansen ,J. and Brunak, S. (1997). Protein distance constraints predicted by neural networks and probability density functions. Protein Engg.10, 1241-1248.

26. Marti-Renom, M.A., Stuart, A., Fiser, A., Sanchez, R., Melo, F. and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. 29, 291-325.

27. Guex, N., Diemand, A. and Peitsch, M.C. (1999). Protein modelling for all. TiBS. 24,364-367.

28. Laskowski, R. A., MacArthur, M. W., Moss, D. S. and Thornton, J.M. (1993), PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst. 26, 283-291.

29. Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. J. Mol. Graph. 8, 52-56.

30. Colovos, C., Yeates, T.O.(1993). Verification of protein structures: patterns of non-bonded atomic interactions. Protein Sci. 2,1511-1519.

31. Eisenberg, D., Luthy, R. and Bowie, J. U. (1997). VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol. 277, 396-404.

32. Pontius, J., Richelle, J. and Wodak, S.J. (1996). Quality assessment of protein 3D structures using standard atomic volumes. J. Mol. Biol. 264, 121-136.

33. Brady, G. P, Jr. and Stouten, P. F.W. (2000). Fast Prediction and Visualization of Protein Binding Pockets With PASS. Journal of Computer-Aided Molecular Design, 14, 383-401.

34. Binkowski, T. A., Naghibzadeh, S. and Liang, J. (2003), CASTp: Computed Atlas of Surface Topography of proteins. Nucleic Acids Res. 31,3352-3355.

35. Kalman, S., Mitchell, W., Marathe, R., Lammel C, Fan, J., Hyman, R.W., Olinger, L, Grimwood, J., Davis, R.W. and Stephens, R.S. (1999). Comparative genomes of Chlamydophila pneumoniae and C. trachomatis. Genet. 21,385-389.

36. Dutta, A., Singh, S. K., Ghosh, P., Mukherjee, R., Mitter, S. and Bandyopadhyay, D. (2006). In silico identification of potential therapeutic targets in the human pathogen Helicobacter pylori. In Silico Biol. 6, 0005.

37. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. and Yamanishi, Y. (2008). KEGG for linking genomes to life and the environment. Nucleic Acids Res. 36, D480-D484.

38. Tuteja, R. and Pradhan, A. (2006). Unraveling the 'DEAD-box' helicases of Plasmodium falciparum. Gene. 376:1,1-12.

39. Frick, D.N. (2003). Helicases as Antiviral Drug Targets Drug. News Perspect. 16(6), 355.

40. Lipinski, C.A., Lombardo, F., Dominy, B. W. and Feeney, P.J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Delivery Reviews. 46:1-3,3-26.

41. Toronto, O.N. (2006). ACD/ChemSketch Freeware, version 10.00, Adv. Chemistry Development, Inc., Canada.

42. Morris, G. M., Goodsell, D. S., Halliday, R.S., Huey, R., Hart, W. E., Belew, R. K. and Olson, A. J. (1998). Automated Docking Using a Lamarckian Genetic Algorithm and and Empirical Binding Free Energy Function. J. Computational Chemistry. 19,1639-1662.

43. Yang, S.I. et al . (1991). Control of protein phosphatase 2A by simian virus 40 small-tantigen. Mol. Cell Biol. 11(4): 1988-1995.

44. Yamada, K., Kunishima, N., Mayanagi, K., Ohnishi, T., Nishino, T., Iwasaki, H., Shinagawa, H.and Morikawa, K. (2001). Crystal structure of the Holliday junction migration motor protein RuvB from Thermus thermophilus HB8. Proc.Natl.Acad.Sci. 98,1442-1447.

45. Putnam, C.D., Clancy, S.B., Tsuruta, H., Gonzalez, S., Wetmur, J.G .and Tainer, J.A. (2001). Structure and mechanism of the RuvB Holliday junction branch migration motor. J. Molecular Biol. 311,297-310.

46. Oyama, T., Ishino, Y., Cann, I.K.O., Ishino, S. and Morikawa, K. (2001). Atomic Structure of the Clamp Loader Small Subunit from Pyrococcus furiosus. Mol. Cell. 8, 455-463.

47. Bowman, G.D., ODonnell, M. and Kuriyan, J. (2004) Structural analysis of a eukaryotic sliding DNA clamp-clamp loader complex. Nature. 429,724-730.

48. Davies, J.M., Delabarre, B., Brunger, A.T. and Weis, W.I. (2008). ATPase: implications for mechanisms of nucleotide-dependent conformational change Structure. Improved structures of full-length. 97.

49. Putnam, C.D., Clancy, S.B., Tsuruta, H., Gonzalez, S., Wetmur, J.G. and Tainer, J.A. (2001) .Structure and mechanism of the RuvB Holliday junction branch migration motor. J.Mol. Biol. 311, 297-310.