Sequence Analysis and Rational Drug Design Studies of Selected Target Proteins from *Mycobacterium tuberculosis* and *Homo sapiens*

A Thesis

Submitted for the Degree of

DOCTOR OF PHILOSOPHY

By

KRISHNA KISHORE INAMPUDI

SCHOOL OF CHEMISTRY
UNIVERSITY OF HYDERABAD
HYDERABAD 500 046
INDIA

February 2008

Dedicated to...

My Beloved Father and Mother
And
My Teachers

CONTENTS

School of Chemistry
University of Hyderabad
Central University P. O.
Hyderabad 500 046
India

Statement

I hereby declare that the matter embodied in this thesis is the result of investigations carried out by me in the School of Chemistry, University of Hyderabad, Hyderabad, under the supervision of **Dr. Lalitha Guruprasad**.

In keeping with the general practice of reporting scientific observations, due acknowledgement has been made wherever the work described is based on the findings of other investigators.

Krishna Kishore Inampudi

School of Chemistry
University of Hyderabad
Central University P. O.
Hyderabad 500 046
India

Certificate

Certified that the work embodied in this thesis entitled **"Sequence Analysis and Rational Drug Design Studies of Selected Target Proteins from *Mycobacterium tuberculosis* and *Homo sapiens*"** has been carried out by Mr. KRISHNA KISHORE INAMPUDI, under my supervision and the same has not been submitted elsewhere for a Degree.

Dr.LALITHA GURUPRASAD
(THESIS SUPERVISOR)

DEAN
SCHOOL OF CHEMISTRY

Acknowledgements

It gives me immense pleasure to express my deep sense of gratitude and profound respect to my teacher Dr. Lalitha Guruprasad for her inspiring guidance, valuable advice and personal motivation. She has been always helpful, approachable and extremely patient through out my tenure for which I am grateful to her.

I would like to acknowledge the suggestions and help of Dr. Guruprasad, CCMB.

I am thankful to Prof. G. R. Desiraju, Prof. E. D. Jemmis and Prof. M. Periasamy, Deans, School of Chemistry during my stay here, for providing the necessary facilities to carry out my research work. I thank all the faculty members of the School of Chemistry for their help and inspiring teaching.

I thank the funding bodies of CMSD for providing excellent computational facilities in the University. I thank Directors of CCMB and IICT for allowing me to use some of their computational and library facilities.

I thank Prof. P.B. Kirthi for helping me in learning gene expression techniques. I am happy to express my gratitude to Dr. Abani K. Bhuyan and his family for their affection. I also thank Dr. Samar K. Das and Prof. D. Basavaiah for their concern on various occasions.

I would like to acknowledge the suggestions and help of Dr. M. Rami Reddy, Dr. P.V.Bharatam and Prof. Reddanna.

I would like to thank my colleagues Dr. Swathi, G. R. Hemalatha, Srinivas, Karunakar and project students Nethu Singh, Satyanarayana, Manoj Kumar for creating a pleasant working atmosphere in the lab.

I would like to thank my project juniors Kalyan, Arun, Vijay, Sudheer, LSrinu, Chanu, Sankaraih, Sthish, Srinu (Chiru), Om, Subhashini and Narayan.

All the non-teaching staff of the School is acknowledged for their help. I also thank Vinod Kumar, Rajender Reddy and other CMSD staff for their cooperation.

I express my warmest heartfelt thanks to Mr. Jagan and Lakshmi, Dr. Kalapala and Suneeta, and my uncle Ravi Kumar for their emotional support and timely help.

I express my warmest heartfelt thanks to P. Charu Sheela, who deserves special mention for her endless affection, encouragement and continued moral support.

I would like to thank my friends P.S.Satya, Venkat annaiah, NageswarRao, K.Ravi, Sivaranjan, Tammulam Garu, GVRamesh, Rambomb, Srinu, GDP, SekarReddy, RameshReddy and Armugam.

I also acknowledge Dr. Mukkanti, Dr. Karunakar, Dr. Shivaiah, Dr. Supriya, Dr. Aparna, Dr. Anamika, Dr.Latheef, Saritha, Bhargavi and Yadu bava, D.K, Bhuvan mama, Anwar baba, JP, DLVK, Narahari, Kavitha and Veerendar for their healthy company through out my tenure.

My special thanks are due to my teachers K.L.N.Varma and late Dr. M. Gopal Rao, Yoganand Rao, Subba Rao, Sheshagiri Rao and Dr. Ashok Kumar for their superior principles and teaching.

I take this opportunity to extend my sincere thanks to my teacher G.Pulla Rao for his constant support through out my studies.

I also thank my friends Venu, Chari, Madhu, Sasi, Aparoy, VB Reddy, Dina, Dr.RamaKrishna, Dr. Honey, Ram, Ashok, Sammeta, Suneel, Raju, Bhusankar, Suresh, Siva, Sathish and Murali.

All the research scholars of the School of Chemistry have been helpful and I thank them all. I also acknowledge Aruna, Harish, Vasan, Sravan, Rajesh, Bashid, RamGopal, Ramesh, Vijay, Vijaybasker, Lakshmi Prasad, Seshu and SRGadda.

I am glad to remember all my friends particularly PVN Reddy, Raghu, Malleswari, MVSRN, Girish, Subbu, Ravi, Kiran, LKbava, banda Murthy, banda Kishore, late Chinni Krishna, Chinna, Mahitha and Pavani whose presence left some unforgettable memories. I also thank my cousins Indira, Chinni, Chanti and Krishna and RaviKumar, Pavani and Harinath for their love and well wishes.

I am wordless to express my gratitude to my Mother and Father who have been the source of encouragement and guiding spirit and for their unfailing love all these years. I owe a great deal of loving thanks to my Brothers, sisters and their kids for their support through out my career.

Financial support from the CSIR New Delhi is greatly acknowledged.

Krishna Kishore Inampudi

Abbreviations

PDB	:	Protein data Bank
1-D	:	One-dimensional
2-D	:	Two-dimensional
3-D	:	Three-dimensional
TMM	:	6-trehalose monomycolate
TDM	:	6, 6'-trehalose dimycolate
PDB-BLAST	:	Blast search against protein data bank
RMSD	:	Root Mean Square Deviation
PFAM	:	Protein families
DNA	:	Deoxyribo Nucleic Acid
RNA	:	Ribo Nucleic Acid
NMR	:	Nuclear Magnetic Resonance
B	:	Beta
A	:	Alfa
TB	:	Tuberculosis
HSP	:	High Segment Pair
E	:	Expectation value
P	:	Probability score
PIR	:	Protein Information Resource
HSP	:	High Scoring Pairs
CADD	:	Computer-Aided Drug Design
CAMD	:	Computer-Aided Molecular Design
CAMM	:	Computer-Aided Molecular Modeling
RDD	:	Rational drug design

LBDD	:	Ligand-Based Drug Design
SBDD	:	Structure-Based Drug Design
QSAR	:	Quantitative Structure-Activity Relationship
VS	:	Virtual Screening
RCSB	:	Research Collaboratory for Structural Bioinformatics
IC_{50}	:	Median Inhibitory Concentration
MIC	:	Minimal Inhibitory Concentration
ATP	:	Adenosine triphosphate
ADP	:	Adenosine diphosphate
PTK	:	Protein tyrosine kinases
NRTK	:	Non-receptor protein tyrosine kinase
CML	:	Chronic myelogenous leukemia
HTS	:	High throughput screening
Å	:	Ångström
MMP	:	Matrix metalloproteinases

Abstract

This thesis describes **Sequence Analysis and Rational Drug Design Studies of Selected Target Proteins from *Mycobacterium tuberculosis* and *Homo sapiens*.** It consists of six chapters 1) Introduction to bioinformatics tools in genomic data analysis and rational drug design 2) Docking of phosphonate and trehalose analog inhibitors into *M. tuberculosis* mycolyltransferase Ag85C: Comparison of the two scoring fitness functions GoldScore and ChemScore, in the GOLD software. 3) Chemical function based virtual screening: Discovery of potent lead molecules for the *Bcr-Abl* tyrosine kinase using VX-680. 4) The identification of new Aurora A kinase inhibitors by pharmacophore modeling, virtual screening and molecular docking. 5) Comparative studies of the ADAM and ADAMTS protein family members in human, frog, fly and worm genomes: A Bioinformatics Approach. 6) Diversity of Ser/Thr kinases in the genomes of various *Mycobacterium* species. The work described in this thesis is exploratory in nature and is arranged in the order the investigations were carried out. Except the first chapter, all chapters are divided into Introduction, Methods, Results and Discussion, Conclusions, followed by References.

In the first chapter, a brief overview of the tools used in bioinformatics to characterize the protein sequences resulting from the genome sequencing projects is provided. Some commonly used programs such as BLASTP, FASTA, CLUSTALW, T-Coffee and PHI-BLAST are described. The databases such as GenBank, nr, SMART, PFAM and INTERPRO are also described. A brief overview of phylogenetic analysis, rational drug design, pharmacophore modeling, QSAR, virtual screening and docking which includes both structure-based and ligand-based drug design methods are discussed.

The second chapter deals with the docking analysis of phosphonate and trehalose analog inhibitors into the active site of Ag85C to identify the inhibitor binding position and affinity, using the Gold software. We compared the GoldScore with the ChemScore in the GOLD software. Tuberculosis (TB) is an infection caused by the bacterium *Mycobacterium tuberculosis*. It is a major disease infecting two billion people, or approximately one-third of the world's population. *M. tuberculosis* is surrounded by a complex envelope of unusually low permeability, which contributes to the resistance of this bacterium to host defense mechanisms. The mycobacterial cell wall consists of three major components forming the mycolyl-arabinogalactan-peptidoglycan (mAGP) complex, among which mycolic acids constitute the outermost layer. Mycolic acids are high molecular weight α- alkyl, β-hydroxy fatty acids unique to *Mycobacterium* and related genera. In the mycobacterial cell wall envelope, they are present as free glycolipids, mainly α, α' trehalosemonomycolate (TMM) and α, α' trehalosedimycolate (TDM), and as esters of the terminal pentaarabinofuranosyl units of arabinogalactan.

In *M. tuberculosis*, a major secreted protein complex, antigen 85, constitutes three proteins antigen 85A, 85B and 85C that are responsible for the synthesis of cell envelope. These enzymes catalyze the transfer of mycolyl residue from one molecule of TMM to another TMM leading to the formation of TDM and are hence termed mycolyltransferases. Mycolic acid biosynthesis is known to be essential for mycobacterium growth, in particular trehalose mycolates aid in virulence of the organism and the structure of mycolates has been found to be important for initial replication and persistence *in vivo.* A mutated *M. tuberculosis* strain lacking the functional Ag85C gene showed a 40% decrease in the amount of cell wall linked mycolic acids indicating its role in cell wall synthesis. Structural comparison of mycolyltransferases revealed that their backbone superimposes with

an overall RMSD of 0.577 Å. The catalytic residues are highly superimposed in these structures indicating that the structure and function of these isozymes are highly similar.

In this work we have carried out the docking of phoshonate and trehalose analog inhibitors into the 3-D structure of mycolyltransferase enzyme, Ag85C of *M. tuberculosis* using the GOLD software. The inhibitor binding positions and affinity were evaluated using both the scoring fitness functions, GoldScore and ChemScore. We observed that the inhibitor binding position identified using the GoldScore was marginally better than the ChemScore. A qualitative agreement between the reported experimental biological activities (IC_{50}) and the GoldScore were observed. We identified that amino acid residues Arg541, Trp762 are important for inhibitor recognition via hydrogen bonding interactions. The alkyl chains of mycolic acid bind to the proposed hydrophobic pockets that contribute to the van der Waals energy. Information obtained from this study will be helpful in designing mutational experiments as well as in designing potent new inhibitors for the mycolyltransferases.

The third chapter deals with the building of a Pharmacophore, virtual screening and docking studies for *Bcr-Abl* kinase protein from human genome. *Bcr–Abl* is an oncogene that arises from fusion of the *Bcr* (breakpoint cluster region) gene with the *c-Abl* proto-oncogene. *Bcr-Abl* is non-receptor protein tyrosine kinase (NRTK), which is expressed in a wide range of cells and it is localized at several sub-cellular sites, including the nucleus, cytoplasm, mitochondria, endoplasmic reticulum and cell cortex, where *Bcr-Abl* interacts with a large variety of cellular proteins, including signaling adaptors, kinases, phosphatases, cell-cycle regulators, transcription factors and cytoskeleton proteins. The *Bcr-Abl* gene was first identified as the cellular homolog of the transforming gene of Abelson murine leukemia and was subsequently found to be involved in

the Philadelphia chromosome translocation in human leukemia and to encode a NRTK.

The NRTK *Bcr-Abl* is a causative agent of chronic myelogenous leukemia (CML) and inhibiting the *Bcr-Abl* kinase might induce apoptosis of the diseased cells from the patient's body. Imatinib is a specific inhibitor that binds with high affinity to the inactive conformation of the *Bcr-Abl* tyrosine kinase and has been shown to be effective in the treatment of CML. But *Bcr-Abl* kinase binding site residues oppose the binding through the mutation-induced resistance to imatinib. In order to overcome the resistance to imatinib, a number of new inhibitors have been synthesized. VX-680 is the Aurora A kinase inhibitor which inhibits the *Bcr-Abl* mutants also. The crystal structure of VX-680 bound to the catalytic domain of *Bcr-Abl* (PDB_ID: 2F4J) containing a mutation (H396P) has been solved. This mutation confers imatinib resistance in *Bcr-Abl* kinase but is inhibited by VX-680 *in vitro*.

This 3-D crystal structure is the source for the virtual screening strategy used to discover novel inhibitors to *Bcr-Abl* kinase. In this work we have generated a chemical function based pharmacophore of VX-680 using Hypogen module in catalyst software. In our hypotheses, two hydrogen bond donors (HD), two hydrogen bond acceptors (HA) and one hydrophobic interaction (HP) were allowed, since these are observed as crucial interactions in the protein structure. This pharmacophore was used for the screening of databases such as, NCI, Maybridge and Derwent-WDI2005 and the obtained hits were docked into the *Bcr-Abl* kinase crystal structure using GOLD software for two mutants (H396P) and (T315I). We have identified some useful molecules for the drug resistant *Bcr-Abl* kinase. Further modifications and addition of suitable functional groups to these new scaffolds will generate high affinity *Bcr-Abl* kinase specific inhibitors.

Our results confirm that, chemical function based virtual screening is a powerful tool to discover novel inhibitors of the protein kinase family and further validate virtual screening as an inexpensive and efficient means for lead discovery.

The fourth chapter deals with the generation of Pharmacophore for Aurora A kinase. Aurora kinases are non-receptor serine/threonine kinases. Mammals express three Aurora kinase paralogues A, B and C each of which is thought to play vital role in regulating mitosis. Aurora kinases have been found to be overexpressed in a number of tumor cell lines and human primary tumors. Therefore, one of the promising targets in cancer drug discovery is represented by Aurora A, B and C kinases. Aurora A itself has been identified as a predominantly attractive drug target through observations that it can act as an oncogene and transform cells when ectopically expressed. For example, Aurora A is overexpressed in primary colorectal cancers, breast tumours, ovarian tumour and cell lines from breast, ovarian, colon, prostate, neuroblastoma and cervical. The expression profile of Aurora A in carcinoma suggests that inhibitors of this kinase may have inherent potential as therapeutic agents.

The pharmacophore model is a good approach to quantitatively explore common chemical characteristics among a considerable number of structures with great diversity. A good pharmacophore model could be used as query for searching small molecule databases in order to discover novel chemical entities. In this work we have generated a hypothetical model of the primary pharmacophore features responsible for the bioactivity of various classes of Aurora A inhibitors using HypoGen module in the Catalyst suite of software. The best pharmacophore model, Hypo1 has been validated and used to screen NCI and Maybridge databases using chemical function descriptors to identify lead molecules as novel inhibitors. These potential inhibitors were docked into the active site of Aurora A using the GOLD software to understand their mode of binding to Aurora A kinase.

Catalyst commonly produces 10 hypotheses for a list of molecules in the training set chosen based on the structural diversity and broad range of affinity for Aurora A. The null cost of the 10 hypotheses is 177.47, the fixed cost value is 98.29 and configuration cost value is 16.44. All the 10 hypotheses have a total cost, close to the cost of the fixed hypotheses. The difference between the fixed cost and the null cost is 79.18 bits. The cost range (Δ cost) between these hypotheses and the null hypotheses varies between 68.04 and 65.39 bits with a low cost range, 2.65 bits. Therefore, we can approximate that for all these hypotheses, there is more than 90% chance of representing a true correlation in the data.

We considered the first pharmacophore model, Hypo1, as the best pharmacophore hypotheses in this study, characterized by the highest cost difference, lowest error cost, closest weight to 2, lowest RMSD and the best correlation coefficient. All hypotheses with the exception of hypotheses 5 and 7 have the same features; one Hydrogen bond acceptor (HA), two Hydrophobic groups (HP) and one Ring aromatic group (R). In this pharmacophore modeling, for the best hypotheses Hypo1, the RMSD value 0.927, signifies a good quality for Hypo1 and correlation coefficient 0.946, shows a good linear regression of the geometric fit index. About 84% of the molecules in the training set were predicted within an error less than 2 units.

The best pharmacophore model, Hypo1 was validated by a test set comprising 21 molecules. To check if the hypotheses can also predict the activity of compounds that are different from those included in the training set, we created a test set comprising 21 molecules with different structural information and activity. Hypo1 shows a good correlation between actual and estimated IC_{50} values. As per the statistical analysis, our pharmacophore hypotheses, Hypo1 is valid. Further evaluation of Hypo1 using the Fischer method was applied to validate the strength of correlation between the chemical structures and their biological activity. The 19

spreadsheets obtained using the CatScramble program by erratically scrambling the binding affinity data were used for the HypoGen run. The low Cost differences, high RMSD values, and low correlation values indicated that the data of cross validation generated after randomization produced hypotheses with no predictive values and we therefore believe that Hypo1 could be used for further database screening. The pharmacophore model, Hypo1, was used to screen NCI and Maybridge databases. In all, about 600 molecules were obtained as hits from *in silico* screening. To assess the drug-likeliness of these hits, a second screen, incorporating Lipinski's rule of 5 was used. A total of 99 molecules were obtained as hits after this screen. This second screen selects only those molecules that possess drug like properties. Finally we conclude that VS methods are becoming an integral part of the drug discovery process. In this chapter, a strategy for the screening of large compound libraries to obtain a limited set of prospective hits against Aurora A Kinase has been suggested. Using pharmacophore modeling and virtual screening, we have identified new lead molecules as Aurora A kinase inhibitors. Further modifications and addition of suitable functional groups to these new scaffolds will generate high affinity Aurora A kinase specific inhibitors.

The fifth chapter deals with the comprehensive study and bioinformatics overview of the ADAM (A Disintegrin And Metalloproteinase) and ADAMTS (A Disintegrin And Metalloproteinase with ThromboSpondin motifs) protein family members in the *Homo sapiens*, *Xenopus laevis*, *Drosophila melanogaster* and *Caenorhabitis elegans* genomes. ADAM family proteins are characterized by the presence of both disintegrin and metalloproteinase domains. Another closely related protein family is the ADAMTS. ADAMs and ADAMTSs are multi-domain protein families and play multiple roles in cell signaling, cell fusion, and cell-cell interactions. Members of ADAM family comprise a C- terminal transmembrane segment and are therefore cell surface proteins.

At the N- terminus, these proteins comprise a propeptide domain that contains a sequence motif similar to the "cysteine switch" of the matrixins. A zinc dependent proteinase domain follows this region, the proteolytic activity of ADAMs is due to the zinc proteinase domain and its activity is directed towards the extracellular domains of the transmembrane proteins. The mechanism of activation of these proteins involves the cleavage of the pro-part followed by conformational changes in the protein. Activation of metalloproteinases is an additional important mechanism for regulating activity of these enzymes. The adjacent disintegrin domain is responsible for the adhesive properties of the protein, thus mediating cell-cell and cell-matrix interactions. This is followed by the cysteine rich domain that supports cell adhesion. In addition to these four domains at the C-terminus, ADAMs also comprise an EGF domain, transmembrane segment and cytoplasmic tail responsible for signaling.

In this work, we have identified and analyzed ADAM & ADAMTS family proteins from four representative genomes to understand the distribution of the members and domain architecture. We report that, the human genome is encoded by 90 ADAMs and 92 ADAMTS genes. 11 ADAMs and 2 ADAMTS genes encode the frog genome. 19 ADAM and 6 ADAMTS genes encode the fly genome and 7 ADAMs and 8 ADAMTS genes encode the worm genome. We identified a different domain architecture pattern in ADAMTS protein family which is not similar to the previous report. The phylogenetic tree of ADAM and ADAMTS is organized into 6 Clades and the phylogenetic tree of the corresponding metalloproteinase domain is organized into 9 Clades. Classification of these proteins on the basis of domain architecture and cellular localization helps us to associate the proteins to the various biochemical pathways and the different cellular niches. Rearrangement and insertion of domains among the various ADAM and ADAMTS seems to be required for adapting them to diverse biological roles. Such

whole genome surveys and cross-genome comparisons using computations should be useful to design rational experiments and enhance our understanding of the specific biological roles of the ADAM and ADAMTS family proteins.

The sixth chapter deals with the diversity of Ser/Thr kinases in the genomes of various *Mycobacterium* species. Ser/Thr kinases (STKs) play a key role in cellular signal transduction and their biological functions have established their roles in cell growth and differentiation. These kinases were originally identified in eukaryotes and later identified in several prokaryotes. Kinases can be classified into two types based on their cellular location, as receptor or non-receptor kinases. Receptor kinases have a membrane spanning region and hence bound to the membrane, and the non-receptor kinases are cytosolic. The kinase domain of STKs has similar 3-D structures, and their catalytic domain comprises 270 amino acid residues. The 3-D structure comprises two domains, the N-terminal domain comprises mainly β-strands and the C-terminal domain comprises α-helices.

In this work we have identified 100 STKs and homologs in 10 completed mycobacterial genomes using profile based search methods adapted in PSI-BLAST, available at NCBI. The domain organization of these proteins has been studied in order to identify other coexisting domains. We observed that the kinase domain is highly conserved among all members of mycobacterial species. Kinase domain coexists with PASTA, PBPb, peptidyl-prolyl cis-trans isomerase and NERD domains, as well as NHL and Kelch repeats. We also observed that certain STKs consists of a conserved domain specifically present in mycobacterial species, while some STKs consist domains at N- and C-terminus that are also present in several actinomycetales indicating the restriction in the divergent evolution of STKs. We conclude that diversity in the STKs is evident from the fact that some STKs are present only in mycobacterial species while some STKs are common to

other members of Actinobacteria. This analysis will aid decipher the various biological functions of these proteins alongside the kinase activity.

CHAPTER 1

Introduction to Bioinformatics Tools in Genomic Data Analysis and Rational Drug Design

1.1 Bioinformatics tools in genomic data analysis

The information from the completed genome sequence projects stored in the databanks such as EMBL, GenBank for nucleotides and SWISSPROT, UNIPROT, NRDB for proteins are freely available to the public. Millions of sequences are available in these databanks that provide basic information about the respective proteins. Characterization of whole genomes is important to understand the structural and functional principles of living organisms. Whole genome comparisons provide clues on evolutionary relationships (Griffiths *et al.,* 1999). The wealth of sequence information brought about by the genome sequencing projects has led to the discovery of several computational tools, which enables the researchers to analyze the genes and proteins in whole genomes. These computational methods have been developed to solve the biological problems, using nucleotide and amino acid sequences and other related information.

Though DNA is the genetic material, it does not carry out the processes of life. This genetic code is transcribed and translated in the synthesis of protein molecules, which are present as the structures and molecular machines that make the cell function. Proteins contribute to almost all the events in the cells of a living organism. The polypeptide chain of a protein folds into a specific 3-D structure, which governs its function. Recent developments in the techniques of structure determination at atomic resolution, X-ray diffraction and nuclear magnetic resonance (NMR) spectroscopy, have enhanced the quality and speed of structural studies (Zhang & Kim, 2003). Nevertheless, current statistics still show that the known protein sequences vastly outnumber the available protein structures (48,778) deposited in protein data bank (PDB) so far. This is due to the inability to express, purify and crystallize some proteins as well as the intrinsic limitations of the structure determination techniques.

It becomes a challenge for the researchers to annotate this huge genomic data. About 40-50% of proteins in each genome are novel and are not bio-chemically and structurally characterized. Experimental characterization of each sequence is however time consuming. Therefore, adding value to the structure and function of these novel proteins by means of comparative studies, using computational tools is one of the challenges to the researchers worldwide. Sophisticated mathematical, statistical and computational techniques are developed to handle, analyze and add value to this flood of data. These studies have become one of the frontier areas of research in modern biology.

1.1.1 Nucleotide and protein databases:

Nucleotide sequence databases were first assembled at Los Alamos National Laboratory (LANL), by Walter Goad and colleagues in the GenBank database and at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. GenBank is now under the auspices of the National Center for Biotechnology Information (NCBI) (http://ncbi.nlm.nih.gov). The EMBL Data Library was founded in 1980 (http://www.ebi.ac.uk). The EMBL maintains DNA and protein sequence databases. In 1984 the DNA DataBank of Japan (DDBJ) came into existence (http://ddbj.nig.ac.jp). GenBank, EMBL and DDBJ have now formed the International Nucleotide Sequence Database Collaboration (http://www.ncbi.nlm.nih.gov/collab), which acts to facilitate exchange of data on a daily basis. Translated nucleotide sequence information is included in the Protein Information Resource (PIR) database at the National Biomedical Research Foundation in Washington, DC. GenBank entries provide a large amount of information describing the entry of each sequence. SwissProt is a protein sequence database and it is similar to the EMBL format. It contains more information about

the physical and biochemical properties of the protein. Researchers are encouraged to submit their newly obtained sequences directly to the various types of nucleotide or protein databases. UniProt is a comprehensive resource for protein sequence and annotation data. UniProt is a result of collaboration between the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the PIR. UniProt is the central hub for the collection of functional information on proteins, with accurate, consistent, and rich annotation. In addition to capturing the core data, each UniProt entry contains the amino acid sequence, protein name or description, taxonomic data and citation information. The "nr" database is the largest nucleotide database available through NCBI. It includes all GenBank, RefSeq Nucleotides, EMBL, DDBJ and PDB sequences.

The format of a database entry is such that each sequence file contains the information about the assigned accession number, source organism, function of the sequence, literature references, location of mRNAs, coding regions, positions of important mutations and sequence.

1.1.2 Sequence analysis tools:

Database searching

Comparison of a sequence with entries in a database is required to identify similar sequences that share homology. This can be done at both nucleotide and protein level. After proper validation of the results, multiple sequence alignments of these related sequences can built using consensus sequences of protein families that help in the identification of domains, motifs or functional sites. Detection of sequence similarity among different proteins has led to the classification of proteins on the basis of structure and function. It has been observed that most often similar sequences share similar structure and function. In addition, database searches are also used as primary requirement in identifying a structural homolog for an

unknown sequence. The most widely used programs for database searching are BLAST and FASTA.

1.1.2.1 Basic Local Alignment Search Tool (BLAST):

The BLAST program is used to identify sequence similar homologs from nucleotide or protein databases. The program takes a query sequence and searches it against the database selected by the user. It aligns the query sequence against every subject sequence in the database and the results are reported in the form of a ranked list followed by a series of individual sequence alignments, plus various statistics and score parameters. Every hit in that list is assigned with a similarity score S. Further, this score is analyzed to calculate the extent of such matching to occur by chance. For that purpose E-value is calculated for every hit. BLAST program finds regions of local similarity and calculates the statistical significance of matches (Altschul *et al.,* 1990) (http://www.ebi.ac.uk/blast2) and (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi/).

The BLAST program first dissects the query sequence into words of length k (3 for proteins and 11 for nucleotides). These words are searched against the database for matches, and scores are assigned with either BLOSUM (Henikoff & Henikoff, 1992) or PAM (Dayhoff, 1978) scoring matrices. Word hits that score more than T (neighborhood word score threshold) are extended in both directions to generate an alignment between segment pairs. The "T" parameter dictates the speed and sensitivity of the search. The extension process is stopped when the scores drop from its maximum achieved score and the segment pairs are referred to high scoring pairs (HSP). The next step is to determine those HSPs of sequences, which have score greater than a cut off score (S). S is determined empirically by examining a range of scores found by comparing random sequences and by choosing a value that is significantly greater. BLAST determines the statistical

significance of HSPs and generates sequence hits in the descending order of E (expectation value) and P (probability score) values. E and P values are different ways of representing the significance of the alignment. These values are the number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. The highly significant E or P values will be those close to 0 and lower values. BLAST also filters the low-complexity regions. Filtering is done by SEG and XNU filters and applied to the query sequence alone to make the search focus on more important parts of the sequence. These regions are marked with X in protein sequences and N in nucleotide sequences and are then ignored by BLAST.

The BLASTP offers various user defined options. A choice can be made on database to be searched. Based on the requirement, a user can switch to PDB or SWISSPROT database or a specific organism. Other options include selection of matrices, filters, adjustment of sensitivity and number of alignments etc. The default parameters for BLASTP include BLOSUM62 scoring matrix, a value of 11 is assigned for gap opening and a value of 1 for gap extension.

BLAST uses Smith-Waterman dynamic programming algorithm (Smith & Waterman, 1981a, 1981b). It detects local as well as global alignments using a heuristic approach. The exhaustive Smith-Waterman approach is too slow for searching large genomic databases such as GenBank. Therefore, the BLAST algorithm uses a heuristic approach that is slightly less accurate than Smith-Waterman but over 50 times faster. There are five different BLAST programs, which can be distinguished by the type of the query sequence (DNA or protein) and the type of the subject database.

BLASTP-compares an amino acid query sequence against a protein sequence database.
BLASTN-compares a nucleotide query sequence against a nucleotide sequence database.
BLASTX-compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.
TBLASTN-compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands).
TBLASTX-compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

1.1.2.2 Position-Specific Iterative BLAST (PSI-BLAST):

PSI-BLAST program is used for finding distant relatives of a protein. The program makes a list of all closely related proteins. These proteins are then combined into a "profile" that is a sort of average sequence. A query against the protein database is then run using this profile, and a larger group of proteins are found. This larger group is used to construct another profile, and the process is repeated (Altschul *et al.,* 1997) till one finds all related proteins in the database. This method is more reliable and used in several other programs such as PSI-PRED, PHD- secondary structure prediction methods. By including related proteins in the search, PSI-BLAST is much more sensitive in picking up distantly related proteins than using the standard protein-protein BLAST (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi/).

1.1.2.3 Pattern Hit Initiated BLAST (PHI-BLAST):

PHI-BLAST is a search program that combines matching of regular expressions with local alignments surrounding the match. The calculation of local alignments is done using a method very similar to gapped BLAST (Zhang, 1998). The most important features of the program have been incorporated into the BLAST framework partially for user convenience and partly so that PHI-BLAST may be combined seamlessly with PSI-BLAST. PHI-BLAST is most preferred to search for pattern occurrences because it filters out those cases where the pattern occurrence is probably random and not indicative of homology. PHI-BLAST may be preferable to other types of BLAST programs because it is faster and allows the user to express a rigid pattern occurrence requirement. PHI-BLAST uses Baeza-Yates and Gonnet (Baeza, 1992), Wu and Manber, 1992 algorithm, permits simple patterns to be represented in a single computer word and matches to be found very efficiently. PHI-BLAST was specifically designed to combine pattern search with the search for statistically significant sequence similarity (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi/).

1.1.2.4 FASTA:

The sequence similarity searching against nucleotide and protein databases are also carried out using the FASTA program. Pearson and Lipman (Pearson & Lipman, 1988) developed this program to achieve good sensitivity for similarity searching at high speed. This is achieved by performing optimized searches for local alignments using a substitution matrix. The search algorithm FASTA proceeds through four steps in determining a score for pair-wise similarity. FASTA searches for the matching sequence patterns called k-tup. Using k-tup FASTA builds a local alignment, scores this alignment and generates a list of sequences similar to a query sequence in the descending order. The high speed of

this program is achieved by using the observed pattern of word hits to identify potential matches before attempting to carry out the more time consuming optimized search. The speed and sensitivity is controlled by the k-tup parameter, which specifies the size of the word. Increasing the value of k-tup decreases the number of background hits. Not every word hit is investigated but instead initially looks for segments containing several nearby hits. This performs a database scan for similarity in a short time, so as to make such scans routinely possible (http://www.ebi.ac.uk/fasta33/).

1.1.3 Multiple sequence alignment:

In bioinformatics, a sequence alignment is the way of arranging the primary sequences of DNA, RNA or proteins in order to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. A multiple sequence alignment arranges more than two sequences such that residues with common structural positions or ancestral residues are aligned in the same column. If two sequences in an alignment share a common ancestor, mismatches can be interpreted as point mutation, and gaps as insertion or deletion mutations that are introduced in one or both lineages in the time since they diverged from one another. In protein sequence alignment, the degree of similarity between amino acid occupying a particular position in the sequence can be interpreted as a rough measure of how conserved a particular region or sequence motif is among lineage. The most similar regions in the multiple sequence alignment may represent structural domains or regions of functional importance. Multiple sequence alignments often provide an understanding of evolutionary history of sequences. If the sequences in the alignment are very well conserved, then it implies that these sequences are recently derived from a common ancestor sequence. Conversely, a group of poorly aligned sequences share a more complex

and distant evolutionary relationship. Multiple sequence alignments of related sequences can build consensus sequences of known families, domains, motifs or sites. These are useful in predicting the function and structure of proteins, and also in identifying new members of protein families. Combining these predictions with primary biochemical data can provide valuable insights into protein structure and function.

1.1.3.1 T-Coffee:

T-Coffee is a multiple sequence alignment program, which pre-processes a dataset of all pair-wise alignments between the sequences. This provides us with a library of alignment information that can be used to guide the progressive alignment. Intermediate alignments are then based not only on the sequences to be aligned next but also on how all of the sequences align with each other. This alignment information can be derived from heterogeneous sources such as a mixture of alignment programs and/or structure superposition (Notredame *et al.,* 2000). T-Coffee will compare all sequences two by two, producing a global alignment and a series of local alignments. The program will then combine all these alignments into a multiple alignment. The main characteristic of T-Coffee is that it allows one to combine results obtained with several alignment methods (http://www.ebi.ac.uk/t-coffee/).

1.1.3.2 CLUSTALW:

CLUSTALW is a fully automated program for global multiple alignment of nucleotide and protein sequences. This is very useful in designing experiments to test the function of specific proteins, in predicting the function and structure of proteins and in identifying new members of protein families. CLUSTALW

generates multiple sequence alignments of divergent sequences, and a phylogenetic tree based on a multiple alignment of sequences. It can manipulate existing alignments and carry out profile analysis (Thompson *et al.,* 1994). The majority of the automated multiple sequence alignments are based on the progressive approach of the Feng and Doolittle (Feng & Doolittle, 1987). CLUSTALW, developed by Thompson *et al.,* 1994, incorporated a number of improvements to the alignment algorithm, including sequence weighting, position-specific gap penalties and the choice of a suitable residue comparison matrix at each stage in the multiple alignments (http://www.ebi.ac.uk/Tools/clustalw2/).

CLUSTALW produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and aligns them up so that the identities, similarities and differences can be seen. The alignment in CLUSTALW is achieved via three steps: 1) pair-wise alignment, 2) guide-tree generation and 3) progressive alignment. Evolutionary relationships can be observed in a diagrammatic form by viewing Cladograms or Phylograms which will be discussed in detail under the section, "Phylogenetic analysis".

In CLUSTALW alignment, scores can be calculated by two methods, slow / accurate and fast / approximate, that use dynamic programming (Smith & Waterman, 1981a; 1981b) and Wilbur and Lipman methods (Wilbur & Lipman, 1983) respectively. CLUSTALW provides several options, such as use of slow or fast pair-wise alignments, nucleotide or protein sequences, protein weight matrix, gap open, gap extension, end gaps and gap distances. The default parameters for protein sequences are: Protein Gap Extension Penalty = 0.2; Protein matrix = Gonnet; Protein ENDGAP = -1; Protein GAPDIST = 4.

1.1.3.3 Phylogenetic analysis:

Phylogenetic analysis of a family of proteins or nucleic acids is the determination of how the family might have been derived during evolution. When the sequences found in two different organisms are similar, then they are likely to have been common ancestor. Phylogenetic analysis is an important area of sequence analysis. There are three main steps in phylogenetic analysis, these are 1) Multiple sequence alignment, 2) Distance calculation and 3) Tree construction. Multiple sequence alignment method first aligns the most closely related sequences and then sequentially adds more distantly related sequences or sets of sequences to these initial alignments. After obtaining the multiple sequence alignment each column is assumed to correspond to an individual site that has been evolving according to the observed sequence variation in the column. Distance methods build trees by grouping them according to their overall similarity. After calculating the distance, one can cluster the data together in a tree. Using the multiple sequence alignment method CLUSTALW, from the input sequences, the program calculates the pair-wise alignments and degree of similarity between all the pairs followed by the calculation of distance. Distance is commonly calculated by number of mismatches in the non-gapped positions between the two sequences. This value is divided with the number of non-gapped pairs. Thus, a distance matrix is generated for all the sequence pairs. Using the distance matrix and neighbor-joining method, CLUSTALW constructs the similarity tree. The root is placed in the middle of the longest chain of consecutive edges. For generating the phylogenetic trees one can use Bootstrapping method to obtain support values for each cluster. Pair-wise distances can be determined with protein parsimony method (Felsenstein, 1996). Representations of the calculated trees can be constructed using TreeView (Page, 1996) (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html).

1.1.4 Motif/pattern:

A sequence motif is a short conserved region found in a number of related protein sequences. Motifs often correspond to core structural and functional elements of the proteins. Their conserved nature allows them to be used to diagnose family membership and predict function. Genome sequencing provides the basis for a systematic analysis of all motifs that are present in a particular organism. Protein sequences can be searched for known motifs in databases such as PROSITE (http://www.expasy.org/prosite/), ProDom (http://prodom.prabi.fr/prodom/current/html/form.php), Pfam (http://www.sanger.ac.uk/Software/Pfam/) and PRINTS (http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS). These pattern and profile searches constitute an important resource for the classification of majority of the newly appearing protein sequences into known families.

1.1.5 Protein families and protein domains:

A protein family is a group of evolutionarily related proteins, and is often nearly synonymous with gene family. Proteins in a family descend from a common ancestor and typically have similar 3-D structure, function and significant sequence similarity. Many proteins comprise multiple independent structural and functional units or domains. Domains are structural and functional units that have specific biochemical activities. Due to evolutionary shuffling, different domains in a protein have evolved independently. A brief description of protein domains is discussed in this section.

1.1.5.1 Simple Modular Architecture Research Tool (SMART):

The SMART is an online resource (http://smart.embl.de/) used for protein domain identification and the analysis of protein domain architectures (Schultz *et al.,* 1998; Letunic *et al.,* 2006). SMART offers a high level of sensitivity and

specificity coupled with ease of use. It contains several unique aspects, including automatic seed alignment generation, detection of repeated motifs or domains and a protocol for combining domain predictions from homologous subfamilies. Visualization tools have been developed to allow analysis of gene intron-exon structure within the context of protein domain structure, and to align these displays to provide schematic comparisons of orthologous genes, or multiple transcripts from the same gene. It also allows batch retrieval of multiple entries.

1.1.5.2 INTERPRO:

INTERPRO is a database of protein families, domains and functional sites (Mulder *et al.,* 2005). It can be applied to predict the function and structure of unknown protein sequences (Zdobnov & Apweiler, 2001). INTERPRO provides an integrated view of the commonly used signature databases such as Pfam, PROSITE, PRINTS, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, Gene3D, and PANTHER. Signatures are manually integrated into INTERPRO entries that are curated to provide biological and functional information. INTERPRO covers over 78% of all proteins in the SWISSPROT and TrEMBL. The database is available for text and sequence based searches via a web server (http://www.ebi.ac.uk/InterProScan/).

1.1.5.3 PFAM:

PFAM is a comprehensive collection of protein domains and families and helps in the genome annotation (Bateman *et al.,* 2004). Each family in PFAM is represented by multiple sequence alignments and Hidden Markov Model (HMM) profile and can be used to view the domain organisation of proteins.

Structural data has been utilised to ensure that families in PFAM correspond to structural domains, and to improve domain based annotation. Predictions of

non-domain regions are also included. In addition to secondary structure, PFAM multiple sequence alignments now contain active site residues highlighted. New search tools, including taxonomy search and domain query, greatly add to the functionality and usability of the PFAM resource. Apart from the well known annotated domains, PFAM also provides the information of functionally uncharacterized families, known as Domains of Unknown Function (DUFs) and Uncharacterized Protein Families (UPFs). DUFs are families that have been created by PFAM and UPFs are those created by SWISSPROT and added to PFAM database (http://pfam.janelia.org/search).

1. 2 Drug design

The numbers of disease target proteins are expected to increase considerably due to the completion of many genome sequencing projects. With the arrival of high-throughput protein purification and 3-D structure determination methods, the importance of computational strategies to focus drug discovery efforts will be more useful and will facilitate the rapid development of novel therapies. Computational models provide an effective path to test hypotheses regarding mechanisms that simulate the behavior of biological systems at all levels, including molecular and cellular systems, and can provide models to such hypotheses.

In silico design of potential drugs for a given protein target can involve change of existing lead compounds, or *de novo* design of entirely new compounds. A lead compound is described as a compound that binds to the target protein and inhibits the activity of protein at a certain level. Lead compounds can be found through experimental high-throughput screening as well as virtual screening. Combinatorial libraries of analogs can be computationally designed and screened and a subset of identified compounds can be synthesized for validation.

Use of computational techniques in drug discovery and development process is highly attractive due to the ease of implementation and reliability. Different nomenclature is being applied to this area, including computer-aided drug design (CADD), computational drug design, computer-aided molecular design (CAMD), computer-aided molecular modeling (CAMM), rational drug design, *in silico* drug design, computer-aided rational drug design.

1.2.1 Computer-Aided Drug Design (CADD):

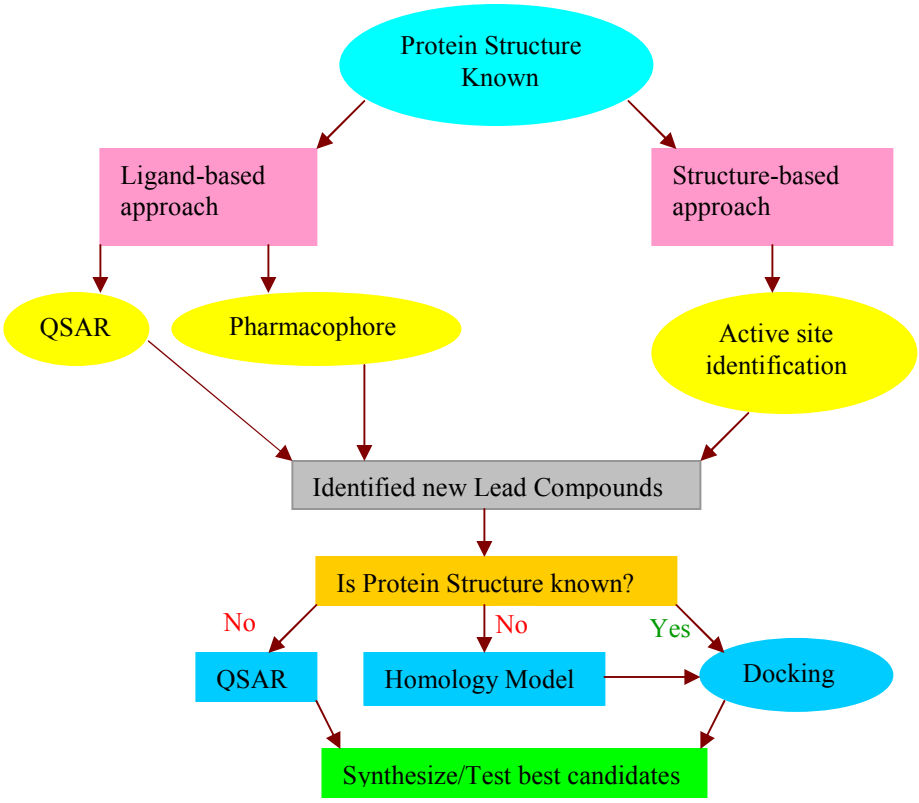
Computational chemistry methods were applied to aid the understanding of theoretical chemistry and pharmaceutical drug discovery. A variety of techniques for similarity searching aid scientists to find potentially active molecules and docking methods are helpful to model the binding of these molecules to desired protein targets. This application, known as CADD or molecular modeling, rapidly became an essential part of modern drug discovery, and thus the pharmaceutical industry became a strong supporter of the field (David & Gary, 2006). It was familiar since the 1960s, that computer-based methods can be useful in the discovery of new lead molecules and can eliminate chemical synthesis and screening of many irrelevant compounds. The rapid advance of computer technology and the development of new modeling software have made CADD an increasingly useful tool in drug design. An ideal computational method for lead molecule discovery should be able to generate structurally diverse lead molecules rapidly and should give an estimate of the binding affinities that would correlate with experimental values and ideally go on to become viable drugs (Mohan *et al.,* 2005).

1.2.2 Rational Drug Design (RDD):

The advent of molecular biology, coupled with advances in screening and synthetic chemistry technologies, has allowed a combination of both knowledge around the receptor and random screening to be used for drug discovery. Rational drug design is a process used in the biopharmaceutical industry to discover and develop new lead molecules. RDD uses a variety of computational methods to identify novel compounds, design compounds for selectivity, efficacy and safety, and thus develop compounds into clinical trial candidates. These methods fall into several natural categories such as, structure-based drug design, ligand-based drug

design, *de novo* design and homology modeling depending on how much information is available about drug targets and potential drug compounds (Jurgen, 2000). Advances in molecular biology, protein crystallography and computational chemistry since the 1980s have greatly aided the RDD paradigms. Figure 1.1 shows a flow chart that describes various approaches that enable RDD.

Figure 1.1: A Flow Chart Indicating Various Approaches to Rational Drug Design (Adopted and modified from ref. Parrill & Reddy 1999).



1.2.3 Computer-Aided Molecular Modeling (CAMM):

Identification of lead molecules with selective bioactivity, whether intended as potential therapeutics or as tools for experimental research, is essential in medicine and the life sciences. The discovery of a new drug to combat a disease takes years to decade and costs are too high (DiMasi *et al.,* 2003; Dickson & Gagnon, 2004). CAMM represents a potentially useful tool for this purpose, and it has made great advances into improving the odds of finding bioactive lead molecules (Burley & Park, 2005, Blundell *et al.,* 2002). CAMM is a truly successful tool and is provided in an easily available and usable format, for the benefit of the wider scientific community. The state of the art CAMM can be divided into two broad categories: ligand-based drug design and structure-based drug design depending upon the availability of 3-D structure of the target protein.

1.2.3.1 Ligand-Based Drug Design (LBDD):

Many receptors are not readily amenable to structure-based drug design. For example, many important receptors are membrane-bound proteins, which are often difficult to crystallize. In such cases, a lead compound or active ligand must be found, and then the structure of the ligand guides the drug design process in the LBDD. Structure-based drug design studies assemble information from already existing lead molecules or drugs that are active against the target biological molecule of interest. Based on the known information, a set of rules are framed to design either a new ligand or modify an existing ligand in order to improve its biological activity.

1.2.3.1.1 Quantitative Structure-Activity Relationship (QSAR):

The QSAR paradigm has evolved over the last hundred years to embody many quantitative approaches to structure-property correlations in physical organic

chemistry, biochemistry and molecular design. QSAR represents an attempt to correlate structural or property descriptors of compounds with activities. These physicochemical descriptors, which include parameters to account for lipophilicity, hydrophobicity, topology, electronic properties, and steric effects, are determined empirically by computational methods. The QSAR models are useful for various purposes including the prediction of activities of untested molecules.

Early QSAR methods related biological activity to the presence (or absence) of functional groups in a series of structurally related compounds (Free-Wilson model). Later the concept of quantitative correlation of physicochemical properties of molecules with their biological activities termed as QSAR was initiated by Corwin Hansch and coworkers during early 1960. Several 3D-QSAR modeling approaches have emerged in 1980s such as, active analog approach, molecular shape analysis, distance geometry and CoMFA. QSARs attempt to correlate physical and chemical properties of molecules to their biological activities. This can be achieved by simply using easily calculable descriptors (for example, molecular weight, number of rotatable bonds, LogP) and simple statistical methods such as multiple linear regression to build a model which describes the activity of the dataset and predict the activities for untested sets of compounds. These types of descriptors are simple to calculate and allow a relatively fast analysis, but often fail to take into account the 3-D nature of chemical structures (which obviously play a major role in ligand-receptor binding, and hence activity). 3D-QSAR uses probe-based sampling within a molecular lattice to determine 3-D properties of molecules (particularly steric and electrostatic values) and can then correlate these 3-D descriptors with biological activity. Hopfinger *et al*., 1997, introduced a fourth dimension to the 3D-QSAR modeling and termed it as 4D-QSAR analysis (Hopfinger *et al.,* 1997).

Hologram QSAR (HQSAR) is a relatively new technique, which does not require any physicochemical descriptors or 3-D structure to generate structure-activity models (Naumann & Lowis, 1997). It needs only 2-D structures and activity as input. HQSAR converts the molecules of a dataset into counts of their constituent fragments. These fragment counts are then related to biological data using partial least square analysis. HQSAR is a rapid, highly predictive QSAR technique. Results reported earlier show that HQSAR can readily produce highly predictive QSAR models over a wide variety of datasets.

1.2.3.1.2 Pharmacophore:

Pharmacophore generation is another method for LBDD. A pharmacophore is the spatial arrangement of key chemical features that are recognized by a receptor and are thus responsible for ligand-receptor binding (Gund & Güner, 2000). A pharmacophore is the ensemble of steric and electronic features that are necessary to ensure the optimal supramolecular interactions with a specific biological target (protein or DNA) structure and to trigger (or to block) its biological response. Pharmacophore models are constructed based on molecules of known biological activity and refined as more data is acquired in an iterative process. Alternatively, a pharmacophore can also be generated from the receptor structure. These models can be used for optimizing known ligands or for screening databases to find potential novel lead molecules suitable for further development (Renner *et al.,* 2004, Singh *et al.,* 2002).

1.2.3.2 Structure-Based Drug Design (SBDD):

In SBDD, the 3-D structure of a receptor (drug target) interacting with small molecules is used to guide drug discovery. The active site of a receptor is the area into which a chemical or biological molecule binds in order to initiate a

biochemical reaction. SBDD aims to create a molecule that will bind to the active site of a targeted receptor, thereby preventing the normal chemical reaction and ultimately halting the progression of the disease. Much of the work in the drug design is now based on the structure of the target and virtual screening of libraries. Captopril is the first success drug that came from the SBDD (Cushman *et al.,* 1977). After that several drugs came to the market, Carbonic anhydrase I and II targeted, Dorzolamide for glaucoma (Tsukamoto & Larsson, 2004), *Bcr-Abl* kinase targeted Imatinib for cancer (Druker *et al.,* 1996) and HIV protease targeted Lopinavir, Indinavir, Nelfinavir, Saquinavir and Ritonavir for AIDS (Verbesselt *et al.,* 2007) . Various approaches used for the SBDD are as follows.

1.2.3.2.1 Docking:

Protein-ligand docking is a molecular modeling technique. The goal of protein-ligand docking is to predict the position and orientation of a ligand when it is bound to a receptor. Docking is frequently used to predict the binding orientation of small molecule drug candidates to their protein targets in order to be able to in turn predict the affinity and activity of the small molecule. Hence docking plays an important role in the RDD studies (Kitchen *et al.,* 2004). Theoretically, docking is an energy optimization process concerned with the search of the lowest free energy binding mode of the ligand within the receptor binding site. In addition, protein flexibility is computationally expensive; therefore many of the existing docking programs treat the protein either as a rigid structure or allow flexibility only to the protein side chain functional groups. A good docking method places the ligand appropriately in the active site and then estimates the forces involved in the receptor-ligand recognition (electrostatic, van der Waals and hydrogen bonding). Docking comprises two components 1. Conformational searching and 2. Scoring.

Table 1.1 lists some of the existing docking methodologies and the strategies they use.

Table 1.1: List of Some of the Available Docking Methodologies and Their Strategies.

Searching Algorithm	Brief Description of Methodology	Examples of software
Monte Carlo (MC)	Stochastic method of generating conformations. Selection based on Metropolis criterion	Ligand Fit
Simulated Annealing (SA)	Random thermal motions are induced, through high temperatures, to explore the local search space. System is driven to a minimum energy conformation by decreasing temperature. SA usually combined with MC	MC-DOCK, AutoDock
Genetic Algorithm	Based on Darwin principles of evolution. 'Chromosome' encoding model parameters (like torsion angles) are varied stochastically. Populations are generated through genetic operations (crossover, mutation, migration). The fittest survives in the population.	GOLD, AutoDock
Matching Methods	Based on clique detection technique from graph theory. Ligand atoms are matched to the complimentary atoms in the receptor	FLOG, DOCK
Simulation Methods	Molecular dynamics simulations are used to generate conformations	DOCK, AutoDock, FlexX
QM-Polarized Ligand Docking algorithm	Quantum mechanic simulations are used to generate conformations	Glide

1.2.3.2.1.1 Conformation generation:

Drug molecules (ligands) usually bind to protein at a cavity of the receptor, which is called the binding site. It is usually assumed that the geometric constraints are the main determinants in this process. The energetic factors are also important, since molecules in nature are usually found in their low energy conformation. Ligand molecules have many degrees of freedom due to the rotatable bonds. Proteins, however, are much bigger molecules, with several hundreds of atoms. Given a ligand and a protein, finding whether they will bind to each other, and if they do bind their configuration in bound state is a difficult problem to predict as they involve many degrees of freedom of rotation. The conformations sought should each be with energy, less than a given threshold, and possibly with spatial features at specific positions in 3-D space.

1.2.3.2.1.2 Scoring:

The process of evaluating the particular conformation of molecule when bound to protein uses a number of descriptive features such as, number of intermolecular interactions including hydrogen bonds, hydrophobic contacts and van der Waals energy. Scoring function used in docking is a mathematical function whose values are proportional to the binding affinities of the lead molecules. A good scoring function should be able to give reliable estimates of binding affinities of structurally diverse lead molecules for different protein targets while considering the thermodynamic aspects of binding (Ajay & Murko, 1995).

Essentially, three types or classes of scoring functions are currently applied. Force field based empirical and knowledge-based scoring functions (Kitchen *et al.,* 2004). 1) Force field based methods are first principle methods that use force field parameters to score the van der Waals and electrostatic interactions between

receptor and ligand. The score includes receptor–ligand interaction energy and internal ligand energy. These methods do not require calibration or training with experimental binding data. 2) Empirical scoring functions are regression based functions derived from a large sample of crystal structures with known affinities for the bound ligands. These functions reflect a best fit with respect to the training set used, but rarely achieve generality. 3) Knowledge- based scoring functions are designed to reproduce experimental structures rather than binding energies. It evaluates the frequencies of particular type of interaction, the mutual distance between particular types of atoms across the interface, in databases of protein–ligand complexes.

1.2.3.2.2 *De Novo* ligand design:

De novo design uses structural information to develop a molecule that can fit into the active site by consecutively adding or joining molecular fragments instead of using libraries of existing compounds (Honma, 2003). Structure sampling is carried out by different methods such as: linking, growing, lattice-based sampling, random structure mutation, transitions driven by molecular dynamics simulations, and graph-based sampling. Apart from these, the ligand can also be built from recombination of bioactive conformations of known ligands for a particular target. Recombination is carried out by overlaying the known ligands and swapping the fragments of different ligands. This procedure is carried out recursively, so that the compounds that emerge from recombination are added to the pool of known active molecules and participate in subsequent cycles of recombination. The largest advantage of *de novo* design is its ability to develop novel scaffolds utilizing the whole chemical space (Schneider & Fechner, 2005). However, this method also suffers limitations such as: 1) synthetic feasibility is not

considered while constructing structures, and 2) the prediction of binding affinities for the designed structures is not accurate.

1.2.3.2.3 Virtual Screening (VS):

There is a growing pressure on the pharmaceutical industry to reduce the cost of drugs and the time taken to market them. The large number of targets made available in the last decade has created a new area for technologies that can rapidly identify quality lead candidates. VS is one such technology that is gaining increasing importance in the drug discovery process. VS is a reliable and inexpensive method currently being employed as a complementary approach to high-throughput screening. VS can be adopted irrespective of the structural information of the target receptor. In the absence of structural data of the receptor, VS using pharmacophore-based search is a major *in silico* tool. However, when the structure of the receptor is available, VS using both pharmacophore-based and docking techniques can be employed. VS is used as an initial screen for large databases to reduce the number of compounds that are to be screened experimentally (Lyne, 2002). VS protocols include ligand-based screens such as: 1-D filters (e.g. molecular weight), 2-D filters (similarity, substructure fingerprints) and 3-D filters (3-D pharmacophore, 3-D shape matching) and docking based on structure-based screening methods (Sirois *et al.,* 2004). The potential sources of error contributing to the identification of false positives and false negatives in VS include: 1) approximations in the scoring functions employed; 2) improper solvation terms; 3) neglect of protein flexibility and; 4) poor assessment of the protonation states of active site residues or ligands (Lyne *et al.,* 2004). Significant improvements in VS have been made by consensus scoring (Bissantz *et al.,* 2000) of multiple scoring functions and by clustering docking poses, from multiple docking tools before scoring (Paul & Rognan, 2002).

A more sophisticated approach in the current drug design process is "Docking based virtual screening" that allows the user to quickly screen large databases of potential drugs and score the ligand-protein interactions. Docking based virtual screening typically involves fast docking of a large number of chemical compounds against a protein binding site. But the accuracy of these screening approaches are underpinned by the molecular-docking methods, which in turn, depend on the computational algorithms for conformational sampling and scoring of different ligand binding conformations.

1.2.4 Protein structure and small molecule databases:

One of the most important and difficult problems in molecular biology is the protein folding problem. The structure-function relationship in proteins is directly concerned with correlating the 3-D structure of a protein to the primary sequence. The richest source of information about protein structure is the Protein Data Bank (PDB). The development of chemoinformatics has been hampered by the lack of large, publicly available, comprehensive repositories of molecules, in particular small molecules. They can be used as combinatorial building blocks for chemical synthesis (Schreiber, 2000), as molecular probes for perturbing and analyzing biological systems in chemical genomics and systems biology (Stockwell, 2004).

1.2.4.1 Protein Data Bank (PDB):

The PDB is a collection of individual "flat" text files, each of which contains the 3-D co-ordinates of one of the several thousands of protein structures determined by various experimental techniques such as X-ray crystallography and NMR. A variety of information associated with each structure is available through the RCSB PDB, including sequence details, atomic co-ordinates, crystallization

conditions, bound cofactors, metal ions or inhibitors, 3-D structural neighbors computed using various methods, derived geometric data, structure factors, 3-D images and a variety of links to other resources. Information about DNA and RNA structures is also available in these databases. Protein Data Bank maintained by the Rutgers, the State University of New Jersey, San Diego Supercomputer Center (SDSC), Skaggs School of Pharmacy and Pharmaceutical Sciences. It is available to researchers worldwide via the website www.rcsb.org/ (Berman *et al*., 2000). Till the year 2008, 48,778 experimentally determined structures have been deposited from scientists all over the world.

1.2.4.2 Small molecule databases:

Small molecules mainly comprise atoms such as carbon, hydrogen, nitrogen, oxygen, sulfur, halogens and phosphorus. These molecules play a fundamental role in organic chemistry and biology. There are various types of small molecule databases such as NCI (http://dtp.nci.nih.gov/), CSD (CCDC) (http://relibase.ccdc.cam.ac.uk/), Maybridge and Derwent (Accelrys), ACD (http://www.chemweb.com/databases/) and ChemBank (http://chembank.broad.harvard.edu/). All these databases contain millions of compounds and have useful information for each molecule, including its, structural, physical, chemical and biological properties. These databases are of great value for the screening, design and discovery of useful compounds.

1.2.5 The Lipniski rule of 5:

Experimental and computational approaches to estimate the solubility and permeability of small molecules for the drug discovery and development are described in the Lipinski "the rule of 5". In the discovery process 'the rule of 5' predicts that, the likelihood of poor absorption or permeation of the molecule is

greater when there are more than, 1) 5 H-bond donors (expressed as the sum of OHs and NHs) 2) 10 H-bond acceptors (expressed as the sum of Os and Ns) 3) The molecular weight is greater than 500, 4) The calculated Log P is greater than 5 and 5) The calculated Moriguchi octanol-water partition coefficient (MlogP) is greater than 4.15.

The problems and methods introduced in this chapter have been instrumental in the advance of our understanding of protein function, organization and structure. These computational analyses are aimed at speeding up the process for identification of protein structure, function and drug design. The computational tools aimed at analyzing the protein data are useful in supporting and explaining the experimental findings, assisting in the design of experiments and creating hypotheses.

1.3 References

Ajay & Murko, M. A. (1995). Computational methods to predict binding free energy in ligand-receptor complexes. *J. Med. Chem.* **38,** 4953-4967.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215,** 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25,** 3389-3402.

Baeza-Yates, R. & Gonnet, G. (1992). A New Approach to Text Searching. *Commun. Assoc. Comp. Mach*. **35,** 74–82.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C. & Eddy, S. R. (2004). The Pfam protein families database. *Nucl. Acids Res.* **32**, D138-D141.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Bourne, P. E. & Shindyalov, I. N. (2000).The Protein Data Bank. *Nucl. Acids Res.* **28,** 235-242.

Bissantz, C., Folkers, G. & Rognan, D. (2000). Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **43,** 4759-4767.

Blundell, T. L., Jhoti, H. & Abell, C. (2002). High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Discov*. **1**, 45–54.

Burley, S. K. & Park, F. (2005). Meeting the challenges of drug discovery: a multidisciplinary re-evaluation of current practices. *Genome Biol*. **6**, 330.

Cushman, D. W., Cheung, H. S., Sabo, E. F. & Ondetti, M. A. (1977). Design of potent competitive inhibitors of angiotensin-converting enzyme. Carboxyalkanoyl and mercaptoalkanoyl amino acids. *Biochemistry.* **16**, 5484-5491.

David, J. W. & Wiggins, G. D. (2006). Challenges for chemoinformatics education in drug discovery. *Drug discovery today.* **11**, 436-439.

Dayhoff, M. O. (1978). Survey of new data and computer methods of analysis. In *Atlas of protein sequence and structure*, pp.1-8, *Nat. Biomed. Res. Found*., Washington D.C.

Dickson, M. & Gagnon, P. (2004). The cost of new drug discovery and development. *Discovery Medicine*. **4**, 172-179.

DiMasi, J. A., Hansen, R. W. & Grabowski, H. (2003). The price of innovation: new estimates of drug development costs. *J. Health Econ.* **22**, 151-185.

Druker, B. J., Tamura, S., Buchdunger, E., Ohno, S., Segal, G. M., Fanning, S., Zimmermann, J. & Lydon, N. B. (1996). Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. *Nat. Med.* **2**, 561–66.

Felsenstein, J. (1996). Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.* **266**, 418-27.

Feng, D. F. & Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25**, 351-360.

Gund, P. & In Güner, O. F. (2000). Pharmacophore Perception, Development, and Use in Drug Design. *International University Line, La Jolla,* CA. 3-11.

Griffiths, A. J. F., Gelbart, W. M., Miller, J. H. & Lewontin, R. C. (1999). Modern Genetic Analysis. W. H. Freeman and Company.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci.* **89,** 10915-10919.

Honma, T. (2003). Recent advances in de novo design strategy for practical lead identification. *Med. Res. Rev.* **23**, 606-632.

Hopfinger, A. J., Wang, S., Tokarski, J. S., Jin, B., Albuquerque, M., Madhav, P. J. & Duraswami, C. (1997). Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **119,** 10509.

Jurgen, D. (2000). Drug Discovery: A Historical Perspective. *Science.* **287**, 1960-64.

Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **3,** 935-949.

Letunic, I., Copley, R. R., Pils, B., Pinkert, S., Schultz, J. & Bork, P. (2006). SMART 5: domains in the context of genomes and networks. *Nucleic Acids Research*. **34,** D257-D260.

Lyne, P. D. (2002). Structure-based virtual screening: an overview. *Drug Discov. Today*. **7**, 1047-1055.

Lyne, P. D., Kenny, P. W., Cosgrove, D. A., Deng, C., Zabludoff, S., Wendoloski, J. J. & Ashwell, S. (2004). Identification of compounds with nanomolar binding affinity for checkpoint kinase-1 using knowledge-based virtual screening. *J. Med. Chem.* **47,** 1962-1968.

Mohan, V., Gibbs, A. C., Cummings, M. D., Jaeger, E. P. & DesJarlais, R. L. (2005). Docking: successes and challenges. *Curr. Pharm. Des*. **11,** 323–333.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., Copley, R., Courcelle, E., Das, U., Durbin, R., Fleischmann, W., Gough, J., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McDowall, J., Mitchell, A., Nikolskaya, A. N., Orchard, S., Pagni, M., Ponting, C. P., Quevillon, E., Selengut, J., Sigrist, C. J., Silventoinen, V., Studholme, D. J., Vaughan, R. & Wu, C. H. (2005). InterPro, progress and status in 2005. *Nucl. Acids Res.* **33,** D201-D205.

Naumann, T. & Lowis, D. R. (1997). First International Electronic Conference on Synthetic organic Chemistry (ECSOC-1).

Notredame, C., Higgins, D. G. & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology.* **302**, 205-217.

Page, R. D. (1996). TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **4**, 357-8.

Parrill, A. L. & Reddy, M. R. Overview of rational drug design. (1999). In *Rational Drug Design: novel methodology and practical applications*, Rami Reddy, M., Parrill, A. L., Ed., American Chemical Society: Washington, DC. **719**, 1-11.

Pearson, W. R. & Lipman, D. J. (1988). Improved Tools for Biological Sequence Analysis. *Proc. Natl Acad. Sci. USA*. **85,** 2444-2448.

Paul, N. & Rognan, D. (2002). ConsDock: A new program for the consensus analysis of protein-ligand interactions. *Proteins.* **47,** 521-533.

Renner, S. & Schneider, G. (2004). Fuzzy pharmacophore models from molecular alignments for correlation-vector-based virtual screening. *J. Med. Chem.* **47**, 4653-4664.

Schneider, G. & Fechner, U. (2005). Computer-based *de novo* design of drug-like molecules. *Nat. Rev. Drug Discov.* **4,** 649-663.

Schreiber, S. L. (2000). Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science*. **287**, 1964–1969.

Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci.* USA. **95**, 5857–5864.

Singh, J., van Vlijmen, H., Liao, Y., Lee, W., Cornebise, M., Harris, M., Shu, I., Gill, A., Cuervo, J. H., Abraham, W. M. & Adams, S. P. (2002). Identification of potent and novel alpha4beta1 antagonists using *in silico* screening. *J. Med. Chem.* **45,** 2988-2993.

Sirois, S., Wei, D. Q., Du, Q. & Chou, K. C. (2004). Virtual screening for SARS-CoV protease based on KZ7088 pharmacophore points. *J. Chem. Inf. Comput. Sci.* **44,** 1111-1122.

Smith, T. F. & Waterman, M. S. (1981a). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-197.

Smith, T. F. & Waterman, M. S. (1981b). Comparison of biosequences. *Adv. Appl. Math.* **2**, 482-489.

Stockwell, B. R. (2004). Exploring biology with small organic molecules. *Nature*. **432**, 846–854.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22,** 4673-4680.

Tsukamoto & Larsson. (2004). Aqueous Humor Flow in Normal Human Eyes Treated With Brimonidine and Dorzolamide, Alone and in Combination. *Arch. Ophthalmol.* **122**, 190-193.

Verbesselt, R., Van Wijngaerden, E. & J. de Hoon. (2007). Simultaneous determination of 8 HIV protease inhibitors in human plasma by isocratic high-performance liquid chromatography with combined use of UV and fluorescence detection: Amprenavir, indinavir, atazanavir, ritonavir, lopinavir, saquinavir, nelfinavir and M8-nelfinavir metabolite. *Journal of Chromatography.* **845***,* 51-60.

Wilbur, W. J. & Lipman, D. J. (1983). Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl Acad. Sci. USA*. **80**,726-730.

Wu, S. & Manber, U. (1992). Fast Text Searching Allowing Errors. *Commun. Assoc. Comp. Mach*. **35**, 83–91.

Zdobnov, E. M. & Apweiler, R. (2001)."InterProScan - an integration platform for the signature-recognition methods in InterPro". *Bioinformatics.* **17**, 847-8.

Zhang, C. & Kim, S. H. (2003). Overview of structural genomics: from structure to function. *Curr. Opin. Chem. Biol.* **7**, 28–32.

Zhang, Z., Schäffer, A. A., Miller, W., Madden, T. L., Lipman, D. J., Koonin, E. V. & Altschul, S. F. (1998). Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res*. **26,** 3986-90.

CHAPTER 2

Docking of Phosphonate and Trehalose Analog Inhibitors Into *M. tuberculosis* Mycolyltransferase Ag85C Comparison of the Two Scoring Fitness Functions GoldScore and ChemScore, in the GOLD Software

2.1 Introduction

Tuberculosis (TB) is an infection caused by the bacterium *Mycobacterium tuberculosis*. It is a major disease infecting two billion people, or approximately one-third of the world's population. These numbers are rising also in the developed countries due to the compromised immune systems, typically as a result of immunosuppressive drugs. This population is at particular risk of infection and active tuberculosis disease. Truly new TB drugs have been developed for nearly 40 years since the introduction of rifampicin in 1965. It is thus necessary to search for new and more effective antimycobacterial agents with novel mechanisms of action to combat the emergence of drug resistance and to shorten the duration of therapy (Duncan, 2004). *M. tuberculosis* is surrounded by a complex envelope of unusually low permeability, which contributes to the resistance of this bacteria to host defense mechanisms (Daffe & Draper, 1998; Jarlier *et al.,* 1994).

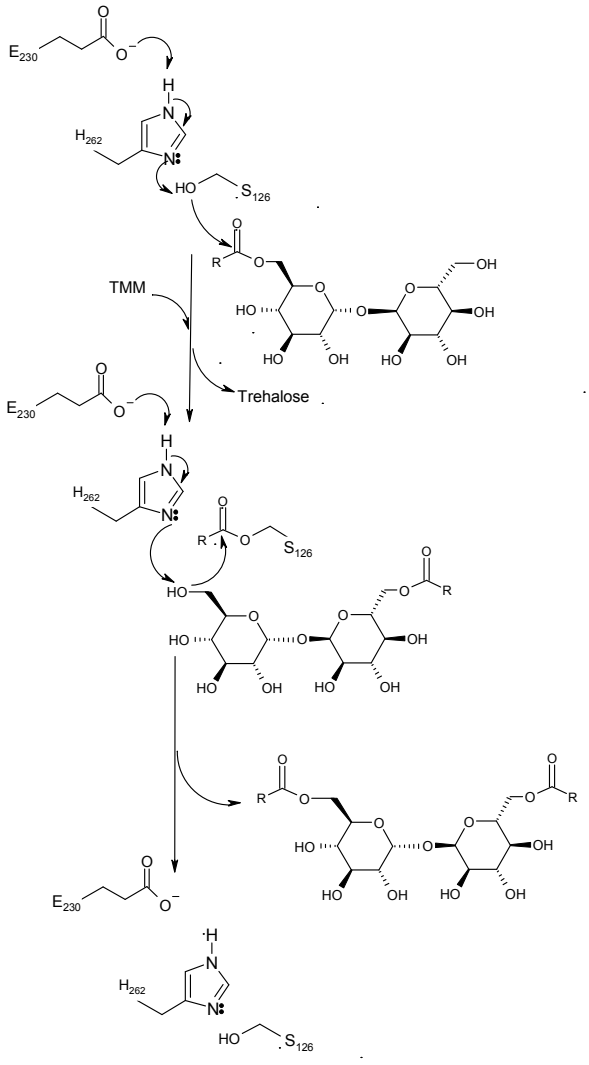
Since several important TB drugs such as isoniazid, ethambutol and ethionamide target mycobacterial cell wall biosynthesis, enzymes involved in this pathway remain the preferred targets in anti-TB drug research (Zhang & Amzel, 2002). The mycobacterial cell wall consists of three major components forming the mycolyl-arabinogalactan-peptidoglycan (mAGP) complex, among which mycolic acids constitute the outermost layer (Brennan & Nikaido, 1995). Mycolic acids are high molecular weight α- alkyl, β-hydroxy fatty acids unique to Mycobacterium and related genera. In the mycobacterial cell wall envelope, they are present as free glycolipids, mainly α, α' trehalosemonomycolate (TMM) and α, α' trehalosedimycolate (TDM) and as esters of the terminal pentaarabinofuranosyl units of arabinogalactan.

Mycolyltransferases are the most abundantly expressed proteins by intracellular mycobacteria within infected human monocytes, and the proteins are localized to the phagosomal space and mycobacterial cell wall (Harth *et al*., 1996). In *M. tuberculosis*, a major secreted protein complex, antigen 85, constitutes three

proteins antigen 85A, 85B and 85C (Wiker & Harboe 1992), that are responsible for the synthesis of cell envelope. These enzymes catalyze the transfer of mycolyl residue from one molecule of TMM to another TMM leading to the formation of TDM (Belisle *et al.,* 1997) and are hence termed mycolyltransferases. These enzymes have the ability to bind the human fibronectin (Abou-Zeid *et al.,* 1998) through which the bacterium gets attached to host cells. Due to their immunodominant and secretory nature, the components of Ag85 complex represent promising protective antigen candidates. Also, the cell envelope of *M. tuberculosis* has been one of the targets of antimycobacterial agents. The pyrazinamide was shown to inhibit fatty acid synthase type I, which in turn, provides precursors for fatty acid elongation to long-chain mycolic acids by fatty acid synthase II (Schroeder *et al.,* 2002). Mycolic acid biosynthesis is known to be essential for mycobacterium growth, in particular trehalose mycolates aid in virulence of the organism and the structure of mycolates has been found to be important for initial replication and persistence *in vivo* (Glickman *et al.,* 2000). A mutated *M. tuberculosis* strain lacking the functional Ag85C gene showed a 40% decrease in the amount of cell wall linked mycolic acids indicating its role in cell wall synthesis (Jackson *et al.,* 1999).

The importance of mycolyltransferases in pathogenesis and cell wall formation makes them an attractive target for drug design studies. A series of 6, 6'-bis(sulfonamido), N,N'-dialkylamino and related derivatives of 6,6'-di deoxytrehalose were designed and synthesized to inhibit the Ag85 complex (Rose *et al.,* 2002). Due to their location on the cell wall and involvement in the cell wall biogenesis, these enzymes represent useful drug targets for the development of new antitubercular agents. Inhibition of these proteins may block the synthesis of the cell wall and therefore inhibit the growth of *M. tuberculosis*.

Scheme 2.1. Schematic Representation of the Catalytic Mechanism of the Mycolyl Transfer Reaction by Mycolyltransferases.



The 3-D crystal structures of Ag85A (PDB_ID: 1SFR) (Ronning *et al.*, 2004) Ag85B (PDB_IDs: 1F0N, 1F0P) (Anderson *et al.,* 2001) and Ag85C (PDB_IDs: 1DQZ, 1DQY, 1VA5) (Ronning *et al.,* 2000) were determined for both native and substrate bound forms. The protein structure corresponds to a α/β hydrolase fold and the catalytic triad responsible for the mycolyltransferase activity

comprises the amino acid residues Ser624, Glu728 and His760 (numbering according to the PDB_ID: 1DQZ, B chain). This active site is typical of a serine protease family member.

Structural comparison of mycolyltransferases revealed that their backbone superimposes with an overall RMSD of 0.577 Å. The catalytic residues are highly superimposed in these structures indicating that the structure and function of these isozymes are highly similar. Minor deviations are observed in the loop region connecting strand 4 and helix 3, comprising the sequence motif, 85Q-S-N-G-Q-N90 in Ag85C. This region is away from the substrate binding site and therefore will not affect substrate or inhibitor binding to the enzyme. A putative picture of the catalytic mechanism of the mycolyl transfer reaction has been proposed by Ronning *et al.,* 2000 as shown in Scheme 2.1. In the first step, catalytic serine attacks the carboxyl carbon of TMM molecule to give a mycolyl-enzyme intermediate and a free trehalose. In the next step, the 6'-OH group of the second TMM molecule attacks the carboxylate carbon of the acyl-enzyme intermediate to yield TDM. Both acylation and deacylation of the enzyme, proceed via a high-energy tetrahedral transition state. It is known that substituted tetrahedral phosphorus (V) species like phosphonates, phosphonamidates and phosphinates represent good tetrahedral transition state analogous of both amide and ester bond cleavage or formation. Incorporation of the phosphorus based transition state mimetics into the substrate or product analogs generally leads to useful enzyme inhibitors. Gobec *et al.,* 2004 synthesized a series of phosphonate inhibitors and reported the inhibition of Ag85C mycolyltransferase activity in the presence of inhibitors. Rose *et al.,* 2002 studied the antimycobacterial activity of trehalose analogs against *M. tuberculosis* H37Ra and clinical isolates of *M. avium.*

Docking in a true sense is the formation of non-covalent protein–ligand complexes *in silico*. Given the structure of a protein and a ligand, the task is to predict the structure of the complex. Conceptually, docking is an energy optimization process concerned with the search of the lowest free energy binding

mode of a ligand within a protein binding site. Docking constitutes two components, conformation searching and scoring. The scoring function used in docking is a mathematical function whose values are proportional to the binding affinities of the lead molecules. A good scoring function should be able to give reliable estimates of binding affinities of structurally diverse lead molecules for different protein targets while considering the thermodynamic aspects of binding (Ajay & Murko, 1995).

Several groups have earlier carried out a comparative evaluation of the docking methods and scoring functions (Perola *et al.,* 2004, Wang *et al.,* 2003, Wang *et al.,* 2004). Results from Verdonk *et al.,* 2003 indicated that for "drug like" and "fragment like" ligands, the docking accuracies obtained from GoldScore and ChemScore are similar, while for larger ligands, GoldScore gave better results. Docking involves the identification of ligand conformation and orientation in the protein binding pockets. The scoring functions are helpful to predict the biological activity of the ligand. The aim of the present work is to study the docking of phosphonate and trehalose analog inhibitors into the active site of Ag85C using the GOLD software (Jones *et al.,* 1997)**.** In this work, we compared the scoring functions, GoldScore and ChemScore that are available in the GOLD docking software. The three mycolyltransferases have highly similar structure, we have therefore chosen the 3-D structure of Ag85C (PDB_ID: 1DQZ) as a representative structure for docking studies. Also, the inhibition of Ag85C by the phosphonate inhibitors by Gobec *et al.,* 2004 has been studied and the experimental median Inhibitory Concentration (IC_{50}) values are reported. Since this enzyme functions as a serine protease, these docking studies will help characterize the inhibitor binding site. The information about the inhibitor binding site can be used to design Ag85C specific inhibitors that would not interfere with the other physiologically important ubiquitous serine proteases in humans that act as host organism for *M. tuberculosis* pathogen.

2.2 Methods

2.2.1 Preparation of the Protein:

The crystal structure co-ordinates of the Ag85C (PDB_ID: 1DQZ) were obtained from the PDB (http://www.rcsb.org/) and the B chain was selected for docking studies. All hydrogen atoms were added to the protein, including those necessary to define the correct ionization and tautomeric states of amino acid residues such as Asp, Ser, Glu, Arg and His using Cache software (www.cachesoftware.com/cache). A two stage protocol was set up for the energy minimization of protein using Hyperchem7 (Hypercube Inc.), molecular modeling software. In the first stage, all hydrogen atoms in the protein were allowed to optimize. The hydrogen locations are not specified by the X-ray structure but these are necessary to improve the hydrogen bond geometries. In the second stage, all protein atoms were allowed to relax. Minimization in both stages was performed using 100 steps of steepest descent and 2000 steps of conjugate gradient algorithm.

2.2.2 Binding site analysis:

The binding site identification of mycolyltransferases, Ag85A, Ag85B and Ag85C was carried out using the Binding Site Analysis module available in InsightII 2005. The default parameters for grid size 1.00 Å and site open size 7.00 Å were used in binding site calculation.

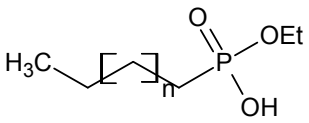
2.2.3 Selection of docking molecules:

A set of 17 phosphonate inhibitors were taken from the reported literature (Gobec *et al.,* 2004) for the docking studies. A list of these molecules is provided in Figure 2.1a and a schematic representation of the molecules with least IC_{50} values is provided in Figure 2.1b. Similarly, a set of 9 trehalose analogs are taken from the reported literature (Rose *et al.,* 2002) for docking studies. A list of these molecules is provided in Figure 2.1c and a schematic representation of the molecules with

least Minimal Inhibitory Concentration (MIC) values are provided in Figure 2.1d. We have adapted the numbering of these inhibitors as represented in the original papers (Gobec *et al.,* 2004, Rose *et al.,* 2002).

Figure 2.1a. A list of 17 Phosphonate Inhibitors Docked Into the Active Site of Mycolyltransferase, Ag85C. The Reported IC_{50} Values of the Corresponding Molecules are also Indicated.

3rd group



IC_{50} μM

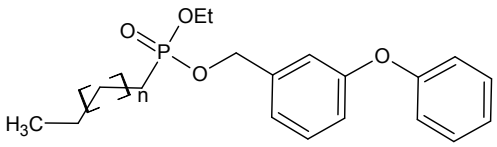
3a, n=1= 430.72

3b, n=3= 3.56

3c, n=4= 1.06

3d, n=6= 471.31

4th group



IC_{50} μM

4a, n=1 =2.01

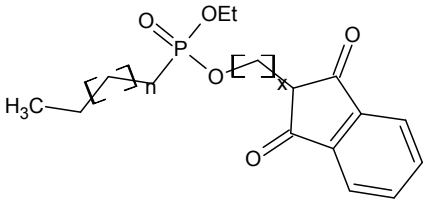
4b, n= 3=42.55

4c, n= 4=21.03

4d, n=6=14.83

4e, =11=862.6

5th group



IC_{50} μM

5a, n=1, x= 1 = 10.0

5b, n= 1, x= 2 = 50.74

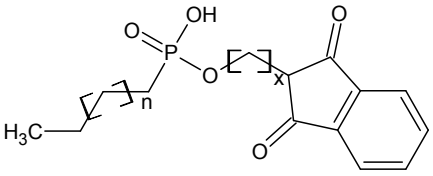
5e, n= 4, x= 1 = 1.34

5g, n= 6, x= 1 = 87.09

5h, n= 6, x =2 = 25.67

5i, n= 11, x=1 = 40.99

6th group



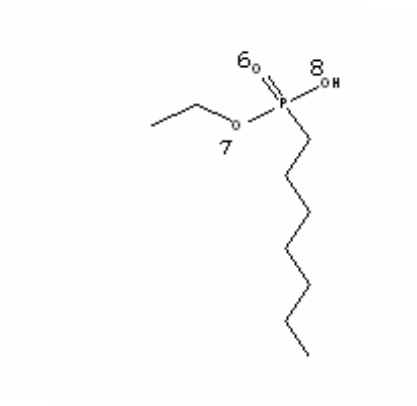
IC_{50} μM

6a, n=3, x= 2 = 4.39

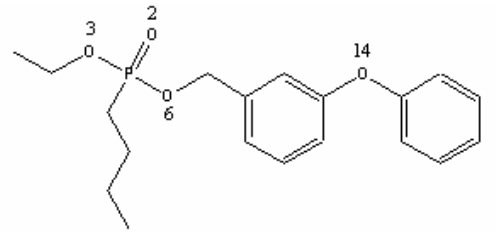
6b, n= 4, x= 2 = 1.47

Figure 2.1b. A Schematic Representation of the Phosphonate Inhibitors with Least IC_{50} Values.

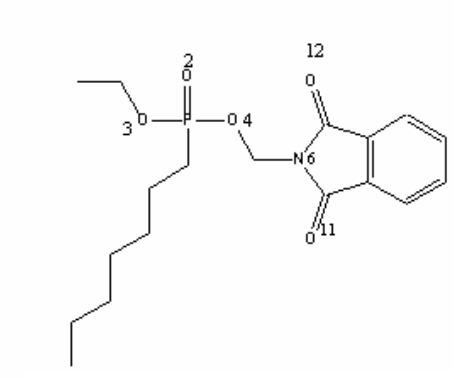
3c



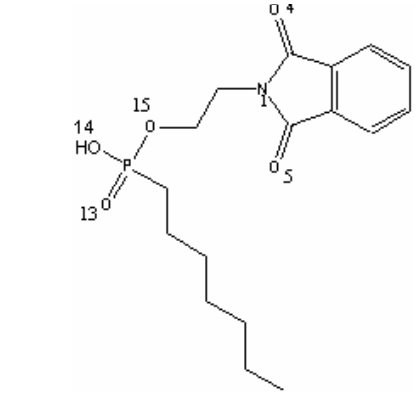
4a



5e



6b



2.2.4 Molecular Modeling:

The phosphonate and trehalose analog inhibitors were built using the Hyperchem7 (Hypercube, Inc). The structures were energy minimized using the steepest descent algorithm with a convergence gradient value of 0.001 kcal/mol. Further, the geometry optimization was carried out for each compound using the MOPAC 6 package and the semi empirical AM1 Hamiltonian.

2.2.5 Docking:

Docking was carried out using GOLD (Genetic Optimization for Ligand Docking) software, that uses the Genetic algorithm (GA). This method allows a partial flexibility of protein and full flexibility of ligand. All water molecules and hetero atoms were removed from the protein to evaluate the two scoring functions in GOLD software. For each of the 25 independent GA runs, a maximum number of 100000 GA operations were performed on a set of five groups with a population size of 100 individuals. Operator weights for crossover, mutation, and migration were set to 95, 95 and 10 respectively. Default cutoff values of 2.5 Å (dH-X) for hydrogen bonds and 4.0 Å for van der Waals distance were employed. When the top three solutions attained RMSD values within 1.5 Å, GA docking was terminated. The RMSD values for the docking calculations were based on the RMSD matrix of the ranked solutions. We observed that the best ranked solutions were always among the first 10 GA runs, and the conformation of molecules based on the best fitness score was further analyzed. During docking, the default algorithm speed was selected, and the ligand binding site in the Ag85C, was defined within 10 Å residues with the centroid as Ser624 side chain OG atom. All the 17 phosphonate and 9 trehalose analog inhibitors were analyzed using docking studies and the molecules with least IC_{50} and MIC values respectively in each group are described in detail in the results and discussion section.

A good docking program also takes the binding affinity of atoms into account. To assess this, a scoring function has to be taken into consideration. On

viewing results of fast-working algorithms, a trend of disagreement between model and actual structure is seen with regards to the binding affinity. GOLD offers a choice of fitness functions, GoldScore, ChemScore and also a user defined score. GoldScore and ChemScore are both equally reliable, but they may give different predictions depending upon the problem.

2.2.6 GoldScore fitness function:

GoldScore performs a force field based scoring function and is made up of four components: 1. Protein-ligand hydrogen bond energy (external H-bond); 2. Protein-ligand van der Waals energy (external vdw); 3. Ligand internal van der Waals energy (internal vdw); 4. Ligand intramolecular hydrogen bond energy (internal- H- bond). The external vdw score is multiplied by a factor of 1.375 when the total fitness score is computed. This is an empirical correction to encourage protein-ligand hydrophobic contact. The fitness function has been optimized for the prediction of ligand binding positions.

$$\text{GoldScore} = S_{(hb\_ext)} + S_{(vdw\_ext)} + S_{(hb\_int)} + S_{(vdw\_int)}$$

Where $S_{(hb\_ext)}$ is the protein-ligand hydrogen bond score, $S_{(vdw\_ext)}$ is the protein-ligand van der Waals score, $S_{(hb\_int)}$ is the score from intramolecular hydrogen bond in the ligand and $S_{(vdw\_int)}$ is the score from intramolecular strain in the ligand.

2.2.7 ChemScore fitness function:

ChemScore is an empirical scoring function to estimate the free energy of ligand binding to protein. It uses simple contact terms to estimate lipophilic and metal-ligand binding contributions, including hydrogen bonding interactions. It does not differentiate between different types of hydrogen bonds based on the nature and geometry of the interaction. It adds a clash penalty and internal torsion terms, which influence against close contacts in docking and poor internal conformations. Covalent and constraint scores may also be included. ChemScore estimates the total free energy change that occurs on ligand binding to protein as per the equation given below

$$\Delta G \text{ binding} = \Delta G_o + \Delta G_{hbond} + \Delta G_{metal} + \Delta G_{lipo} + \Delta G_{rot}$$

$$\Delta G_0 = \nu_0$$
$$\Delta G_{hbond} = \nu_1 P_{hbond}$$
$$\Delta G_{metal} = \nu_2 P_{metal}$$
$$\Delta G_{lipo} = \nu_3 P_{lipo}$$
$$\Delta G_{rot} = \nu_4 P_{rot}$$

Each component of this equation is the product of a term dependent on the magnitude of a particular physical contribution to free energy (e.g. hydrogen bonding) and a scale factor determined by regression.

$$\text{ChemScore} = \Delta G_{\text{binding}} + P_{clash} + C_{internal} P_{internal} + (C_{covalent} P_{covalent} + P_{constraint})$$

Here the *v* terms are regression coefficients and the *P* terms represent the various types of physical contributions to binding. The final ChemScore value is obtained by adding in a clash penalty and internal torsion terms, which influence against close contacts in docking and poor internal conformations. Covalent and constraint scores may also be included in the ChemScore.

2.3 Results and Discussion

Binding site analysis of mycolyltransferases, Ag85A, Ag85B and Ag85C identified that, their binding pockets are identical and the largest binding pocket overlaps with the trehalose binding position in the PDB_ID: 1F0P, see Figure 2.1e. This further increased our confidence that the structures of mycolyltransferases are highly similar and we have therefore taken Ag85C as the representative structure for docking studies. The docking of phosphonate and trehalose analog inhibitors into the active site of mycolyltransferase, Ag85C was carried out using the GOLD software and the docking evaluations were made on the basis of two fitness functions, GoldScore and ChemScore. For both the scoring functions, the docking conformation generated with best fitness score has been analyzed using two different criteria. These are 1) ligand binding position and 2) fitness function score comparison.

2.3.1 Phosphonate inhibitors:

2.3.1.1 Criteria-1:

Ligand binding position: From the docking of molecule 3c into the active site of Ag85C, we observed that, the hydrogen bond acceptor atoms are 6, 7 and atom 8 is also a hydrogen bond donor, as indicated in Figure 2.1b.

GoldScore fitness function: One hydrogen bond was observed when the molecule 3c was docked into the active site of Ag85C, see Figure 2.2a and Table 2.1a. The alkyl chain of inhibitor entered into hydrophobic region of the protein comprising amino acid residues indicated in Table 2.2a. The best RMSD was found to be 0.28 Å.

Figure 2.1e. Structural overlay of the Trehalose Binding Site in Ag85 A, B and C. Trehalose Molecule Indicated in Ball and Stick Model. Ag85A is Represented in Violet, Ag85B is Represented in Orange and Ag85C is Represented in Light Blue.

ChemScore fitness function: Two hydrogen bonds were observed when the molecule 3c was docked into active site of Ag85C, see Figure 2.3a and Table 2.1a. The alkyl chain of inhibitor entered into hydrophobic region of protein comprising amino acid residues indicated in Table 2.2a. The best RMSD was found to be 0.62 Å.

Table 2.1a. List of Hydrogen Bonding Interactions Between the Phosphonate Inhibitors and Mycolyltransferase, Ag85C.

Molecule	Goldscore				ChemScore			
	No. of H- bonds	Protein residue atom	Ligand atom	H- bond distance Å	No. of H- bonds	Protein Residue atom	Ligand atom	H- bond distance Å
3c	1	Arg541 (NH)	6(O)	2.78	2	Arg541 (NH) Ser624 (OG)	6(O) 8(O)	3.04 2.94
4a	1	Arg541 (NH)	14(O)	2.71	1	His760 (C=O)	2(O)	3.08
5e	2	Arg541 (NH) Trp762 (NE2)	6(O) 11(O)	2.81 3.12	1	Ser624 (OG)	12(O)	2.65
6b	3	Arg541 (NH) Trp762 (NE1) Ser624 (OG)	5(O) 4(O) 13(O)	2.64 3.11 2.93	1	Ser624 (OG)	5(O)	2.41

From the docking of molecule 4a into the active site of Ag85C, we observed that the hydrogen bond acceptor atoms are 2, 3, 6 and 14 as indicated in Figure 2.1b.

GoldScore fitness function: One hydrogen bond was observed when the molecule 4a was docked into the active site of Ag85C, see Figure 2.2b and Table 2.1a. The alkyl chain of inhibitor entered into hydrophobic region of protein comprising amino acid residues indicated in Table 2.2a. The best RMSD was found to be 0.90 Å.

ChemScore fitness function: One hydrogen bond was observed when the molecule 4a was docked into active site of Ag85C, see Figure 2.3b and Table 2.1a. The alkyl chain of inhibitor entered into hydrophobic region of protein comprising amino acid residues, as shown in Table 2.2a. The best RMSD was found to be 0.80 Å.

From the docking of molecule 5e into the active site of Ag85C, we observed that, the hydrogen bond acceptor atoms are 2, 3, 4, 6, 11 and 12 as indicated in Figure 2.1b.

GoldScore fitness function: Two hydrogen bonds were observed when the molecule 5e was docked into active site of Ag85C, see Figure 2.2c and Table 2.1a. The alkyl chain of inhibitor entered into hydrophobic region of protein comprising amino acid residues, as indicated in Table 2.2a. The best RMSD was found to be 0.61 Å.

ChemScore fitness function: One hydrogen bonding interaction was observed when the molecule 5e was docked into the active site of Ag85C, see Figure 2.3c and Table 2.1a. The inhibitor alkyl chain entered into hydrophobic region of protein comprising amino acid residues as indicated in Table 2.2a. The best RMSD was found to be 0.82 Å.

From the docking of molecule 6b into the active site of Ag85C, we observed that, the hydrogen bond acceptor atoms are 1, 4, 5, 13, 14 and 15 and the atom 14 is also a hydrogen bond donor, as indicated in Figure 2.1b.

GoldScore fitness function: Three hydrogen bonds were observed when molecule 6b was docked into the active site of Ag85C, see Figure 2.2d and Table 2.1a. The alkyl chain of ligand entered into hydrophobic region of protein comprising residues as indicated in Table 2.2a. The best RMSD was found to be 1.04 Å.

ChemScore fitness function: One hydrogen bonding interaction was observed when molecule 6b was docked into the active site of Ag85C with a distance of 2.41 Å, see Figure 2.3d and Table 2.1a. The alkyl chain entered into hydrophobic region of protein comprising amino acid residues as indicated in Table 2.2a. The best RMSD was found to be 1.01 Å.

Figure 2.2. GoldScore Based Interactions of Molecules (a) 3c, (b) 4a (c) 5e and (d) 6b, Docked Into the Active Site of Mycolyltransferase, Ag85C.

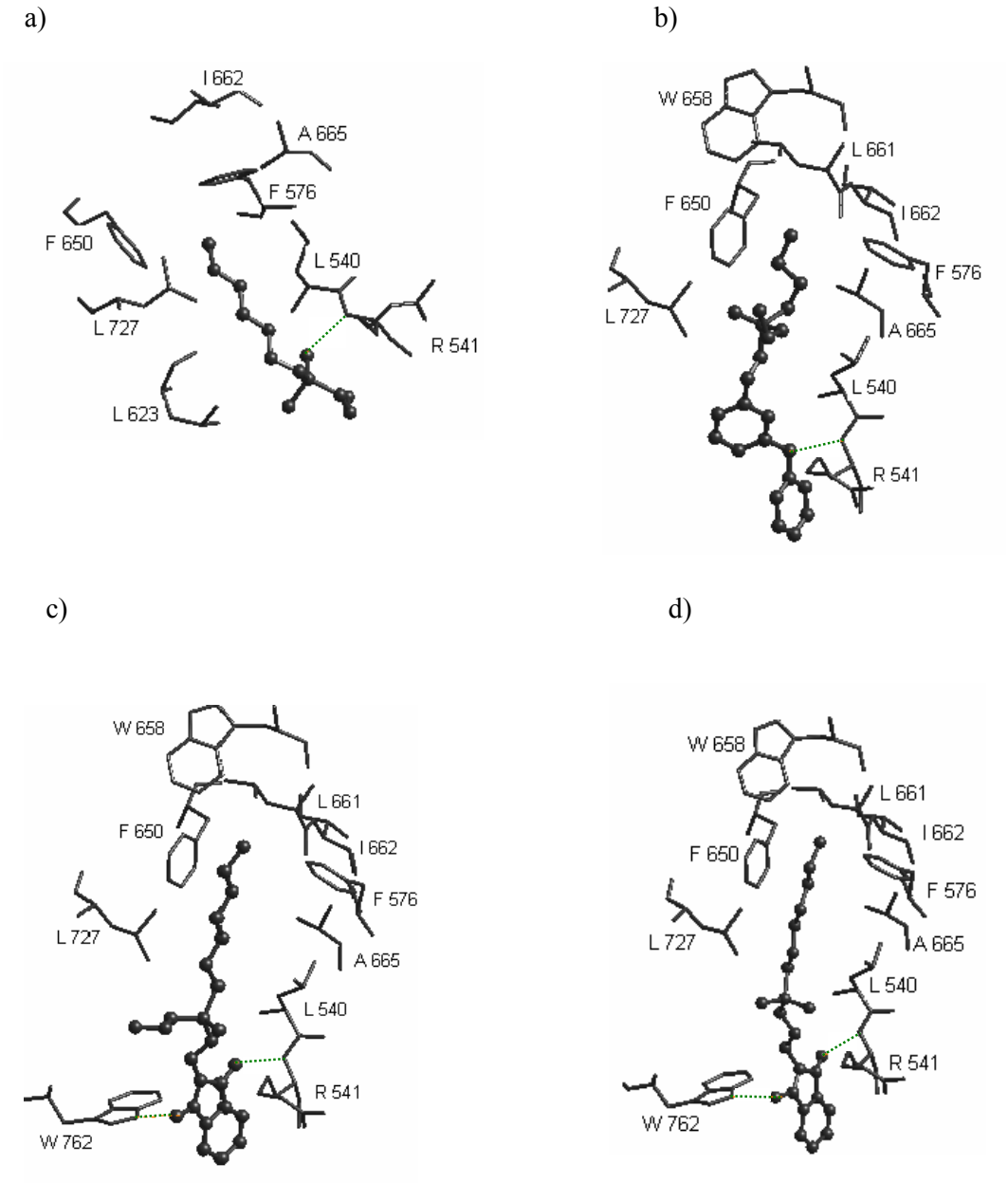
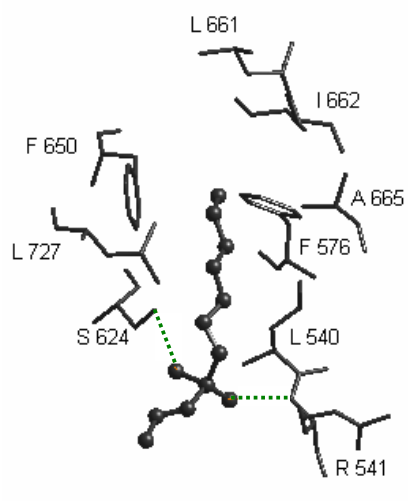
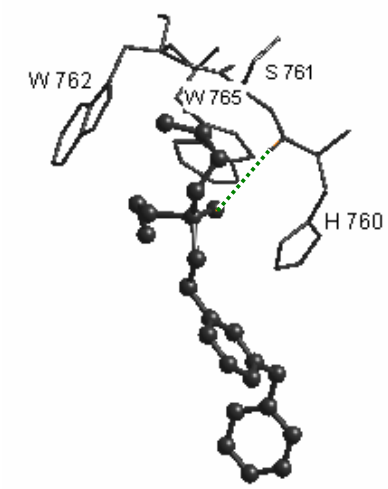


Figure 2.3. ChemScore Based Interactions of Molecules (a) 3c, (b) 4a (c) 5e and (d) 6b, Docked Into the Active Site of Mycolyltransferase, Ag85C.

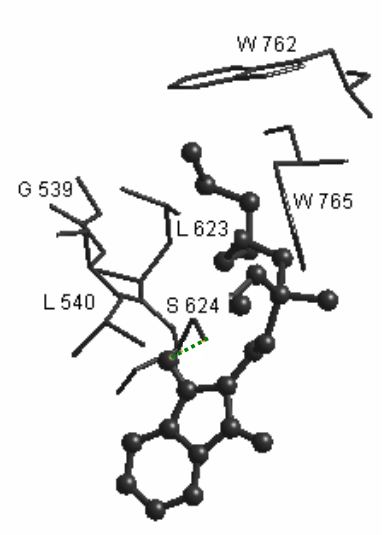
a)



b)



c)



d)

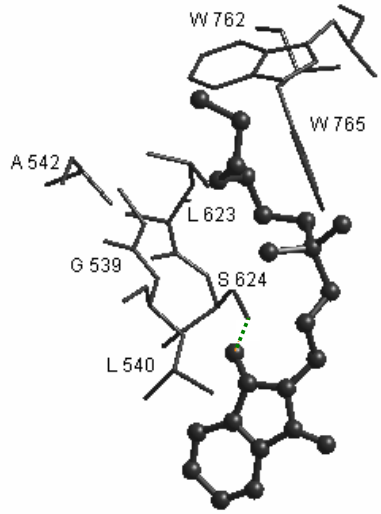


Table 2.2a. List of Amino Acid Residues Contributing to the Hydrophobic Pocket in Docking of Phosphonate Inhibitors Into the Mycolyltransferase, Ag85C.

Phosphonate inhibitor	GoldScore	ChemScore
3c	Gly539, Leu540, Ala665, Phe576, Phe650, Leu727, Leu623, Ser624 and Met625	Gly539, Leu540, Ala665, Phe576, Phe650, Leu727, Ile662, Ser624, Met625 and Trp658
4a	Leu540, Met625, Phe576, Leu727, Phe650, Leu661, Ile662, Ala665, Trp658 and Ser624	Trp762, Ala724 and His760
5e	Leu540, Leu727, Phe650, Leu661, Ile662, Trp658 and Ser624	Leu540, Asp538, Gly539, Leu623 and Ser624
6b	Leu727, Phe650, Leu661, Ile662, Trp658 and His760	Gly548, Gly539, Leu623, Asp538 and Trp762

2.3.1.2 Criteria-2:

Fitness function score comparison: The phosphonate inhibitors were docked into the active site of Ag85C and evaluated based on both the fitness scoring functions-GoldScore and ChemScore available in the GOLD software. We observed that the GoldScore correlated well with the reported IC_{50} values as shown in Figure 2.4a. There are no compounds with high GoldScore and low activity (high IC_{50} values). The exceptions are molecules 3b and 3c that bind the Ag85C with a GoldScore of 50.30 and 51.18, while the IC_{50} values are 3.5 μM and 1.06 μM, respectively. We hypothesize that the low GoldScore for these molecules is due to the presence of OH substitution alone on the phosphate head group. When the docked molecules were evaluated based on the ChemScore, there is no such correlation with the IC_{50} values, see Figure 2.4b. These results indicate that the GoldScore is a better parameter to assess the binding of phosphonate inhibitors to Ag85C and there is an overall 70-80 % correlation between the GoldScore and IC_{50} values.

Figure 2.4a. Correlation Between IC50 and GoldScore values for GoldScore Based Docking of Phosphonate Inhibitors Into the Mycolyltransferase, Ag85C.

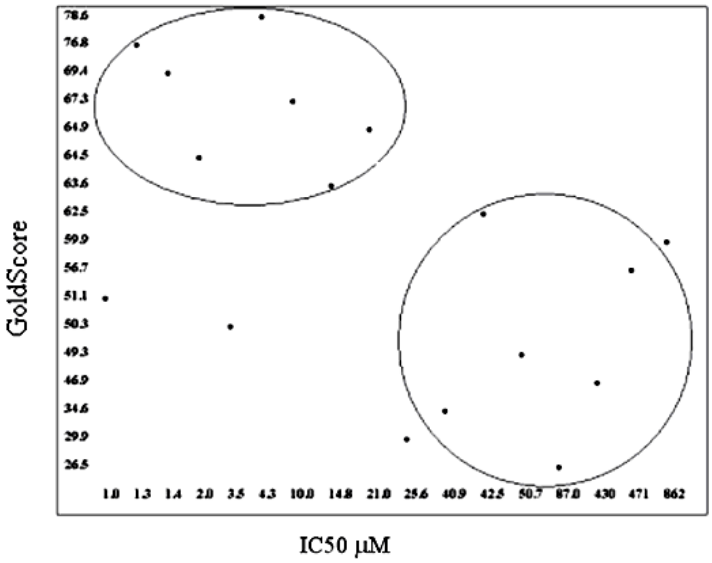
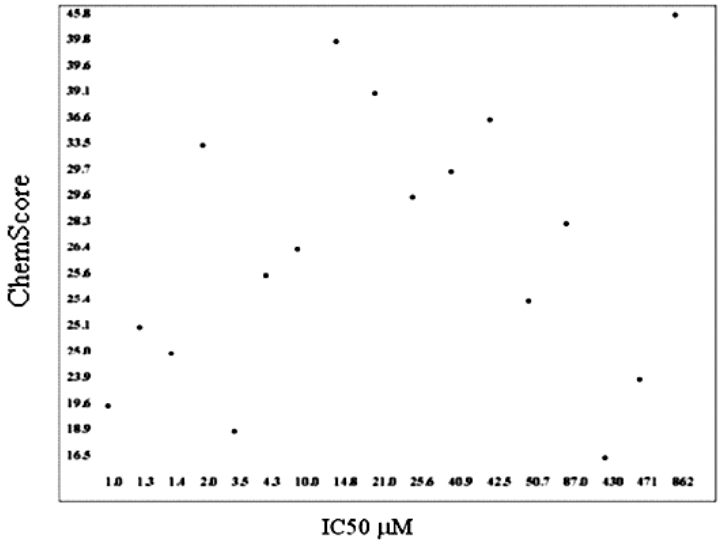


Figure 2.4b. Correlation Between IC50 and GoldScore values for ChemScore Based Docking of Phosphonate Inhibitors Into the Mycolyltransferase, Ag85C.



2.3.2 Trehalose analogs:

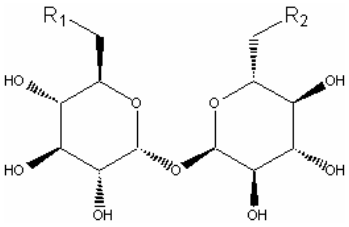
The trehalose analog inhibitors were docked into the active site of mycolyltransferase, Ag85C and we analyzed the conformations of protein-ligand complexes that resulted in highest value of the GoldScore and ChemScore.

Ligand binding position: From the docking of molecule 11b into the active site of Ag85C, we observed that, the hydrogen bond acceptor atoms are 7-12, 19-23, 25, 26, 54 and 55. The atoms 8-10, 12, 19-21, 23 are also hydrogen bond donors, as indicated in Figure 2.1d.

GoldScore fitness function: Seven hydrogen bonds were observed when the molecule 11b was docked into the active site of Ag85C, see Figure 2.5a and Table 2.1b. The alkyl chain of the inhibitor entered into the hydrophobic region of the protein comprising amino acid residues indicated in Table 2.2b. The best RMSD was found to be 2.16 Å.

ChemScore fitness function: Four hydrogen bonds were observed when the molecule 11b was docked into active site of Ag85C, see Figure 2.6a and Table 2.1b. The alkyl chain of the inhibitor entered into hydrophobic region of protein comprising amino acid residues, as indicated in Table 2.2b. The best RMSD was found to be 2.64 Å.

Figure 2.1c. A List of 9 Trehalose Analog Inhibitors Docked Into the Active Site of Mycolyltransferase, Ag85C. The Reported MIC Values of the Corresponding Molecules are Also Indicated.

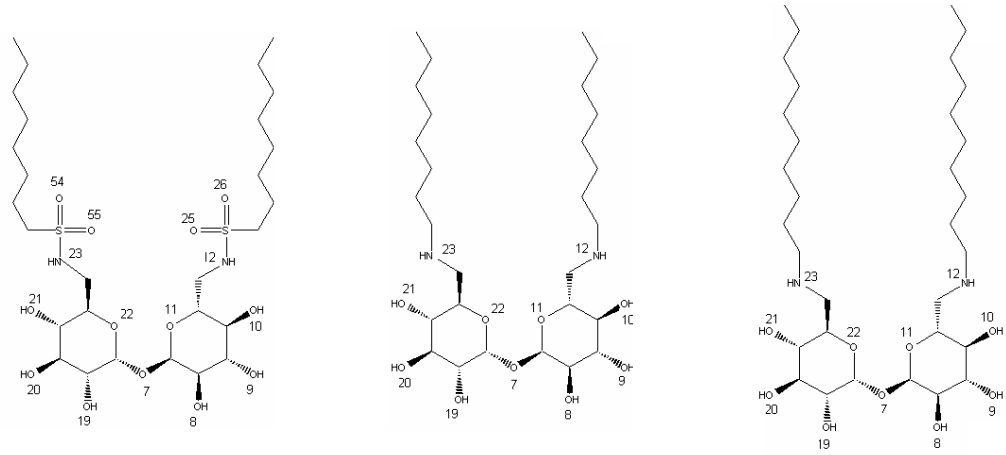


11th group	15th group
MIC (μg/mL)	MIC (μg/mL)
11a) R1=R2= $CH_3(CH_2)_3SO_2NH$- >128	15a) R1=R2=$CH_3(CH_2)_7NH$- 4
11b) R1=R2= $CH_3(CH_2)_7SO_2NH$- 16-32	15b) R1=R2= $CH_3(CH_2)_{11}NH$- 8
11c) R1=R2= $CH_3(CH_2)_{15}SO_2NH$- >32	15c) R1=R2= $CH_3(CH_2)_{17}N(CH3)$- >128

18th group
MIC (μg/mL)
18a) R1=R2= $CH3(CH_2)_5NH$- 128
18b) R1=R2= $CH_3(CH_2)_9NH$- >1.3<13
18c) R1=R2=$CH_3(CH_2)_3CH(Et)CH_2NH$- 32

Figure 2.1d. A Schematic Representation of the Trehalose Analog Inhibitors with Least MIC Values.

11b 15a 18b



From the docking of molecule 15a into the active site of Ag85C, we observed that, the hydrogen bond acceptor atoms are 7-12, 19-23. The hydrogen bond donor atoms are 8-10, 12, 19-21 and 23 as indicated in Figure 2.1d.

GoldScore fitness function: Four hydrogen bonds were observed when the molecule 15a was docked into the active site of Ag85C, see Figure 2.5b and Table 2.1b. The alkyl chain of the inhibitor entered into the hydrophobic region of the protein comprising amino acid residues as indicated in Table 2.2b. The best RMSD was found to be 0.78 Å.

ChemScore fitness function: Three hydrogen bonds were observed when the molecule 15a was docked into active site of Ag85C, see Figure 2.6b and Table 2.1b. The alkyl chain of inhibitor entered into hydrophobic region of protein comprising amino acid residues, as indicated in Table 2.2b. The best RMSD was found to be 1.60 Å.

Table 2.1b. List of Hydrogen Bonding Interactions Between the Trehalose Analog Inhibitors and Mycolyltransferase, Ag85C.

	GoldScore				ChemScore			
Mol Name	No of H bonds	Protein residue atom	Ligand atom	H Bond distance Å	No of H bonds	Protein residue atom	Ligand atom	H bond distance Å
11b	7	Arg541 (N)	25 (O)	2.51	4	Arg541 (O)	25 (O)	3.13
		Trp762 (NE1)	26 (O)	3.63		Trp762 (NE1)	26 (O)	2.80
		Gln543 (NE2)	9 (O)	3.04		Gln543 (NE2)	11 (O)	2.95
		Gln543 (NE2)	10 (O)	2.52		Asn552 (O)	19 (O)	3.07
		Asn552 (OD1)	23 (N)	3.28				
		Asn552 (O)	54 (O)	3.21				
15a	4	Arg541 (N)	22 (O)	3.85	3	Arg541 (N)	20 (O)	3.19
		Ser624 (OG)	19 (O)	3.21		Ser624 (O)	12 (N)	3.34
		His760 (O)	20 (O)	3.02		His760 (O)	23 (N)	3.03
		Trp762 (NE1)	23 (N)	3.43				
18b	3	Ser624 (OG)	20 (O)	3.30	2	His760 (NE2)	19 (O)	3.30
		His760 (O)	10 (O)	3.19		His760 (NE2)	20 (O)	2.76
		Trp762 (NE1)	23 (N)	2.94				

Table 2.2b. List of Amino Acid Residues Contributing to the Hydrophobic Pockets in Docking of Trehalose Analog Inhibitors Into Mycolyltransferase, Ag85C.

Trehalose inhibitors	GoldScore		ChemScore	
	R1	R2	R1	R2
11b	Ala665, Leu661, Ile662, Leu540, Gly539, Met625 Leu727, Phe650, Leu623	Pro763, Thr553, Ala555, Pro554, Ile551, Asp550	Ala542, Gle539, Leu540, Leu727, Leu727, Trp765	Tyr546, Gln525, Tyr510, Asp550, Ile551
15a	Leu540, Ala665, Ile662, Leu661, Trp658, Leu730, Phe650, Leu727	Ala542, Gln543, Gly548, Asn552	Leu540, Phe576, Ala665, Ile662, Leu661, Trp658, Phe650, Leu730	Gly548, Asn552, Pro763, Gly719, Trp762
18b	Leu540, Ala665, Ile662, Leu661, Trp658, Leu730, Phe650, Leu727	Ala542, Gln543, Gly548, Asn547, Ile551, Asn552	Ala665, Ile662, Leu661, Trp658, Leu730, Phe650, Leu727	Ala542, Gln543, Asp545, Ile551, Gly548

From the docking of molecule 18b into the active site of Ag85C, we observed that, the hydrogen bond acceptor atoms are 7-12, 19-23, the hydrogen bond donor atoms are 8-10, 12, 19-21, 23 as indicated in Figure 2.1d.

GoldScore fitness function: Three hydrogen bonds were observed when the molecule 18b was docked into the active site of Ag85C, see Figure 2.5c and Table 2.1b. The alkyl chain of the inhibitor entered into the hydrophobic region of the protein comprising amino acid residues, as indicated in Table 2.2b. The best RMSD was found to be 2.34 Å.

ChemScore fitness function: Two hydrogen bonds were observed when the molecule 18b was docked into active site of Ag85C, see Figure 2.6c and Table 2.1b. The alkyl chain of inhibitor entered into hydrophobic region of protein comprising amino acid residues, as indicated in Table 2.2b. The best RMSD was found to be 2.28 Å.

For the trehalose analog inhibitors, Rose *et al.,* 2002 have reported the MIC values in order to estimate their activity as antimycobacterial agents; therefore, we have not correlated them with the GoldScore and ChemScore.

In the crystal structure of Ag85B complexed with trehalose substrate (PDB_ID: 1F0P), it has been shown that Ser124 (equivalent of Ser624 in Ag85C, B chain) forms a hydrogen bonding interaction with atom O6 of trehalose (Ronning *et al.,* 2004) (Figure 2.1f). From our docking studies, we have shown that this serine forms a hydrogen bonding interactions with phosphonate and trehalose analog inhibitors (see Tables 2.1a and 2.1b). Further, we also identified additional amino acid residues, Arg541 and Trp762 important for inhibitor recognition mediated via non-bonding interactions. Sequence analysis identified that Arg541, Ser624 and Trp762 are highly conserved in mycolyltransferases (Adindla *et al.,* 2004) of *M. tuberculosis*. Current docking studies using the GoldScore fitness function reported a hydrophobic tunnel to accommodate the hydrocarbon alkyl chain of phosphonate and R1 alkyl chain of trehalose analog inhibitors. This hydrophobic tunnel comprises the residues Leu540, Phe576, Phe650, Trp658, Leu661, Ile662, Leu623, Met625, Ala665 and Leu727. In the reported crystal structures of Ag85B and Ag85C (Ronning *et al.,* 2004; Anderson *et al.,* 2001), the authors have described a hydrophobic tunnel comprising these residues to accommodate α- alkyl chain of mycolic acids. It was also proposed that β- alkyl chain of mycolic acids would fit the trough on the surface of the protein. Docking of trehalose analog inhibitors identified the R2 alkyl chain to bind the region comprising amino acid residues Tyr510, Gln525, Ala542, Gln543, Asp545, Gly548, Asp550, Ile551, Asn552, Pro554 that are located on the surface of the protein. These results indicate that the predicted nonbonding interactions and hydrophobic region to accommodate the alkyl chains predicted in this work agree with the experimental data obtained from crystal structures.

Figure 2.1f. A Schematic View of the Ag85B Complexed with Trehalose Substrate. Trehalose Molecule is Indicated in Ball and Stick.

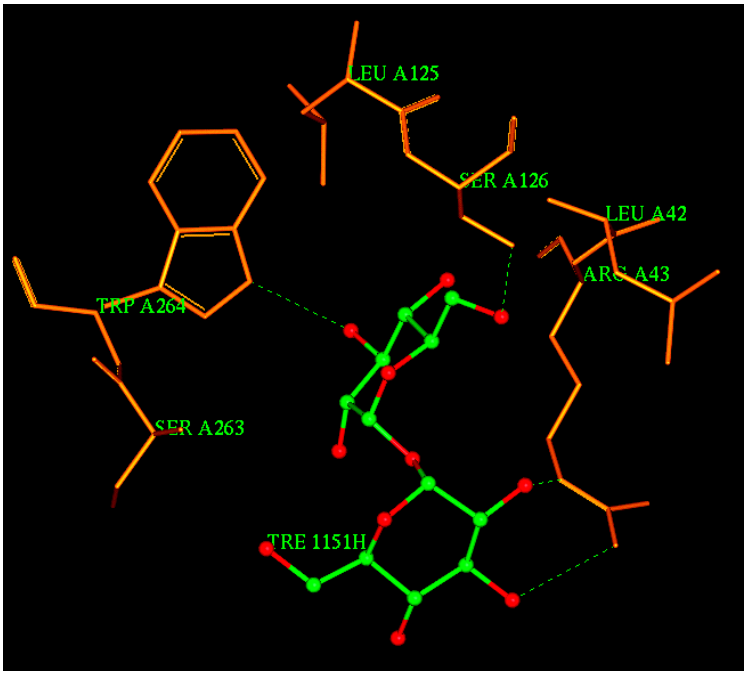
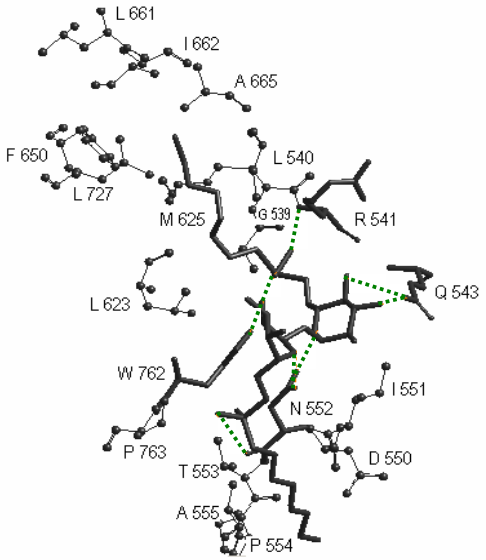
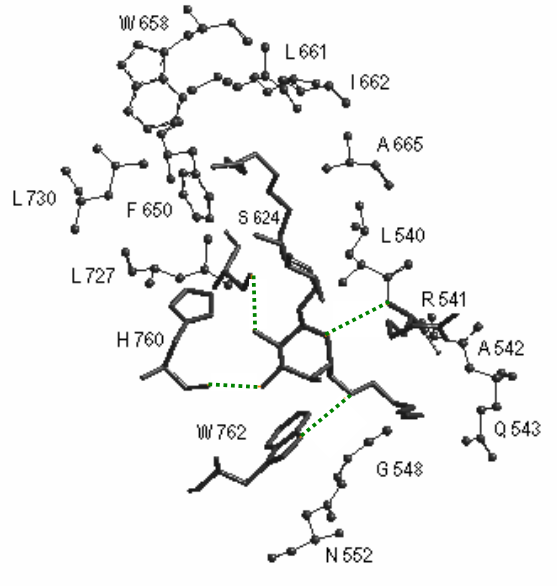


Figure 2.5. GoldScore Based Interactions of Molecules (a) 11b, (b) 15a and (c) 18b Docked Into the Active Site of Mycolyltransferase, Ag85C.

a)



b)



c)

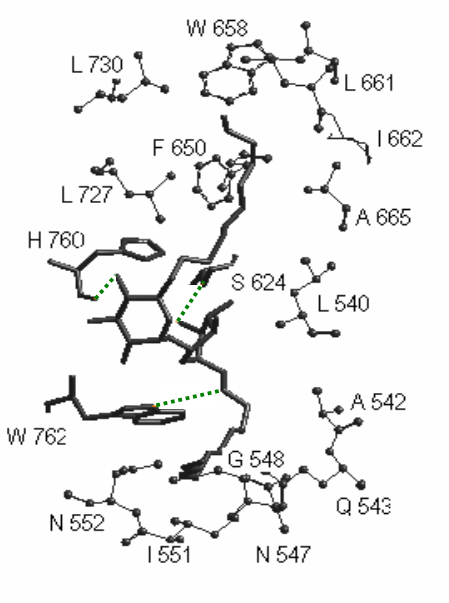
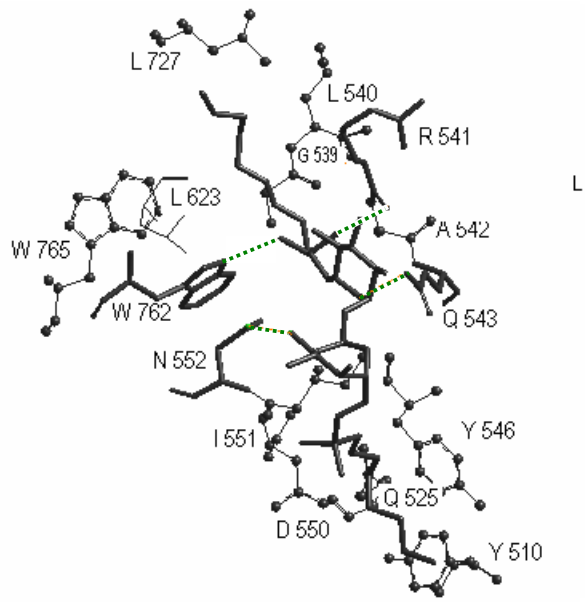
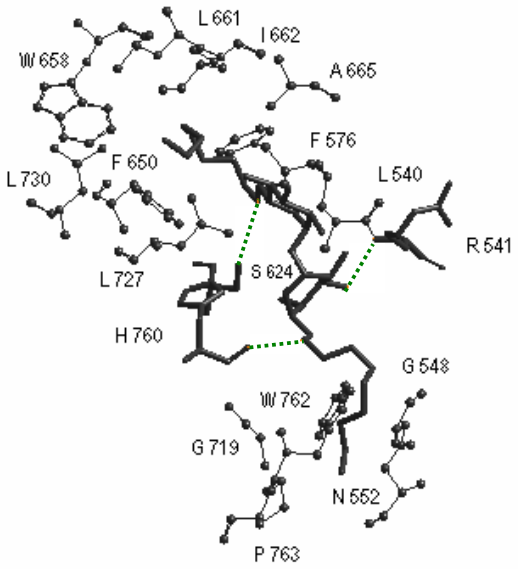


Figure 2.6. ChemScore Based Interactions of Molecules (a) 11b (b) 15a and (c) 18b Docked Into the Active Site of Mycolyltransferase, Ag85C.

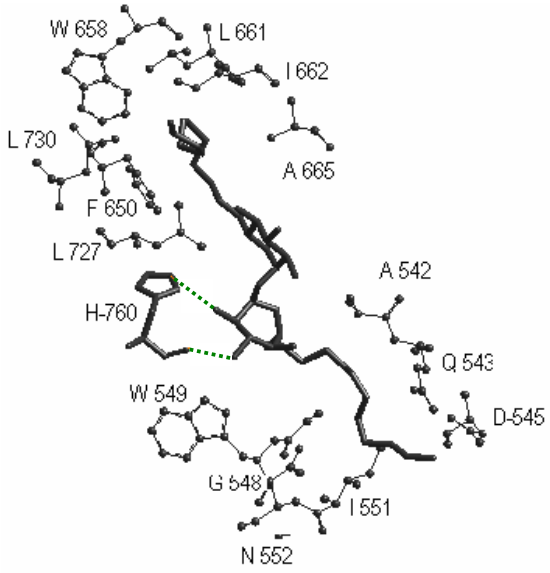
a)



b)



c)



2.4 Conclusions

1. A detailed docking analysis of phosphonate and trehalose analog inhibitors into the active site of Ag85C has been studied in the present work, to identify the inhibitor binding position and affinity to Ag85C using the Gold software.
2. We report that the GoldScore fitness function is marginally better than the ChemScore fitness function, to understand the binding conformation.
3. GoldScore provides a qualitative agreement with the reported IC_{50} values of phosphonate inhibitors.
4. And also we identified that amino acid residues Arg541, Trp762 are important for inhibitor recognition via hydrogen bonding interactions.
5. Phosphonate and trehalose analog alkyl chains are binding in the hydrophobic pockets of the enzyme.
6. Information obtained in this study will be used for designing potent new inhibitors for the mycolyltransferases.

2.5. References

Abou-Zeid, C., Ratliff, T. L.,Wiker, H. G., Harboe, M., Bennedsen, J. & Rook, G. A. (1988). Characterization of fibronectin-binding antigens released by *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG. *Infect Immun*. **12**, 3046-51.

Adindla, S., Guruprasad, K. & Guruprasad, L. (2004). Three-dimensional models and structure analysis of corynemycolyltransferases in Corynebacterium glutamicum and Corynebacterium efficiens. *Int. J. Biol. Macromol.* **34,** 181-9.

Ajay & Murko, M. A. (1995). Computational methods to predict binding free energy in ligand-receptor complexes. *J. Med. Chem.* **38,** 4953-4967.

Anderson, D. H., Harth, G., Horwitz, M. A. & Eisenberg, D. (2001). An interfacial mechanism and a class of inhibitors inferred from two crystal structures of the *Mycobacterium tuberculosis* 30 kDa major secretory protein (Antigen 85B), a mycolyl transferase. *J. Mol. Biol*. **307,** 671-81.

Belisle, J. T., Vissa, V. D., Sievert, T., Takayama, K., Brennan, P. J. & Besra, G. S. (1997). Role of the major antigen of *Mycobacterium tuberculosis* in cell wall biogenesis. *Science*. **276,** 1420-2.

Brennan, P. J. & Nikaido, H. (1995). The envelope of mycobacteria. *Annu Rev Biochem*. **64**, 29-63.

Daffe, M. & Draper. P. (1998). The envelope layers of mycobacteria with reference to their pathogenicity. *Adv. Microb. Physiol*. **39**, 131-203.

Duncan, K. (2004). Identification and validation of novel drug targets in tuberculosis. *Curr. Pharm. Des*. **10,** 3185-3194.

Glickman, M. S., Cox, J. S. & Jacobs, W. R. (2000). A novel mycolic acid cyclopropane synthetase is required for cording, persistence, and virulence of *Mycobacterium tuberculosis*. *Mol Cell.* **5,** 717-27.

Gobec, S., Plantan, I., Mravljak, J., Wilson, R. A., Besra, G. S. & Kikelj, D. (2004). Phosphonate inhibitors of antigen 85C, a crucial enzyme involved in the biosynthesis of the *Mycobacterium tuberculosis* cell wall. *Bioorg. Med. Chem. Lett.* **14**, 3559-62.

Harth, G., Lee, B. Y., Wang, J., Clemens, D. L. & Horwitz, M. A. (1996). Novel insights into the genetics, biochemistry, and immunocytochemistry of the 30-kilodalton major extracellular protein of *Mycobacterium tuberculosis*. *Infect. Immun*. **64,** 3038-3047.

Hyperchem7, Hypercube, Inc.

Jackson, M., Raynaud, C., Lanéelle, M. A., Guilhot, C., Laurent-Winter, C., Ensergueix, D., Gicquel, B. & Daffé, M. (1999). Inactivation of the antigen 85C gene profoundly affects the mycolate content and alters the permeability of the *Mycobacterium tuberculosis* cell envelope. *Mol. Microbiol.* **31**, 1573-87.

Jarlier, V. & Nikaido, H. (1994). Mycobacterial cell wall: structure and role in natural resistance to antibiotics. *FEMS Microbiol. Lett*. **123**, 11-18.

Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol*. **267**, 727-48.

Perola, E., Walters, W. P. & Charifson, P. S. (2004). A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins.* **56,** 235-49.

Ronning, D. R., Klabunde, T., Besra, G. S., Vissa, V.D., Belisle, J. T. & Sacchettini, J. C. (2000). Crystal structure of the secreted form of antigen 85C reveals potential targets for mycobacterial drugs and vaccines. *Nat. Struct. Biol*. **7,** 141-6.

Ronning, D. R., Vissa, V., Besra, G. S., Belisle, J. T. & Sacchettini, J. C. (2004). *Mycobacterium tuberculosis* antigen 85A and 85C structures confirm binding orientation and conserved substrate specificity. *J. Biol. Chem*. **279**, 36771-7.

Rose, J. D., Maddry, J. A., Comber, R. N., Suling, W. J., Wilson, L. N. & Reynolds, R. C. (2002). Synthesis and biological evaluation of trehalose analogs as potential inhibitors of mycobacterial cell wall biosynthesis. *Carbohydr. Res.* **337,** 105-20.

Schroeder, E. K., de Souza, N., Santos, D. S., Blanchard, J. S. & Basso, L. A. (2002). Drugs that inhibit mycolic acid biosynthesis in *Mycobacterium tuberculosis. Curr. Pharm. Biotechnol.* **3,** 197-225.

Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W. & Taylor, R. D. (2003). Improved protein-ligand docking using GOLD. *Proteins.* **52,** 609-23.

Wang, R., Lu, Y. & Wang, S. (2003). Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem*. **46,** 2287-303.

Wang, R., Lu, Y., Fang, X. & Wang, S. (2004). An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes. *J. Chem. Inf. Comput Sci.* **44,** 2114-25.

Wiker, H. G & Harboe, M. (1992). The antigen 85 complex: a major secretion product of *Mycobacterium tuberculosis*. *Microbiol. Rev*. **4,** 648-61.

Zhang,Y. & Amzel, L. M. (2002). Tuberculosis drug targets. *Curr. Drug. Targets*. **3,** 131-154.

CHAPTER 3

Chemical Function Based Virtual Screening: Discovery of Potent Lead Molecules for the *Bcr-Abl* Tyrosine Kinase Using VX-680

3.1 Introduction

Protein phosphorylation is a central regulatory strategy to alter the cellular functions, and protein kinases catalyze the transfer of the γ-phosphate of adenosine triphosphate (ATP) to acceptor proteins. Protein tyrosine kinases (PTKs) are critical regulators of cell proliferation, invasion, metastasis and cell survival (Edward & Sausville, 1999). Two classes of PTKs are present in cells, the receptor protein tyrosine kinases (RTKs) and the non-receptor protein tyrosine kinases (NRTKs). RTKs are transmembrane glycoproteins that are activated by the binding of their cognate ligands, and transduce the extracellular signal to the cytoplasm by phosphorylating tyrosine residues on the receptors themselves (autophosphorylation) and on downstream signaling proteins. NRTKs are integral components of the signaling cascades triggered by RTKs and by other cell surface receptors such as G protein-coupled receptors and receptors of the immune system.

NRTKs lack receptor-like features such as an extracellular ligand-binding domain and a transmembrane spanning region, and most NRTKs are localized in the cytoplasm (Neet & Hunter, 1996). NRTKs are anchored to the cell membrane through amino terminal modification, such as myristoylation or palmitoylation. In addition to a tyrosine kinase domain, NRTKs possess domains that mediate protein-protein, protein-lipid and protein-DNA interactions. The most commonly found protein-protein interaction domains in NRTKs are the Src homology 2 (SH2) and Src homology 3 (SH3) domains (Kuriyan & Cowburn, 1997). The SH2 domain is a compact domain of 100 amino acid residues that binds phosphotyrosine residues in a sequence-specific manner. The smaller SH3 domain (60 residues) binds proline rich containing sequences capable of forming a polyproline type II helix. Some NRTKs lack SH2 and SH3 domains but possess subfamily-specific domains used for protein-protein interactions. For example, members of the Jak

family contain specific domains that target them to the cytoplasmic portion of cytokine receptors. The NRTK Fak possesses two domains that mediate protein-protein interactions; an integrin-binding domain and a focal adhesion-binding domain. The NRTK *Bcr-Abl* contains a nuclear localization signal but is found in both the nucleus and the cytoplasm. In addition to SH2 and SH3 domains, *Bcr-Abl* possesses an F actin–binding domain and a DNA-binding domain (Stevan & Till, 2000).

A number of diseases, including cancer, diabetes and inflammation, are linked to perturbation of protein kinase mediated cell signaling pathways. Therefore, protein kinases are targets for treatment of a number of diseases. *Bcr-Abl* kinase is NRTK, that is expressed in a wide range of cells and it is localized at several subcellular sites, including the nucleus, cytoplasm, mitochondria, endoplasmic reticulum and cell cortex, where *Bcr-Abl* interacts with a large variety of cellular proteins, including signaling adaptors, kinases, phosphatases, cell-cycle regulators, transcription factors and cytoskeletal proteins (Pendergast, 2002). The *Bcr-Abl* gene was first identified as the cellular homolog of the transforming gene of Abelson murine leukaemia and subsequently found to be involved in the Philadelphia chromosome translocation in human leukaemia and to encode a non-receptor tyrosine kinase (Wong & Witte, 2004; Hantschel & Superti-Furga, 2004; Sawyers & Druker, 1999).

Bcr–Abl is an oncogene that arises from fusion of the *Bcr* (breakpoint cluster region) gene with the *c-Abl* proto-oncogene. The Philadelphia chromosome involves fusion of the *Bcr* gene on chromosome 22 at band q11 with the *Abl* proto-oncogene on chromosome 9 at band q34 (Rowley, 1973). Three different *Bcr–Abl* variants can be formed, depending on the amount of *Bcr* gene included: *p185*, *p210* and *p230*. The three variants are associated with distinct types of human leukemia. *P185* is associated with 20–30% of acute lymphocytic leukemia (ALL), *p210* with

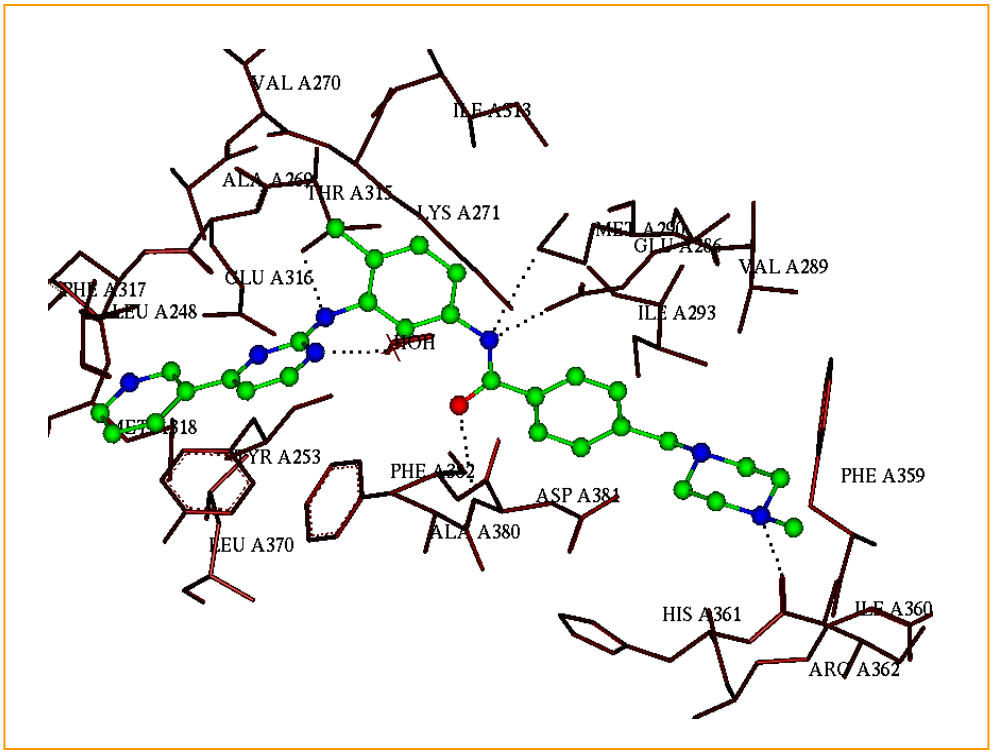
90% of chronic myelogenous leukemia (CML) and *p230* with a subset of patients with chronic neutrophilic leukemia (CNL) (Melo, 1996). The oncogenic ability of *Bcr–Abl* requires deregulated tyrosine kinase activity which leads to the recruitment of adaptor molecules, phosphorylation of signaling molecules and activation of downstream signaling events (Lugo *et al.,* 1990; Daley *et al.,* 1990).

The NRTK *Bcr-Abl* kinase is a causative agent of CML and inhibiting the *Bcr-Abl* kinase enzyme might induce the apoptosis of the diseased cells from the patient's body. In 1996, Novartis team (Druker *et al.,* 1996) reported a successful *Bcr-Abl* inhibitor, CGP57148, which is later renamed STI-571, Gleevec or imatinib.

Imatinib is a specific inhibitor that binds with high affinity to the inactive conformation of the *Bcr-Abl* tyrosine kinase and has been shown to be effective in the treatment of CML with little toxicity, compared to other cancer therapies (Schindler *et al.,* 2000; Druker *et al.,* 2001a; O'Brien *et al.,* 2003). In addition to its ability to block *Bcr-Abl*, imatinib also inhibits the platelet-derived growth factor (PDGF) receptor and the *c-Kit* receptor (Druker *et al.,* 2001b; Buchdunger *et al.,* 2000). *c-Kit* is the cellular homolog of the v-*kit* retroviral oncogene and the c-*Kit* gene product is expressed in hematopoietic progenitor cells, mast cells, germ cells, interstitial cells of cajal and some human tumors (Nocka *et al.,*1989). CML represents the first human malignancy to be successfully treated with a small molecule inhibitor, imatinib. In spite of its several virtues, clinical resistance to imatinib has been reported in small number of patients due to *Bcr-Abl* gene mutation or amplification. Although some of these mutations are located close to the imatinib-binding site, most of the mutations occur at distal positions. A plausible mechanism for the induction of resistance by these mutations involves the destabilization of the inactive conformation, with concomitant preservation of the catalytic capabilities of the kinase domain (Shah *et al.,* 2002).

A significant progress for the treatment of patients with resistance to imatinib is the identification of inhibitors that can bind to both active and inactive conformations of *Bcr-Abl* kinase and those that bind preferably to the active form and provide a way to oppose the mutation-induced resistance to imatinib. Other mechanisms of imatinib resistance include *Bcr-Abl* gene amplification (Gadzicki *et al.,* 2005). In order to overcome the resistance to imatinib, a number of new inhibitors have been synthesized. The most effective ATP mimics are AMN107 (Weisberg *et al.,* 2005) and BMS-354825, which inhibit almost all imatinib resistant forms of *Bcr-Abl* (Shah *et al.,* 2004; Doggrell, 2005) but are not effective against the T315I mutant. The T315I mutation is the most common mutation found in patients undergoing imatinib therapy (Shah *et al.,* 2002) and this is responsible for nearly 15% of resistant cases. A single nucleotide change at the genetic level, replaces threonine with isoleucine in the protein product at position 315 thus causing this mutation (T315I). The side chain hydroxyl group of Thr315 forms critical hydrogen bonds with imatinib (see Figure 3.1a) and Thr315 is located at the center of the imatinib binding site in *Bcr-Ab*l kinase. This residue separates the ATP binding site from an internal cavity that is of variable size in different protein kinases (*Bcr-Abl* kinase, *c-Kit* receptor), and this gatekeeper residue plays a vital role in determination of the inhibitor specificity (Liu, *et al.,* 1999) and regulates the binding of inhibitors. Thr315 opens up an auxiliary binding site, which is occupied by the piperazinyl-substituted benzamide moiety of imatinib and participates through the hydroxymethylene side chain, in a crucial H bonding interaction between imatinib and *Abl* (Schindler *et al.,* 2000), as well as *Bcr-Abl* (Manley *et al.,* 2002; Nagar *et al.,* 2002). Mutation to isoleucine abrogates the possibility of this H bonding interaction, which, combined with the additional bulk of the isoleucine side chain, sterically hinders imatinib binding and leads to imatinib insensitivity and consequently resistance in patients.

Figure 3.1a. A Schematic View of the Inhibitor, Imatinib Bound to the *Bcr-Abl* Kinase. Hydrogen Bonding Interactions in the Protein-Inhibitor Complex are Indicated.



Under these situations, it is essential to find possible molecules that have been developed as drugs for other protein kinases and might also serve to inhibit imatinib resistant forms of *Bcr-Abl* kinase. Similar to the behavior of imatinib, dasatinib (BMS-354825) (Shah *et al.,* 2004) and other *Bcr-Abl* inhibitors, exhibit a significant loss of affinity for *Bcr-Abl* (T315I) relative to other *Bcr-Abl* variants. This implies that it is particularly difficult to inhibit *Bcr-Abl* (T315I) with an ATP-competitive compound (Carter *et al.,* 2005). The possibility of ATP competitive compounds is to bind either the wild-type *Bcr-Abl* or T315I mutant *Bcr-Abl*, but not both. Imatinib and dasatinib are two clinically valuable *Bcr-Abl* kinase

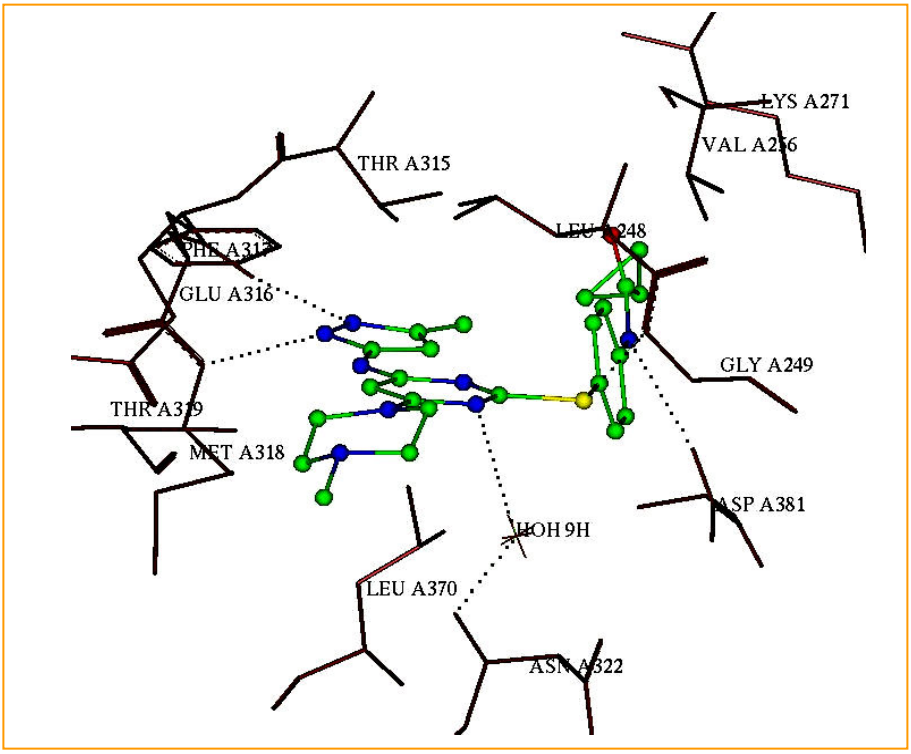
inhibitors that serve as a paradigm for the study of emergence of resistance in targeted cancer therapy.

In order to test the existing inhibitors against drug-resistant mutants of *Bcr-Abl*, Todd and co-workers (von *et al.,* 2003) developed competition binding assays for a panel of clinically important mutants. In this study they have used various types of kinase inhibitors and found that VX-680 is binding with high affinity to *Bcr-Abl* kinase (T315I) mutant. VX-680 has been previously reported as a potent inhibitor of all three Aurora kinases A, B and C with apparent inhibition constant (K_i) values of 0.6, 18 and 4.6 nM for Aurora A, Aurora B and Aurora C respectively (Harrington *et al.,* 2004). VX-680, not only blocks cell proliferation but can also induce cell death by apoptosis in multiple tumor types, both *in vitro* and *in vivo*. VX-680 also blocks the phosphorylation of a direct downstream substrate of the Aurora kinases, histone H3, in tumor tissue *in vivo.* The VX-680 molecule binds tightly to the wild type *Bcr-Abl* kinase, (*K*d of ~20 nM or lower), and most of the *Bcr-Abl* mutants, including T315I (*K*d 5-20 nM). In the enzyme activity assays, VX-680 potently inhibited wild-type *Bcr-Abl* with an IC_{50} value of 10 nM and *Bcr-Abl* (T315I) with an IC_{50} value of 30 nM (Carter *et al.,* 2005).

The crystal structure of VX-680 bound to the catalytic domain of *Bcr-Abl* (PDB_ID: 2F4J) containing a mutation (H396P) has been solved (Young *et al.,* 2006). This mutation confers imatinib resistance in *Bcr-Abl* kinase but is inhibited by VX-680 *in vitro*. It has been shown that VX-680 inhibits *Bcr-Abl* kinase activity in cells derived from patients carrying the T315I mutation in the kinase domain of *Bcr-Abl*, and that it retains activity towards purified T315I mutation *in vitro*. These results provide a structural explanation for the retention of inhibitory activity of VX-680 towards mutant proteins, which are no longer inhibited by imatinib. The structure of the kinase domain of *Bcr-Abl* (H396P) bound to VX-680 is shown in Figure 3.1b.

This 3-D crystal structure (PDB_ID: 2F4J) is the source for the virtual screening strategy used to discover novel inhibitors to *Bcr-Abl* kinase. Virtual screening provides assurance to an inexpensive and fast alternative, to high throughput screening (HTS) in order to discover useful lead compounds for drug discovery projects (Jürgen, 2002). Drug discovery methods such as structure-based virtual screening focuses on using the protein crystal structure and is exemplified by receptor-based docking methods such as affinity docking, FlexX, Autodock and GOLD.

Figure 3.1b. A Schematic View of the Inhibitor, VX-680 Bound to the *Bcr-Abl* (H396P) kinase. Hydrogen Bonding Interactions in the Protein-Inhibitor Complex are Indicated.



Another virtual screening approach is to generate a pharmacophore which represents the 3-D arrangement of a set of chemical features, functional groups from an inhibitor that have critical interactions with the receptor (Mason *et al.,* 2001). The chemical features of the inhibitor are critical for its biological activity. More recently, several new approaches have been described for pharmacophore screening that enable pharmacophore information to be included in the search query (Hahn, 1997; Putta *et al.,* 2002).

In this work we have generated a chemical function based pharmacophore of VX-680 using "View Hypotheses" module in catalyst software. This pharmacophore was used for the screening of databases such as, NCI, Maybridge and Derwent-WDI2005 and the obtained hits were docked into the *Bcr-Abl* kinase crystal structure using GOLD software (Jones *et al.,* 1997). Our goal is to search for molecules with alternative leads to the VX-680 in a commercially available database that would inhibit wild type and mutant *Bcr-Abl* kinases. The chemical component pharmacophoric hypotheses was built using the "View Hypotheses" workbench within Catalyst using the conformation of VX-680- *Bcr-Abl* kinase interactions reported in the X-ray complex. This pharmacophoric query was used to search a multi-conformational databases using Catalyst. Finally, the molecules identified from virtual screening were docked into the *Bcr-Abl* kinase using GOLD software to observe the key interactions between the screened molecules and *Bcr-Abl* kinase and thus validate the hits as useful and novel *Bcr-Abl* kinase inhibitors.

3.2 Methods

3.2.1 Protein preparation:

The 3-D co-ordinates of *Bcr-Abl* kinase complexed with VX-680 (PDB_ID: 2F4J) (Young *et al.,* 2006) was downloaded from protein structure databank, (http://www.rcsb.org/). Hydrogen atoms were added to the protein using Biopolymer module in InsightII 2005 (InsightII 2005, Accelrys) keeping all the residues in their charged form. In the first step all the hydrogen atoms were minimized, keeping the other atoms fixed. In the second step whole protein complex including crystal water was energy minimized by the steepest descent followed by conjugate gradient methods to achieve a convergence gradient of 0.01 kcal/mol using CVFF force fields in InsightII 2005. Crystallographic waters were retained for docking studies. In addition to this, we have mutated T315I in Biopolymer module and used the same methods for energy minimization in order to study the binding of these molecules to *Bcr-Abl* kinase (T315I) mutant.

3.2.2 Pharmacophore model generation:

From the crystal structure (PDB_ID: 2F4J), a ligand-based (VX-680) pharmacophore query was generated for *Bcr-Abl* kinase using View Hypotheses workbench module in Catalyst (Catalyst 4.11, Accelrys) using the conformation of VX-680 reported in the X-ray complex with *Bcr-Abl* kinase. In the hypotheses, two hydrogen bond donors (HD), two hydrogen bond acceptors (HA) and one hydrophobic interaction (HP) were allowed as observed in the protein structure. A maximum of 5 features were selected to construct the pharmacophore hypotheses. This pharmacophore query was used for the virtual screening of small molecule databases and we identified 289 molecules form the NCI, Maybridge and Derwent-

WDI2005 databases. These databases are multi-conformational Catalyst databases, which were built using the best option with the MAXCONFS option set to 250 and the energy threshold set to 15 kcal/mol.

3.2.3 Virtual screening:

The Pharmacophore query was used as a 3-D structural query in the screening of NCI, Maybridge and Derwent-WDI2005 databases. NCI, Maybridge and Derwent-WDI2005 databases comprise 2,38,819, 59,652 and 67,050 molecules respectively. The chemical function based pharmacophore model was used for database searching by the best flexible search method in Catalyst. The molecules obtained were further filtered using Lipinski's rule of 5 (Lipinski *et al.,* 1997).

3.2.4 Docking:

The new lead molecules identified from virtual screening, were docked into the crystal structure of *Bcr-Abl* kinase (PDB_ID: 2F4J) using GOLD (GOLD 3.10, CCDC, UK) software. GOLD (Genetic Optimization of Ligand Docking) is a genetic algorithm for docking flexible ligands into protein binding site. The details were discussed in the section 2.2.5. During docking, the default algorithm speed was selected, and the ligand binding site in the *Bcr-Abl* kinase, was defined within a 10 Å radius with the centroid as Glu 316 main chain carbonyl oxygen atom. For docking, the number of poses for each inhibitor was set to ten, and early termination was allowed if the top five bound conformations of a ligand were within 1.5 Å (RMSD). After docking, the individual binding poses of each ligand were re-ranked according to the GOLD score. The top ranked conformation of each ligand was selected and analysed using SILVER (SILVER 1.1.1, CCDC, UK) to examine the mode of protein-inhibitor binding.

3.2.5 Hardware and software:

InsightII 2005 was used for energy minimization of *Bcr-Abl* kinase, and Catalyst 4.11 was used for pharmacophore generation and virtual screening on SGI Octane2 workstation equipped with 2600 MHz MIPS R14000 processors. The docking calculations using GOLD software and docking analysis using SILVER (Nissink *et al.,* 2002) were carried out on an Intel P4-based windows system.

3.3 Results and Discussion

The aim of the present work is to identify novel lead molecules as inhibitors for *Bcr-Abl* kinase and its mutant (T315I). We have achieved this using pharmacophore model generation, virtual screening of small molecule databases and molecular docking studies.

3.3.1 Generation of pharmacophore model:

Our choice of pharmacophore features was based upon the conformation of VX-680 reported in the X-ray complex with *Bcr-Abl* kinase. Two hydrogen bond donor features were predicted to interact with the side chain amino group of Asp381 and the main chain carbonyl oxygen of Glu316. The pyrazole group N_{20} of HD1 accepts a hydrogen bond from Met318 and N_{30} atom of VX-680 as HD2 accepts a hydrogen bond from Asp381. Two hydrogen bond acceptor features were predicted to interact with the main chain nitrogen of Met318 and N_{13} atom of VX-680 with water molecule. This water acts as a bridge molecule between N_{13} atom of VX-680 and main chain nitrogen of Asn322. The pharmacophore feature HA_1 is complementary to the pyrazole group N_{19} of Glu316, while pyrimidine group of N_{13} as HA_2 accepts a hydrogen bond from Wat9 that in turn forms a hydrogen bond with the main chain NH of Asn322. The positions of these glutamic acid and metionine are strictly conserved across the *Bcr-Abl* kinase family and are involved in binding to the adenine moiety of ATP. The cyclopropane ring in VX-680 was selected as the required group for hydrophobic interaction (HP) (Figure 3.1c).

The majority of kinase inhibitors that have been deposited in the protein databank form interactions with amino acid residues at these positions. These interactions are shown in Figure 3.1b. A Schematic representation of

pharmacophore model is shown in Figure 3.1c. The mapping of Pharmacophore with VX-680 molecule is shown in Figure 3.1d.

Figure 3.1c. A Schematic Representation of Pharmacophore Model. Distances Between the Characters are in Å Units.

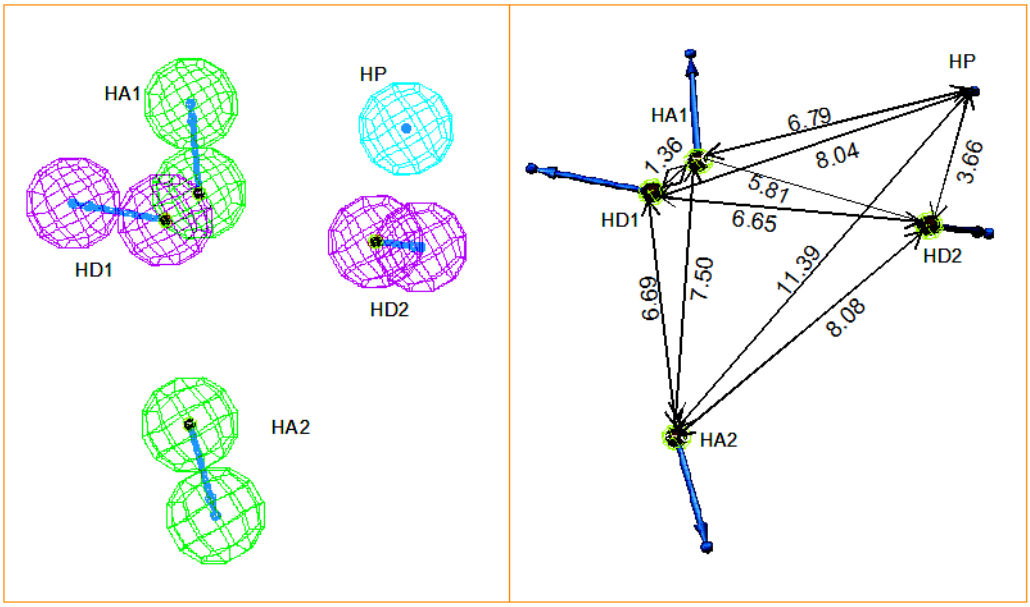
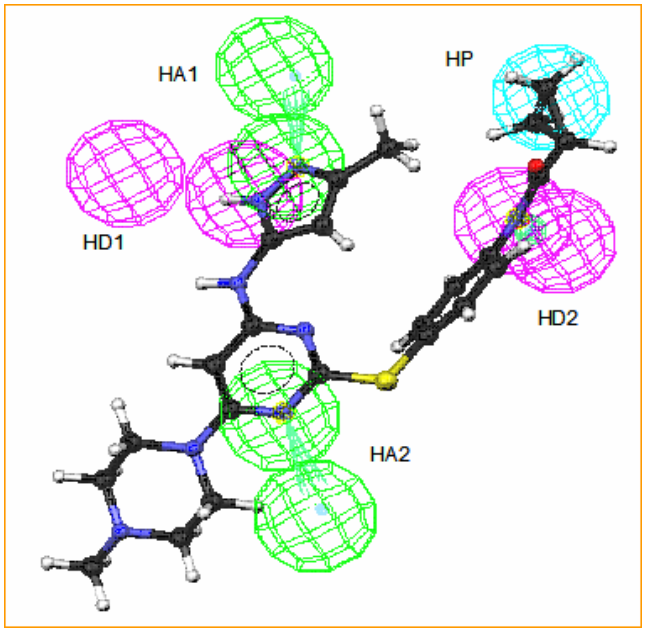


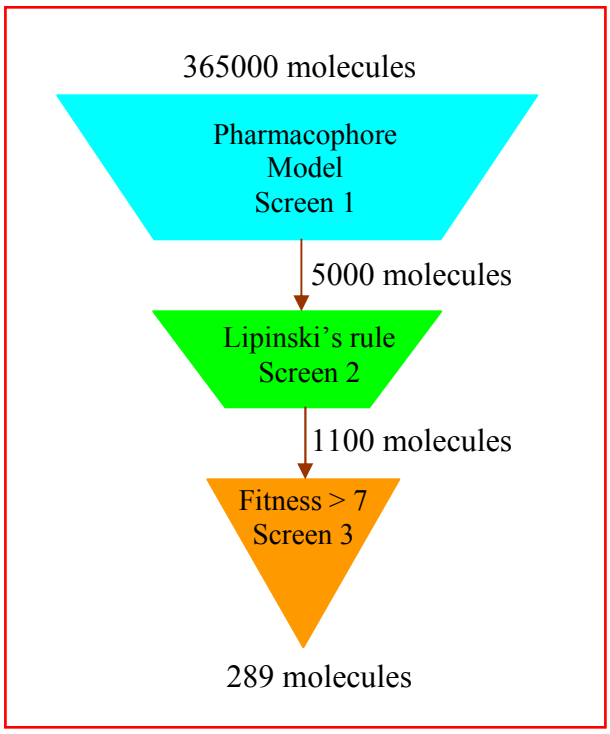
Figure 3.1d. The Mapping of Pharmacophore with VX-680 Molecule.



3.3.2 Database screening:

The pharmacophore query generated above was used to screen NCI, Maybridge and Derwent-WDI2005 databases. In all, about 5000 molecules were obtained as hits from *in silico* screening (screen 1). To assess the drug-likeliness of these hits, a second screen, incorporating Lipinski's rule of 5 was used. A total of 1100 molecules were obtained as hits from this screen (Screen 2). To further increase the probability of the hit to be a lead, a fitness score >7.00 was used as the third screen (Screen 3). The fitness score indicates how well the features in the pharmacophore overlap with the chemical features in the ligand. A total of 289 molecules were obtained as hits from this screen. In Figure 3.1d shows a Schematic representation of VS strategy.

Figure 3.1d. Schematic Representation of VS Strategy.



3.3.3 GOLD docking:

The crystal structure of *Bcr-Abl* kinase bound to substrate VX-680 (PDB_ID: 2F4J) was used for the docking studies. All amino acids within 10 Å radius from the Glu316 main chain carbonyl oxygen atom were considered to comprise the active site. Docking was carried out using GOLD 3.10 software.

The inhibitor VX-680 was docked into the *Bcr-Abl* kinase and the following interactions between VX-680 and the *Bcr-Abl* kinase have been observed. (i) A hydrogen bond interaction between the pyrazole ring $N_{19}H$ and Glu316 carbonyl oxygen ($N_{19}H\cdots O{=}C$, 2.32 Å). (ii) A hydrogen bond between pyrazole ring N_{20} and the main chain NH of Met318 ($N_{20}\cdots NH$, 3.17 Å). (iii) A hydrogen bond between N_{30} and the side chain carbonyl oxygen of Asp381

($N_{30}\cdots O{=}C$, 2.35 Å). (iv) One bridge water molecule (H_2O) is in between N13of VX-680 and main chain NH of Asn322 ($N_{13}\cdots H_2O\cdots NH$). (V) A hydrogen bond between S_{23} and the main chain NH of Gly249 ($S_{23}\cdots H$-N, 3.66 Å). The RMSD between the docked pose of VX-680 and its bound conformation in the crystal structure 2F4J is 0.56 Å, indicating that GOLD software was able to reproduce the correct pose and is a reliable method for these docking studies.

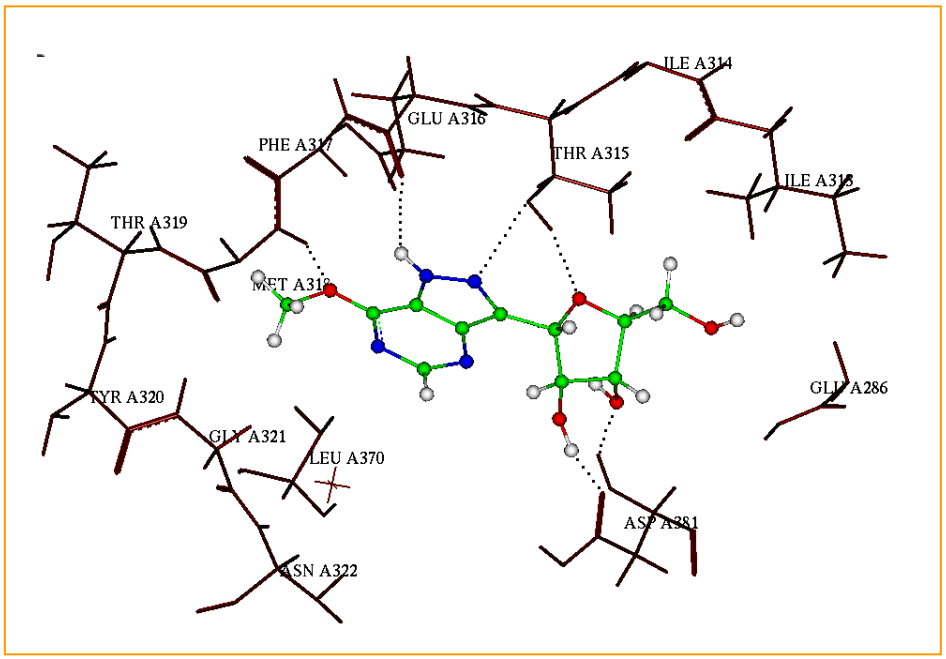
Docking calculations were carried out using two types of mutant proteins; these are *Bcr-Abl* (H396P) kinase and *Bcr-Abl* (T315I) kinase.

3.3.3.1 *Bcr-Abl* (H396P) kinase docking:

The molecules obtained from virtual screening were docked into the *Bcr-Abl* kinase (H396P) crystal structure. All molecules fit into the VX-680 binding site of the enzyme. The binding of these docked molecules to *Bcr-Abl* kinase was examined on graphics. Based on the values of GoldScore, and the protein-ligand binding interactions, some molecules were selected to have better binding and these are described below. We have given the numbering of atoms according to those databases.

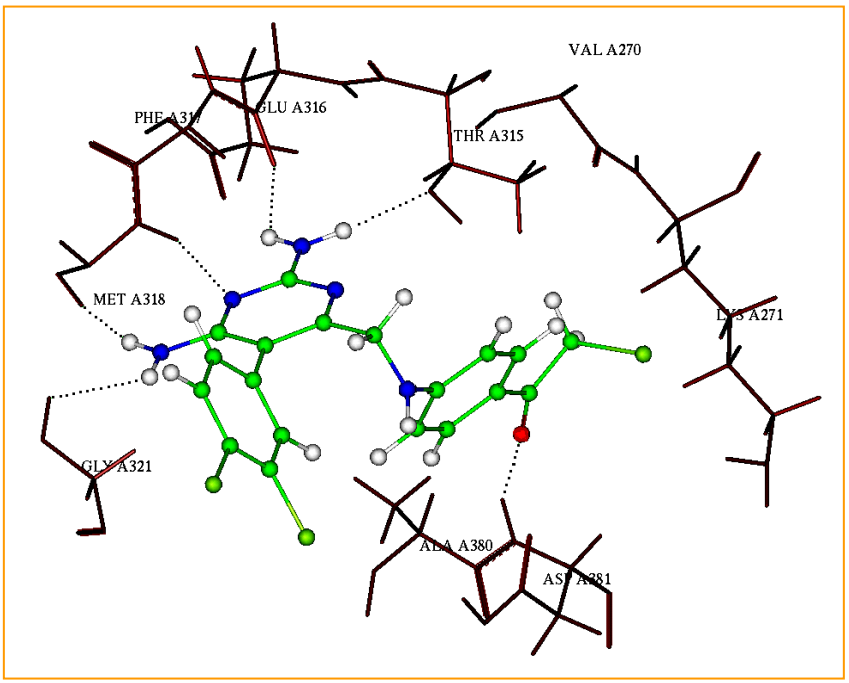
The interaction of Hit NCI0166619 is shown in Figure 3.2a. In the molecule NCI0166619, O_{11} makes hydrogen bonding interactions with the main chain NH of Met318 (O_{11}····HN, 3.28 Å). Further, the N_{19} makes hydrogen bonding interactions with the main chain carbonyl oxygen of Glu316 (N_{19}····O=C, 2.48 Å). A bifurcated hydrogen bond between O_3 and N_{20}, with the side chain oxygen of Thr315 (O_3····HO, 2.94, N_{20}····O 2.63 Å). The O_6H makes hydrogen bonding interactions with the side chain carbonyl oxygen of Asp381 (O_6H ····O=C, 2.51 Å), O_7 makes hydrogen bonding interactions with the main chain NH of Asp381 (O_7 ····HN, 2.58 Å).

Figure 3.2a. A Schematic View of the Inhibitor, NCI0166619 Bound to the *Bcr-Abl* (H396P) kinase. Hydrogen Bonding Interactions in the Protein-Inhibitor Complex. Inhibitor is Indicated in Ball and Stick.



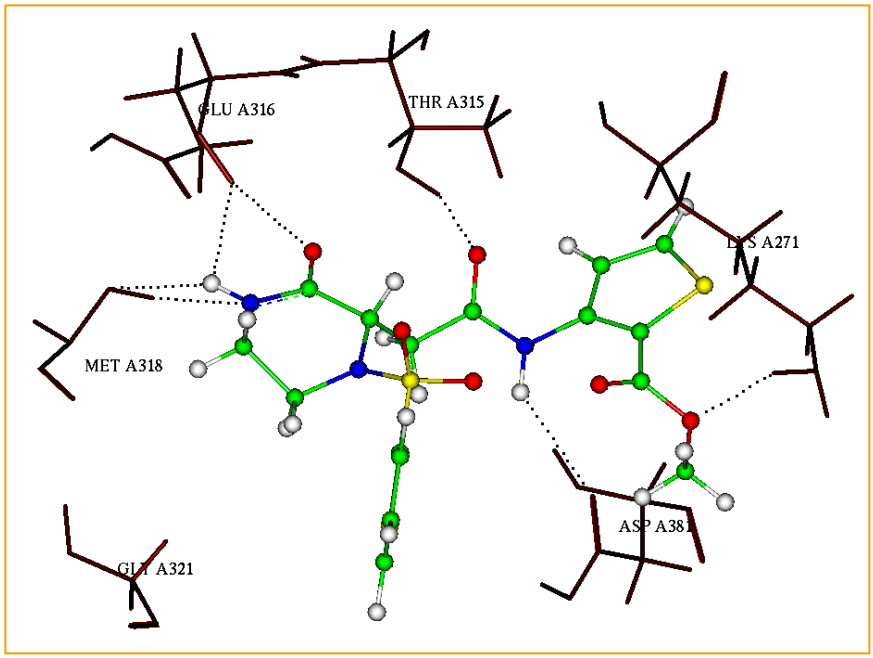
The interaction of Hit NCI0210892 is shown in Figure 3.2b. In the molecule NCI0210892, N_5 makes hydrogen bonding interactions with the main chain NH of Gly321 (N_5 ····HN, 3.39 Å) and N_5H makes hydrogen bonding interactions with the main chain carbonyl oxygen of Met318 (N_5H····O=C, 2.07 Å). The N_{20} makes hydrogen bonding interactions with the main chain NH of Met318 (N_{20} ····HN, 2.83 Å). The $N_{17}H_2$ makes hydrogen bonding interactions with the main chain carbonyl oxygen of Glu316 ($N_{17}H_2$····O=C, 2.32 Å) and The $N_{17}H_1$ makes hydrogen bonding interactions with the side chain oxygen of Thr315 ($N_{17}H_1$····O, 3.28 Å). The O_{12} makes hydrogen bonding interactions with the main chain HN of Asp381 (O_{12}····HN, 2.59 Å).

Figure 3.2b. A Schematic View of the Inhibitor, NCI0210892 Bound to the *Bcr-Abl* (H396P) kinase. Hydrogen Bonding Interactions in the Protein-Inhibitor Complex. Inhibitor is Indicated in Ball and Stick.



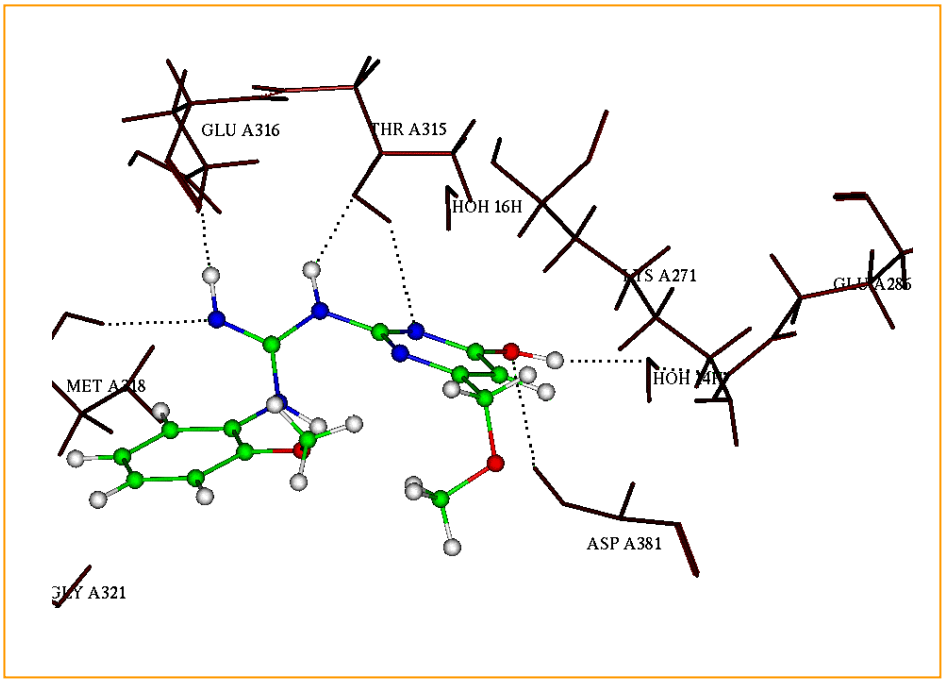
The interaction of Hit HTS07964 is shown in Figure 3.2c. In the molecule HTS07964, N_{13} makes hydrogen bonding interactions with the main chain HN of Met318 (N_{13} ····HN, 3.01 Å) and $N_{13}H$ makes hydrogen bonding interactions with the main chain carbonyl oxygen of Glu316 ($N_{13}H$ ····O=C, 3.09 Å). The O_{28} makes hydrogen bonding interactions with the side chain OH of Thr315 (O_{28}····HO, 2.48 Å). The $N_{18}H$ makes hydrogen bonding interactions with the main chain carbonyl oxygen of Asp381 ($N_{18}H$···O=C, 3.46 Å). Further, O_{23} forms a bifurcated hydrogen bond with the side chain oxygen of Glu286 and the side chain NH of Lys271 (O_{23}···HO, 3.47 Å; O_{23}···NH 2.87Å).

Figure 3.2c. A Schematic View of the Inhibitor, HTS07964 Bound to the *Bcr-Abl* (H396P) kinase. Hydrogen Bonding Interactions in the Protein-Inhibitor Complex. Inhibitor is Indicated in Ball and Stick.



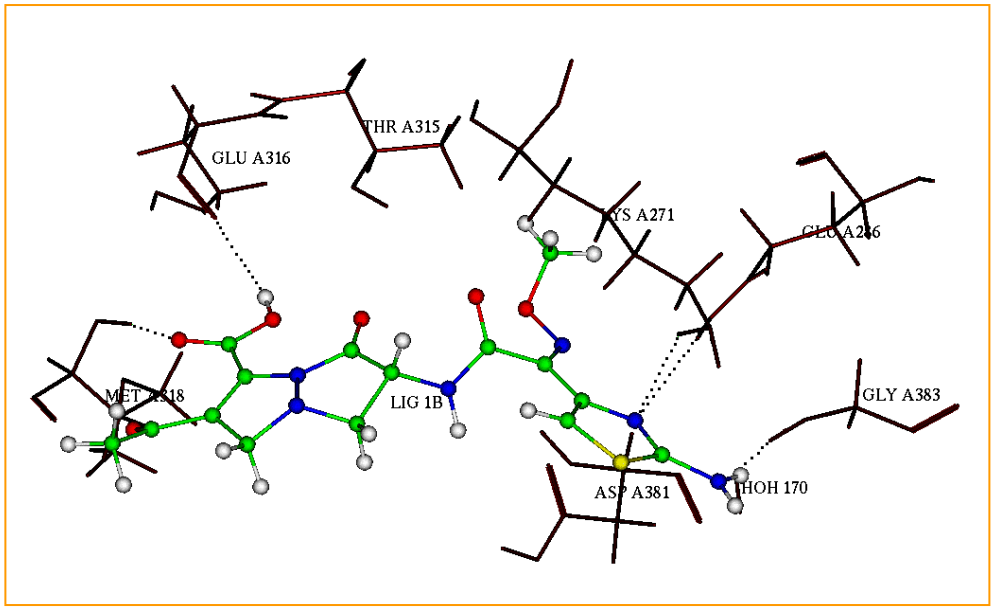
The interaction of Hit RJF00578 is shown in Figure 3.2d. In the molecule RJF00578, N_3 makes hydrogen bonding interactions with the main chain NH of Met318 (N_3 ····HN, 3.48 Å) and N_3H makes hydrogen bonding interactions with the main chain carbonyl oxygen of Glu316 (N_3H ····O=C, 3.11 Å). A bifurcated hydrogen bond between RJF00578 N_1H and side chain oxygen of Thr315 (N_1H····O, 2.58 Å), N_5 and side chain oxygen hydrogen of Thr315 (N_5····HO, 3.28 Å). The O_7 makes hydrogen bonding interactions with main chain NH of Asp381 (O_7 ····HN, 3.45 Å). One bridge water molecule (H_2O) is in between O_7 of RJF00578 and side chain OH of Glu286 (O_7···H_2O···HO 2.95, 2.88 Å).

Figure 3.2d. A Schematic View of the Inhibitor, RJF00578 Bound to the *Bcr-Abl* (H396P) kinase. Hydrogen Bonding Interactions in the Protein-Inhibitor Complex. Inhibitor is Indicated in Ball and Stick.



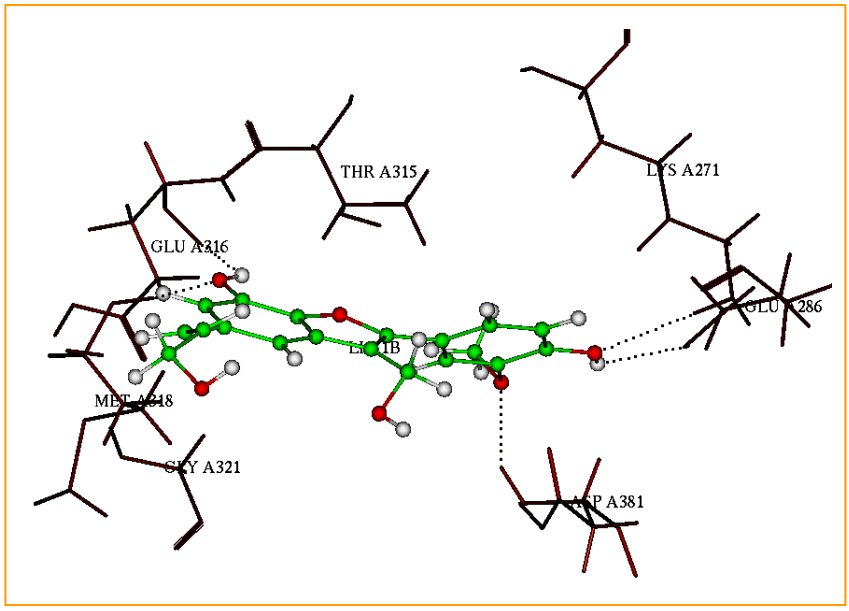
The interaction of Hit LY186826 is shown in Figure 3.2e. In the molecule LY186826, O_{28} makes hydrogen bonding interactions with the main chain NH of Met318 (O_{28} ····HN, 2.63 Å). The $O_{28}H$ makes hydrogen bonding interactions with the main chain carbonyl oxygen of Glu316 ($O_{28}H$ ····O=C, 3.59 Å). A bifurcated hydrogen bond between LY186826 N_{20} and side chain oxygen of Glu286 (N_{20}····O, 3.58 Å), N_{20} and side chain NH of Lys271 (N_{20}····HN, 2.96 Å). Further, a bridge water molecule (H_2O) is in between $N_{18}H$ of LY186826 and main chain NH of Gly383 ($N_{18}H$ ···H_2O···HN 2.92, 2.87 Å).

Figure 3.2e. A Schematic View of the Inhibitor, LY186826 Bound to the *Bcr-Abl* (H396P) kinase. Hydrogen Bonding Interactions in the Protein-Inhibitor Complex. Inhibitor is Indicated in Ball and Stick.



The interaction of Hit Vibsanol is shown in Figure 3.2f. In the molecule Vibsanol, O_{14} makes hydrogen bonding interactions with the main chain NH of Met318 (O_{14} ····HN, 2.66 Å). The $O_{14}H$ makes hydrogen bonding interactions with the main chain carbonyl oxygen of Glu316 ($O_{14}H$ ····O=C, 3.12 Å). The O_{25} makes hydrogen bonding interactions with the main chain NH of Asp381 (O_{25} ····HN, 3.07 Å). Further, $O_{21}H$ makes hydrogen bonding interactions with the side chain OH of Glu286 ($O_{21}H$ ····HO, 3.27 Å) and O_{21} makes hydrogen bonding interactions with the side chain NH of Lys271 (O_{21} ····HN, 3.59 Å)

Figure 3.2f. A Schematic View of the Inhibitor, Vibsanol Bound to the *Bcr-Abl* (H396P) kinase. Hydrogen Bonding Interactions in the Protein-Inhibitor Complex. Inhibitor is Indicated in Ball and Stick.

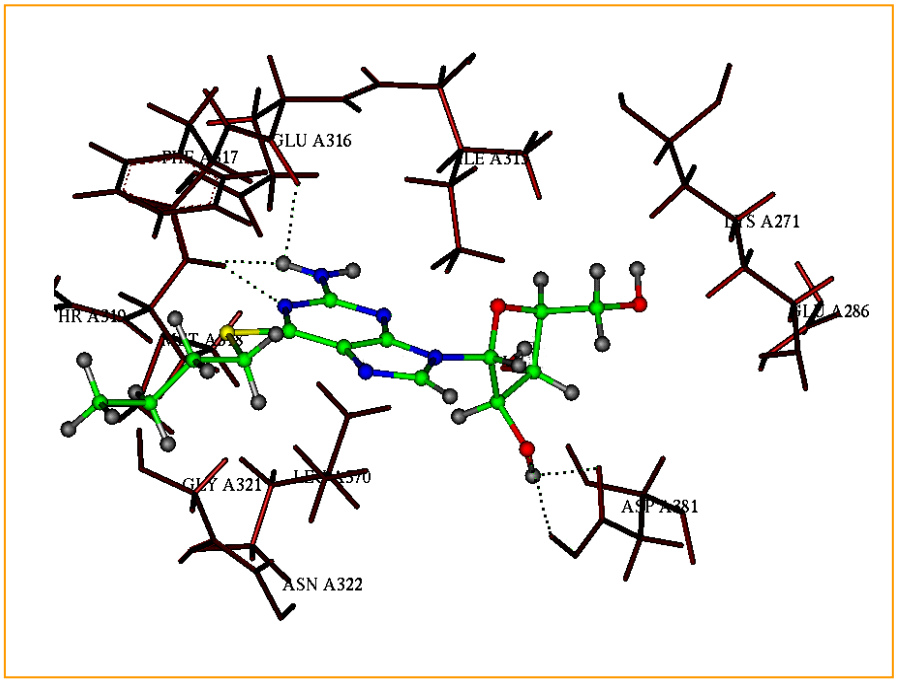


3.3.3.2 *Bcr-Abl* (T315I) kinase docking:

In order to test the docking of screened molecules against drug-resistant mutants of *Bcr-Abl* kinase (T315I), we have mutated T315I by using Biopolymer module in *in silico* modeling. We observed that the side chains of Thr315 and mutated Ile315 are well superimposed. So we consider that, it is a good model to carry out the docking studies. Although we mutated T315I we observe that some molecules have formed very good interactions with the protein active site.

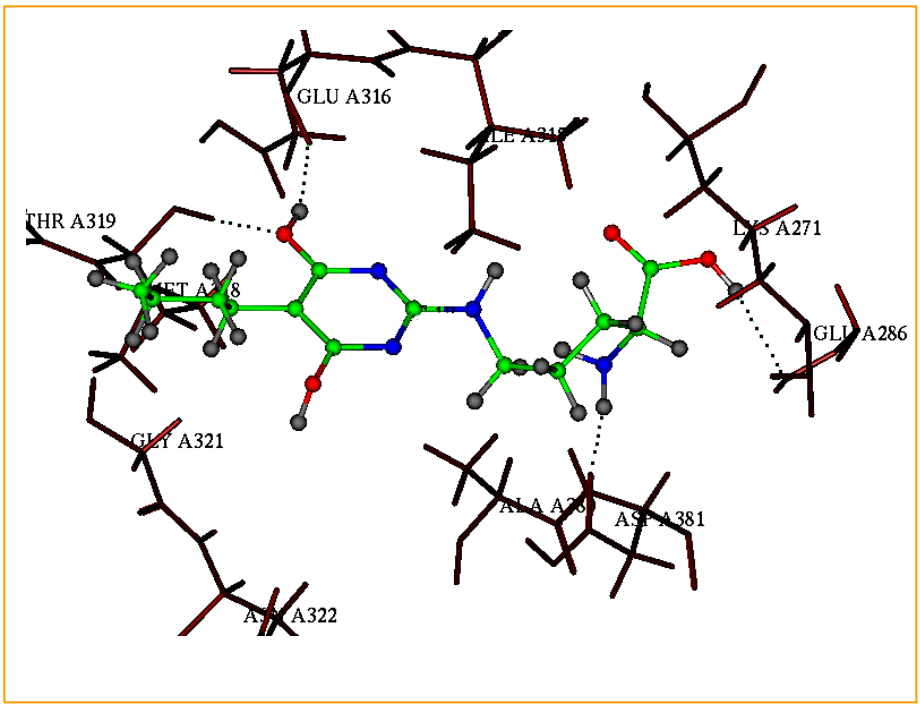
The interaction of Hit NCI0046391 is shown in Figure 3.3a. In the molecule NCI0046391, N_3 makes hydrogen bonding interactions with the main chain NH of Met318 (N_3 ····HN, 3.16 Å). The N_4H makes hydrogen bonding interactions with the main chain carbonyl oxygen of Glu316 (N_4H ····O=C, 2.71 Å). The N_4H makes hydrogen bonding interactions with the main chain N of Met318 (N_4H ····NH, 2.42 Å). Further, O_3H forms a bifurcated hydrogen bond with the side chain carbonyl oxygen of Asp381 and the side chain hydroxyl oxygen of Asp381 (O_3H ····O=C, 2.65 Å; O_3H ····OH, 3.20 Å).

Figure 3.3a. A Schematic View of the Inhibitor, NCI0046391 Bound to the *Bcr-Abl* (T396I) kinase. Hydrogen Bonding Interactions in the Protein-Inhibitor Complex. Inhibitor is Indicated in Ball and Stick.



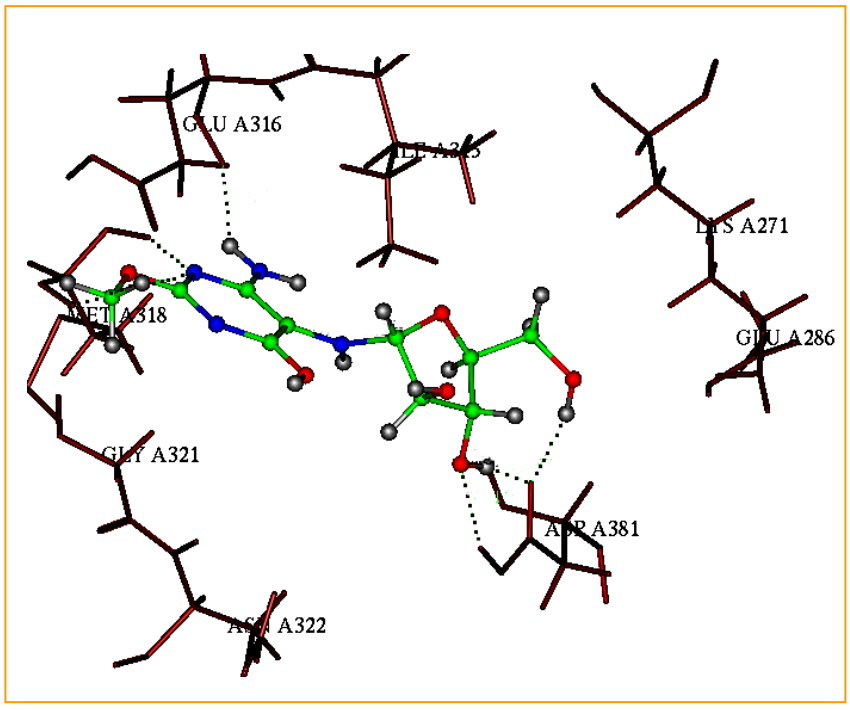
The interaction of Hit NCI0132917 is shown in Figure 3.3b. In the molecule NCI0132917, O_3 makes hydrogen bonding interactions with the main chain NH of Met318 (O_3 ····HN, 2.68 Å). The O_3H makes hydrogen bonding interactions with the main chain carbonyl oxygen of Glu316 (O_3H ····O=C, 3.22 Å). The N_1H makes hydrogen bonding interactions with the main chain carbonyl oxygen of Asp381 (N_1H ····O=C, 2.60 Å). Further, O_2H makes hydrogen bonding interactions with the side chain OH of Glu286 (O_2H ····OH, 3.26 Å).

Figure 3.3b. A Schematic View of the Inhibitor, NCI0132917 Bound to the *Bcr-Abl* (T396I) kinase. Hydrogen Bonding Interactions in the Protein-Inhibitor Complex. Inhibitor is Indicated in Ball and Stick.



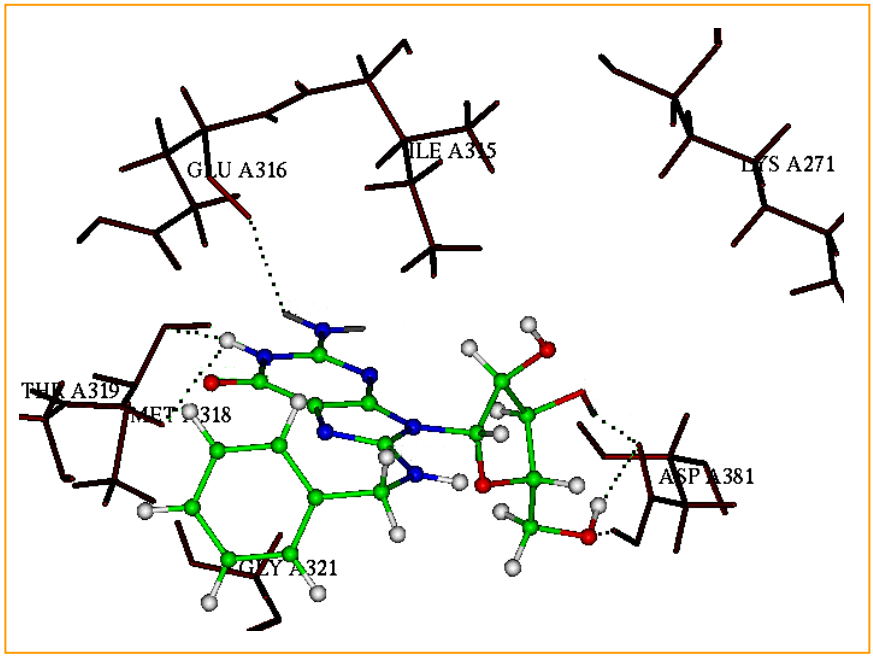
The interaction of Hit NCI0694766 is shown in Figure 3.3c. In the molecule NCI0694766, N_2 makes hydrogen bonding interactions with the main chain NH of Met318 (N_2 ····HN, 2.96 Å). N_2H makes hydrogen bonding interactions with the main chain carbonyl oxygen of Met318 (N_2H ····O=C, 2.85 Å). The N_3H makes hydrogen bonding interactions with the main chain carbonyl oxygen of Glu316 (N_3H ····O=C, 3.48 Å). Further, O_6H froms a bifurcated hydrogen bond with the side chain carbonyl oxygen of Asp381 and the side chain hydroxyl oxygen of Asp381 (O_6H ····O=C, 2.72 Å; O_6H ····OH, 3.04 Å) and O_1H makes hydrogen bonding interactions with the side chain carbonyl oxygen of Asp381 (O_1H ····O=C, 3.06 Å)

Figure 3.3c. A Schematic View of the Inhibitor, NCI0694766 Bound to the *Bcr-Abl* (T396I) kinase. Hydrogen Bonding Interactions in the Protein-Inhibitor Complex. Inhibitor is Indicated in Ball and Stick.



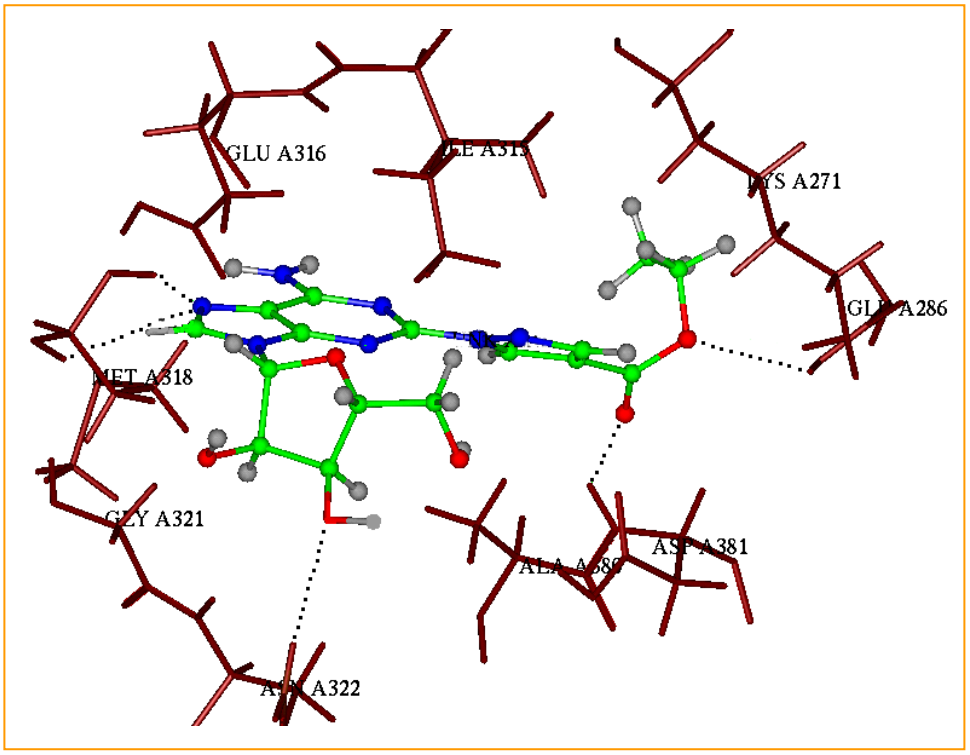
The interaction of Hit HTS 11169 is shown in Figure 3.3d. In the molecule HTS 11169, N_5 makes hydrogen bonding interactions with the main chain NH of Met318 (N_5 ····HN, 2.79 Å). The N_5H makes hydrogen bonding interactions with the main chain carbonyl oxygen of Met318 (N_5H ····O=C, 2.75 Å). The N_6H makes hydrogen bonding interactions with the main chain carbonyl oxygen of Glu316 (N_6H ····O=C, 3.45 Å). O_1H forms a bifurcated hydrogen bond with the side chain carbonyl oxygen of Asp381 and the side chain hydroxyl oxygen of Asp381 (O_1H ····O=C, 3.30 Å; O_1H ····OH, 2.33 Å) and O_5H makes hydrogen bonding interactions with the side chain carbonyl oxygen of Asp381 (O_5H ····O=C, 3.42 Å).

Figure 3.3d. A Schematic View of the Inhibitor, HTS 11169 Bound to the *Bcr-Abl* (T396I) kinase. Hydrogen Bonding Interactions in the Protein-Inhibitor Complex. Inhibitor is Indicated in Ball and Stick.



The interaction of Hit CVT-3127 is shown in Figure 3.3e. In the molecule CVT-3127, N_7 makes hydrogen bonding interactions with the main chain NH of Met318 (N_7 ····HN, 2.69 Å). The N_6H makes hydrogen bonding interactions with the main chain carbonyl oxygen of Met318 (N_6H ····O=C, 2.28 Å). The O_6 makes hydrogen bonding interactions with the side chain NH of Asn322 (O_6 ···· HN, 3.61 Å). A hydrogen bond between O_5 with the main chain NH of Asp381 (O_5 ····HN, 2.62 Å) and O_4 makes hydrogen bonding interactions with the side chain hydroxyl hydrogen of Glu286 (O_4····HO, 2.83 Å).

Figure 3.3e. A Schematic View of the Inhibitor, CVT-3127 Bound to the *Bcr-Abl* (T396I) kinase. Hydrogen Bonding Interactions in the Protein-Inhibitor Complex. Inhibitor is Indicated in Ball and Stick.



These results shows that the new molecules obtained from virtual screening form several non bonding interactions, and bind *Bcr-Abl* kinase in the VX-680 binding site. Further modifications of these lead molecules will generate inhibitors that bind *Bcr-Abl* kinase with high specificity.

3.4 Conclusions

1. Using pharmacophore modeling and virtual screening, we have identified new lead molecules as *Bcr-Abl* kinase inhibitors.

2. We have studied the binding of these inhibitors to *Bcr-Abl* kinase using docking methods and confirm that these molecules bind the VX-680 binding site of the enzyme.

3. Further modifications and addition of suitable functional groups to these new scaffolds will generate high affinity *Bcr-Abl* kinase specific inhibitors.

4. Using two mutant *Bcr-Abl* kinase proteins for docking, we have identified some useful molecules for drug resistant *Bcr-Abl* kinase protein.

5. Our results confirm chemical function based virtual screening as a powerful tool to discover novel inhibitors of the protein kinase family, and further validate virtual screening as an inexpensive and efficient means for lead discovery.

3.5 References

Buchdunger, E., Cioffi, C. L., Law, N., Stover, D., Ohno-Jones, S., Druker, B. J. & Lydon, N. B. (2000). Abl proteintyrosine kinase inhibitor STI571 inhibits *in vitro* signal transduction mediated by c-kit and platelet-derived growth factor receptors. *J. Pharmacol. Exp. Ther*. **295**, 139–45.

Carter, T. A., Wodicka, L. M., Shah, N. P., Velasco, A. M., Fabian, M. A., Treiber, D. K., Milanov, Z. V., Atteridge, C. E., Biggs, W. H., Edeen, P. T., Floyd, M., Ford, J. M., Grotzfeld, R. M., Herrgard, S., Insko, D. E., Mehta, S. A., Patel, H. K., Pao, W., Sawyers, C. L., Varmus, H., Zarrinkar, P. P. & Lockhart, D. J. (2005). Inhibition of drug-resistant mutants of ABL, KIT, and EGF receptor kinases. *Proc. Natl. Acad. Sci.* U S A. **102**, 11011–6.

Catalyst; Accelrys Inc., San Diego, CA.

Daley, G. Q., Van Etten, R. A. & Baltimore, D. (1990). Induction of chronic myelogenous leukemia in mice by the *P210bcr/abl* gene of the Philadelphia chromosome. *Science.* **247**, 824.

Doggrell, S. A. (2005). BMS-354825: a novel drug with potential for the treatment of imatinib-resistant chronic myeloid leukaemia. *Expert Opin. Investig. Drugs*. **14**, 89–91.

Druker, B. J., Tamura, S., Buchdunger, E., Ohno, S., Segal, G. M., Fanning, S., Zimmermann, J. & Lydon, N. B. (1996). Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of *Bcr-Abl* positive cells. *Nat. Med.* **2**, 561–66.

Druker, B. J., Talpaz, M., Resta, D. J., Peng, B., Buchdunger, E., Ford, J. M., Lydon, N. B., Kantarjian, H., Capdeville, R., Ohno-Jones, S. & Sawyers, C. L. (2001a). Efficacy and safety of a specific inhibitor of the *Bcr-Abl* tyrosine kinase in chronic myeloid leukemia. *N. Engl. J. Med.* **344**, 1031–1037.

Druker, B. J., Sawyers, C. L., Kantarjian, H., Resta, D. J., Reese, S. F., Ford, J. M., Capdeville, R. & Talpaz, M. (2001b). Activity of a specific inhibitor of the *Bcr-Abl* tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome. *N. Engl. J. Med*. **344**, 1038.

Edward, A. & Sausville, A. (1999). *Bcr-Abl* Kinase Antagonist for Chronic Myelogenous Leukemia: a Promising Path for Progress Emerges. *Journal of the National Cancer Institute.* **91,** 102-103.

Gadzicki, D., von Neuhoff, N., Steinemann, D., Just, M., Busche, G., Kreipe, H., Wilkens, L. & Schlegelberger, B. (2005). *Bcr-Abl* gene amplification and overexpression in a patient with chronic myeloid leukemia treated with imatinib. *Cancer Genet.Cytogenet*. **159**, 164–67.

GOLD 3.10, CCDC, UK

Hahn, M. (1997). Three-dimensional shape-based searching of conformationally flexible compounds. *J. Chem. Inf. Comput. Sci.* **37**, 80-86.

Hantschel, O. & Superti-Furga, G. (2004). Regulation of the c-Abl and *Bcr-Abl* tyrosine kinases. *Nat. Rev. Mol. Cell Biol.* **5,** 33–44.

Harrington, E. A., Bebbington, D., Moore. J., Rasmussen, R. K., Ajose-Adeogun, A. O., Nakayama, T., Graham, J. A., Demur, C., Hercend, T., Diu-Hercend, A., Su, M., Golec, J. M. & Miller, K. M. (2004). VX-680, a potent and selective small-molecule inhibitor of the Aurora kinases, suppresses tumor growth *in vivo*. *Nat. Med.* **10**, 262-7.

InsightII 2005, Accelrys, USA

Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R.(1997). Development and validation of a genetic algorithm for flexible docking. *J. Mol.Biol.* **267,** 727-748.

Jürgen Bajorath. (2002). Integration of virtual and high-throughput screening. *Nature Reviews Drug Discovery.* **1**, 882-894.

Kuriyan, J. & Cowburn, D. (1997). Modular peptide recognition domains in eukaryotic signaling. *Annu. Rev. Biophys. Biomol. Struct.* **26**, 259–88.

Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery ReV.* **23**, 3-25.

Liu, Y., Bishop, A., Witucki, L., Kraybill, B., Shimizu, E., Tsien, J., Ubersax, J., Blethrow, J., Morgan, D. O. & Shokat, K. M. (1999). Structural basis for selective inhibition of Src family kinases by PP1. *Chem. Biol.* **6**, 671–8.

Lugo, T. G., Pendergast, A. M., Muller, A. J. & Witte, O. N. (1990). Tyrosine kinase activity and transformation potency of *Bcr–Abl* oncogene products. *Science.* **247,** 1079.

Manley, P. W., Cowan-Jacob, S. W., Buchdunger, E., Fabbro, D., Fendrich, G., Furet, P., Meyer, T. & Zimmermann, J. (2002). Imatinib: a selective tyrosine kinase inhibitor. *Eur. J. Cancer.* **38,** S19–S27.

Mason, J. S., Good, A. C. & Martin, E. J. (2001). 3-D pharmacophores in drug discovery. *Curr. Pharm. Des.* **7,** 567-97.

Melo, J. V. (1996). The diversity of *Bcr–Abl* fusion proteins and their relationship to leukemia phenotype. *Blood*. **88**, 2375.

Nagar, B., Bornmann, W. G., Pellicena, P., Schindler, T., Veach, D. R., Miller, W. T., Clarkson, B. & Kuriyan, J. (2002). Crystal structures of the kinase domain of c-Abl in complex with the small molecule inhibitors PD173955 and imatinib (STI-571). *Cancer Res*. **62,** 4236–4243.

Neet, K. & Hunter, T. (1996). Vertebrate non-receptor protein-tyrosine kinase families. *Genes Cells.* **1,** 147–69.

Nissink, J. W., Murray, C., Hartshorn, M., Verdonk, M. L., Cole, J. C. & Taylor, R. A. (2002). New test set for validating predictions of protein-ligand interaction. *Proteins*. **49,** 457-471.

Nocka, K., Majumder, S., Chabot, B., Ray, P., Cervone, M., Bernstein, A. & Besmer, P. (1989). Expression of c-kit gene products in known cellular targets of W mutations in normal and W mutant mice Evidence for an impaired c-kit kinase in mutant mice. *Genes Dev.* **3,** 816–826.

O'Brien, S. G., Guilhot, F., Larson, R. A., Gathmann, I., Baccarani, M., Cervantes, F., Cornelissen, J. J., Fischer, T., Hochhaus, A., Hughes, T., *et al.,* (2003). Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. *N. Engl. J. Med.* **348**, 994–1004.

Pendergast, A. M. (2002). The *Abl* family kinases: mechanisms of regulation and signaling. *Adv. Cancer. Res*. **85**, 51–100.

Putta, S., Lemmen, C., Beroza, P. & Greene, J. (2002). A novel shape-feature based approach to virtual library screening. *J. Chem. Inf. Comput. Sci.* **42,** 1230.

Rowley, J. D. (1973). A new consistent chromosomal abnormality in chronic myelogenous leukemia identified by quinacrine fluorescence and Giesma staining. *Nature*. **243**, 290.

Sawyers, C. L. & Druker, B. (1999). Tyrosine kinase inhibitors in chronic myeloid leukemia. *Cancer J. Sci. Am.* **5**, 63–9.

Schindler, T., Bornmann, W., Pellicena, P., Miller, W. T., Clarkson, B. & Kuriyan, J. (2000). Structural mechanism for STI-571 inhibition of Abelson tyrosine kinase. *Science.* **289**, 1938–1942.

Shah, N. P., Nicoll, J. M., Nagar, B., *et al.,* (2002). Multiple *Bcr-Abl* kinase domain mutations confer polyclonal resistance to the tyrosine kinase inhibitor imatinib (STI571) in chronic phase and blast crisis chronic myeloid leukemia. *Cancer Cell.* **2,** 117–25.

Shah, N. P., Tran, C., Lee, F, Y., Chen, P., Norris, D. & Sawyers, C. L. (2004). Overriding Imatinib Resistance with a Novel ABL Kinase Inhibitor. *Science.* **305**, 399–401.

SILVER 1.1.1, CCDC, UK.

Stevan, R. H. & Till, J. H. (2000). Protein tyrosine kinase structure and function. *Annu. Rev. Biochem*. **69,** 373–98.

von, B. N., Veach, D. R., Miller, W. T., Li, W., Sänger, J., Peschel, C., Bornmann, W. G., Clarkson, B. & Duyster, J. (2003). Inhibition of wild-type and mutant Bcr-Abl by pyrido-pyrimidine-type small molecule kinase inhibitors. *Cancer Res.* **63**, 6395-404.

Weisberg, E., Manley, P. W., Breitenstein, W., Bruggen, J., Cowan-Jacob, S. W., *et al.* (2005). Characterization of AMN107, a selective inhibitor of native and mutant *Bcr-Abl*. *Cancer Cell.* **7**, 129–41.

Wong, S. & Witte, O. N. (2004). The *Bcr-Abl* story: bench to bedside and back. *Annu. Rev. Immunol*. **22**, 247–306.

Young, M. A., Shah, N. P., Chao, L. H., Seeliger, M., Milanov, Z. V., Biggs, W. H., Treiber, D. K., Patel, H. K., Zarrinkar, P. P., Lockhart, D. J., Sawyers, C, L. & Kuriyan, J.(2006). Structure of the kinase domain of an imatinib-resistant Abl mutant in complex with the Aurora kinase inhibitor VX-680. *Cancer Res*. **66**, 1007-14.

CHAPTER 4

The Identification of New Aurora A Kinase Inhibitors by Pharmacophore Modeling, Virtual Screening and Molecular Docking

4.1 Introduction

Aurora kinases are NRTK proteins. Mammals express three Aurora kinase paralogues A, B and C each of which is thought to play vital role in regulating mitosis (Bischoff & Plowman, 1999). Besides playing a crucial role in mitosis for G2/M check point where they have been implicated in centrosome maturation, chromosome segregation and cytokinesis (Meraldi *et al.,* 2004; Doggrell, 2004; Sasai *et al.,* 2004), Aurora kinases have been found to be overexpressed in a number of tumor cell lines and human primary tumors (Bischoff *et al.,* 1998; Warner *et al.,* 2003). Therefore, one of the promising targets in cancer drug discovery is represented by Aurora A, B and C kinases. Aurora A itself has been identified as a predominantly attractive drug target through observations that it can act as an oncogene and transform cells when ectopically expressed. For example, Aurora A is overexpressed in primary colorectal cancers (Bischoff *et al.,* 1998), breast tumours (Zhou *et al.,* 1998; Tanaka *et al.,* 1999), ovarian tumour (Tanner *et al.,* 2000) and cell lines from breast, ovarian, colon, prostate, neuroblastoma and cervical (Zhou *et al.,* 1998; Tanner *et al.,* 2000; Sen *et al.,* 1997). Experimental evidence suggests that the oncogenic transformation of Aurora A is mediated through centrosome amplification, resulting in chromosomal instability (Giet & Prigent, 1999; Miyoshi *et al.,* 2001). The expression profile of Aurora A in carcinoma suggests that inhibitors of this kinase may have inherent potential as therapeutic agents. Several groups have designed kinase inhibitors based on the ADP binding interactions with the catalytic residues of kinase (Pevarello *et al.,* 2006; Moriarty *et al.,* 2006; Mortlock *et al.,* 2005; Fancelli *et al.,* 2005).

The structure of a kinase comprises two domains, the smaller N-terminus has an antiparallel β sheet and the larger C-terminus mainly comprises α helices and an activation loop that undergoes conformational changes upon substrate

binding. The substrate/inhibitor binding site is located between the two domains. The 3-D crystal structure of Aurora A kinase bound to ADP has been determined (Nowakowski *et al.,* 2002).

When the 3-D structure of a substrate bound target protein is known, it is possible to design novel inhibitors through computational SBDD methods. In the RDD process many computer aided techniques have emerged to increase the efficiency of finding new lead molecules (Bajorath, 2002; Eckert & Bajorath, 2007). Further, the time and costs for biological assays can be reduced using *in silico* studies and by examining the likeliness of a molecule to exhibit activity towards the target of interest in advance.

The pharmacophore model is a good approach to quantitatively explore common chemical characteristics among a considerable number of structures with great diversity. A good pharmacophore model could be used as query for searching small molecule databases in order to discover novel chemical entities. A reliable pharmacophore model can be built using HypoGen (a part of the Catalyst suite of software), provided, we have experimental affinities of protein-ligand interactions. The ligands should have a variety of scaffolds and the range of affinity should vary at least a thousand fold. The pharmacophore model that represents a QSAR can be used to search small molecule databases in order to identify molecules that represent similar shape and chemical features from a given database. The best pharmacophore must identify most of the known inhibitors and almost no false positives while searching databases. Docking presents a widely applied, computer assisted approach to predict the conformation of a ligand when bound to protein and requires 3-D structural knowledge of the target protein. The molecules obtained from virtual screening can be examined through docking studies to examine the binding mode of protein-ligand complex.

In this work we have generated a hypothetical model of the primary pharmacophore features responsible for the bioactivity of various classes of Aurora A inhibitors using HypoGen module in the Catalyst suite of software. The best pharmacophore model, Hypo1 has been validated and used to screen NCI and Maybridge databases using chemical function descriptors to identify lead molecules as novel inhibitors. These potential inhibitors were docked into the active site of Aurora A using the GOLD software (Jones *et al.,* 1997) to understand their mode of binding to Aurora A kinase.

4.2 Methods

4.2.1 Pharmacophore model Generation:

A ligand-based pharmacophore model was generated for Aurora A using HypoGen module in Catalyst (Catalyst 4.11, Accelrys), for training set consisting 24 compounds with their corresponding affinity data (Moriarty *et al.,* 2006; Mortlock *et al.,* 2005). These molecules were selected by considering the structural diversity and wide coverage of activity range shown in Figure 4.1. Activities are reported as IC_{50} values straddling from 0.6nM to 10 μM range and also listed in Figure 4.1. The molecules were built using catalyst 2D-3D sketcher and a family of representative conformations were generated for each compound using the "best conformational analysis" method with Poling algorithm and CHARMM force field parameters (Smellie *et al.,* 1995). A maximum number of 250 conformations for each compound were selected within a constraint of 20 kcal/mol energy threshold, above the minimum conformer searched, in order to ensure maximum coverage of the conformational space. In the hypotheses generation process, a default uncertainty factor of 3 for each compound was defined, and all possible combinations of features types; hydrogen bond donors (HD), hydrogen bond acceptors (HA), hydrophobic (HP) and ring aromatic (R) were allowed. A maximum of 5 features per hypotheses were selected to construct the pharmacophore hypotheses.

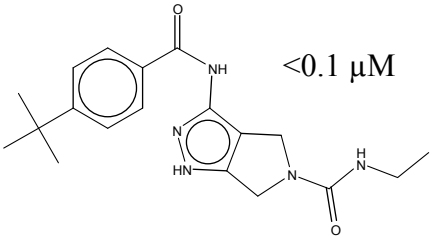
4.2.2 Validation of the pharmacophore model:

4.2.2.1 Test set validation:

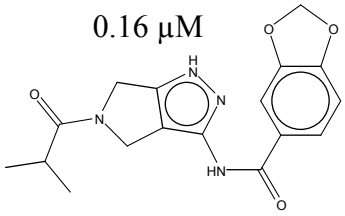
A test set consisting 21 compounds (Moriarty *et al.,* 2006; Mortlock *et al.,* 2005) were selected by considering the structural diversity and wide coverage of activity range. These molecules were built and conformational analysis was carried out similar to the molecules from training set. For hypotheses validation, we have considered Hypo1 and test set molecules to estimate the biological activity and compare with the experimental data.

Figure 4.1. Schematic Representation of the Molecules in the Training set and Their Corresponding IC_{50} values.

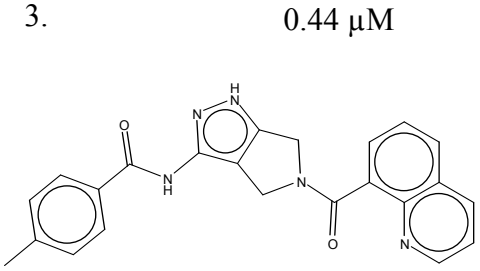
1.



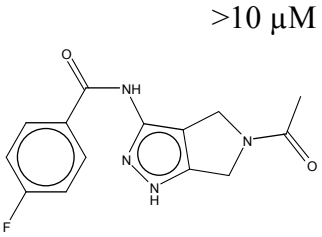
2.



3.



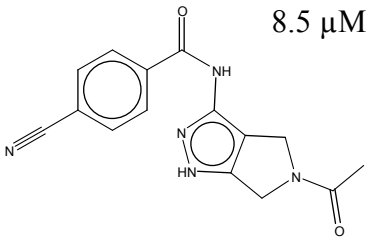
4.



…… continued

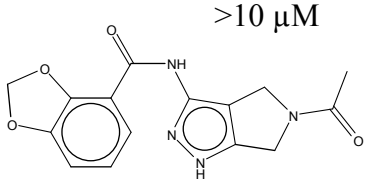
5.

8.5 μM



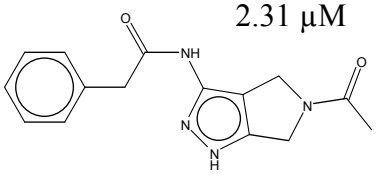
6.

>10 μM



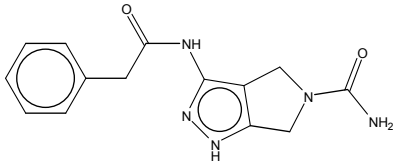
7.

2.31 μM



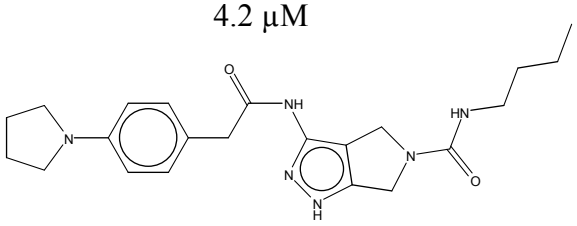
8.

>10 μM



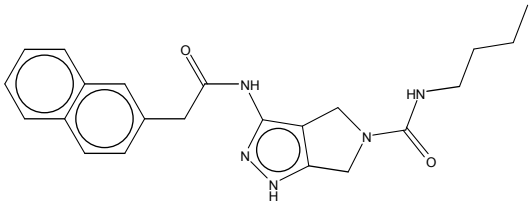
9.

4.2 μM



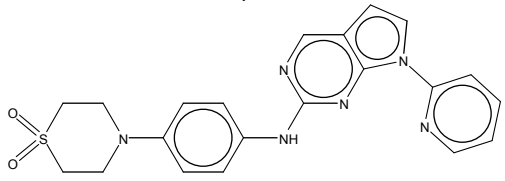
10.

>10 μM



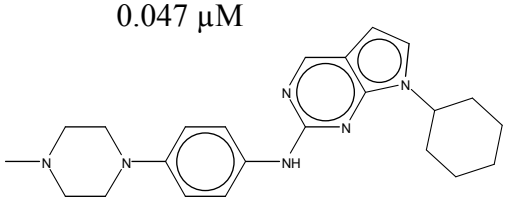
11.

0.0008 μM



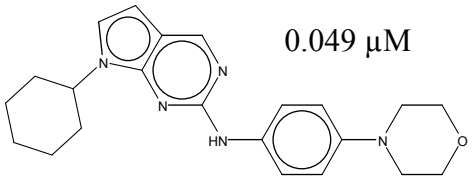
12.

0.047 μM

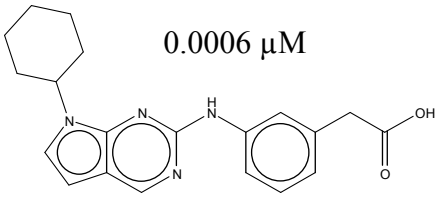


…… continued

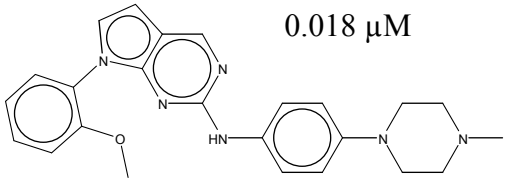
13.



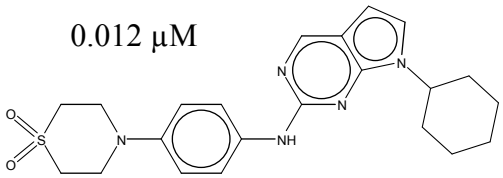
14.



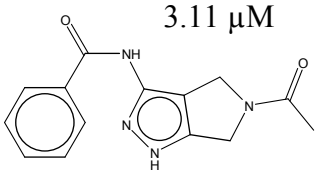
15.



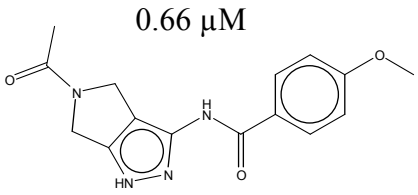
16.



17.

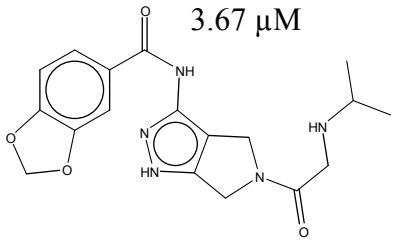


18.

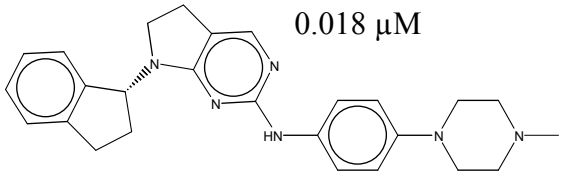


…… continued

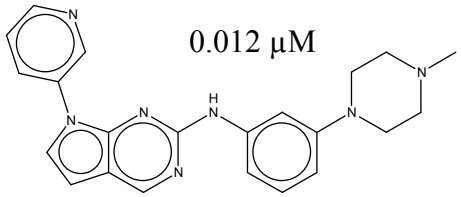
19.



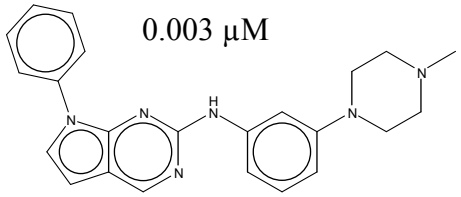
20.



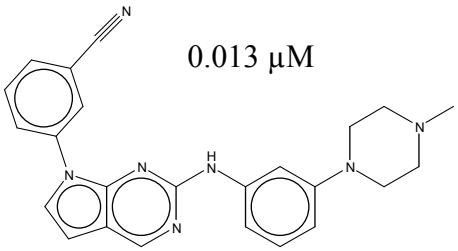
21.



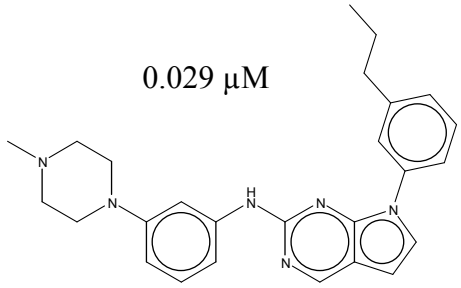
22.



23.



24.



4.2.2.2 CatScramble validation:

To further validate the statistical relevance of hypotheses, Hypo1, Fischer method (Fischer, 1966) was applied. Using the CatScramble program, the experimental values were scrambled erratically and the training sets obtained were used for the HypoGen run. All parameters similar to the initial HypoGen run for the training set have been used for the scrambled sets and a 95% confidence level was selected, as a result, 19 spreadsheets were generated using the CatScramble command.

4.2.2.3 Generation of new database and screening:

To further corroborate our pharmacophore, we have generated a new database of 200 molecules. This new database consists of 24 molecules from training set, 21 molecules from test set, 106 random molecules and 45 molecules from NCI and 4 molecules from Maybridge databases that were obtained using pharmacophore screening. We have generated 250 conformations for each molecule in this new database using "best flexible search" method and finally we have screened this database using our Pharmacophore generated (Hypo1).

4.2.3 Virtual screening:

The best pharmacophore model developed using HypoGen, Hypo1, was used as a 3-D structural query in the *in silico* screening of NCI and Maybridge databases**.** NCI and Maybridge databases comprise 238,819 and 59,652 molecules respectively. For each molecule in the database, up to 100 conformers were generated using the "fast fit" method in Catalyst. The chemical function based pharmacophore model was used for database searching by the "best flexible search" method in Catalyst. The hits obtained were further filtered using Lipinski's rule of 5 (Lipinski *et al.,* 1997).

4.2.4 Protein preparation:

The 3-D co-ordinates of Aurora A kinase complexed with ADP (PDB_ID: 1MQ4) (Pevarello *et al.,* 2006) was downloaded from protein structure databank (http://www.rcsb.org/). Hydrogen atoms were added to the protein using Biopolymer module in InsightII 2005 (InsightII 2005, Accelrys) keeping all the residues in their charged form. Primarily, all the hydrogen atoms were minimized, keeping all other atoms fixed. The whole protein complex including crystal water was energy minimized by the steepest descent followed by conjugate gradient methods to achieve a convergence gradient of 0.01 kcal/mol using CVFF force fields in InsightII. Crystallographic waters were retained for docking studies.

4.2.5 Docking: discussed in the section 2.2.5 previous chapter

The natural substrate ADP and the new lead molecules identified from virtual screening, were docked into the crystal structure of Aurora A (PDB_ID: 1MQ4) using GOLD (GOLD 3.10, CCDC, UK) software. GOLD (Genetic Optimization of Ligand Docking) is a genetic algorithm for docking flexible ligands into protein binding sites. The details were discussed in the section 2.2.5. During docking, the default algorithm speed was selected, and the ligand binding site in the Aurora A, was defined within a 10 Å radius with the centroid as Glu211 main chain carbonyl oxygen atom. For docking, the number of poses for each inhibitor was set to 10, and early termination was allowed if the top 5 bound conformations of a ligand were within 1.5 Å RMSD. After docking, the individual binding poses of each ligand were re-ranked according to the GOLD score. The top ranked conformation of each ligand was selected and analyzed using SILVER (SILVER 1.1.1, CCDC, UK) to understand the mode of protein-inhibitor binding.

4.2.6 Hardware and software:

InsightII 2005 was used for energy minimization of Aurora A and Catalyst 4.11 was used for pharmacophore generation and virtual screening on SGI Octane2 workstation equipped with 2600 MHz MIPS R14000 processors. The docking calculations using GOLD 3.1 software (Jones *et al.,* 1995) and docking analysis using SILVER 1.1.1 (Nissink *et al.,* 2002) were carried out on an Intel P4-based windows system.

4.3 Results and Discussion

The aim of the present work is to identify novel lead molecules as inhibitors of Aurora A kinase and we have achieved this using pharmacophore model generation, virtual screening of databases and docking studies.

4.3.1 Generation of pharmacophore model:

Catalyst commonly produces 10 hypotheses for a list of molecules in the training set chosen based on the structural diversity and broad range of affinity for Aurora A. The null cost of the 10 hypotheses is 177.47 and the fixed cost value is 98.29. Configuration cost value is 16.44. All the 10 hypotheses have a total cost close to the cost of the fixed hypotheses. The difference between the fixed cost and the null cost is 79.18 bits. The cost range (Δ cost) between these hypotheses and the null hypotheses varies between 68.04 and 65.39 bits with a low cost range, 2.65 bits. Therefore, we can approximate that for all these hypotheses, there is more than 90% chance of representing a true correlation in the data.

We considered the first pharmacophore model, Hypo1, as the best pharmacophore hypotheses in this study, characterized by the highest cost difference, lowest error cost, closest weight to 2, lowest RMSD and the best correlation coefficient. All hypotheses with the exception of hypotheses 5 and 7 have the same features; one Hydrogen bond acceptor (HA), two Hydrophobic groups (HP) and one Ring aromatic group (R). In this pharmacophore modeling, for the best hypotheses Hypo1, the RMSD value 0.927, signifies a good quality for Hypo1 and correlation coefficient 0.946, shows a good linear regression of the geometric fit index. Table 4.1 shows these parameters of statistical significance and the prediction power. About 84% of the molecules in the training set were

predicted within an error less than 2 units. The mapping of Hypo1 on inhibitors with highest and lowest affinity is shown in Figures 4.2a and b. Null cost of top 10 score hypotheses is 177.470 bits. Fixed cost is 98.292 bits. Configuration cost is 16.4406 bits. Abbreviation used for features: HA, Hydrogen bond acceptor; HP, hydrophobic; R, aromatic ring; HD, hydrogen bond donor.

Table 4.1. Statistical Parameters and Composition Features of Pharmacophore Models for the Training Set Molecules.

Hypotheses	Features	Total cost	Δ cost	RMSD	Correlation (r)
1	HA HP HP R	109.431	68.039	0.927	0.946
2	HA HP HP R	109.594	67.876	0.957	0.942
3	HA HP HP R	110.618	66.852	0.959	0.942
4	HA HP HP R	110.921	66.549	1.019	0.933
5	HA HD HP HP HP	111.154	66.316	1.020	0.933
6	HA HP HP R	111.42	66.050	1.011	0.935
7	HA HD HP HP HP	111.635	65.835	1.052	0.929
8	HA HP HP R	111.708	65.762	1.038	0.931
9	HA HP R R	111.885	65.585	1.047	0.929
10	HA HP HP R	112.076	65.394	1.042	0.931

Figure 4.2. Pharmacophore Mapping of the Training Set Molecules. (a) Molecule 11 with High Affinity (b) Molecule 10 with Low Affinity. Green, Blue, Orange Spheres Represent Hydrogen Bond Acceptor, Hydrophobic Groups and Ring Aromatic Group Respectively.

Figure 4.2a

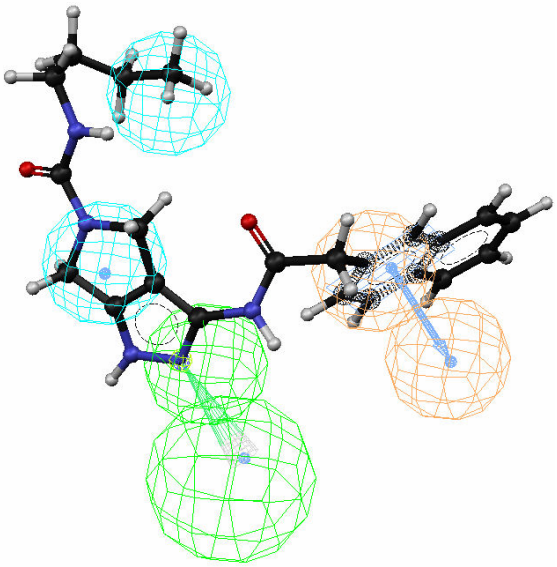
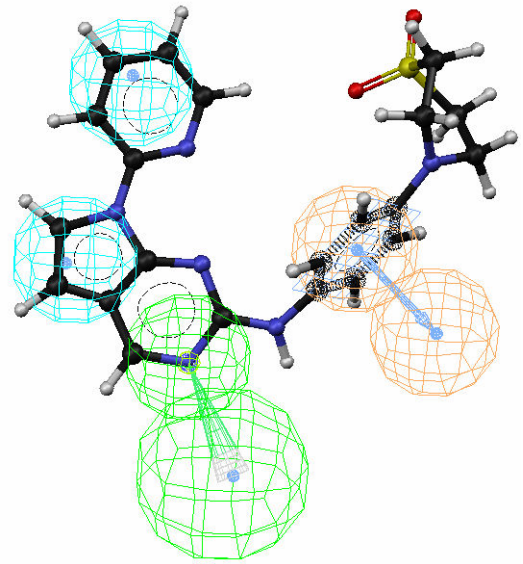


Figure 4.2b



4.3.2 Validation of pharmacophore model:

The main purpose of a quantitative pharmacophore model is to identify lead molecules that are potential inhibitors and to predict their activity accurately. The best pharmacophore model, Hypo1, was validated by a test set comprising 21 molecules. To check if the hypotheses can also predict the activity of compounds that are different from those included in the training set, we created a test set comprising 21 molecules with different structural information and activity. Molecules in the test set were built, energy minimized and their conformations were generated similar to the molecules in the training set. Hypo1 shows a good

correlation between actual and estimated IC_{50} values. As per the statistical analysis, our pharmacophore hypotheses, Hypo1 is valid, see Table 4.2. The error values of 19 compounds were found to be less than 7 indicating that for 90% molecules, the activity prediction is either 7 fold higher or lower than the actual activity. The pharmacophore mapping of the molecules with highest and lowest affinity is shown in Figures 4.3a and b.

Figure 4.3. Pharmacophore Mapping of the Test Set Molecules. (a) Molecule 19 with High Affinity (b) Molecule 11 with Low Affinity. Green, Blue, Orange Spheres Represent Hydrogen Bond Acceptor, Hydrophobic Groups and Ring Aromatic Group Respectively.

Figure 4.3a

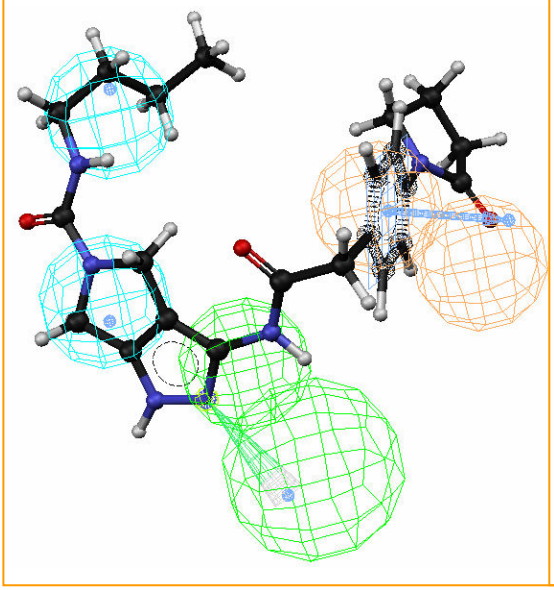
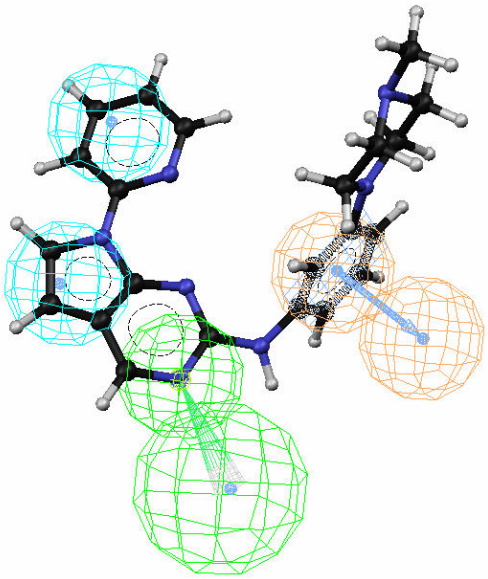


Figure 4.3b



Further evaluation of Hypo1 using the Fischer method was applied to validate the strength of correlation between the chemical structures and their biological activity. The 19 spreadsheets obtained using the CatScramble program by erratically scrambling the binding affinity data was used for the HypoGen run. In Table 4.3, we have shown the statistical significance of 19 hypotheses. The low Cost differences, high RMSD values and low correlation values indicate that the data of cross validation generated after randomization produced hypotheses with no predictive values and we therefore believe that Hypo1 could be used for further database screening.

The screening of new database comprising 200 molecules using the best pharmacophore, Hypo1 identified, 33 molecules from our test and training sets, negligible number of molecules from random set and 40 molecules from the screened set. These validation results indicate that the hypotheses Hypo1, generated by HypoGen using the training set molecules is significant and not random. We further propose that it is specific to Aurora A inhibitors.

4.3.3 Database screening:

The pharmacophore model, Hypo1, was used to screen NCI and Maybridge databases. In all, about 600 molecules were obtained as hits from *in silico* screening. To assess the drug-likeliness of these hits, a second screen, incorporating Lipinski's rule of 5 was used. A total of 99 molecules were obtained as hits after this screen. This second screen selects only those molecules that possess drug like properties.

Table 4.2. Actual and Predicted Activities of the Test Set Molecules.

Molecule	Actual IC_{50} (μM)	Estimated IC_{50} (μM)	Error
1	2.14	1.9	-1.1
2	3.67	2.8	-1.3
3	3.67	5.6	+1.5
4	0.16	0.092	+1.7
5	0.1	0.9	+9
6	2.31	2.4	1
7	10	5.8	-1.7
8	10	1.4	-7.1
9	10	2.6	-3.8
10	10	3	-3.3
11	10	6.7	-1.4
12	0.049	0.68	+13
13	0.039	0.019	-2
14	0.013	0.0064	-2
15	0.009	0.005	-1.8
16	0.01	0.0042	-2.3
17	0.006	0.0078	+1.3
18	0.011	0.005	-2.2
19	0.005	0.005	+1
20	0.0006	0.0043	+7.1
21	0.0008	0.0049	+6.1

Table 4.3. Statistical Parameters of the 10 Best Pharmacophore Models Obtained from Cross-validation Using CatScramble.

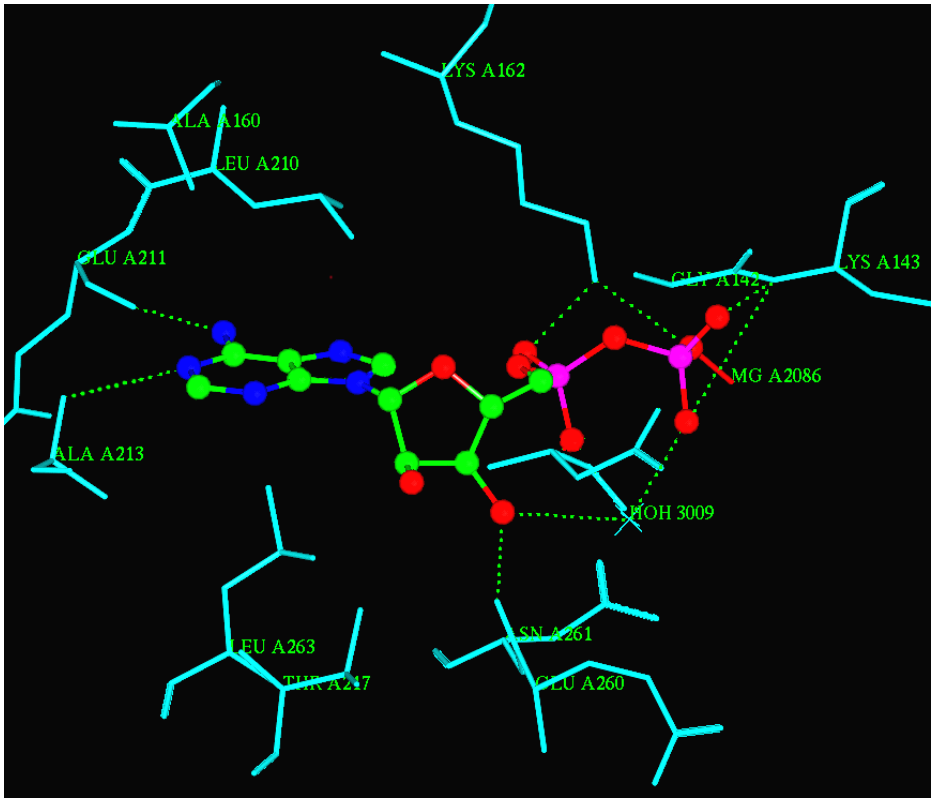
Trial No	Total cost	Fixed cost	RMSD	Correlation (r)	Configuration cost
Results for original data					
Hypo1	109.431	98.292	0.927	0.946	16.440
Results for scrambled data					
1	132.527	95.354	1.653	0.819	13.503
2	132.393	94.609	1.695	0.807	12.758
3	145.649	96.423	1.987	0.717	14.571
4	157.447	97.552	2.163	0.660	15.701
5	152.698	95.084	2.158	0.654	13.233
6	147.313	95.464	2.048	0.695	13.612
7	132.055	96.122	1.608	0.831	14.270
8	148.613	93.653	2.119	0.672	11.802
9	130.47	97.732	1.516	0.853	15.881
10	158.468	93.669	2.308	0.585	11.818
11	138.167	97.273	1.791	0.779	15.422
12	135.698	95.472	1.569	0.858	13.620
13	151.018	96.763	2.064	0.694	14.912
14	145.303	97.224	1.992	0.713	15.373
15	132.086	95.852	1.654	0.817	14.001
16	137.995	96.741	1.592	0.855	14.889
17	151.236	97.629	2.067	0.691	15.778
18	139.882	97.293	1.815	0.774	15.441
19	154.741	95.395	2.137	0.674	13.543

4.3.4 GOLD docking:

The crystal structure of Aurora A bound to substrate ADP (PDB_ID: 1MQ4) was used for the docking studies. All amino acids within 10 Å radius from the Glu211 main chain carbonyl oxygen atom were considered to comprise the active site. Docking was carried out using GOLD 3.10 software.

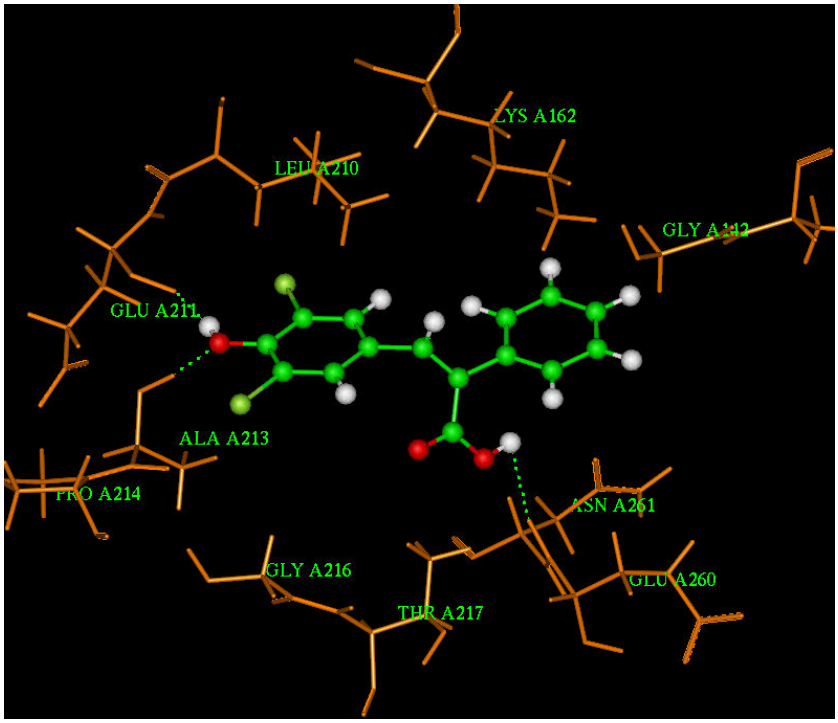
The substrate ADP was docked into the Aurora A kinase and the following interactions between ADP and the Aurora A have been observed as shown in Figure 4.4a. (i) a hydrogen bond interaction between the adenine ring N_6H and Glu211 carbonyl oxygen ($N_6H^{\cdots}O$=C, 2.69 Å). (ii) A hydrogen bond between adenine ring N_1 and the main chain NH of Ala213 ($N_1^{\cdots}$NH, 3.16 Å). (iii) A hydrogen bond between ribose ring O_3H and the main chain carbonyl oxygen of Glu260 ($O_3H^{\cdots}O$=C, 2.60 Å). (iv) A hydrogen bond between the ribose ring O_3H and water molecule ($O_3H^{\cdots}$ O, 2.60 Å). (v) Two hydrogen bonds between ADP 1OA, ADP 1OB and Lys162 (1OA$^{\cdots}NZH_1$, 2.30 Å; 1OB$^{\cdots}NZH_2$, 2.18 Å). (vi) A bifurcated hydrogen bond between ADP 2OB, 3OB and Lys143 NH (2OB$^{\cdots}$NH, 2.67 Å; 3OB$^{\cdots}$NH Å). The RMSD between the docked pose of ADP and its bound conformation in the crystal structure 1MQ4 is 0.55 Å, indicating that GOLD software was able to reproduce the correct pose and is a reliable method for these docking studies.

Figure 4.4a. The Interaction of ADP Molecule with Aurora A Kinase. Hydrogen Bonding Interactions in the Protein-ADP Complex. ADP is Indicated in Ball and Stick.



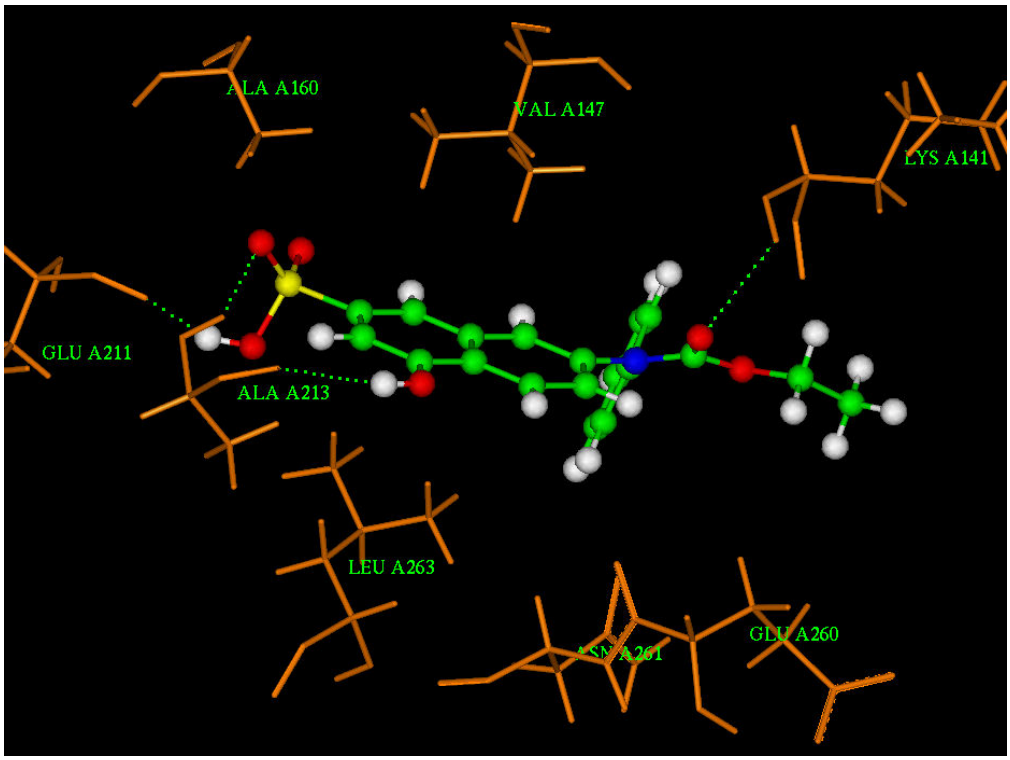
The molecules obtained from virtual screening were docked into the Aurora A crystal structure. All molecules fit into the ADP binding site of the enzyme. The binding of some molecules to Aurora A is described below. We have given the numbering of atoms according to those databases. The interaction of Hit NCI0000161 is shown in Figure 4.4b. In the molecule NCI0000161, O_9H makes hydrogen bonding interactions with the main chain carbonyl oxygen of Glu211 (O_9H ····O=C, 2.30 Å). Further, the O_9 makes hydrogen bonding interactions with the main chain NH of Ala213 (O_9····HN, 2.63 Å). The $O_{20}H$ makes hydrogen bonding interactions with the main chain carbonyl oxygen of Glu260 ($O_{20}H$ ····O=C, 2.30 Å).

Figure 4.4b. The Interaction of NCI0000161 Molecule with Aurora A Kinase. Inhibitor is Indicated in Ball and Stick.



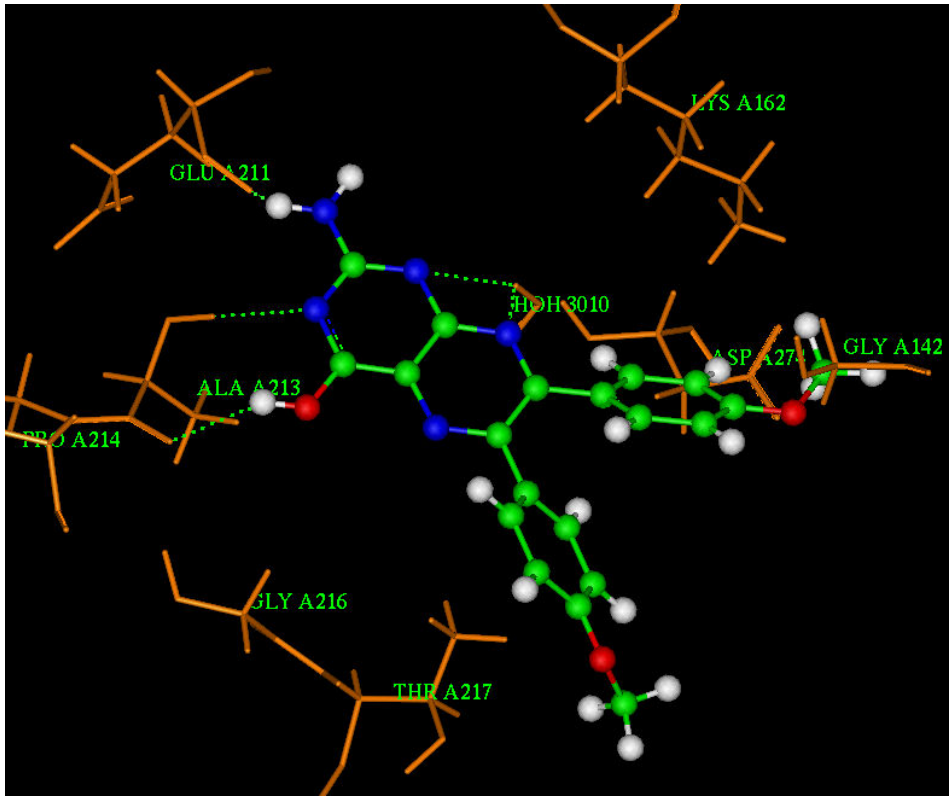
The interaction of Hit NCI0000169 is shown in Figure 4.4c. In the molecule NCI0000169, $O_{15}H$ makes hydrogen bonding interactions with the main chain carbonyl oxygen of Glu211 ($O_{15}H$ ····O=C, 2.49 Å). Further, the O_{14} makes hydrogen bonding interactions with the main chain NH of Ala213 (O_{14}····H-N, 2.44 Å). The $O_{16}H$ makes hydrogen bonding interactions with the main chain carbonyl oxygen of Ala213 ($O_{16}H$ ····O=C, 2.53 Å). The O_1 makes hydrogen bonding interactions with the main chain NH of Lys141 (O_1····HN, 2.60 Å).

Figure 4.4c. The Interaction of NCI0000169 Molecule with Aurora A Kinase. Inhibitor is Indicated in Ball and Stick.



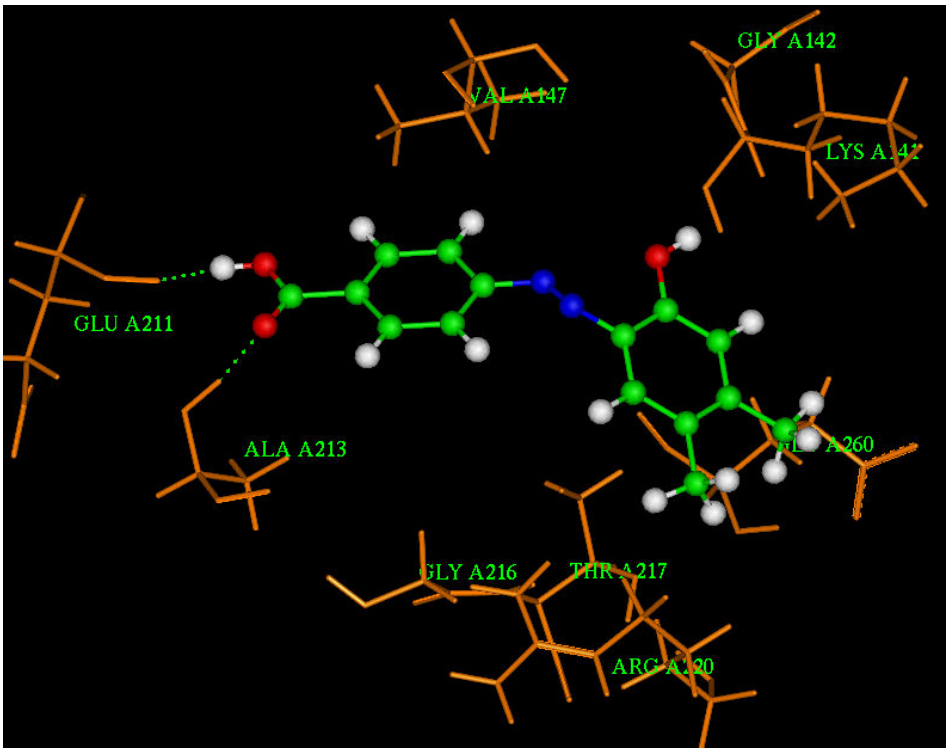
The interaction of Hit NCI00001568 is shown in Figure 4.4d. In the molecule NCI0001568, $N_{20}H$ makes hydrogen bonding interactions with the main chain carbonyl oxygen of Glu211 ($N_{20}H$ ····O=C, 1.43 Å). The N_{19} makes hydrogen bonding interactions with the main chain NH of Ala213 (N_{19}····HN, 2.23 Å). The O_1H makes hydrogen bonding interactions with the main chain carbonyl oxygen of Ala213 (O_1H ····O=C, 2.16 Å). Further, the N_{15} makes hydrogen bonding interactions with the water molecule 3010 (N_{15}····HOH, 2.83 Å) and the N_{17} makes hydrogen bonding interactions with the water molecule 3010 (N_{17}····HOH, 2.73 Å).

Figure 4.4d. The Interaction of NCI00001568 Molecule with Aurora A Kinase. Inhibitor is Indicated in Ball and Stick.



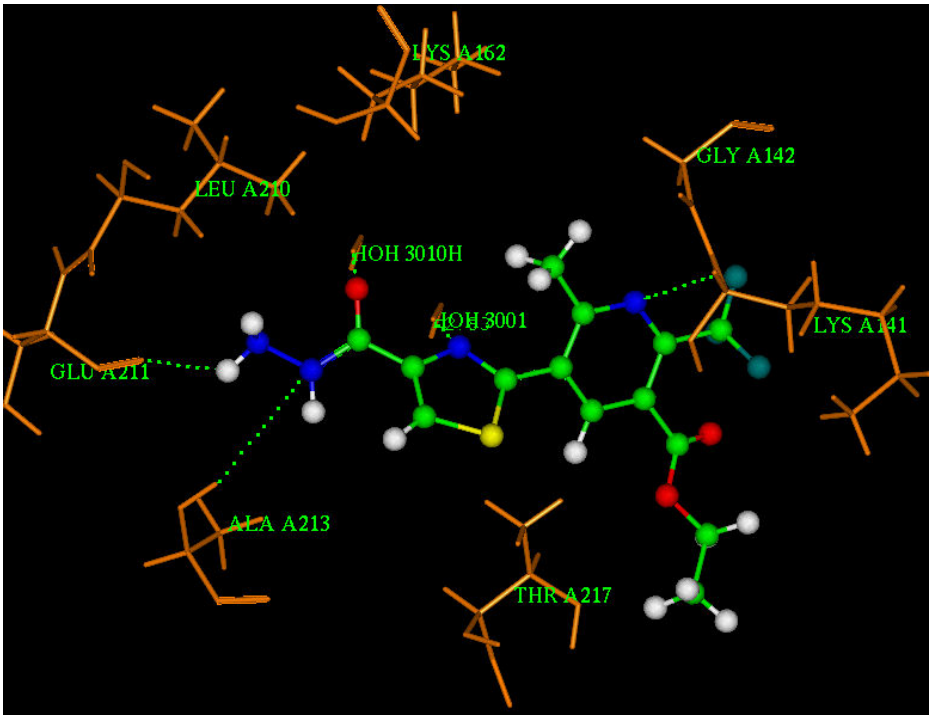
The interaction of Hit NCI00001576 is shown in Figure 4.4e. In the molecule NCI0001576, $O_{18}H$ makes hydrogen bonding interactions with the main chain carbonyl oxygen of Glu211 ($O_{18}H$ ····O=C, 1.57 Å). Further, the O_{16} makes hydrogen bonding interactions with the main chain NH of Ala213 (O_{16}····HN, 2.65 Å).

Figure 4.4e. The Interaction of NCI00001576 Molecule with Aurora A Kinase. Inhibitor is Indicated in Ball and Stick.



The interaction of Hit AW_00732 is shown in Figure 4.4f. In the molecule AW_00732, $N_{24}H$ makes hydrogen bonding interactions with the main chain carbonyl oxygen of Glu211 ($N_{24}H$ ····O=C, 2.13 Å). The N25 makes hydrogen boning with main chain NH of Ala213 (N_{25}····HN, 3.64 Å). The $N_{10}H$ makes hydrogen bonding interactions with the main chain carbonyl oxygen of Lys 141 ($N_{10}H$····O=C, 1.85 Å) The O_1 makes hydrogen bonding interactions with the water molecule 3010 (O_1····HOH, 2.54 Å). Further, the N_4 makes hydrogen bonding interactions with the water molecule 3001 (N_4····HOH, 2.83 Å).

Figure 4.4f. The Interaction of AW_00732 Molecule with Aurora A Kinase. Inhibitor is Indicated in Ball and Stick.



These results show that the new molecules obtained from virtual screening form several non bonding interactions, and bind Aurora A in the ADP binding site. Hence we believe that the molecules obtained from the virtual screening of the databases will inhibit Aurora A with high affinity. Further modifications of these lead molecules will generate inhibitors that bind Aurora A kinase with high specificity.

4.4 Conclusions

1. VS methods are becoming an integral part of the drug discovery process. In this chapter, a strategy for the screening of large compound libraries to obtain a limited set of prospective hits against Aurora A Kinase has been suggested.
2. Using pharmacophore modeling and virtual screening, we have identified new lead molecules as Aurora A kinase inhibitors.
3. We have studied the binding of these inhibitors to Aurora A kinase using docking methods and confirm that these molecules bind the ADP binding site of the enzyme.
4. This pharmacophore was used to screen a database of about 300,000 compounds. Application of subsequent filters ensures that the hit list comprises of drug like molecules. The hits obtained from the VS include molecules with diverse scaffolds that could be possible leads for further development as novel Aurora A kinase inhibitors.
5. Further modifications and addition of suitable functional groups to these new scaffolds will generate high affinity Aurora A kinase specific inhibitors.

4.5 References

Bajorath, J. (2002). Integration of virtual and high-throughput screening. *Nature Reviews Drug Discovery.* **1***,* 882-894.

Bischoff, J. R., Anderson, L., Zhu, Y., Mossie, K., Ng, L., Souza, B., Schryver, B., Flanagan, P., Clairvoyant, F., Ginther, C., Chan, C. S., Novotny, M., Slamon, D. J. & Plowman, G. D. (1998). A homolog of Drosophila aurora kinase is oncogenic and amplified in human colorectal cancers. *EMBO J.* **17**, 3052-3065.

Bischoff, J. R. & Plowman, G. D. (1999). The Aurora/Ipl1p kinase family: regulators of chromosome segregation and cytokinesis. *Trends Cell Biol.* **9**, 454-459.

Catalyst, Version 4.11, Accelrys, USA.

Doggrell, S. A. (2004). Dawn of Aurora kinase inhibitors as anticancer drugs. *Expert Opin. Investig. Drugs.* **13,** 1199-1201.

Eckert, H. & Bajorath, J. (2007). Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov Today.* **12,** 225-33.

Fancelli, D., Berta, D., Bindi, S., Cameron, A., Cappella, P., Carpinelli, P., Catana, C., Forte, B., Giordano, P., Giorgini, M. L., Mantegani, S., Marsiglio, A., Meroni, M., Moll, J., Pittalà, V., Roletto, F., Severino, D., Soncini, C., Storici, P., Tonani, R., Varasi, M., Vulpetti, A. & Vianello, P. (2005). Potent and selective Aurora inhibitors identified by the expansion of a novel scaffold for protein kinase inhibition. *J. Med. Chem.* **48**, 3080-3084.

Fischer, R. (1966). Chapter2, The Design of Experiments, Hafner Publishing: New York.

Giet, R. & Prigent, C. (1999). Aurora/Ipl1 p-related kinases, a new oncogenic family of mitotic serine-threonine kinases. *Journal of Cell Science*. **112,** 3591-3601.

InsightII 2005, Accelrys, USA.

Jones, G., Willett, P. & Glen, R. C. (1995). Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **245,** 43-53.

Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *J. Mol.Biol.* **267,** 727-748.

GOLD 3.10, CCDC, UK.

Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery ReV.* **23,** 3-25.

Meraldi, P., Honda, R. & Nigg, E. A. (2004). Aurora kinases link chromosome segregation and cell division to cancer susceptibility. *Curr. Opin. Genet., Dev.* **14,** 29-36.

Miyoshi, Y., Iwao, K., Egawa, C. & Noguchi, S. (2001). Association of centrosomal kinase STK15/BTAK mRNA expression with chromosomal instability in human breast cancers.*Int. J. of Cancer*. **92,** 370-373.

Moriarty, K. J., Koblish, H. K., Garrabrant, T., Maisuria, J., Khalil, E., Ali, F., Petrounia, I. P., Crysler, C. S., Maroney, A. C., Johnson, D. L. & Galemmo, R. A. (2006). The synthesis and SAR of 2-amino-pyrrolo[2,3-d]pyrimidines: a new class of Aurora-A kinase inhibitors. *Bioorg. and Med. Chem. Lett.* **16**, 5778-5783.

Mortlock, A. A., Keen, N. J., Jung, F. H., Heron, N. M., Foote, K. M., Wilkinson, R. W. & Green, S. (2005). Progress in the development of selective inhibitors of Aurora kinases. *Curr. Top. Med. Chem.* **5,** 807-821.

Nissink, J. W., Murray, C., Hartshorn, M., Verdonk, M. L., Cole, J. C. & Taylor, R. (2002). A new test set for validating predictions of protein-ligand interaction. *Proteins*. **49**, 457-471.

Nowakowski, J., Cronin, C. N., McRee, D. E., Knuth, M. W., Nelson, C. G., Pavletich, N. P., Rogers, J., Sang, B. C., Scheibe, D. N., Swanson, R. V. & Thompson, D. (2002). Structures of the cancer-related Aurora-A, FAK, and EphA2 protein kinases from nanovolume crystallography. *Structure.* **10,** 1659-1667.

Pevarello, P., Fancelli, D., Vulpetti, A., Amici, R., Villa, M., Pittalà, V., Vianello, P., Cameron, A., Ciomei, M., Mercurio, C., Bischoff, J. R., Roletto, F., Varasi, M. & Brasca, M. G. (2006). 3-Amino-1,4,5,6-tetrahydropyrrolo[3,4-c]pyrazoles: a new class of CDK2 inhibitors. *Bioorg. and Med. Chem. Lett.* **16***,* 1084–1090.

Sasai, K., Katayama, H., Stenoien, D. L., Fujii, S., Honda, R., Kimura, M., Okano, Y., Tatsuka, M., Suzuki, F., Nigg, E. A., Earnshaw, W. C., Brinkley, W. R. & Sen, S. (2004). Aurora-C kinase is a novel chromosomal passenger protein that can complement Aurora-B kinase function in mitotic cells. *Cell Motil. Cytoskeleton.* **59**, 249-263.

Sen, S., Zhou, H. & White, R. A. (1997). A putative serine/threonine kinase encoding gene BTAK on chromosome 20q13 is amplified and overexpressed in human breast cancer cell lines. *Oncogene*. **14,** 2195-2200.

SILVER 1.1.1, CCDC, UK.

Smellie, A., Teig, S. L. & Towbin, P. P. (1995). Promoting conformational coverage. *J. Comput. Chem.* **16,** 171-187.

Tanaka, T., Kimura, M., Matsunaga, K., Fukada, D., Mori, H. & Okano, Y. (1999). Centrosomal kinase AIK1 is overexpressed in invasive ductal carcinoma of the breast. *Cancer Res.* **59,** 2041-2044.

Tanner, M. M., Grenman, S., Koul, A., Johannsson, O., Meltzer, P., Pejovic, Borg, A. & Isola, J. J. (2000). Frequent amplification of chromosomal region 20q12-q13 in ovarian cancer. *Clinical Cancer Research*. **6,** 1833-1839.

Warner, S. L., Bearss, D. J., Han, H. & Von Hoff, D. D. (2003). Targeting Aurora-2 kinase in cancer. *Mol. Cancer Ther.* **2,** 589-595.

Zhou, H., Kuang, J., Zhong, L., Kuo, W. L., Gray, J. W., Sahin, A., Brinkley, B. R. & Sen, S. (1998). Tumour amplified kinase STK15/BTAK induces centrosome amplification, aneuploidy and transformation. *Nature Genetics*. **20**, 189-193.

CHAPTER 5

Comparative Studies of the ADAM and ADAMTS Protein Family Members in Human, Frog, Fly and Worm Genomes: A Bioinformatics Approach

5.1 Introduction

ADAM (A Disintegrin And Metalloproteinase) family proteins are characterized by the presence of both disintegrin and metalloproteinase domains. Another closely related protein family is the ADAMTS (A Disintegrin And Metalloproteinase with ThromboSpondin motifs). Several members of these two protein families are known in humans. In the recent years, a number of publications report the importance of ADAMs and ADAMTSs in diseases such as prostate, breast and non-small-cell lung cancers, arthritis and alzheimer. These proteins are also of great physiological importance in regulating cell-cell and cell-matrix interactions. Keeping in view the variety of functions regulated by these enzymes in humans, and the completed genome sequencing of several organisms, we intended to analyse the ADAM and ADAMTS protein families in the completed genomes of mammals- *Homo sapiens* (human), amphibian- *Xenopus laevis* (frog), arthropoda *Drosophila melanogaster* (fly) and nematode- *Caenorhabitis elegans* (worm). These four genomes represent all the organisms according to phylogeny, in which the ADAM and ADAMTS protein families are present.

ADAMs and ADAMTSs are multi-domain protein families and play multiple roles in cell signaling, cell fusion and cell-cell interactions. Members of ADAM family comprise a C- terminal transmembrane segment and are therefore cell surface proteins. At the N- terminus, these proteins comprise a propeptide domain that contains a sequence motif similar to the "cysteine switch" of the matrixins. A zinc dependent proteinase domain follows this region, the proteolytic activity of ADAMs is due to the zinc proteinase domain and its activity is directed towards the extracellular domains of the transmembrane proteins (Wolfsberg & White, 1996). The mechanism of activation of these proteins involves the cleavage

of the pro-part followed by conformational changes in the protein. Activation of metalloproteinases is an additional important mechanism for regulating activity of these enzymes. The adjacent disintegrin domain is responsible for the adhesive properties of the protein, thus mediating cell-cell and cell-matrix interactions. This is followed by the cysteine rich domain that supports cell adhesion. In addition to these four domains, at the C-terminus, ADAMs also comprises of an EGF domain, transmembrane segment and cytoplasmic tail responsible for signaling.

ADAMTS protein family members are soluble, multi-domain proteins. The first two domains in ADAMTS are similar to that in ADAMs, i.e. they comprise a pro-domain and zinc dependent proteinase domain. In addition they also comprise the thrombospondin repeats. Thrombospondin repeats are 50 amino acid residues long and are responsible for binding to extracellular matrix ligands including fibrinogen, fibronectin, some collagens, latent and active transforming growth factor-beta-1, TSG6, heparin, plasmin, cathepsin G, neutrophil elastase, some MMPs, tissue factor pathway inhibitor and heparan sulfate proteoglycans.

The zinc proteinase domain is also common to matrix metalloproteinases (MMPs) superfamily, known as the "metzincins" (Bode *et al.,* 1993). The name is derived from consensus sequence and structural features involving a methionine residue, which forms a conserved structure known as the "metturn" and "zinc binding motif" (zincin). The essential components necessary for the catalytic proteinase mechanism are a glutamic acid, three histidine residues, and a water molecule (Skiles *et al.,* 2001; Lovejoy *et al.,* 1994). The key amino acid residues are arranged in a highly characteristic sequence; HExxHxxGxxH forming a zinc binding motif. The triad of histidine residues co-ordinates the zinc ion, which in combination with the glutamic acid forms the critical sequence components of the catalytic mechanism in the proteins of MMPs, ADAM and ADAMTS families. The first two histidines are separated by a single turn of a helix, and allow the side chain

imidazole ring to be positioned toward the catalytic zinc. The carboxylate group of glutamic acid residue acts as a neutrophile, with which an associated water molecule promotes cleavage of the substrate peptide scissile bond. The conserved glycine residue in the sequence motif allows a sharp turn, permitting the third histidine in the triad to associate with the zinc ion. On the C-terminal to the zinc ion binding motif is the conserved methionine residue which is responsible for the "metturn" in metzincin structures and provides a hydrophobic base for the histidine triad (Bode *et al.,* 1993). In addition to zinc, these enzymes also require calcium for stabilization of the protein tertiary structure.

The identification of these enzymes under normal physiological and disease states has led to the functional characterization. For example, ADAM12-S stimulates bone growth by modulating chondrocyte proliferation and maturation through mechanisms probably involving both metalloproteinase and adhesion activities (Kveiborg *et al.,* 2006). ADAM33 has been identified as a susceptible gene for asthama (Holgate *et al.,* 2006). ADAM19 has a constitutive alpha-secretase activity for amyloid precursor protein (Tanabe *et al.,* 2007) and it has been suggested that ADAM19 may have a modulatory role in the dysfunctional renal allograft state (Melenhorst *et al.,* 2006). ADAM12 and ADAMTS1 are implicated in non-small-cell lung cancer (Rocks *et al.,* 2006). ADAM23 is frequently silenced in gastric cancers by homozygous deletion or aberrant promoter methylation (Takada *et al.,* 2005). ADAM15 generally overexpressed in adenocarcinoma is associated with metastatic progression of prostate and breast cancers (Kuefer *et al.,* 2006) and ADAM28 is overexpressed in an activated form in breast carcinoma cells. ADAM29 expression ratio is a novel prognosis indicator in chronic lymphocytic leukemia (Oppezzo *et al.,* 2005). ADAM9 is induced in human prostate cancer cells (Shigemura *et al.,* 2007). ADAMTS8 and ADAMTS15 have emerged as novel predictors of survival in patients with breast carcinoma

(Porter *et al.,* 2002). One of the factors responsible for asthama is ADAM8. ADAM20 and ADAM21 are human testis-specific membrane metalloproteinases and ADAM30 shows testis-specific gene expression (Cerretti *et al.,* 1999). ADAM18 is a sperm surface protein for oocyte recognition (Frayne *et al.,* 2002). ADAM2 is required for sperm egg fusion. ADAM22, a brain-specific cell surface protein, mediates growth inhibition using an integrin dependent pathway. It is expressed in normal brain but not in high grade gliomas (D'Abaco *et al.,* 2006). ADAMTS10 plays a major role in growth, skin, lens and heart development in humans. ADAM17 is TNF-alpha convertase enzyme (TACE) from human arthritis affected cartilage (Solomon *et al.,* 2007). ADAM10 represents an important molecular modulator of FasL-mediated cell death. It has been shown that there is genetic association between polymorphisms in the ADAMTS14 gene and multiple sclerosis (Goertsches *et al.,* 2005). ADAMTS1 expression is associated with decidualization of the endometrial stroma *in vivo* (Ng *et al.,* 2006). ADAMTS8 has a role in brain tumorigenesis. ADAMTS12 is important for the initiation and progression of arthritis. The L1565 variant of von Willebrand factor is susceptible to proteolysis by ADAMTS13 (Davies *et al.,* 2007). ADAMTS5 mainly contributes to ECM (extra cellular matrix) metabolism in growth plate and condylar cartilage during growth. ADAMTS1 and ADAMTS4 may be involved in ECM turn over in articular cartilage (Mitani *et al.,* 2006). The importance of the ADAMs and ADAMTSs in these disease states makes them important drug targets and comparative genome studies of these protein families will help validate the drug targets. These comparative studies will help estimate the divergence of these protein families during evolution as required for their adaptation.

5.2 Methods

5.2.1 Search for ADAM and ADAMTS in the human, frog, fly and worm genomes:

Sequences encoded by the ADAM and ADAMTS were obtained from protein database at NCBI (www.ncbi.nlm.nih.gov/). Preliminary searches for ADAM and ADAMTS were performed individually using BLAST (Altschul *et al.,* 1990). Sequences belonging to these families with greater than 30% identities among themselves were considered as queries. Reciprocal searches were carried out with each of the protein sequence using PSI-BLAST (Altschul *et al.,* 1997), till no new proteins were identified. Non redundant (nr) protein sequence database was chosen. BLOSUM62 matrix, with existence 11 and extension 1 as gap penalties, Expect threshold 10 and PSI-BLAST threshold 0.005 was chosen for all the PSI-BLAST searches. To identify human proteins, the Organism option in PSI-BLAST was chosen as *Homo sapiens* (taxid: 9606). Similarly, for proteins from frog, fly and worm, the corresponding Organism options were; *Xenopus laevis* (taxid: 8355), *Drosophila melanogaster* (taxid: 7227) and *Caenorhabditis elegans* (taxid:6239). Thus, the mammalian genome, *Homo sapiens*, the amphibian genome *Xenopus laevis*, the fly genome *Drosphila melanogaster* and the nematode genome *Caenorhabditis elegans* were analysed using PSI-BLAST searches to identify all ADAMs and ADAMTSs. The PSI BLAST hits were further scanned using the online SMART database (Schultz *et al.,* 1998, Letunic *et al.,* 2006) in the batch mode and INTERPRO database (Mulder & Apweiler, 2007). The E values from the BLAST output and the SMART or INTERPRO annotation of each BLAST hit were verified before including a protein into the superfamily. The hits obtained from these methods were merged together after removing the redundant proteins.

The proteins present as fragments or identical to larger proteins were also discarded from the dataset. The hits sharing 100% sequence identity with other proteins were considered as redundant and hence only the representative sequence that had the longest amino acid sequence length was included in further steps of analysis.

5.2.2 Multiple sequence alignment and phylogenetic tree analysis:

Multiple sequence alignment of the ADAM and ADAMTS proteins from human, frog, fly and worm was performed using CLUSTALW (Thompson *et al.,* 1994). BLOSUM62 matrix, an open gap penalty of 10 and an extension penalty of 0.05 were the parameters employed for multiple sequence alignment. For generating the phylogenetic trees, Bootstrapping was performed 1000 times to obtain support values for each internal branch. Pairwise distances were determined with protpars protein parsimony method (Felsenstein, 1996). Representations of the calculated trees were constructed using TreeView (Page, 1996). The multiple sequence alignments and the phylogenetic trees were constructed separately for the full length proteins as well as the amino acid sequence region corresponding to the zinc proteinase domain alone.

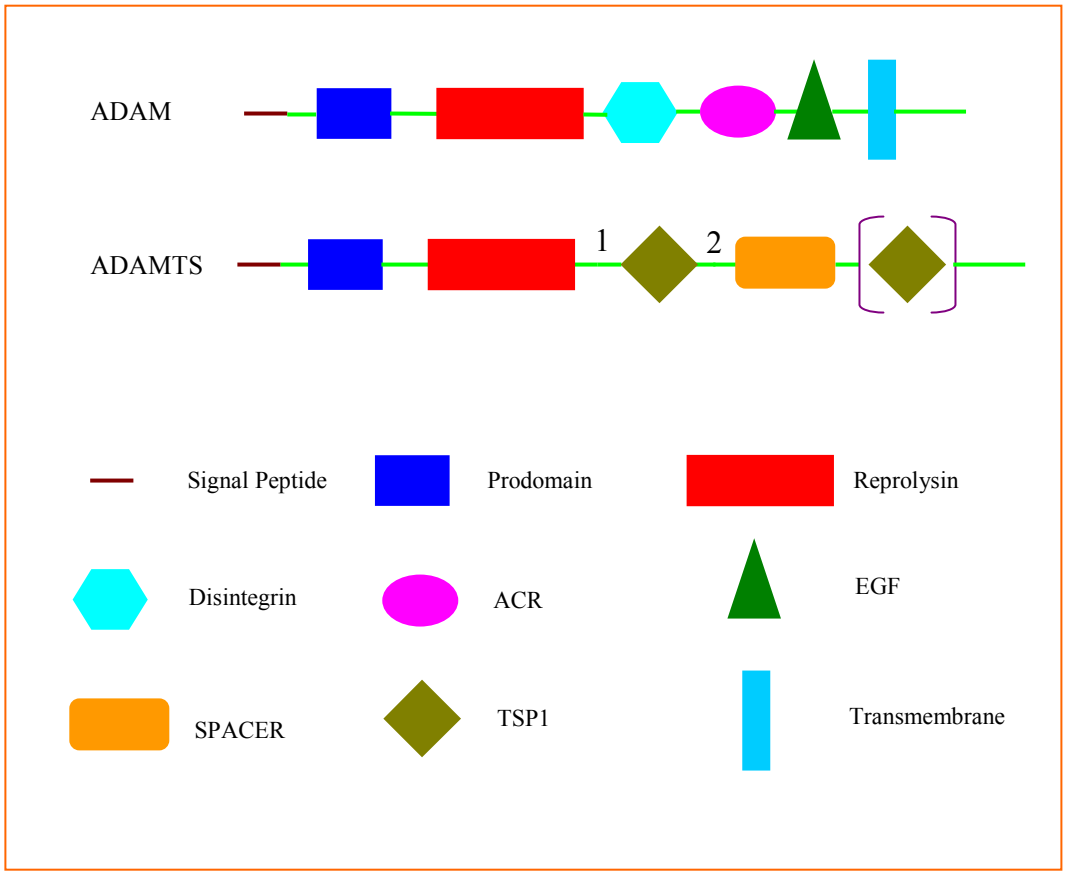
5.3 Results and Discussion

The genome sizes of human, fly and worm correspond to 3100 Mb, 180 Mb and 100 Mb, respectively and are encoded by corresponding number of genes (28920, 19778, and >20,000 [www.ncbi.nlm.nih.gov/]). The frog genome, *Xenopus laevis* is an ongoing project at DDBJ (www.ddbj.nig.ac.jp) with a 3100 Mb genome size. PSI-BLAST searches identified 235 proteins belonging to the ADAM and ADAMTS protein families in human proteome and after removing redundancy, there were 182 proteins. Among these, 90 are ADAMs and 92 are ADAMTS proteins. Similarly, the frog proteome is encoded by 11 ADAMs and 2 ADAMTS proteins. The fly proteome is encoded by 19 ADAM and 6 ADAMTS proteins and the worm genome is encode by 7 ADAMs and 8 ADAMTS proteins. These figures indicate that the relative ratio of the numbers ADAM and ADAMTS are comparable in human and worm, while the relative ratio of the numbers of ADAMTS are far fewer compared to ADAMs in frog and fly. A list of these proteins is provided as Appendix-1. The above mentioned proteins are unique and distinct and when present as isoforms share high sequence homology but are not identical. We discuss below the occurrence, domain organization and phylogenetic analysis of these two protein families in four distinctly different representative organisms.

5.3.1 Domain organization of ADAM and ADAMTS:

The ADAM family members comprise variable sequence length, between 375 and 1538 amino acid residues and the ADAMTS family members also comprise a variable sequence length, between 344 and 2165 amino acid residues. The domain organization of these two protein families have been analysed using the SMART and INTERPRO database. Most ADAMs follow a similar domain architecture pattern as shown in Figure 5.1. A typical ADAM comprises pro-domain, zinc dependent metalloproteinase, Disintegrin, ACR region followed by the EGF domain and a transmembrane segment at the C-terminus that makes ADAMs membrane bound. These domains are followed by a C-terminal cytoplasmic tail. While the order of domains remains unchanged in all the proteins, depending upon their amino acid sequence length, some of the domains either from N- or C- terminus are absent. In some instances, SMART did not identify some domains corresponding to a region and we attribute this to the low sequence homology shared in these regions.

Figure 5.1. A Representative Domain Organization Diagram of the ADAM and ADAMTS Protein Family Members.



Note: The regions 1 and 2 indicated in the figure are specifically present in ADAMTS and comprise cysteine residues at conserved positions.

At the N-terminus, the soluble ADAMTS proteins comprise pro-domain and zinc dependent metalloproteinase. These two domains are followed by region that varies in length between 90-100 amino acid residues. This region identifies the ACR region of ADAMs with low sequence homology (from the second iteration of PSI-BLAST searches). Also, this region is characterized by the presence of

conserved cysteine residues at equivalent positions in all the ADAMTS proteins. This region is followed by a single TSP1 repeat, which is followed by amino acid sequence region that comprises 100-110 amino acid residues. PSI-BLAST searches identified this region to be unique to ADAMTSs and only in few very instances, ACR domain of ADAM is identified with very low sequence homology (from the third iteration of PSI-BLAST searches). Further, this region is also characterized by the presence of conserved cysteine residues at equivalent positions in all the ADAMTS proteins. This region is followed by ADAM_spacer1 domain comprising 110-120 amino acid residues and TSP1 repeats. Typically, the number of TSP1 repeats varies from 0 to 17 according to SMART. The number of TSP1 repeats depends on the length of the protein. Similar domain architecture pattern is present in ADAMTSs and is shown in Figure 5.1. Few exceptions to the standard domain architecture of ADAMs and ADAMTSs have been noticed. Some notable exceptions are in the drosophila proteins. For example, the protein NP_996218 has an N-terminal 165 amino acid residue insertion. Also proteins, AAC47275, ABV53679, ABV53680, AAQ22412, NP_001014481, NP_651716 and AAS48649 have specific inserts that vary in length between 155 and 264 amino acid residues. In both the cases, the insert regions are restricted to proteins from drosophila alone and may be responsible for mediating specific interactions required by the organism.

5.3.2 Multiple sequence alignment:

The multiple sequence alignment of entire ADAMs and ADAMTSs was generated using CLUSTALW. According to CLUSTALW, the sequence homology between ADAMs and ADAMTSs across the four genomes analysed is as low as 1%. We explain that, this is partly due to the low sequence homology shared between similar domains in the proteins and mainly due to the significant variation

in the lengths of protein sequences. Therefore, we propose that in spite of low sequence homology, the proteins that belong to either ADAM or ADAMTS protein family are highly similar in terms of 3-D structure and function. We observe that the members of ADAM and ADAMTS are distinctly different and share similar domains only at the N-terminus (pro-domain and metalloproteinase domains).

The zinc binding catalytic region represented by REPRO domain mostly varies in length between 210 and 225 amino acid residues, and is responsible for proteolytic activity. The zinc binding motif HExxHxxGxxH is present in most ADAMs and ADAMTS with the exception of few ADAMs in which mutations have been observed.

5.3.3 Phylogenetic analysis:

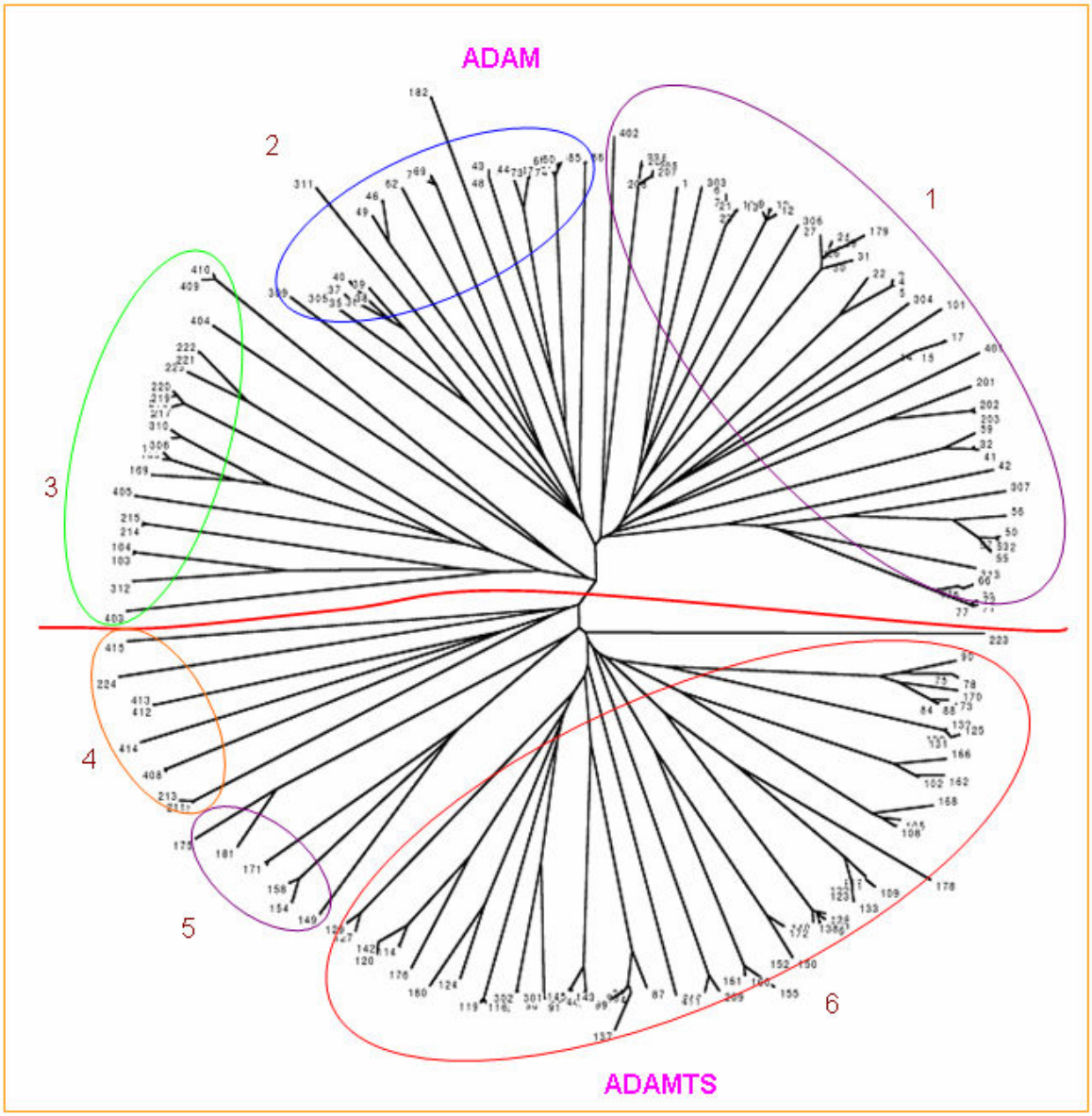
The phylogenetic trees were constructed for proteins belonging to the four representative organisms. To observe the differences between the ADAM and ADAMTS full length proteins, and their corresponding metalloproteinase domain alone, we have built two phylogenetic trees. One, using full length proteins and the second, is corresponding to their domain sequence regions alone. Figure 5.2a indicates the tree generated for the full length ADAM and ADAMTS members, Figure 5.2b indicates the tree generated for the region corresponding to the zinc dependent metalloproteinase domain for ADAM and ADAMTS members. From the examination of the phylogenetic trees shown in Figure 5.2a and 5.2b, it is obvious that the members of ADAM and ADAMTS fall into two distinct clusters. We also observe that there are some distinct differences between the trees built for full length ADAMs and ADAMTSs and the corresponding metalloproteinase domain alone. The proteins can be organized into 6 Clades according to the phylogenetic tree of full length ADAM and ADAMTS proteins. These are termed as Clades 1 to 6. Clades 1, 2 and 3 comprise ADAMs and Clades 4, 5 and 6

comprise ADAMTS. The phylogenetic tree corresponding to the metalloproteinase domain is organized into 9 Clades and these are termed as Clades I to IX. The Clades I to V comprise ADAMs and Clades VI to IX comprise ADAMTSs. The increased number of Clades in this phylogenetic tree implies that the degree of divergence during evolution is greater within the metalloproteinase domain than compared to the full length ADAMs and ADAMTSs.

5.3.3.1 Clade1:

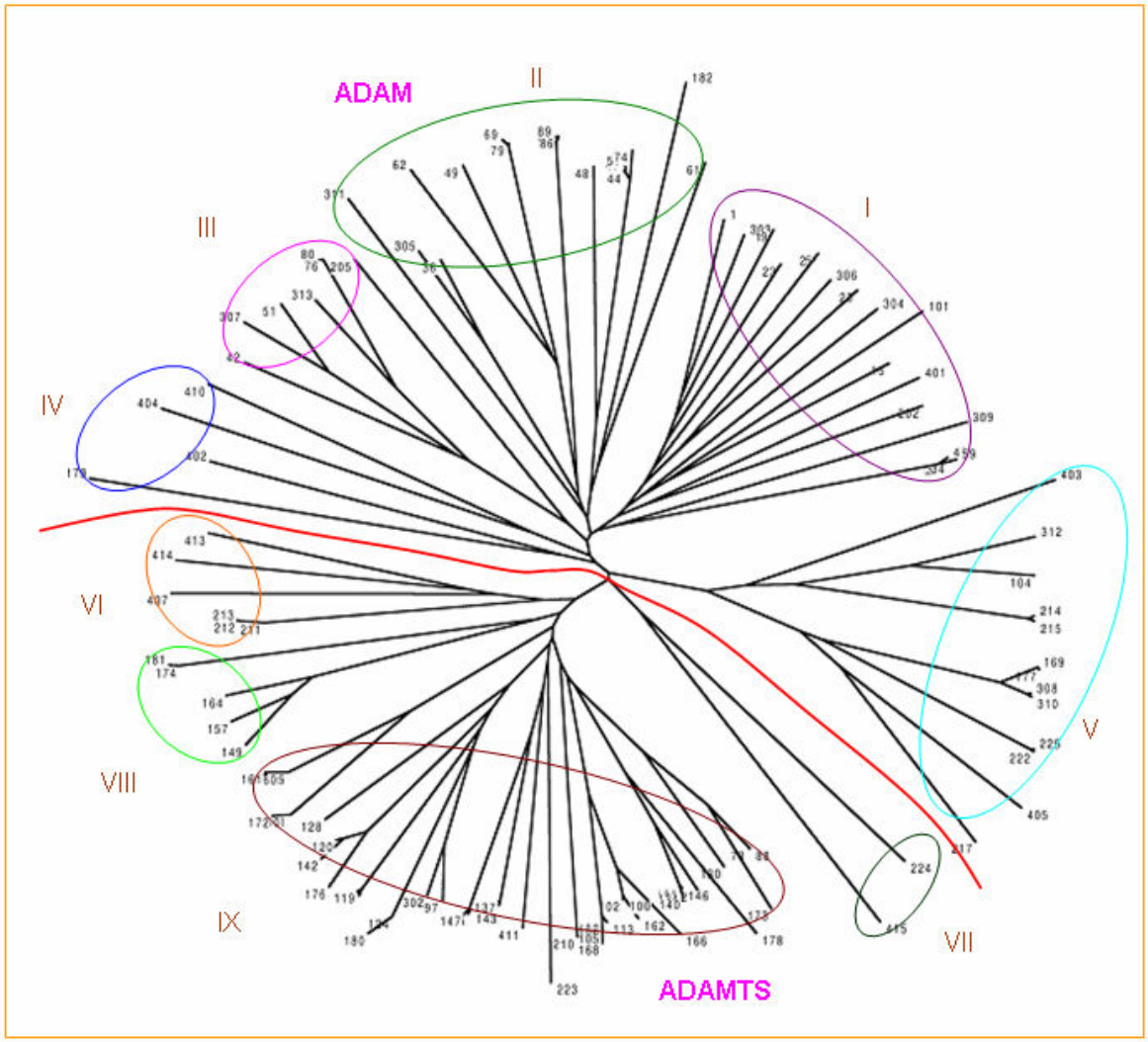
The ADAMs that belong to this Clade are 11, 22, 23, 7, neu3, 28, 8, 15, 19, 12, 13, 33, mind meld and ADAM decysin. This Clade constitutes proteins from all four organisms. ADAM decysin is a secreted protein belonging to the disintegrin metalloproteinase family and its expression is upregulated during dendritic cells maturation (Mueller, 1997). ADAMs in drosophila are encoded by 3 isoforms of neu3 genes. Using whole-genome microarray assays, Stathopoulos *et al.,* 2002 have showed that neu3 is expressed in broad lateral stripes in wild-type embryos, but is expressed throughout the dorsal-ventral axis of mutant embryos derived from *Tollrm9/Tollrm10* females. Another ADAM gene in drosophila, mind meld, is neuronally expressed. Neu3 and mind meld are restricted to drosophila alone. ADAM13 homolog is restricted to frog alone, while ADAM11, 22 and 15 homologs are present in both humans and frog. Two members of worm ADAM11 and ADAM22 belong to this Clade. ADAM19, 15, 8 and 12 are closely related members. ADAM28 types of proteins are lymphocyte-expressed proteins and their alternative splicing results in two transcript variants. The shorter version encodes a secreted isoform, while the longer version encodes a transmembrane isoform. ADAM33 proteins are implicated in asthma and bronchial hyper-responsiveness. ADAM19 proteins serve as a marker for dendritic cell differentiation.

Figure 5.2a. The Phylogenetic Tree Generated for the Full Length ADAM and ADAMTS Protein Family Members. Clades 1 to 6 are Indicated. Red Horizontal Line Across the Figure Divides ADAM and ADAMTS Proteins.



ADAM8 proteins may be involved in cell adhesion during neurodegeneration. ADAM15 family members are type I transmembrane glycoproteins known to be involved in cell adhesion and proteolytic ectodomain processing of cytokines and adhesion molecules. Through its disintegrin-like domain, these proteins specifically interact with the integrin beta 3 chain. It also interacts with Src family protein-tyrosine kinases in a phosphorylation dependent manner, suggesting that these proteins may function in cell-cell adhesion as well as in cellular signaling. ADAM7 family is composed of zinc-binding proteins that can function as adhesion proteins and/or endopeptidases. They are involved in a number of biological processes, including fertilization, neurogenesis, muscle development, and immune response. ADAM11 genes represent candidate tumor supressor genes for human breast cancer based on its location within a minimal region of chromosome 17q21 defined by tumor deletion mapping. In frog, this acts as a probable ligand for integrin in the brain and it can be detected in testis and barely expressed in heart and muscle. In developing embryos, this expression is restricted to neural crest derivatives. ADAM12 has an important role to play in myoblast differentiation. ADAM22 and 23 family members are highly expressed in the brain and may function as an integrin ligand in the brain.

Figure 5.2b. The Phylogenetic Tree Generated for the Region Corresponding to the Zinc Dependent Metalloproteinase Domain Region for ADAMs and ADAMTSs Protein Family Members. Clades I to IX are Indicated. Red Line Across the Figure Divides ADAM and ADAMTS Proteins.



The members of Clade1 are divided into Clades I, III and as some members of Clade IV. The members of Clade I are present in the four representative organisms and are active with the exception of ADAM7. The members of ADAM7 have the mutation HQ (HE) in the zinc binding sequence motif and it is unlikely that the activity is retained. The members of Clade III are present in human, frog and drosophila and are all inactive. The members of this Clade, ADAM11 and 23 families have all three histidines mutated and members of ADAM22 have the histidines, H1 and H3 mutated and these proteins therefore lack the zinc proteinase activity. Further, the members of mind meld proteins from drosophila have the mutation HM (HE) in the zinc binding sequence motif and it is unlikely that the activity is retained. One ADAM15 member of Clade 1 now belongs to Clade IV. This protein has the mutation HL (HE) in the zinc binding sequence motif and it is unlikely that its activity is retained.

5.3.3.2 Clade 2:

The ADAMs that belong to this Clade are ADAM30, 29, 21, 20, 18, 2, 4, 9, 32. These proteins correspond to the members of human and frog. ADAM9 proteins interact with SH3 domain-containing proteins, bind mitotic arrest deficient 2 beta protein, and are also involved in TPA-induced ectodomain shedding of membrane-anchored heparin-binding EGF-like growth factor. The expression of ADAM20 and 21 is testis-specific. ADAM29 proteins encoded genes are highly expressed in testis and may be involved in human spermatogenesis. ADAM24 and 25 are expressed exclusively in testis and more specifically on the surface of mature sperm. ADAM30 is testis-specific and contains a polymorphic region, resulting in isoforms with varying numbers of C-terminal repeats. ADAM3 and 5 are germ-cell specific that may play a role in cell-cell and cell-matrix interactions during spermatogenesis. ADAM18 is a sperm surface protein. ADAM32 is similar

to ADAM2. ADAM18 is expressed on cell-cell and cell-matrix interactions, including fertilization, muscle development, and neurogenesis. ADAM1 and 2 are expressed on the plasma membrane of developing spermatogenic cells and sperm (this member is a subunit of an integral sperm membrane glycoprotein called fertilin, which plays an important role in sperm-egg interactions). ADAM4 and 6 have high similarity to human metargidin and it may participate in dual proteolysis and integrin-mediated cell-cell, cell-matrix interactions. We therefore conclude that all members of this Clade are important for mediating cell-cell and cell-matrix interactions thus play a major role in fertilization.

The members of ADAM2, 18 and 32 have all three histidines mutated and members of ADAM21 have the histidine, H2 mutated and these proteins therefore lack the zinc proteinase activity. The members of ADAM29 and a member of ADAM4 have the mutation HH/N (HE) in the zinc binding sequence motif respectively and this makes them unlikely to function as a proteinase. All members of Clade 2 are present in Clade II.

5.3.3.3 Clade 3:

The ADAMs that belong to this Clade are ADAM10, 17. Members of this Clade are present in all four organisms. Members of ADAM10 cleave many proteins including TNF-alpha, E-cadherin and myelin basic protein. The homolog of ADAM10 in drosophila is Kuz and is required for proper development of peripheral and central nervous system. Kuz is also required for axon extension, vein formation and wing margin formation and is involved in NOTCH mediated lateral inhibition in drosophila. Any mutations in Kuz result in production of excessive number of neurons and bristles in the central nervous system. Kuz homolog in worm is SUP-17 and is also known to modify Notch-mediated cell fare decisions in the organism. Recently it has been shown that ADAM10 is required

for the formation of optic projection by xenopus retinal ganglion cell (RGC) axons and its mRNA is expressed in the dorsal neuroepithelium through which RGC axons extend (Chen *et al.,* 2007). In Humans, a similar role for ADAM10 in the vertebrate development is speculated. All members of ADAM10 are active.

ADAM17 proteins function as a TACE and are important for the normal release of soluble TNF in humans. Their homologs are also present drosophila, frog and worm and are active with the exception of two members from worm, that have the mutation HQ (HE) in the zinc binding sequence motif and it is unlikely that the zinc metalloproteinase activity is retained. Members of Clade 3 are present in Clade V and some members in Clade IV. The members present in Clade IV are four ADAMs from worm, among which only one represents an active metalloproteinase. The members of Clade V are present in all organisms and are active proteins.

5.3.3.4 Clade 4:

This Clade constitutes ADAMTS proteins from drosophila and worm. These proteins do not belong to any known ADAMTS types of proteins, but are described as ADAMTS like subgroup members. These proteins belong to angiogenesis inhibitor homologs, ADT-2, MIG-17 and T19D2. The MIG-17 protein is secreted from muscle cells of the body wall and localizes in the basement membranes of gonad. MIG-17 is essential for its function in controlling distal tip cells (DTC) migration. The MIG-17 is expressed by the body wall muscles and then localizes to the DTCs where its activity is sufficient for guiding DTC migration. This expression is initially seen on the pseudocoelomic face of body wall muscles and then on the surface of the gonad. All members of this Clade are active. Members of Clade 4 are distributed in Clades VI and VII. The members in

this Clade are AAF46905, NP_611718, AAL90078, BAC41253, T18856 and NP_508681. The MIG-17 members in drosophila and worm belong to Clade VII.

Table 5.1. List of ADAM and ADAMTS Protein Family Members with their NCBI_IDs from Human, Drosophila, Frog and Worm.

Human				Drosophila	Frog	Worm
1.NP_079496	47.NP_003804	93.Q9P2N4	139.AAH63283	201.NP_001027575	301.BAE94917	401.NP_510291
2.NP_001100	48.O43506	94.NP_891550	140.Q8TE60	202.NP_001027576	302.NP_001088627	402.NP_499680
3.EAW61326	49.XP_001131733	95.CAI46043	141.EAW57863	203.AAO39439	303.NP_001080914	403.NP_509318
4.AAI15405	50.EAW76927	96.AAF15317	142.AAI13875	204.AAF48548	304.AAH91726	404..NP_509031
5.P78325	51.AAF73288	97.NP_008919	143.P59510	205.AAS65376	305.NP_001079073	405.NP_492377
6.O43184	52.NP_068368	98.AAH36515	144.CAD56160	206.NP_523358	306.AAI46627	406.NP_001024534
7.NP_003465	53.NP_068369	99.BAA92550	145.NP_079279	207.NP_996475	307.NP_001080913	407.BAC41253
8.Q9H013	54.AAD55251	100.NP_922932	146.CAC83612	208.AAF98331	308.AAH77950	408.NP_001024532
9.NP_075525	55.EAW76919	101.NP_055294	147.CAD56159	209.AAM50192	309.NP_001079284	409.T16189
10.EAW61597	56.EAW76929	102.AAW47397	148.O15072	210.NP_572247	310.NP_001083912	410.NP_509295
11.CAC20585	57.AAF22476	103.NP_003174	149.NP_055058	211.AAF46905	311.NP_001089707	411.NP_501792
12.NP_150377	58.AAC52042	104.AAC39721	150.NP_620688	212.NP_611718	312.NP_001089130	412.NP_510116
13.EAW61595	59.AAH43207	105.EAW68908	151.ABB70740	213.AAL90078	313.Q9PSZ3	413.T18856
14.AAD25100	60.NP_055084	106.NP_112219	152.Q8TE56	214.AAS48650		414.NP_508681
15.NP_068547	61.AAF03777	107.EAW68907	153.EAW53816	215.NP_733334		415.NP_505901
16.NP_055080	62.NP_055052	108.Q9H324	154.NP_067610	216.AAC47275		
17.CAB42085	63.NP_001455	109.ABB70405	155.EAW62382	217.ABV53679		
18.Q9UKQ2	64.AAD04206	110.CAC86015	156.O95450	218.ABV53680		
19.AAC08703	65.AAF03779	111.EAX08119	157.EAW53815	219.AAQ22412		
20.NP_067673	66.EAW51579	112.Q9UP79	158.NP_055059	220.NP_001014481		
21.AAH60804	67.AAF03778	113.AAG35563	159.NP_598377	221.NP_651716		
22.EAW61325	68.AAF22163	114.NP_008968	160.Q8TE59	222.AAS48649		
23.BAF84318	69.AAC51110	115.BAF84262	161.EAW62383	223.NP_996218		
24.BAD92394	70.EAW51578	116.AAQ89245	162.AAW47398	224.NP_611827		
25.NP_997080	71.EAW51580	117.AAL02262	163.NP_631894	225.XP_001357772		
26.AAS48595	72.NP_002381	118.O75173	164.AAL79814			
27.AAS48597	73.ABA43715	119.NP_005090	165.NP_001101			
28.AAC50404	74.EAW81029	120.AAH89435	166.Q9UKP5			
29.AAC51112	75.EAW99151	121.BAA31663	167.CAC87943			
30.NP_997074	76.BAA06671	122.Q8TE57	168.BAD92752			
31.AAS48591	77.BAA06670	123.NP_620687	169.CAA88463			
32.NP_003808	78.Q9UKP4	124.NP_620686	170.EAW99150			
33.Q9H2U9	79.CAA67753	125.AAI31734	171.Q8WXS8			
34.EAW63606	80.AAB29191	126.AAH63293	172.BAD18500			
35.NP_003807	81.AAF03780	127.EAX09948	173.EAW99148			
36.EAW63283	82.EAW99149	128.NP_008969	174.NP_620594			
37.BAA03499	83.AAQ94616	129.EAX09949	175.AAL17652			
38.NP_001005845	84.NP_055087	130.P58397	176.EAW67784			
39.AAI26407	85.NP_068566	131.NP_112217	177.AAI26254			
40.AAB46867	86.AAH28372	132.EAX10809	178.EAX08118			
41.AAC36742	87.AAF89106	133.BAC23125	179.EAW53144			
42.NP_003803	88.AAH61631	134.EAW95600	180.EAW67787			
43.NP_003805	89.AAF03781	135.EAW95599	181.CAD12730			
44.EAW81035	90.BAD92734	136.EAW95598	182.EAW81036			
45.AAH26085	91.BAA92584	137.EAW65434				
46.NP_659441	92.AAO15765	138.NP_955387				

5.3.3.5 Clade 5:

The ADAMTSs that belong to this Clade are ADAMTS2, 3, 13, 14. This Clade constitutes proteins only from the human genome. ADAMTS14 mainly expressed in lung, is highly similar to ADAMTS2 and 3, and possess the aminoprocollagen peptidase activity. The ADAMTS2 gene provides instructions for making an enzyme that processes several types of procollagen molecules. Procollagens are the precursors of collagens, which are complex molecules that add strength, support, and elasticity to many body tissues. Specifically, the ADAMTS2 enzyme clips a short chain of protein building blocks off one end of procollagens. This clipping step is necessary for the resulting collagen molecules to assemble into strong, slender fibrils outside cells. ADAMTS13 is a plasma metalloproteinase that cleaves von Willebrand factor to smaller, less thrombogenic forms. This protein is mainly expressed in liver. All members of this Clade are active. Members of Clade 5 are present in Clade VIII.

5.3.3.6 Clade 6:

The ADAMTSs that belong to this Clade are ADAMTS7, 9, 1, 12, 6, 10, 16, 18, 19, 17, 20, 4, 15, 8 and 5. Members of this Clade are present in all four organisms. ADAMTS12 is an important enzyme that causes cartilage degradation in arthritic disorders and plays important role in the development and progression of inflammatory and tumor processes. ADAMTS7 exhibited higher expression in musculoskeletal tissues and its concentration was found to be up-regulated in the cartilage and synovium of patients with rheumatoid arthritis. ADAMTS10 and 6 have a role in inflammatory eye disease. ADAMTS10 plays a vital role in growth and skin, lens and heart development. ADAMTS16 is specific to human and highly expressed both in the kidney and in the ovary, where they are predominantly expressed in the parietal granulose cells of pre-ovulatory follicles but only slightly

expressed in cells of the cumulus oophorus. ADAMTS16 is capable of cleaving a2-macroglobulin MG, a common substrate for proteinases, which is present at high concentrations in the follicular fluid of ovarian follicles. ADAMTS16 plays a physiological role of ovarian follicles, during the pre-ovulatory phase. ADAMTS18 functions as a tumor suppressor. ADAMTS1 has anti-angiogenic activity. The expression of these proteins is associated with various inflammatory processes as well as development of cancer cachexia. ADAMTS20 protein is also overexpressed in some human malignant tumors, including brain, colon and breast carcinomas. ADAMTS15 is very similar to ADAMTS1 and 8. ADAMTS4 and 8 are inflammatory regulated enzymes expressed in macrophage-rich areas of atherosclerotic plaques and the ADAMTS5 proteins cleave the aggrecan interglobular domain. ADAMTS9 is a secreted, cell-surface-binding metalloproteinase that cleaves the proteoglycans, versican and aggrecan. ADAMTS17 is mainly expressed in fetal tissues, especially in lung, ADAMTS19 is in kidney. ADAMTS19 is virtually undetectable in adult tissues, suggesting that this functional role is specifically restricted to processes occurring during human fetal development. The members of this Clade appear to participate in vital roles in the developmental biology of various organs in the four representative organs. One member of ADAMTS16 (EAX08118) has the histidine, H3 mutated and this protein will therefore not have the zinc proteinase activity. We report that this is the only inactive ADAMTS in the dataset analysed. Members of Clade 6 are present in Clade IX.

5.4 Conclusions

1. In this work, we have presented a comprehensive study and bioinformatics overview of the ADAM and ADAMTS protein family members in the representative *Homo sapiens*, *Xenopus laevis*, *Drosophila melanogaster* and *Caenorhabitis elegans* genomes.
2. Classification of these proteins on the basis of domain architecture and cellular localization helps us to associate the proteins to the various biochemical pathways and the different cellular niches.
3. Rearrangement and insertion of domains among the various ADAM and ADAMTS seems to be required for adapting to diverse biological roles.
4. We identified a different domain architecture pattern in ADAMTS protein family which is not as similar as to the previous report (Claudia *et al.,* 2005).
5. Such whole genome surveys and cross-genome comparisons using computations should be useful to design rational experiments and enhance our understanding of the specific biological roles of the ADAM and ADAMTS family proteins.

5.5 References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). "Basic local alignment search tool." *J. Mol. Biol.* **215,** 403-410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*. **25**, 3389-3402.

Bode, W., Gomis-Ruth, F. X. & Stockler, W. (1993). Astacins, serralysins, snake venom and matrix metalloproteinases exhibit identical zinc-binding environments (HEXXHXXGXXH and Met-turn) and topologies and should be grouped into a common family, the 'metzincins'. *FEBS Lett.* **331**, 134–40.

Cerretti, D. P., DuBose, R. F., Black, R. A. & Nelson, N. (1999). Isolation of two novel metalloproteinase-disintegrin (ADAM) cDNAs that show testis-specific gene expression. *Biochem Biophys Res. Commun.* **3**, 810-5.

Chen, Y. Y., Hehr, C. L., Atkinson-Leadbeater, K., Hocking, J. C. & McFarlane, S. (2007). Targeting of retinal axons requires the metalloproteinase ADAM10. *J Neurosci.* **27**, 8448-56.

Claudia, A., Lucia, B., Ivano, B., Sara, E. & Antonio, R. (2005). Comparative Analysis of the ADAM and ADAMTS Families. *Journal of Proteome Research.* **4,** 881- 888.

Davies, J. A. & Bowen, D. J. (2007). The association between the L1565 variant of von Willebrand factor and susceptibility to proteolysis by ADAMTS13. *Haematologica*. **2**, 240-3.

D'Abaco, G. M., Ng, K., Paradiso, L., Godde, N. J., Kaye, A. & Novak, U. (2006). ADAM22, expressed in normal brain but not in high-grade gliomas, inhibits cellular proliferation via the disintegrin domain. *Neurosurgery.* **1,** 179-86.

Felsenstein, J. (1996). Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.* **266,** 418-27.

Frayne, J., Hurd, E. A. & Hall, L. (2002). Human tMDC III: a sperm protein with a potential role in oocyte recognition. *Mol. Hum. Reprod.* **9**, 817-22.

Goertsches, R., Comabella, M., Navarro, A., Perkal, H. & Montalban, X. (2005). Genetic association between polymorphisms in the ADAMTS14 gene and multiple sclerosis. *J. Neuroimmunol.* **164**, 140-7.

Holgate, S. T., Yang, Y., Haitchi, H. M., Powell, R. M., Holloway, J. W., Yoshisue, H., Pang, Y. Y., Cakebread, J. & Davies, D. E. (2006). The genetics of asthma: ADAM33 as an example of susceptibility. *Gene. Proc. Am. Thorac. Soc.* **5**, 440-443.

Kuefer, R., Day, K. C., Kleer, C. G., Sabel, S., Hofer, M. D., Varambally, S., Zorn, C. S., Chinnaiyan, A. M., Rubin, M. A. & Day, M. L. (2006). ADAM15 disintegrin is associated with aggressive prostate and breast cancer disease. *Neoplasa.* **4,** 319-29.

Kveiborg, M., Albrechtsen, R., Rudkjaer, L., Wen, G. & Damgaard-Pedersen, K. (2006). Wewer UM. ADAM12-S stimulates bone growth in transgenic mice by modulating chondrocyte proliferation and maturation. *J Bone Miner Res.* **8**, 1288-96.

Letunic, I., Copley, R. R., Pils, B., Pinkert, S., Schultz, J. & Bork, P. (2006). SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.* **34,** D257-60.

Lovejoy, B., Hassell, A. M., Luther, M. A., Weigl, D. & Jordan, S. R. (1994). Crystal structures of recombinant 19-kDa human fibroblast collagenase complexed to itself. *Biochemistry*. **33**, 8207–17.

Melenhorst, W. B., van den Heuvel, M. C., Stegeman, C. A., van der Leij, J., Huitema, S., van den Berg, A. & van Goor, H. (2006). Upregulation of ADAM19 in chronic allograft nephropathy. *Am. J. Transplant*. **7**, 1673-81.

Mitani, H., Takahashi, I., Onodera, K., Bae, J. W., Sato, T., Takahashi, N., Sasano, Y., Igarashi, K. & Mitani, H. (2006). Comparison of age-dependent expression of aggrecan and ADAMTSs in mandibular condylar cartilage, tibial growth plate, and articular cartilage in rats. *Histochem Cell Biol.* **3,** 371-80.

Mueller, C. G., Rissoan, M. C., Salinas, B., Ait-Yahia, S., Ravel, O., Bridon, J. M., Briere, F., Lebecque, S. & Liu, Y.J. (1997). Polymerase chain reaction selects a novel disintegrin proteinase from CD40-activated germinal center dendritic cells. *J Exp Med.* **186,** 655-63.

Mulder, N. & Apweiler, R. (2007). InterPro and InterProScan: Tools for Protein Sequence Classification and Comparison. *Methods Mol Biol*. **396**, 59-70.

Ng, Y. H., Zhu, H., Pallen, C. J., Leung, P. C. & MacCalman, C. D. (2006). Differential effects of interleukin-1beta and transforming growth factor-beta1 on the expression of the inflammation-associated protein, ADAMTS-1, in human decidual stromal cells *in vitro*. *Hum. Reprod.* **8,** 1990-9.

Oppezzo, P., Vasconcelos, Y., Settegrana, C., Jeannel, D., Vuillier, F., Legarff-Tavernier, M., Kimura, E. Y., Bechet, S., Dumas, G., Brissard, M., Merle-Béral, H., Yamamoto, M., Dighiero, G., Davi, F. & French Cooperative Group on CLL. (2005). The LPL/ADAM29 expression ratio is a novel prognosis indicator in chronic lymphocytic leukemia. *Blood.* **2,** 650-7.

Page, R. D. (1996). TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **4,** 357-8.

Porter, S., Span, P. N., Sweep, F. C., Tjan-Heijnen, V. C., Pennington, C. J., Pedersen, T. X., Johnsen, M., Lund, L. R., Rømer, J. & Edwards, D. R. (2002). ADAMTS8 and ADAMTS15 expression predicts survival in human breast carcinoma. *Int. J. Cancer.* **5**, 1241-7.

Rocks, N., Paulissen, G., Quesada, C. F., Polette, M., Gueders, M., Munaut, C., Foidart, J. M., Noel, A., Birembaut, P. & Cataldo, D. (2006). Expression of a disintegrin and metalloprotease (ADAM and ADAMTS) enzymes in human non-small-cell lung carcinomas. *Br. J. Cancer*. **5,** 724-30.

Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci* USA. **95,** 5857-64.

Shigemura, K., Sung, S. Y., Kubo, H., Arnold, R. S., Fujisawa, M., Gotoh, A., Zhau, H. E. & Chung, L. W. (2007). Reactive oxygen species mediate androgen receptor- and serum starvation-elicited downstream signaling of ADAM9 expression in human prostate cancer cells. *Prostate*. **7,** 722-31.

Skiles, J. W., Gonnella, N. C. & Jeng, A. Y. (2001). The design, structure, and therapeutic application of matrix metalloproteinase inhibitors. *Current Med. Chem.* **8**, 425–74.

Solomon, D. H., Stedman, M., Licari, A., Weinblatt, M. E., Maher, N. & Shadick, N. (2007). Agreement between patient report and medical record review for medications used for rheumatoid arthritis: the accuracy of self-reported medication information in patient registries. *Arthritis Rheum*. **2,** 234-9.

Takada, H., Imoto, I., Tsuda, H., Nakanishi, Y., Ichikura, T., Mochizuki, H., Mitsufuji, S., Hosoda, F., Hirohashi, S., Ohki, M. & Inazawa, J. (2005). ADAM23, a possible tumor suppressor gene, is frequently silenced in gastric cancers by homozygous deletion or aberrant promoter hypermethylation. *Oncogene.* **54,** 8051-60.

Tanabe, C., Hotoda, N., Sasagawa, N., Sehara-Fujisawa, A., Maruyama, K. & Ishiura, S. (2007). ADAM19 is tightly associated with constitutive Alzheimer's disease APP alpha-secretase in A172 cells. *Biochem Biophys Res. Commun.* **1,** 111-7.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673-80.

Wolfsberg, T. G. & White, J. M. (1996). ADAMs in fertilization and development. *Dev BiD.* **180,** 389-401.

CHAPTER 6

Diversity of Ser/Thr Kinases in the Genomes of Various *Mycobacterium* Species

6.1 Introduction

Ser/Thr kinases (STKs) play a key role in cellular signal transduction and their biological functions have established their roles in cell growth and differentiation. These kinases were originally identified in eukaryotes and later identified in several prokaryotes. Kinases can be classified into two types based on their cellular location, as receptor or non-receptor kinases. Receptor kinases have a membrane spanning region and hence bound to the membrane, and the non-receptor kinases are cytosolic. The other types of kinases are tyrosine and histidine kinases. These proteins function by phosphorylating the cellular target proteins, thereby bringing about a conformational change in them, such that, they either covalently or non-covalently interact with signaling partners to decide the fate of the cell.

The kinase domain of STKs has similar 3-D structures, and their catalytic domain comprises 270 amino acid residues. The 3-D structure comprises two domains, the N-terminal domain comprises mainly β-strands and the C-terminal domain comprises α-helices (Zheng *et al.,* 1993). Previous studies have shown that protein phosphorylation and regulation of kinase activity is sufficiently, divergently evolved in eukaryotes and prokaryotes (Han & Zheng, 2001). Several studies on the evolution of eukaryotic and prokaryotic STKs have been addressed in order to assess the essential features in the regulation of their cellular activities (Hanks & Hunter, 1995; Av-Gay & Everett, 2002; Petrickova & Petricek, 2003). It has been observed that STKs occur as multi-domain proteins viz, a kinase domain is present as combination with other protein domains. Such coexistence of domains in proteins is an essential feature observed in some proteins since: 1) the different domains in proteins are involved in cascade of related events, 2) the activation of one domain by some cofactors brings about an allosteric change, thus leading to the activity of other domains and 3) these domains might be required for the proper positioning of the protein and then translocate it to the site of activity.

The organization of domains in a large dataset of bacterial STKs has been investigated in order to recognize variety in domain combinations which determine the functions of bacterial STKs. Previous studies (Krupa & Srinivasan, 2005) have shown that STKs in prokaryotic genomes have diverse domain arrangements which are different from that of eukaryotes. Similar results were observed by Zhang *et al*., 2007 in the genome wide survey of cyanobacteria.

The Mycobacterium genus comprises a number of Gram-positive, acid-fast, rod-shaped aerobic bacteria and is the only member of the family *Mycobacteriaceae* within the order *Actinomycetales*. Like other closely related *Actinomycetales*, such as *Nocardia* and *Corynebacterium*, Mycobacteria also have unusually high genomic DNA GC content and are capable of producing mycolic acids as major components of their cell wall.

M. tuberculosis is the causative agent of tuberculosis, a chronic infectious disease with a growing incidence worldwide. This species is responsible for more morbidity in humans than any other bacterial disease. It infects 1.7 billion people a year (~33% of the entire world population) and causes over 3 million deaths per year. The *M. tuberculosis* H37Rv genome comprises 3989 proteins (Cole *et al.,* 1998). *M. avium* 104 was earlier thought to cause tuberculosis in birds, but it now proved that it causes infections in immuno-compromized humans, such as the elderly, children and especially patients with AIDS. This genome codes for 5120 proteins (genome project; txid: 243243). *M. bovis* is the causative agent of classic bovine tuberculosis, but it can also cause the disease in humans, especially if contaminated milk is consumed without prior pasteurization. This genome comprises 3920 proteins (Garnier *et al.,* 2003). *M. bovis* strain BCG was used to produce BCG (Bacille de Calmette et Guan) vaccine, a well-known tuberculosis vaccine, originally developed by Calmette and Guan in the 1920's by multiple subculturings that resulted in attenuation or weakened virulence of the strain. This genome comprises 3952 proteins (Brosch *et al.,* 2007). *M. leprae* is the causative agent of leprosy or Hanson's disease. The infection is thought to be spread by the

respiratory route because lepromatous patients harbour bacilli in their nasal passages. This genome consists of 1605 proteins (Geluk *et al.,* 2005). The genome of *M. leprae* is smaller due to reductive genome evolution, with many important metabolic activities and their regulatory circuits eliminated due to extensive recombination events between dispersed repetitive sequences. The bacterium *M. smegmatis* was initially isolated from human smegma. It is associated with soft tissue lesions following trauma or surgery and is also reported as a possible factor in penile carcinogenesis. This genome comprises 6716 proteins (genome project; txid: 246196). *M. ulcerans* Agy99 causes Buruli ulcer, the third most common mycobacterial pathogen after *M. tuberculosis and M. leprae*. This genome comprises 4160 genes (Stinear *et al.,* 2007). *M. gilvum* PYR-GCK was isolated from river sediment and is able to degrade pyrene and other aromatic hydrocarbons. This genome consists of 5579 proteins (genome project; txid: 350054). *M. vanbaalenii* PYR-1 strain was isolated from contaminated sites exposed to petrogenic chemicals in the watershed of Redfish Bay, Texas, in 1986. It can degrade polycyclic aromatic hydrocarbons such as fluoranthene, pyrene, phenanthrene. This genome comprises 5979 proteins (genome project; txid: 350058). *Mycobacterium* sp. JLS. was isolated from creosote-contaminated soil and this microbe, along with some others collected at this site, is able to rapidly mineralize ^{14}C-labeled pyrene. This genome comprises 5739 genes (Miller *et al.,* 2004).

The original genus mycobacterium has been speciated with several variations within the bacterium such that the resultant genomes have diverse functions and variable genome size. While some species are pathogenic to humans under normal conditions, other species are disease causing under immunocompromized conditions and some bacteria have a hydrocarbon processing ability. Due to this diversity in the chemical functions of these species and the recently availability of their complete proteome information has prompted

us to carry out *in silico* analysis of the Ser/Thr kinases and study their domain architecture in these genomes.

In the current study, we have identified 100 STKs and homologs in 10 completed representative mycobacterial genomes using profile based search methods adapted in PSI-BLAST searches available at NCBI. The domain organization of these proteins has been studied in order to identify other coexisting domains. This analysis will aid decipher the various biological functions of these proteins alongside the kinase activity.

6.2 Methods

Initially sequences encoded by the Ser/Thr kinases were obtained from protein database at NCBI (www.ncbi.nlm.nih.gov/) using keyword searches. These proteins were used by PSI-BLAST searches as queries in order to search ten mycobacterium genomes namely; *M. tuberculosis* H37Rv, *M. avium* 104, *M. bovis, M. bovis* strain BCG, *M. leprae, M. smegmatis*, *M. ulcerans* Agy99, *M. gilvum* PYR-GCK, *M. vanbaalenii* PYR-1, *Mycobacterium* sp. JLS. The identified proteins were validated using reciprocal BLAST searches, till no new proteins were identified. BLOSUM62 matrix, with existence 11 and extension 1 as gap penalties, Expect threshold 10 and PSI-BLAST threshold 0.005 was chosen for all the PSI-BLAST searches. Multiple sequence alignment was generated to confirm that these proteins do belong to the STK family.

These proteins confirmed as Ser/Thr kinases were submitted to SMART (Schultz *et al.,* 1998) using the batch mode, in order to analyse the domains present in the proteins. The regions that are not identified by SMART as already known domains were analysed manually using PSI-BLAST against nr database at NCBI.

6.3 Results and Discussion

From the analysis of ten complete mycobacterial genomes, we identified 120 STKs and their homologs using profile based database searching methods incorporated in PSI-BLAST. Some of these proteins share 100% sequence identity and hence removed as being redundant. The final database comprises 100 proteins. Multiple sequence alignment of these proteins indicated that the similarity among the kinase domain regions of these proteins is high (>40% sequence homology), compared to the non-kinase domain regions within the proteins (data not shown). The domain organization of these proteins was identified using SMART analysis and we observed that all the proteins comprise a STK domain. Further, we also identified that nearly 75% of the STKs have a membrane spanning region positioned C-terminus to the kinase domain indicating that these are receptor STKs.

The PSI-BLAST searches of the regions, that were not identified by SMART as already known domains revealed the extent of similarity or variation among the STK family members of mycobacterial species. Further, based on the extent of homology in the nonkinase domain regions, these proteins have been classified into subclasses and proteins from each subclass were examined. Proteins in the subclass 1 comprise representative members from all the mycobacterial genomes analysed in this work. The STK domain is followed by 4 tandem PASTA domains at the C-terminus. These PASTA domains were earlier reported by Krupa and Srinivasan, 2005 to be present in bacterial STKs. The presence of PASTA domains strongly suggests that it is a signal-binding sensor domain.

Members of this subclass 2 are present in *M. smegmatis*, *Mycobacterium* sp. *JLS*, *M. vanbaalenii* and *M. gilvum.* These proteins comprise a bacterial periplasmic substrate-binding protein domain (PBPb), C-terminus to the STKs. It is known that bacterial high affinity transport systems are involved in active transport of solutes across the cytoplasmic membrane. The protein components of these traffic systems include one or two transmembrane protein components, one or two membrane-

associated ATP-binding proteins and a high affinity periplasmic solute-binding protein. The latter are thought to bind the substrate in the vicinity of the inner membrane and to transfer it to a complex of inner membrane proteins for concentration into the cytoplasm. It is known that some solute-binding proteins function in the initiation of sensory transduction pathways. Proteins of the subclass 3 comprise NHL repeats, positioned C-terminus to the STK domain. The members of this subclass are proteins from *M. avium* 104, *M. tuberculosis* H37Rv, *M. bovis* BCG. NHL repeats are known to be present in cell surface proteins and these tandem repeats fold into beta-propeller architecture and it is possible that these proteins function by ligand binding to the extracellular beta-propeller structure. The members of subclasses 2 and 3 were earlier also reported by Krupa and Srinivasan, 2005 but, we have identified these proteins in more genomes due to the larger dataset understudy.

Our analysis has further identified novel domain architectures in mycobacterial STKs that have not been reported earlier. Proteins in the subclass 4 comprise two distinct domains, one each at the N- and C-termini, that sandwich the central STK domain. The sequence analysis of N- and C-terminal regions using PSI-BLAST identified STKs from actinobacterial genomes such as *Rhodococcus, Janibacter, Streptomyces, Corynebacterium* and *Nocardia*. This indicates that these two domains are responsible for a function, conserved in all species from actinobacteria. The multiple sequence alignments of N and C-terminal regions are indicated in Figures 6.1a and 6.1b. We do not know the structure or function of these domains, but the presence of conservatively varied residues and cysteines at conserved positions in the multiple sequence alignment (Figure 6.1a) and several conserved sequence motifs in the multiple sequence alignment (Figure 6.1b) implies a common function mediated by these novel domains.

Members of the subclass 5 are present in *Mycobacterium* sp. *JLS* (YP_001072984) and *M. vanbaalenii* (YP_951920) genomes. They comprise Kelch repeats, positioned C-terminus to the STK domain. Kelch is a 50 amino acid residue

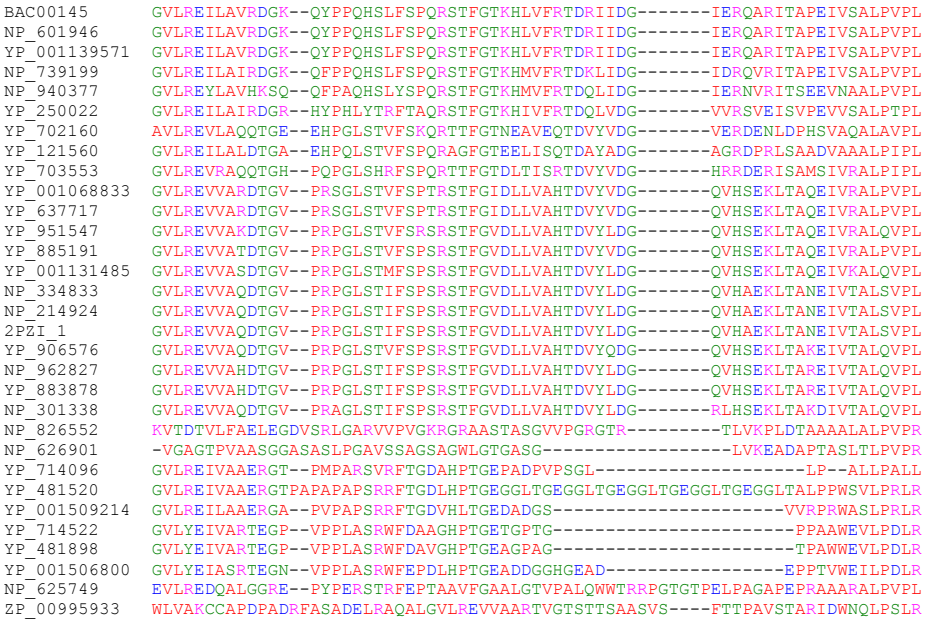
sequence motif that represents one beta-sheet blade and several of these repeats can associate to form a beta-propeller. The funnel of the beta propeller structure is often important for ligand recognition and this might be important for the activation of these kinases.

Members of the subclass 6 are present in all mycobacterial species analysed. They comprise a conserved C-terminal domain as indicated in Figure 6.2. Also from this multiple sequence alignment it can be seen that several sequence motifs are conserved, indicating a common function mediated by these proteins. One protein from *M. avium* (YP_025533), comprises Peptidyl-prolyl cis-trans isomerase domain. One protein from *M. gilvum* (YP_001134469) comprises NERD domain positioned N-terminus to the kinase domain. The NERD (nuclease-related domain) is found in a range of bacterial as well as archaeal and plant proteins. NERD domain has distant similarity to endonucleases. The representation of STKs in these subclasses is shown in Figure 6.3.

From the analysis of STKs in ten representative mycobacterial genomes, we observed that the kinase domain is highly conserved among all members of mycobacterial species. Kinase domain coexists with PASTA, PBPb, peptidyl-prolyl cis-trans isomerase and NERD domains, as well as NHL and Kelch repeats. Certain STKs (from subclass 6) consists of a conserved domain specifically present in mycobacterial species while STKs (from subclass 4) consists of domains at N- and C-terminus that are also present in several actinomycetales indicating the restriction in the divergent evolution of STKs.

Figure 6.1a. The Multiple Sequence Alignment of N-terminal Regions of Some STKs. (These proteins also occur in *Rhodococcus* sp, *Janibacter* sp, *Streptomyces coelicolor*, *Streptomyces avermitilis*, *Frankia alni* and *Nocardia farcinica* Apart from mycobacterium species. * indicates a conserved residue).

Figure 6.1b. The Multiple Sequence Alignment of C-terminal Regions of Some STKs. (These proteins also occur in *Rhodococcus* sp, *Janibacter* sp, *Streptomyces coelicolor*, *Streptomyces avermitilis*, *Frankia alni, Corynebacterium glutamicum*, *Corynebacterium efficiens*, *Corynebacterium diphtheriae*, *Corynebacterium jeikeium*, *Streptomyces avermitilis*, *Janibacter* sp and *Nocardia farcinica* apart from mycobacterium species. * indicates a conserved residue).

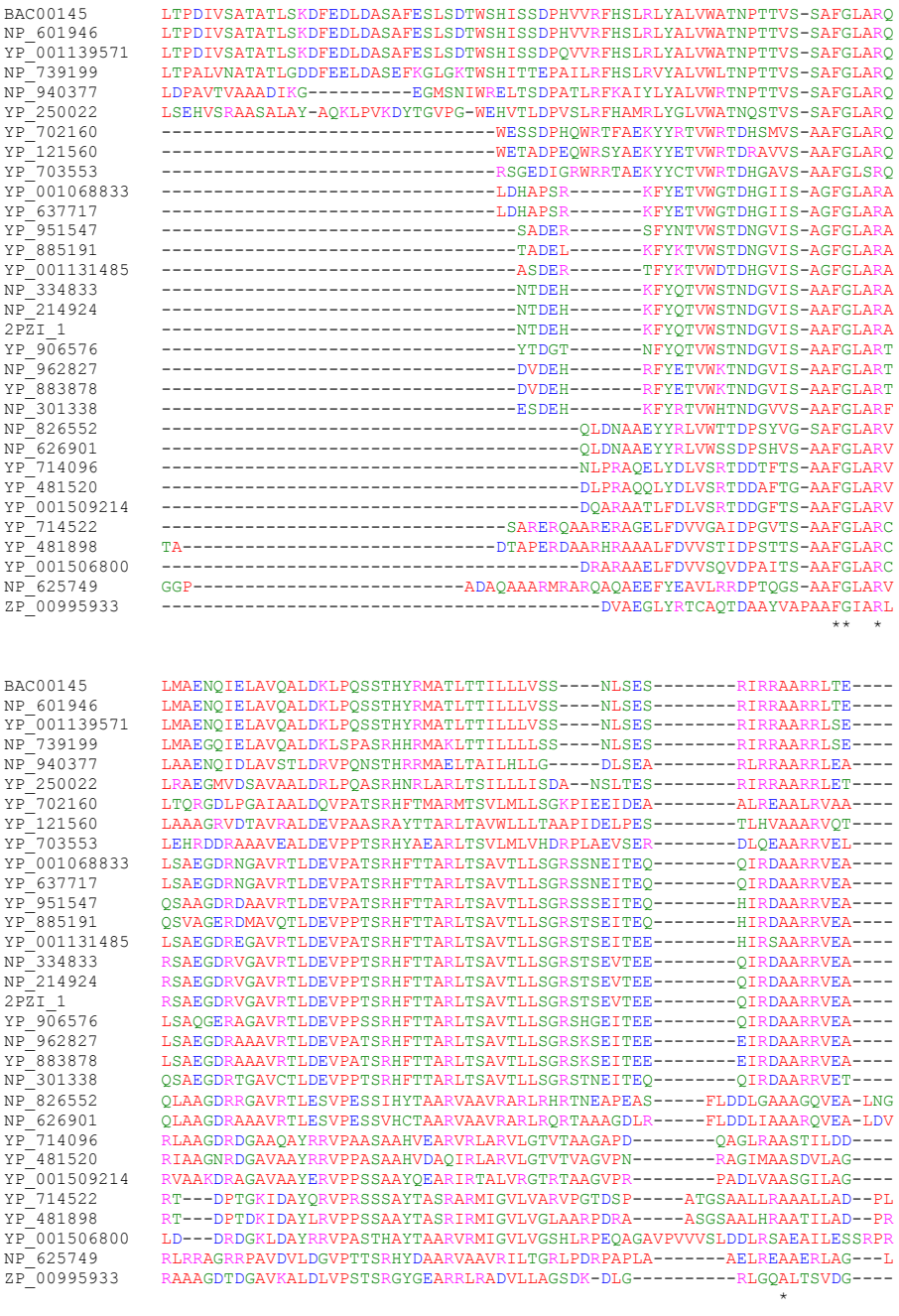


…… continued

```
BAC00145       IDRTDPGARMLSGSSYAEPSETLETLRNSMED--------EQYRQSIEIPLGVVRALLDLGFTTEARQWLETL
NP_601946      IDRTDPGARMLSGSSYAEPSETLETLRNSMED--------EQYRQSIEIPLGVVRALLDLGFTTEARQWLETL
YP_001139571   IDRTDPGARMLSGSSYAEPSETLETLRNSMED--------EQYRQSIEIPLGVVRALLDLGFTTEARQWLETL
NP_739199      IDRTDPGARMLSGSSYAEASETLENLRTSMEN--------EKYRDSIEIPLGVVRALLDLGFTAEARAWLDTL
NP_940377      LDRQDPGAILISGSSYTEPSEALQTMREAMTQ--------EKFSGSVEIPLGIVRALLDLGFTDEAASWLEDL
YP_250022      ADSQDPGYGLLSATSFTEAGDLLDTLRAAYAQ--------PELKHSVEIPLTMVRALLDVGQTREAHDLLDEL
YP_702160      IDPTDPNAPLLAAAVHSEPQQTLESIRHARENGIERVVGDLDVSFSRELTLAEIKAHLDLGDAATATELLRAV
YP_121560      IDPGEPAAALLAAAMQPDPARALDALGEARAR-AEADPDTAPDTLGVELTLAEVRIRLDMGDAAAALRLLARL
YP_703553      VDPDDPAAPLLS-VVHSRPRELLDSLHRAKEV-----ADDGVASTSIEIPLAEVRAHLDLGQPRDAATILESL
YP_001068833   VDPTDVGAAVLSASVLSEPVQTLDQLRAARHG--SLDSEGIDLSESVELPLMEARALLDLGDVAKATRKLDDL
YP_637717      VDPTDVGAAVLSASVLSEPVQTLDQLRAARHG--SLDSEGIDLSESVELPLMEARALLDLGDVAKATRKLDDL
YP_951547      VDPTDVGATILSAIVLSQPVQTLDSLRAARHG--TLDSEGIDLSESVELPLMEVRALLDLGDVAKATRKLDDL
YP_885191      VDRTDVGAPMLVASVLSEPVHTLDQLRAARHG--ALDTEGIDLNESVELPLMEVRALLDLGDVAKATRKLEDL
YP_001131485   VDPTDVGATVLSATLLSQPVQTLDSLRAVRHG--ALDSDGVDLSQSVELPLMEVRALLDLGDVAKATRKLDDL
NP_334833      VDPTDVAASVLQATVLSQPVQTLDSLRAARHG--ALDADGVDFSESVELPLMEVRALLDLGDVAKATRKLDDL
NP_214924      VDPTDVAASVLQATVLSQPVQTLDSLRAARHG--ALDADGVDFSESVELPLMEVRALLDLGDVAKATRKLDDL
2PZI_1         VDPTDVAASVLQATVLSQPVQTLDSLRAARHG--ALDADGVDFSESVELPLMEVRALLDLGDVAKATRKLDDL
YP_906576      VDRTDVAASVLQATVLSQPVQTLDSLRAARHG--ALDSDGADLSESVELPLMEVRALLDLGDVAKATRKLDDL
NP_962827      VDPADVAAPVLQATVLSQPVQTLDSLRAARHG--TLDADGVELSESIELPLMEVRALLDLGDVAKATRKLDDL
YP_883878      VDPADVAAPVLQATVLSQPVQTLDSLRAARHG--TLDADGVELSESIELPLMEVRALLDLGDVAKATRKLDDL
NP_301338      VDPTDVAAPVLQATVLSQPVQTLDSLRAARHG--MLDAQGIDLAESVELPLMEVRALLDLGDVVKANRKLDDL
NP_826552      VDPGDPNAGFLAGLMTSGPLELVAALGNAPSP-------------SVETRLRQIRARLENGDSHTALEVLAKL
NP_626901      VDAGDPNAGFLAGLMASAPAELLTALAAAPAP-------------STETRLRQIRARLENGDASGALEALGAL
YP_714096      PDPDDPAATALAALPDVSPGQLAELLDAMGAD-------------SAGARLRLADLRTRLGEHDAARELLDAV
YP_481520      ADPDDPAADPLTALPDLAPEQLAELLGAMGTT-------------SVGARLHLADLRMRLGRTAAAREMLAEI
YP_001509214   VDPEDPAAGTLAALPDSSPAQLATLLAAISPA-------------TVEVRLRLARAHLETGDTAAAAAVLDEV
YP_714522      VDQDDPRAPALTAGPEEDPAALAARLAAIVPV-------------TTEVRLALARARIRAGQLGEAARALDAA
YP_481898      IDQDDPWAPGLTAGSDEEPASLATRLAAIVPR-------------TTEVRLSLARAQIRAGQLHEAAGTLDDA
YP_001506800   IDADDPQARVLATTPGEDPAALAERLAEVAQP-------------TRQTRLALARAQMRAGYTGKAEVTVREL
NP_625749      PDASDPAAVLLGGLAADTPDRIAERSAGDPALR-----------TVETALWLCRAYLEAGDAARAEEWVARA
ZP_00995933    PDTTDAQYAWLVSLAPGEPQERLADLAKAPEA-------------TAEVHLARGAEYLLDGNGSATKEQAAQL

BAC00145        EGRI-GDDWRHKWFSGITYLLLD----DYATAQVFFNHVLTILPGEAAPKLALAAVDELILQQIGAESTAY
NP_601946       EGRI-GDDWRHKWFSGITYLLLD----DYATAQVFFNHVLTILPGEAAPKLALAAVDELILQQIGAESTAY
YP_001139571    EGRI-GDDWRHKWFSGITYLLLD----DYATAQVFFNHVLTILPGEAAPKLALAAVDELILQQIGAESTAY
NP_739199       KERM-GRDWRHQWFSGITHLLLD----DYAAAQEYFYTVLTILPGEAAPKLALAAVDELLLQQMGYENTPL
NP_940377       VPRL-GNEWRHQWYSGVTSLLLD----DYLTAQQHFNEVYNILPGESAPKLARAAVCEMLLQEKGLESTAL
YP_250022       KPWL-EKDWRFQWHSGVAALLTG----QFAEAQKFFNRVLYILPGEPAPKLALAATDELLLQQQGVNSTKL
YP_702160       EGDM-GGNWRVDWYAGLATLIDG----EYETAFSRFESVLKAMPGETAPKLALAATAELILQH--------
YP_121560       D----AGDWRVHWYTGLAELREQ----HYESAFAAFEDVLRVLPGEIAPKLALAATAELVLQH--------
YP_703553       DHQR-EDTWRVDWHTGLCALVGG----DFEAAFTRFDSVLTALPGEEAPKLALAATAEMWCEY--------
YP_001068833    ATRV-GWRWRLVWFRAVAELLSA----DYESATKHFTEVLDTVPGELAPKLALAATAELAGSD--------
YP_637717       ATRV-GWRWRLVWFRAVAELLSA----DYESATKHFTEVLDTVPGELAPKLALAATAELAGSD--------
YP_951547       AERV-GWRWRLVWFRAVSELLTA----DYDSATKHFNEVLDTLPGELAPKLALAATAELAG----------
YP_885191       AARV-GWRWRLVWFKAVSEMLSA----DYDSATKHFTEVLDTLPGELAPKLALAATAELAG----------
YP_001131485    AERV-GWRWRLVWFRAVSELLTA----DYDSATKHFTQVLDTLPGELAPKLALAATAELAG----------
NP_334833       AERV-GWRWRLVWYRAVAELLTG----DYDSATKHFTEVLDTFPGELAPKLALAATAELAG----------
NP_214924       AERV-GWRWRLVWYRAVAELLTG----DYDSATKHFTEVLDTFPGELAPKLALAATAELAG----------
2PZI_1          AERV-GWRWRLVWYRAVAELLTG----DYDSATKHFTEVLDTFPGELAPKLALAATAELAG----------
YP_906576       SERV-GWRWRLVWYRAVSELLTG----AYDAAIKHFTEVLDTFPGELAPKLALAATAELAG----------
NP_962827       AERV-GWQWRLVWYKAVAELLTG----DYDSATTHFTEVLDTFPGELAPKLALAATAELAG----------
YP_883878       AERV-GWQWRLVWYKAVAELLTG----DYDSATTHFTEVLDTFPGELAPKLALAATAELAG----------
NP_301338       ADRV-SCQWRLVWYRAVADLLTG----DYASATKHFTEVLNTFPGELAPKLALAATAELAG----------
NP_826552       EDER-PDDWRVVWYRGVAALVTG----DHEVGALSFDAIYDAFPGEPAPKLALGLCAEVLG----------
NP_626901       EGER-PDDWRVVWYRGLAALVTG----AHEDAALAFDAIYDAFPGEIAPKLALGLCAEVLG----------
YP_714096       AAED-PFAWRVHWQRGLLLLTQG----DTAGAVTAFERVYGEVPGELAPKLALARAAEAGG----------
YP_481520       EAED-PFEWRVDWQRGLLALADG----DTAAARGAFDRVYDEVPGELAPKLALARTAEATG----------
YP_001509214    EAED-PFEWRVRWYRGLLALGDGDGDGDTAAAAAVFTEVYAQVPGELAPKLALAVTAEAAG----------
YP_714522       AAAA-PREWRVDWYRGLAALVAR----RPGQAAAAFDRVYSQVPGELAPRLALATALADVATTAG------
YP_481898       AVEQ-PREWRVDWYRGLHALVGG----RPAVAAAAFDRVYSRVPGELAPRLALATALADVATTADTADTAD
YP_001506800    AAAY-PREWRVTWCQGLLDLAAG----RSTQAAEAFEKVYARMPGELAPRLAVATASANAG----------
NP_625749       KGWSGDYDWRIFWHRGLIHLTRD----AVDKAEDEFAATYAALPGEAAPKLALGFCAEYLAERPRESAGEG
ZP_00995933     LALD-PWDWRALWLQGLASLQSS----DWADAQASFSAVYQQVPGELAPKLALAVACEAGGQG--------
                       **  *                       *       *** ** **
```

…… continued


BAC00145 LTPDIVSATATLSKDFEDLDASAFESLSDTWSHISSDPHVVRFHSLRLYALVWATNPTTVS-SAFGLARQ
NP_601946 LTPDIVSATATLSKDFEDLDASAFESLSDTWSHISSDPHVVRFHSLRLYALVWATNPTTVS-SAFGLARQ
YP_001139571 LTPDIVSATATLSKDFEDLDASAFESLSDTWSHISSDPQVVRFHSLRLYALVWATNPTTVS-SAFGLARQ
NP_739199 LTPALVNATATLGDDFEELDASEFKGLGKTWSHITTEPAILRFHSLRVYALVWLTNPTTVS-SAFGLARQ
NP_940377 LDPAVTVAAADIKG----------EGMSNIWRELTSDPATLRFKAIYLYALVWRTNPTTVS-SAFGLARQ
YP_250022 LSEHVSRAASALAY-AQKLPVKDYTGVPG-WEHVTLDPVSLRFHAMRLYGLVWATNQSTVS-SAFGLARQ
YP_702160 ----------WESSDPHQWRTFAEKYYRTVWRTDHSMVS-AAFGLARQ
YP_121560 ----------WETADPEQWRSYAEKYYETVWRTDRAVVS-AAFGLARQ
YP_703553 ----------RSGEDIGRWRRTAEKYYCTVWRTDHGAVS-AAFGLSRQ
YP_001068833 ----------LDHAPSR-------KFYETVWGTDHGIIS-AGFGLARA
YP_637717 ----------LDHAPSR-------KFYETVWGTDHGIIS-AGFGLARA
YP_951547 ----------SADER-------SFYNTVWSTDNGVIS-AGFGLARA
YP_885191 ----------TADEL-------KFYKTVWSTDNGVIS-AGFGLARA
YP_001131485 ----------ASDER-------TFYKTVWDTDHGVIS-AGFGLARA
NP_334833 ----------NTDEH-------KFYQTVWSTNDGVIS-AAFGLARA
NP_214924 ----------NTDEH-------KFYQTVWSTNDGVIS-AAFGLARA
2PZI_1 ----------NTDEH-------KFYQTVWSTNDGVIS-AAFGLARA
YP_906576 ----------YTDGT-------NFYQTVWSTNDGVIS-AAFGLART
NP_962827 ----------DVDEH-------RFYETVWKTNDGVIS-AAFGLART
YP_883878 ----------DVDEH-------RFYETVWKTNDGVIS-AAFGLART
NP_301338 ----------ESDEH-------KFYRTVWHTNDGVVS-AAFGLARF
NP_826552 ----------QLDNAAEYYRLVWTTDPSYVG-SAFGLARV
NP_626901 ----------QLDNAAEYYRLVWSSDPSHVS-AAFGLARV
YP_714096 ----------NLPRAQELYDLVSRTDDTFTS-AAFGLARV
YP_481520 ----------DLPRAQQLYDLVSRTDDAFTG-AAFGLARV
YP_001509214 ----------DQARAATLFDLVSRTDDGFTS-AAFGLARV
YP_714522 ----------SARERQAARERAGELFDVVGAIDPGVTS-AAFGLARC
YP_481898 TA----------DTAPERDAARHRAAALFDVVSTIDPSTTS-AAFGLARC
YP_001506800 ----------DRARAAELFDVVSQVDPAITS-AAFGLARC
NP_625749 GGP----------ADAQAAARMRARQAQAEEFYEAVLRRDPTQGS-AAFGLARV
ZP_00995933 ----------DVAEGLYRTCAQTDAAYVAPAAFGIARL
LMAENQIELAVQALDKLPQSSTHYRMATLTTILLLVSS----NLSES--------RIRRAARRLTE----
LMAENQIELAVQALDKLPQSSTHYRMATLTTILLLVSS----NLSES--------RIRRAARRLTE----
LMAENQIELAVQALDKLPQSSTHYRMATLTTILLLVSS----NLSES--------RIRRAARRLSE----
LMAEGQIELAVQALDKLSPASRHHRMAKLTTILLLLSS----NLSES--------RIRRAARRLSE----
LAAENQIDLAVSTLDRVPQNSTHRRMAELTAILHLLG-----DLSEA--------RLRRAARRLEA----
LRAEGMVDSAVAALDRLPQASRHNRLARLTSILLLISDA--NSLTES--------RIRRAARRLET----
LTQRGDLPGAIAALDQVPATSRHFTMARMTSVLMLLSGKPIEEIDEA--------ALREAALRVAA----
LAAAGRVDTAVRALDEVPAASRAYTTARLTAVWLLLTAAPIDELPES--------TLHVAAARVQT----
LEHRDDRAAAVEALDEVPPTSRHYAEARLTSVLMLVHDRPLAEVSER--------DLQEAARRVEL----
LSAEGDRNGAVRTLDEVPATSRHFTTARLTSAVTLLSGRSSNEITEQ--------QIRDAARRVEA----
LSAEGDRNGAVRTLDEVPATSRHFTTARLTSAVTLLSGRSSNEITEQ--------QIRDAARRVEA----
QSAAGDRDAAVRTLDEVPATSRHFTTARLTSAVTLLSGRSSSEITEQ--------HIRDAARRVEA----
QSVAGERDMAVQTLDEVPPTSRHFTTARLTSAVTLLSGRSTSEITEQ--------HIRDAARRVEA----
LSAEGDREGAVRTLDEVPATSRHFTTARLTSAVTLLSGRSTSEITEE--------HIRSAARRVEA----
RSAEGDRVGAVRTLDEVPPTSRHFTTARLTSAVTLLSGRSTSEVTEE--------QIRDAARRVEA----
RSAEGDRVGAVRTLDEVPPTSRHFTTARLTSAVTLLSGRSTSEVTEE--------QIRDAARRVEA----
RSAEGDRVGAVRTLDEVPPTSRHFTTARLTSAVTLLSGRSTSEVTEE--------QIRDAARRVEA----
LSAQGERAGAVRTLDEVPPSSRHFTTARLTSAVTLLSGRSHGEITEE--------QIRDAARRVEA----
LSAEGDRAAAVRTLDEVPATSRHFTTARLTSAVTLLSGRSKSEITEE--------EIRDAARRVEA----
LSAEGDRAAAVRTLDEVPATSRHFTTARLTSAVTLLSGRSKSEITEE--------EIRDAARRVEA----
QSAEGDRTGAVCTLDEVPPTSRHFTTARLTSAVTLLSGRSTNEITEQ--------QIRDAARRVET----
QLAAGDRRGAVRTLESVPESSIHYTAARVAAVRARLRHRTNEAPEAS-----FLDDLGAAAGQVEA-LNG
QLAAGDRAAAVRTLESVPESSVHCTAARVAAVRARLRQRTAAAGDLR-----FLDDLIAAARQVEA-LDV
RLAAGDRDGAAQAYRRVPAASAAHVEARVRLARVLGTVTAAGAPD--------QAGLRAASTILDD----
RIAAGNRDGAVAAYRRVPPASAAHVDAQIRLARVLGTVTVAGVPN--------RAGIMAASDVLAG----
RVAAKDRAGAVAAYERVPPSSAAYQEARIRTALVRGTRTAAGVPR--------PADLVAASGILAG----
RT---DPTGKIDAYQRVPRSSSAYTASRARMIGVLVARVPGTDSP-----ATGSAALLRAAALLAD--PL
RT---DPTDKIDAYLRVPPSSAAYTASRIRMIGVLVGLAARPDRA-----ASGSAALHRAATILAD--PR
LD---DRDGKLDAYRRVPASTHAYTAARVRMIGVLVGSHLRPEQAGAVPVVVSLDDLRSAEAILESSRPR
RLRRAGRRPAVDVLDGVPTTSRHYDAARVAAVRILTGRLPDRPAPLA--------AELREAAERLAG---L
RAAAGDTDGAVKALDLVPSTSRGYGEARRLRADVLLAGSDK-DLG----------RLGQALTSVDG----

…… continued

```
BAC00145        IPTNEPRFNQIKIAIMSAGLSWLRERKLKASASA--------N---PLFEYPFS----------------
NP_601946       IPTNEPRFNQIKIAIMSAGLSWLRERKLKASASA--------N---PLFEYPFS----------------
YP_001139571    IPTNEPRFNQIKIAIMSAGLSWLRERKLKASASA--------N---PLFEYPFS----------------
NP_739199       IPTNDPRFNQIKIAIMSAGLSWLRDRDLQSAASP--------N---PLFEYPFT----------------
NP_940377       IPTNEPRFLQIQIAIMNAALQWLRTG---GEAAN--------D---PIFEYPFTQRGLRNGLSATLRQ--
YP_250022       LPTNEPRLPQVRIAVLSAGLNWLRGASDGSDGAAGIAGGGAGN----------------------------
YP_702160       LAPEESRALQMRTLVLGTALDWMRAGHTSATERE----------QILGVPFTERGLRKGAEAGLRALA
YP_121560       LPAGEARALQLRVLVLGAALAWLRAGHEPERADA----------TFFGAPFTERDLREGTEAGLRALA
YP_703553       LAHDERRALQTRVLVLGVAVDWLHAGGIPETTR----------ILSVPFTERGLRTGAESALRALA
YP_001068833    LPDTEPRVLQIRALVLGTAMDWLVD---NSANTN----------HILGFPFTEYGLQLGVEASLRALA
YP_637717       LPDTEPRVLQIRALVLGTAMDWLVD---NSANTN----------HILGFPFTEYGLQLGVEASLRALA
YP_951547       LPDSEPRVLQIRALVLGTAMDWLAD---NTASTN----------HILGFPFTEHGLALGVEASLRSLA
YP_885191       LPDSEPRVLQIRALVLGTALDWLAD---NTASSN----------HILGFPFTEHGLKLGVEASLRALA
YP_001131485    LPDTEPRVLQIRALVLGTAMDWLAD---NTASTN----------HILGFPFTEHGLQLGVEASLRSLA
NP_334833       LPPTEPRVLQIRALVLGGALDWLKD---NKASTN----------HILGFPFTSHGLRLGVEASLRSLA
NP_214924       LPPTEPRVLQIRALVLGGALDWLKD---NKASTN----------HILGFPFTSHGLRLGVEASLRSLA
2PZI_1          LPPTEPRVLQIRALVLGGALDWLKD---NKASTN----------HILGFPFTSHGLRLGVEASLRSLA
YP_906576       LPPTEPRVLQIRALVLGGAMDWLQD---NEASTN----------HILGFPFTDHGLRLGVEASLRSLA
NP_962827       LPPTEPRVLQIRALVLGCAMDWLED---NKASTN----------HILGFPFTEHGLRLGVEAALRNLA
YP_883878       LPPTEPRVLQIRALVLGCAMDWLED---NKASTN----------HILGFPFTEHGLRLGVEAALRNLA
NP_301338       LPPTEPRVLQIRALVLGCAMDWLAD---NQASAN---------G--HILGFPFTKHGLRLGVEASLRSLA
NP_826552       YGLDAVRREQLSTEVLGCALDWVLSGSQGSAPPPD--------GGRALLGNELDERGLRFGLERSYRTLA
NP_626901       YGLDPARREQLSAEVLGCALDWVLSGGRGSVPPAA---------GRTLLGSGLDERGLRFGLERSYRTLA
YP_714096       LDLDPAGRATLTRDLLAGALDLVS---TGAMAADP---------QVTIAGSALRPAALRLGLERAYRTLA
YP_481520       LDLDSGRRAALTRDLLTTALDLVA---AGTLPVDP---------GVTVAGAALREADLRFGLERAYRELA
YP_001509214    LDVDRRRRVALTRDLLRCALDLLL---AGDTPPDP--------RDVEVAGTRLREDDLRFGLERAYRELA
YP_714522       LDLGERRRAELRRDIFTAALTLV---PAGHPPPGSP-------RPPTLLGRPMVERELRFGLAETYREMA
YP_481898       LDLGERRRAELHRDLFTAALALVMTYPAAYPAADAP-------GPPTLLSRVMVERDLRFGLAETYRE--
YP_001506800    LGLDRRRRAELRLDILAAALRMVG---AGPPGPAGT--------PKQVFGRRLVPRELRLGLEAAYRDIA
NP_625749       HLDGSGSWDRLVTELREHALACRPPGGWGSGFPAG----------ELCGPQDTEEALRRLLSASLRRLA
ZP_00995933     VRLDTVEREGLTARILEKAIAIVGET--GPQPLN----------IGPYAAQDETLRTALEASYRALA

BAC00145        ------------------------
NP_601946       ------------------------
YP_001139571    ------------------------
NP_739199       ------------------------
NP_940377       ------------------------
YP_250022       ------------------------
YP_702160       RNAPERTHRYTLV-----------
YP_121560       RAAPGRDHRYALVDLAN-------
YP_703553       RKAPERRHRYTLVDLANLIRP---
YP_001068833    RVAPTQAHRYALIDLANSVRPMST
YP_637717       RVAPTQAHRYALIDLANSVRPMST
YP_951547       RVAPTQAHRYALVDLANSVRPMST
YP_885191       RIAPTQSHRYALVDLANSVRPMST
YP_001131485    RVAPTQAHRYALVDLANSVRPMST
NP_334833       RVAPTQRHRYTLVDMANKVRPTST
NP_214924       RVAPTQRHRYTLVDMANKVRPTST
2PZI_1          RVAPTQRHRYTLVDMANKVRPTST
YP_906576       RVAPTQRHRYTLVDMANKVRPTST
NP_962827       RVAPTQRHRYALVDMANKVRPTST
YP_883878       RVAPTQRHRYALVDMANKVRPTST
NP_301338       RVAPTQRHRYTLVDMANKVRPTST
NP_826552       ------------------------
NP_626901       RLARGGEERIDLVERANRYRPRTW
YP_714096       RLAATPDERYALVDLANSVRPRTL
YP_481520       RLAATTDERYALVDLANRVRPRTL
YP_001509214    TLARSAQERYDLVDLANAVRPRTW
YP_714522       RLTPDRASRV--------------
YP_481898       ------------------------
YP_001506800    RLAGDRRERIRYVDAANRIRPRTL
NP_625749       ------------------------
ZP_00995933     REAGTREERIALVDRANTVRPWTS
```

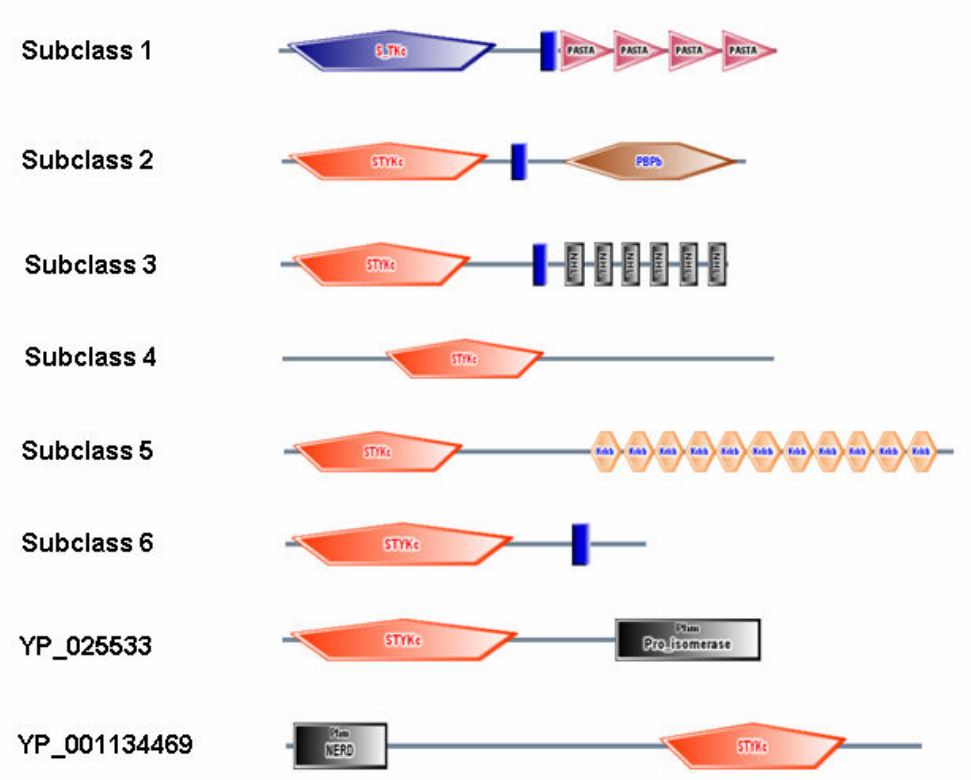
Figure 6.2. The Multiple Sequence Alignment of C-terminal Regions of Some STKs from Mycobacterium Species. (* indicates a conserved residue).

```
NP_214529       VRAGRRPPRPSQTPPPGRAAPAAIPSGTTARVAANSAGRTAASRRSRPATG--GHRPP--RRTFSSGQRAL
YP_904268       VRAGRRPPRPSQSPPAGRAAPAAIPSSAPARIAPAPTTRTTTPRRTRPATG--GHRPPPSRRTFSSGQRAL
YP_879320       VRAGRRPPRPSQSPPPGRASPAAIPSSPTTRAAAVSSGRTAAPRRTRPATG--GHRPPPARRTFSSGQRAL
NP_301143       VRAGHRPPRPNQTPSSGRASPTTIPSSTQARAAVACGTKTPAPRRSRPSTS--GNRPPPARNTFSSGQRAL
YP_001068320    VRSGRRPPRPNQAPSIGRATPAAVPSAAQARAAADLTGRAPVTAARARPTG-AIHRPPPPRRTFSSGQRAL
YP_884450       VRAGRRPPRPNQAPTLGRAAPAAVPSAAQARASADLTGRAPVTAARARPTATAAHRTPPPRRTFSSGQRAL
YP_001132083    VRAGRRPPRPNAAPSIGRAAPAAVPPAIQNRPPAEATGRAPAQTSRTRTTGSHHRSAPPARRTFSSGQRAL
YP_950881       VRAGRRPPRPNAAPSIGRAAPTAVPPAIASRPAADPTGRAPAPTSRTRATGSHHRSAPPPRGGFSSGQRAL
                **:*:*****. :*. ***:*:::*..   * .     ::..      .*.   . .*  *  ********

NP_214529       LWAAGVLGALAIIIAVLLVIKAPGDNSPQQAPTPTVTTTGNPPASNTGG----TDASPRLNWTERGETRHS
YP_904268       LWAAGVLGALAIIIAVLIVINSRAD-SQQQSPPPTVTQTTTAHQTPTG-------QGPRLNWTDGQSIGNP
YP_879320       LWAAGVLGALAIIIAVLIVINSRAD---QQQQPPTVTDTGTPPAS-----------APPPTKTPSGSGAHP
NP_301143       LWAAGMLGALAIIIAVLIVINSYAGNEQHQPPTPTVTDTGTPPATKTLSGFPAAYCEYRVNWTNHKEISNS
YP_001068320    LWAAGVLGALAIVIAILIVLNAQDRKDR-QLPPPTVTNTITET-TPYQS--------PAAMPEWMPDWTSS
YP_884450       LWAAGVLGALAIVIAILIVLNAQDRKDREQSPPPTVTDTVTET-TPYEET-------PAAMPDLMIMLR--
YP_001132083    LWAAGVLGALAIVIAILIVLNHQDQQNS----PNRTTNTVTETPGSPPAA-------PGEGPAETPGAAAP
YP_950881       LWAAGVLGALAIVIAILIVLNAQDRKDQ----PPPTT--VTETPTSQGAP-------P-----ETPGAAAP
                *****:******:**:*:*::           .  .*   .

NP_214529       GLQS----------------------WVVPPTPHSRASL-----ARYEIAQ------
YP_904268       GLQSNQPDLVDDAGRNRPIGNRDSATSWNVTVTPQHRAAL-----ARFEMQR------
YP_879320       GLRLDWPDD-------------------GTIGASGFRDGP-----ARHWTSQ------
NP_301143       GLPK-----------------------QAARAQLAGATD-----ISPVAGQT-----
YP_001068320    AVIGHGGDT-------------------SGAVRPDAVAASAHLRALPAPAQTWLRTQT
YP_884450       -------------------------------AAQPE-----------PPPSEQIQR---
YP_001132083    VKPD-EPGQ-------------------NRVSGPVDTASP---SVMLHVSEQILR---
YP_950881       AFGDHEPDR-------------------YWVLRPVDNLSP---LTTRHVSEQNLR---
```

Figure 6.3. A Representative Domain Architecture Diagram of STKs from Ten Mycobacterial Genomes Analysed.



6.4 Conclusions

1. The kinase domain in STKs is highly conserved compared to the other regions in the proteins.
2. In mycobacterial STKs, the kinase domain coexists with PASTA, PBPb, peptidyl-prolyl cis-trans isomerase and NERD domains. It also coexists with NHL and Kelch repeats.
3. Diversity in the STKs is evident from the fact that some STKs are present only in mycobacterial species while some STKs are common to other members of Actinobacteria.

6.5 References

Av-Gay, Y. & Everett, M. (2002). The eukaryotic-like Ser/Thr protein kinases of *Mycobacterium tuberculosis*. *Trends Microbiol*. **8,** 238-244.

Brosch, R., Gordon, S. V., Garnier, T., Eiglmeier, K., Frigui, W., Valenti, P., Dos Santos, S., Duthoy, S., Lacroix, C., Garcia-Pelayo, C., Inwald, J. K., Golby, P., Garcia, J. N., Hewinson, R. G., Behr, M. A., Quail, M. A., Churcher, C., Barrell, B. G., Parkhill, J. & Cole, S. T. (2007). Genome plasticity of BCG and impact on vaccine efficacy. *Proc. Natl. Acad. Sci. U S A*. **104**, 5596-5601.

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E. 3rd, Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M. A., Rajandream, M. A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J. E, Taylor, K., Whitehead, S. & Barrell BG. (1998). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature.* **393**, 537-544.

Garnier, T., Eiglmeier, K., Camus, J. C., Medina, N., Mansoor, H., Pryor, M., Duthoy, S., Grondin, S., Lacroix, C., Monsempe, C., Simon, S., Harris, B., Atkin, R., Doggett, J., Mayes, R., Keating, L., Wheeler, P. R., Parkhill, J., Barrell, B. G., Cole, S. T., Gordon, S. V. & Hewinson, R. G. (2003). The complete genome sequence of Mycobacterium bovis. *Proc. Natl. Acad. Sci. U S A*. **100**, 7877-7882.

Geluk, A., Klein, M. R., Franken, K. L., van Meijgaarden, K. E., Wieles, B., Pereira, K. C., Bührer-Sékula, S., Klatser, P. R., Brennan, P. J, Spencer, J. S., Williams, D. L., Pessolani, M. C, Sampaio, E. P. & Ottenhoff, T. H. (2005) Postgenomic approach to identify novel Mycobacterium leprae antigens with potential to improve immunodiagnosis of infection. *Infect. Immun.* **273**, 5636-5644.

Han, G. & Zheng, C. C. (2001). On the origin of Ser/Thr kinases in a prokaryote. *FEMS Microbiol lett.* **200**, 79-84.

Hanks, S. K. & Hunter, T. (1995). The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification *FASEB J.* **9**, 576-596.

Krupa, A. & Srinivasan, N (2005) Diversity in domain architectures of Ser/Thr kinases and their homologues in prokaryotes. *BMC Genomics.* **6,** 129-148.

Miller, C. D., Hall, K., Liang, Y. N., Nieman, K., Sorensen, D., Issa, B., Anderson, A. J. & Sims, R. C. (2004). Isolation and characterization of polycyclic aromatic hydrocarbon-degrading Mycobacterium isolates from soil. *Microb Ecol*. **48,** 230-238.

Petrickova, K. & Petricek, M. (2003). Eukaryotic-type protein kinases in *Streptomyces coelicolor*: variations on a common theme. *Microbiol*. **149**,1609-1621.

Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci* USA. **95,** 5857-64.

Stinear, T. P., Seemann, T., Pidot, S., Frigui, W., Reysset, G., Garnier, T., Meurice, G., Simon, D., Bouchier, C., Ma, L., Tichit, M., Porter, J. L., Ryan, J., Johnson, P. D., Davies, J. K. & Jenkin, G. A., Small, P. L., Jones, L.M., Tekaia, F., Laval, F., Daffé, M., Parkhill, J. & Cole, S.T. (2007). Reductive evolution and niche adaptation inferred from the genome of Mycobacterium ulcerans, the causative agent of Buruli ulcer. *Genome Res.* **17**, 192-200.

Zhang, X., Zhao, F., Guan, X., Yang, Y., Liang, C. & Qin, S. (2007). Genome-wide survey of putative Serine/Threonine protein kinases in cyanobacteria. *BMC Genomics.* **8**, 395.

Zheng, J., Knighton, D. R., ten Eyck L. F., Karlsson, R., Xuong, N., Taylor, S. S. & Sowadski, J. M. (1993). Crystal structure of the catalytic subunit of cAMP-dependent protein kinase complexed with MgATP and peptide inhibitor. *Biochemistry.* **32**, 2154-2161.

List of Publications

1) Docking of phosphonate and trehalose analog inhibitors into *M. tuberculosis* mycolyltransferase Ag85C: Comparison of the two scoring fitness functions GoldScore and ChemScore, in the GOLD software. (2006). Manoj Kumar Annamala, **Krishna Kishore Inampudi**, Lalitha Guruprasad. *Bioinformation* **9:** 339-350.

2) Chemical Function Based Virtual Screening: Discovery of potent lead molecules for the *Bcr-Abl* tyrosine kinase using VX-680. **Krishna Kishore Inampudi** and Lalitha Guruprasad (*being communicated*).

3) The Identification of New Aurora A Kinase Inhibitors by Pharmacophore Modeling, Virtual Screening and Molecular Docking. **Krishna Kishore Inampudi** and Lalitha Guruprasad (*being communicated*).

4) Comparative studies of the ADAM and ADAMTS protein family members in human, frog, fly and worm genomes: A Bioinformatics Approach. **Krishna Kishore Inampudi,** G.R. Hema Latha and Lalitha Guruprasad (*being communicated*).

5) Diversity of Ser/Thr kinases in the genomes of various *Mycobacterium* species. **Krishna Kishore Inampudi**, N. Srinivas and Lalitha Guruprasad (*being communicated*).

6) Identification and Analysis of novel tandem repeats in the cell surface proteins of Archaeal and Bacterial genomes using computational tools (2004). Swathi Adindla, **K.K. Inampudi**, K. Guruprasad and L. Guruprasad. *Comp. Func. Genom*. 5:2-16.

7) Cell surface proteins in archaeal and bacterial genomes comprising "LVIVD", "RIVW" and "LGxL" tandem sequence repeats are predicted to fold as beta-propeller. (2007). Swathi Adindla, **Krishna Kishore Inampudi** and Lalitha Guruprasad. *Int J Biol Macromol*. 41: 454-68.

8) Identification and Analysis of Novel Amino Acid Sequence Repeats and Domains in Pyrobaculum aerophilum Using Computational Tools. (2007). Golaconda Hemalatha; **Inampudi Krishna Kishore**; Raghavarapu Srinivasa Rao and Lalitha Guruprasad *Protein & Peptide Letters*. **14:** 692-697.