DEVELOPMENT OF COMPUTATIONAL METHODS FOR DRUG DESIGN BASED ON NATURAL PRODUCTS
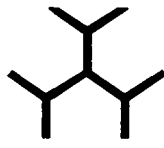
Thesis Submitted to
Jawaharlal Nehru University
for the degree of
DOCTOR OF PHILOSOPHY
September 2011

KATE BHUSHAN NARAYAN
National Institute of Immunology
New Delhi, India

राष्ट्रीय प्रतिरक्षाविज्ञान संस्थान
NATIONAL INSTITUTE OF IMMUNOLOGY

CERTIFICATE

This is to certify that the thesis entitled **"Development of computational methods for drug design based on natural products"** comprises the work done by **Kate Bhushan Narayan** at the National Institute of Immunology, New Delhi for the degree of Doctor of Philosophy from the Jawaharlal Nehru University. This work is original and has not been submitted in part or full for any degree or diploma of any university.

Kate Bhushan Narayan
Bioinformatics Center

Dr. Debasisa Mohanty
Thesis Supervisor
National Institute of Immunology
New Delhi

Director
National Institute of Immunology

Table of Contents

Acknowledgements

Doing a PhD has proved to be a life changing experience for me. I will like to remember this for long period of time so that I can apply the lessons learnt and insight gained during this period for future betterment.

First of all I will like to thank my supervisor Dr. Debasisa Mohanty for his inspiring contribution for the genesis of this thesis both in terms of intellectual input and encouragement. During the time of need, I have always found him patient and supportive.

I will also like to thank Prof. A. Surolia, Director of NII for providing the infrastructure for conducting this research. I also thank all the administrative and technical staff of the institute.

I will also like to thank current and former members of Bioinformatics Laboratory. I thank staff of Bioinformatics Laboratory particularly for the help they have given for timely completion of thesis.

Lastly but truly and honestly I will like to express my gratitude for the teachings of Gautam Buddha and Lord Krishna which enabled me to endure though times.

Abbreviations

ADT	AutoDock Tools
AMBER	Assisted Model Building with Energy Refinement
BLAST	Basic Local Alignment Search Tool
COB	Chloroorientacin B
DVV	DesVancosaminyl Vancomycin
GB	Generalized Born
GTr	Glycosyltransferase
H-bond	Hydrogen Bond
HBP	Halogen Binding Pocket
HMM	Hidden Markovs Models
HTS	High Throughput Screening
LBR	Ligand Binding Regions
LGA	Lamarckian genetic algorithm
MSA	Multiple Sequence Alignment
NDP	Nucleotide DiPhosphate
NP	Natural Products
NRPS	Non-Ribosomal Peptide Synthetase
NSS	Neurotransmitter Sodium Symporters
PB	Poisson Boltzmann
PDB	Protein Data Bank
PKS	Polyketide Synthase
PME	Particle Mesh-Ewald
RMSD	Root Mean Square Deviation
SAH	s-adenosylhomocysteine
SDRs	Specificity Determining Residues
VHTS	Virtual High Throughput Screening

Introduction

The term 'Natural Product' refers to drug like small molecules produced by living organisms with proved or potential application as a medicine in treatment of diseases. Majority of these small molecules are secondary metabolites produced by plants and microorganisms like bacteria and fungi. Enzymes involved in biosynthesis of these secondary metabolite natural products (NPs) can be categorized in two types based on stage of biosynthesis at which they act. Non Ribosomal Peptide Synthases (NRPSs) and PolyKetide Synthases (PKSs) are huge multifunctional modular enzymes, which act at first stage of biosynthesis and produce secondary metabolite scaffolds from simple building blocks. In second stage of biosynthesis, these products of NRPS/PKS enzymes are extensively modified by stand-alone tailoring enzymes specialized in catalyzing different reactions like methylation, epimerization, hydroxylation, heterocyclization, oxidative cross-linking, halogenation and glycosylation. These tailoring enzymes impart additional diversity and in many cases impart biological activity to NP through introduction of key functional groups.

Natural products have been prolific source of new drug candidates. According to recent reviews, 34% of all small-molecule drugs are natural products or their direct semi-synthetic derivatives. The percentage of natural products is even more in case of drugs that are used to treat severe or life threatening diseases. 47% of small molecule anticancer drugs are derived from NPs and remarkably more than 75% of all approved antibacterials are NPs or their semi-synthetic derivatives (1). Literature survey reveals that focus of current drug discovery effort is shifting from synthetic combinatorial libraries of small molecules as a source of new drug entities to NPs as a source of new drug entities. There are two primary causes for this shift. First is the failure of high throughput in-vitro/in-silico methods based on synthetic combinatorial libraries to yield new drug candidates and second is the emerging conscience that NPs are privileged molecules as they have evolved with biology and thus belong to the biologically and medicinally relevant chemical space. Hence, it has been proposed that emphasis on natural products and better design of compound libraries instead of random screening of chemical libraries may help overcome shortcomings of high throughput drug discovery efforts (2).

But NPs is still a largely unexploited source of new drugs. This is because of difficulties in uncovering and obtaining them from natural resources and also because of their complexity, which makes it difficult to synthesize them in laboratory on a commercial basis. Analysis of known natural product biosynthetic pathways and harnessing biosynthetic potential of different enzymes involved in production of NPs by

2

biosynthetic engineering approaches will allow for abundant supply of diverse NPs. Major advances in biosynthetic engineering during the last decade have demonstrated the feasibility of obtaining novel engineered natural products (3). Availability of complete genome sequences of increasing number of organisms has also opened up the possibility of discovering novel natural products by genome mining. Continued sequencing of new genomes of different microorganisms with biosynthetic potential has opened up possibility of identifying new secondary metabolite biosynthetic clusters from these genomes by knowledge based bioinformatics approaches so as to expand the repertoire of this biosynthetic machinery and of NPs they produce. As structures of some biosynthetic enzymes became available, web-based tools evolved alongside expanding genomic data to correlate sequence and structural data using approaches like homology modeling, fold prediction, threading so as to find out specificity determining regions and residues of these biosynthetic enzymes as such data will help biosynthetic engineers in developing enzymes with higher activity and broader substrate specificities. Modified enzymes thus obtained have potential use in building natural product based compound libraries, which can overcome pitfalls of synthetic combinatorial libraries of not being biologically relevant. However, for further progress in natural product based drug discovery efforts, it is also necessary to develop powerful computational methods which can help in discovery of new NPs by genome mining and reprogramming of known biosynthetic pathways for producing rationally designed NPs.

In this thesis, we have used a combination of bioinformatics and computational chemistry approaches to develop and standardize powerful computational methods which can help in discovery of new NPs by genome mining and reprogramming of known biosynthetic pathways for producing novel NPs by rational manipulation of key tailoring enzymes. Several pharmaceutically important secondary metabolite natural products biosynthesized by nonribosomal peptide synthetases (NRPS) and polyketide synthases (PKS) are often extensively modified by specialized tailoring enzymes for imparting additional structural diversity and biological activity through introduction of key functional groups (4). Glycosyltransferases and oxidoreductases are two important families of tailoring enzymes involved in altering structural and functional diversity of PKS/NRPS family of natural products. We have carried out genome mining studies for identification of novel glycosyltransferases involved in tailoring of secondary metabolite NPs. Detailed structural analysis involving docking and molecular dynamics (MD) simulations have been carried out for representative glycosyltransferases and cytochrome

P450 monooxigenase family of oxidoreductases for understanding structural basis of their substrate preference and identifying crucial specificity determining residues.

Chapter 1 of the thesis gives a very brief review of literature on power and utility natural product based drug discovery and various tailoring enzymes which can be exploited in natural product based drug discovery. Detailed review of literature on each of these tailoring enzymes is given in the subsequent chapters. **Chapter 2** describes a computational protocol for identification of novel tailoring enzymes by genome mining and demonstrates this computational approach using antibiotic glycosyltransferase as a test case. **Chapter 3** reports structure based analysis involving docking and MD simulations on few important antibiotic glycosyltransferases from vancomycin family for understanding structural basis of donor substrate selection by these enzymes. **Chapter 4** describes structure based analysis which provides novel insights for substrate selection by Cytochrome P450 monooxygenase family of oxidoreductases involved in hydroxylation of secondary metabolites. **Appendix** describes computational structure based studies performed on structural models of human neurotransmitter transporters based on crystal structure information of related prokaryotic tranporter.

Chapter 1
Review of Literature

Review of Literature gives basic ideas and current trends in research on which this thesis is based and gives brief introduction to themes that form thesis chapters. Detailed literature review, which concerns individual chapters, is covered in introduction section of each chapter.

1.1   Modern Trends in Drug Discovery

Advances in modern biology are providing insights into molecular mechanisms underlying diseases, suggesting rational protein targets for therapeutic intervention. Thus modern drug discovery efforts are largely focused on finding selective small molecule modulators of activity of these target proteins as such molecules have potential to treat diseases. Thus *in vitro* and *in silico* high throughput screening (HTS) approaches have evolved to screen small molecules against target proteins to find potential binders. In-vitro HTS approaches are more direct methods in which small molecules are screened against actual target proteins using techniques of molecular biology(Gidrol *et al.*, 2009). *In silico* methods are indirect cost effective methods based on 3D structures of target proteins(Jorgensen, 2004). Thus prerequisite for these computational methods are reliable crystal structures or structural models of target proteins. Different protein ligand docking tools can then be used to dock small molecules into protein active site. These tools can simulate binding of different small molecules in protein active site and can rank them according to their computational binding affinity in virtual (*in silico*) high throughput screening (VHTS) so as to differentiate between binders and non-binders(Sousa *et al.*, 2006).

1.2   Synthetic Combinatorial Libraries

In order to support advances in *in vitro* HTS and *in silico* VHTS large libraries of synthetic compounds are created using approaches like combinatorial chemistry to supply small molecules for these powerful drug screening methods(Mason *et al.*, 2001). Lipinski's rule of five for drug-likeness was proposed in response to large number of randomly made compound libraries mostly due to feasibility of chemical synthesis. It was an attempt to rationalize compound library design so as not to make too polar, floppy or large molecules, which have lower chance of exhibiting desired pharmaceutical properties and oral bioavailability(Keller *et al.*, 2006). This "rule of five" states that high absorption or permeability of a compound is more likely when there are less than 5 hydrogen bond (H-bond) donors, molecular mass is below 500 daltons, calculated log P is less than 5 and

6

sum of nitrogen and oxygen atoms in a molecule is less than 10. Thus combinatorial libraries are developed under the guidelines of these empirical rules to satisfy demands of HTS methods.

But the downside of adhering to 'rule of five' and ease of synthesis directed combinatorial-library is that it excludes NPs, as they are usually exception to this rule. Also, NPs are more difficult to synthesize and their semi-synthetic derivatives are usually bulkier than parent molecule. For these reasons, synthetic molecules produced by combinatorial chemistry got preference over NPs. Hence exclusion of NPs and their derivatives and deviation of synthetic libraries from NPs which represent biologically relevant chemical space, is considered as major reason for surprisingly small success of major drug discovery programs despite the use of powerful high throughput screening approaches(Shang & Tan, 2005; Zhang & Wilkinson, 2007b).

1.3   Natural Products as Alternative Drug Source

It has been suggested, that a major drawback of using large synthetic libraries is lack of desired diversity needed to cover all possible chemical space. Natural products offer abundant pool of distinct molecular frameworks. Hence they can be ideal starting points for molecular design considerations. Some of the distinctive features of NPs over synthetic drugs are presence of on an average twice as many oxygen atoms and three times lower nitrogen atoms and slightly higher number of hydrogen-bond donors. Further natural products contain approximately four times more chiral centers and far fewer aromatic rings. This fact may reason for natural products higher selectivity when binding to stereo-defined sites. Thus NPs represent the prolific source for identification of novel structural scaffolds. Among the FDA-approved new chemical entities introduced from 1981 to 2002, 49% have natural product origin or are derived from them using computer-based design (Balamurugan et al., 2005; Grabowaski & Schneider, 2007).

Biological relevance of a combinatorial library is also an important issue. NPs are made by biosynthetic enzymes and are evolved to bind and act on protein targets. Therefore NPs are considered biologically more relevant. Results from structural biology and bioinformatics efforts indicate that the number of distinct protein families and folds are fairly limited. Often, same structural domain is used by many proteins, in a more or less modified form, created by divergent evolution (McArdle & Quinn, 2007). Currently SCOP classification contains only 1195 distinct folds, 1962 superfamilies and 3902 families, as compared to total 16014672 entries in TrEMBL protein sequence database

(Andreeva *et al.*, 2007). As a result there is good probability of protein target of interest being consisted of similar building block and having similar structural domain to one of the proteins to which some NP is evolved to bind. Hence combinatorial library made from NPs will have higher probability of containing a drug lead as compared to simple synthetic molecules as scaffold of one NP may preferentially bind to another protein with similar fold and active site architecture to that of its cognate protein (McArdle & Quinn, 2007). Recent experimental studies have demonstrated that, proteins with similar topology and fold tend to bind similar ligands (Koch & Waldmann, 2005). Thus NPs are privileged structures and they are distinct from synthetic molecules as mentioned earlier, this could probably be the reason for limited success of high throughput efforts with synthetic combinatorial libraries. Therefore, if combinatorial libraries derived from NPs are used for high throughput screening against protein targets, chances of finding a drug lead may be higher.

1.4   Biosynthetic Engineering and tailoring enzymes

But one of the major obstacles for NP based drug discovery is their structural complexity and low synthetic feasibility. Total synthesis of complex NPs usually requires many steps and may not be scalable and environment friendly. For example, total synthesis of rapamycin requires more than 50 steps and gives an overall yield of less than 0.5% (Zhang & Wilkinson, 2007a). Therefore, biosynthetic engineering approach has been used for obtaining these products from genetically engineered microorganisms. Genetic modifications are made to gene clusters involved in biosynthesis of NPs to produce analogs with pre-designed structural modifications as in case of synthetic combinatorial library (Baltz, 2006; Gregory *et al.*, 2005). Hence biosynthetic engineering is important not only in terms of obtaining NPs but it also has a potential to be used for generating diverse derivatives of cognate NPs, which will have application in lead optimization.

These NPs are usually secondary metabolites biosynthesized by various microbial, fungal or plant organisms. Polyketide synthases (PKSs) and non-ribosomal peptide synthases (NRPSs) are huge multifunctional modular enzymes involved in production of many of these diverse secondary metabolite scaffolds. Therefore, most biosynthetic engineering studies for obtaining novel natural products have been carried out by reprogramming of PKS or NRPS biosynthetic machinery (Baltz, 2006).

The products of these NRPS/PKS enzymes are often extensively modified through methylation, epimerization, hydroxylation, heterocyclization, oxidative cross-linking, halogenation, and glycosylations. Specialized tailoring enzymes perform these modifications. These enzymes are responsible for additional structural diversity and for imparting biological activity to NP through introduction of key functional groups (Rix *et al.*, 2002). Recent experimental studies have demonstrated the feasibility of altering structural diversity of PKS/NRPS family of natural products by rational manipulation of tailoring enzymes like glycosyltransferases and acyltransferases (Baltz, 2006). Tailoring enzymes constitute mainly the groups of oxidoreductases and transferases. Group transferases catalyze transfer reactions and introduce novel functional groups often having both new reactivity and altered sterio-electronic profiles. This large group of enzymes contains important subgroups as methyltransferases, acyl transferases and glycosyltransferases (GTrs). GTrs are responsible for attachment of sugar groups to the aglycone core. Crystal structure and biochemical information is available for some GTrs (Mulichak *et al.*, 2004; Oberthur *et al.*, 2005). In particular, information is available for the GTrs involved in the biosynthesis of antibiotic vancomycin.

1.4.1 Glycosyltransferases

GTrs are involved in biosynthesis of important antibiotics like erythromycin, doxorubicin, vancomycin etc. They catalyze transfer of activated sugars from NDP-sugar donors to varied acceptor molecules of NRPS-PKS origin to produce active antibiotics. The site of glycosylation, nature of sugar and the number of sugars affect the efficacy of these antibiotics (Schlunzen *et al.*, 2001). The NDP-sugar donors for GTrs are typically TDP-hexoses.

Different modifications of the functional groups on hexoses can lead to enormous variations in donor substrates. Combined with probable variations in acceptor molecules, diverse derivatives of these well known antibiotics can be produced. Several biosynthetic engineering studies have attempted to design novel antibiotics against microbes that are quickly developing resistance to present day antibiotics (Baltz, 2006).

For example, in biosynthesis of chloroeremomycin a vancomycin analogue, glycosyltransferase GtfA transfers 4-*epi*-L-vancosamine to benzylic hydroxyl of amino acid 7 of the vancomycin pseudoaglycone 1 to produce chloroorienticin B (**Figure 1**). GtfC then transfers 4-*epi*-L-vancosamine to the glucose C-2 hydroxyl of compound 2. GtfD is part of the vancomycin biosynthetic cluster and transfers L-vancosamine to the

9
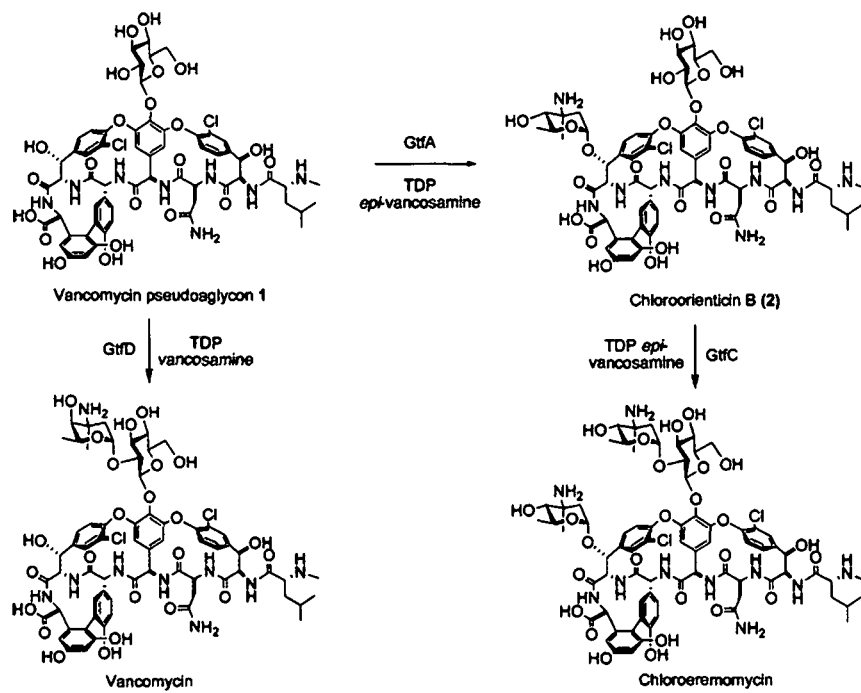
Figure 1: Action of glycosyltransferases GtfA, GtfC, GtfD involved in the biosynthesis of glycopeptide antibiotics

glucose C-2 hydroxyl of pseudoaglycone **1**. Recently it has been shown that GtfD and GtfE are more flexible in donor and acceptor specificities than other glycosyltransferases like GtfA and GtfB. GtfE glycosylates aglycone and GtfD glycosylates monoglycosylated derivatives. As a result this tandem combination is ideal for combinatorial glycosylations of glycopeptide aglycone. These enzymes have been successfully utilized in generating diverse analogs of vancomycin (Baltz, 2006; Oberthur *et al.*, 2005).

Therefore, understanding the specificity, mechanism of action and key specificity determining residues of glycosyltransferases will be an important step forward for harnessing immense metabolic power of these enzymes for introducing diverse modifications on important secondary metabolites.

1.4.2 Oxidoreductases

These are very broad group of enzymes consisting of dehydrogenases, oxygenases, oxidases, peroxidases and reductases. These enzymes either introduce oxygen containing functionalities, such as hydroxyl groups (hydroxylases), aldehyde or keto groups, and epoxides (epoxidases) or modify these functionalities by addition or removal of hydrogen atoms like dehydrogenases (enzymes that oxidizes a substrate by transferring one or more protons and a pair of electrons to an acceptor, usually NAD/NADP or a flavin coenzyme such as FAD or FMN) (Rix *et al.*, 2002). Most common oxidoreductases are Cytochrome P-450 monooxygenases (EryF, EryK, EpoK) (Cupp-Vickery & Poulos, 1995), anthrone oxygenases (Sciara *et al.*, 2003), flavin-dependent mono and dioxygenases (Funa *et al.*, 2005) and recently crystallized aromatic hydroxylases involved in angucycline biosynthesis (Koskiniemi *et al.*, 2007). Very less information is available about these tailoring oxidoreductases.

Importance of oxidoreductases comes from their ability to catalyze oxidation of non-reactive C-H bonds. Hence these enzymes can play a crucial role in generating diverse analogues of NPs. Three such potential enzymes are EryF, EryK and PikC which are involved in biosynthesis of macrolide type of antibiotics(Cupp-Vickery & Poulos, 1995; Savino *et al.*, 2008; Xue *et al.*, 1998). EryF, EryK are part of Erythromycin biosynthetic machinery while PikC is more versatile and involved in biosynthesis of two antibiotics namely Pikromycin and Methymycin (**Figure 2**).

All three enzymes add hydroxyl group on their substrate at specific positions. Crystal structures are available for all three enzymes in complex with their substrates along with specificity data from biochemical studies. Thus just as GTrs GtfD, GtfA and
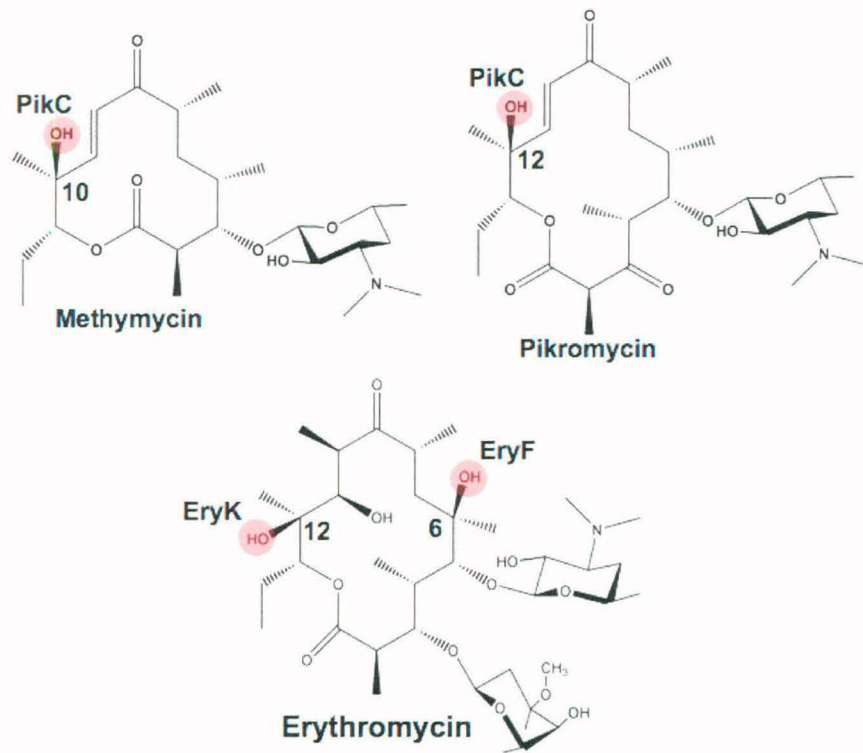
Figure 2 Schematic diagrams of Methymycin and Pikromycin and Erythromycin. PikC is involved in biosynthesis of Methymycin and Pikromycin while EryF and EryK are involved in biosynthesis of Erythromycin. All enzymes perform hydroxylations. Position of hydroxylation is highlighted and enzyme responsible is indicated near it.

GtfC form ideal candidates for detail structural analysis of Glycosyltransferase class of tailoring enzymes due to availability of structural and specificity data, EryF, EryK and PikC make up ideal candidates for detail structural studies on oxidoreductase class of enzymes.

1.5 Computational Methods for Identification and Categorization of Biosynthetic Enzymes

Over the past few decades explosion of genomic and structural data lead to development of various web-based tools for identification and characterization of biosynthetic enzymes so as to allow optimal use of this natural enzyme pool in NP biosynthesis.

Example of such efforts is NRPS-PKS a web-based software for analyzing large multi-enzyme, multi-domain megasynthases involved in biosynthesis of pharmaceutically important secondary metabolite scaffolds(Ansari *et al.*, 2004). It is based on comprehensive analysis of sequence and structural features of experimentally characterized biosynthetic gene clusters. This knowledge base is used for predicting domain organization and substrate specificity of uncharacterized NRPS/PKS clusters. Another example of evolving web-based tools is SEARCHGTr for analysis of Glycosyltransferases (GTrs), which form an important class of tailoring enzymes(Kamra *et al.*, 2005). This software has been developed on the basis of comprehensive analysis of 102 GTrs of known specificity from 52 NP biosynthetic gene clusters. This software can correlate sequence of GTrs to chemical structures of their corresponding substrates. It also indicates donor/acceptor specificity and identifies putative substrate binding residues. Thus together tools like NRPS-PKS and SEARCHGTr facilitate computational analysis of both upstream enzymes involved in biosynthesis of NP scaffold and downstream tailoring enzymes, in this case GTrs. Thus these web-based tools assist in identification of biosynthetic products of secondary metabolite gene clusters found in newly sequenced genomes and present opportunity for rational design of NPs.

1.6 Predicting substrate specificity of novel biosynthetic enzymes by combination of bioinformatics and computational chemistry

In order to generate structurally diverse biosynthetic combinatorial libraries, it will be of interest to search for various tailoring enzymes in genomes of newly sequenced organisms. Structure based approaches can then be used to identify their natural

substrates and their specificity for other ligands and hence applicability in biosynthetic combinatorial library design. It has been shown that structure based models of NRPS adenylation domains can be used to predict amino acids incorporated into putative peptide products, and hence to predict probable final product (Ansari *et al.*, 2004; Stachelhaus *et al.*, 1999).

In contrast to these knowledge based computational methods, some studies have made efforts to identify function and substrates of unknown enzymes by combined use of bioinformatics and energy based docking simulations. These methods can be ideal starting points for studying proteins of unknown function particularly when proteins are unrelated to others of known activity thus making knowledge based bioinformatics inference impossible. Fold prediction and docking studies have been used to find out function of unknown proteins. In one of them, the enzyme structure was modeled by threading the sequence of known structure of a homologous protein (35% sequence identity), and automated ligand docking was used to successfully predict the enzyme substrate and assign function to enzyme (Song *et al.*, 2007). Example of such work is prediction of activity of Tm0936 by Hermann *et al.* They have docked high-energy intermediates of thousands of candidate substrates to identify s-adenosylhomocysteine (SAH) as a substrate and C6-deamination as a function of this enzyme. The results were then validated by biochemical analysis in which rate constant for SAH deamination was found out to be significant thus confirming the results of docking. Further they have solved crystal structure of enzyme with S-inosylhomocysteine, deaminated metabolite of SAH to corroborate binding pose of SAH obtained through docking.

In another study, Song *et al* (Hermann *et al.*, 2006; Hermann *et al.*, 2007) proved application of docking studies in identifying possible substrates for protein of unknown activity. Library of potential ligands were docked on to homology model of protein built from most similar characterized protein (35% sequence identity). Docking results matched closely with enzymatic essays based on same potential substrates. Experimentally determined structure confirmed predicted protein-substrate complex. New function was successfully assigned to the enzyme i.e. N-succinyl arginine/lysine racemase correcting earlier mis-annotation.

1.7   Need for New Methodologies

As described earlier, knowledge based and structure based approaches can help in identifying substrates for newly found biosynthetic enzymes and their specificity

determining residues. Such studies thus can aid biosynthetic engineers in developing enzymes with desired substrate specificity.

But as different biosynthetic enzymes are studied further complexities of enzyme-substrate interactions are becoming apparent. For example, in case of tailoring GTrs of vancomycin biosynthetic pathway, biochemical studies have proved that GtfD transfers sugar vancosamine from donor TDP-vancosamine on to acceptor DVV. In chloroeremomycin biosynthesis (another antibiotic of vancomycin class) GtfA transfers sugar *epi*-vancosamine from TDP-*epi*-vancosamine on to DVV. Biochemical studies have also shown that GtfD is capable of transferring both sugars vancosamine and *epi*-vancosamine at high catalytic rates, while GtfA can transfer only its cognate sugar epi-vancosamine that too at a very low catalytic rate compared to GtfD(Oberthur *et al.*, 2005). This biochemical data expose the limitations of available knowledge based methods as these methods can only imply possible substrate and substrate binding residues but are of little help in generating a detail map of enzyme-substrate interactions which can help in understanding strategies evolved by enzymes to select its substrates from similar contaminants. This necessitates development of novel computational methods which can take into account all available information about specific class of enzymes like sequence, structure and biochemical information to give in-depth understanding of their substrate specificity profiles.

In this thesis, we have attempted to develop novel structure based methodology for detailed analysis of enzyme-substrate interactions and deciphering mechanism of substrate selection. Such methodology development needs a group of well studied related enzymes. This computational approach has been tested on two important class of tailoring enzymes involved in biosynthesis of secondary metabolite natural products.

Chapter 2

Genome Mining for Identification of New Tailoring Enzymes

2.1 INTRODUCTION

Natural products (NPs) produced by microorganism and plants are proving to be valuable source of new drug entities compared to synthetic compounds produced in laboratory by organic chemists(Newman & Cragg, 2007). Common reasoning given for this fact is, as living organisms produce natural products, these NPs have evolved to interact with living organisms and hence have more biological significance than synthetic molecules(McArdle & Quinn, 2007). As these NPs have complex chemical structures compared to synthetic organic molecules, only feasible way of obtaining them in commercial quantities is to harness the biosynthetic potential of the corresponding enzymes, which produce them in various organisms(Baltz, 2006).

Many of these NPs are produced by large multi-enzymes systems like nonribosomal peptide synthetases (NRPSs) and polyketide synthases (PKSs). These multi-enzyme complexes produce distinct scaffolds, which are further modified by stand-alone enzymes called tailoring enzymes, which add different functional groups to these basic scaffolds to complete their biosynthesis(Rix *et al.*, 2002). First step towards exploiting biosynthetic potential of these enzymes is to identify such enzymes from newly sequenced genomes and to group them according to their function and specificity(Ansari *et al.*, 2004). This will help in mapping the diversity of available natural products and thus the potential of various microorganisms to yield new drug candidates. Such analysis can also potentially enable scientists to understand biosynthetic space to which each enzyme belongs so as to enable maximum exploitation of biosynthetic potential of each enzyme.

In this work an attempt has been made to identify, classify and cluster biosynthetic enzymes from newly sequenced genomes as an initial step and carry out detailed bioinformatics analysis to understand their substrate specificities. The knowledge-base generated by such *in silico* analysis can be used to exploit enormous biosynthetic potential of enzymes involved in biosynthesis of NPs, which have potential as future medicines. Even though the computational protocol used in the current study is applicable to tailoring enzymes of natural product biosynthetic pathways in general, in this work glycosyltransferase has been chosen as prototype to demonstrate the power and utility of the genome mining approach. Glycosyltransferases (GTrs) constitute one of the most important classes of tailoring enzymes. These enzymes transfer sugars from sugar donors on to natural product acceptors which are typically scaffolds made by PKSs and

NRPSs. These transferred sugars moieties are important determinants of bioactivity and physiochemical properties of NPs. As most members of these GTrs are involved in biosynthesis of NPs with antibiotic activity these GTrs are also named as antibiotic glycosyltransferases(Kamra *et al.*, 2005).

This study has attempted to identify the putative antibiotic glycosyltransferases present in various genomes by *in silico* analysis and decipher their donor and acceptor specificities. In an earlier study Kamra *et al* have compiled the donor and acceptor specificities of 102 GTrs involved in biosynthesis of total 52 different antibiotics and developed GTrDB, a database on sequence, structure and substrate specificities of glycosyltransferases(Kamra *et al.*, 2005). In GTrDB these 52 antibiotics are clustered further into 20 acceptor families based on structural similarity of acceptor aglycon. Current study uses information from GTrDB as the starting point for this analysis, further expands the knowledge-base by including curated information from published literature on experimentally characterized glycosyltransferases and subsequently uses the expanded knowledge-base for identifying GTrs from newly sequenced genomes and assigning their putative substrate specificities.

2.2   METHODS

The computational protocol for identification of novel glycosyltransferases by genome mining is schematically illustrated in **Figure 1**. The overall protocol consists of three major steps, namely, 1. Identification of experimentally characterized GTrs of known donor and acceptor specificities and clustering them into different acceptor families, 2. Building Hidden Marcov Model (HMM) profiles for each of the acceptor families and 3. Profile HMM searches for identification of putative GTrs belonging to different acceptor families from among experimentally uncharacterized sequence in various genomes. A brief overview of each of these three steps is given below.

2.2.1   Identification of experimentally characterized GTrs and clustering them into different acceptor families

The obvious approach for compilation of experimentally characterized GTrs of known donor and acceptor specificities would be to search the published scientific literature on experimental characterization of antibiotic GTrs and obtain from that the accession numbers for sequences of GTrs, chemical structures of donors and accepts etc. However, such an approach would be entirely manual and would be time consuming.

Figure 1 Protocol for identification and categorization of new GTrs by genome mining.

Collect sequence of 102 characterized
GTrs from GTrDB database

↓

Use each sequence as a query in BLAST against
NCBI non-redundant protein sequence data base

↓

Search for newly characterized GTrs in BLAST results
using accession numbers to find out GTrs with published
literature. Gather donor and acceptor specificity data for
new enzymes

↓

Classify new GTrs along with 102 known GTrs in different
groups based on similarity of acceptor molecules

↓

Build HMM profiles using HMMER for each class of
GTrs and search TrEMBL database of protein
sequences using these HMM profiles to identify and
categorize GTrs from newly sequenced genomes

Therefore, attempt was made to develop a fast and semi-automated approach by which most experimentally characterized GTrs can be easily identified.

The sequences of GTrs from GTrDB with known acceptor and donor specificities are taken as a starting point for the search of new GTrs by using BLAST program of NCBI. GTrDB contains 102 GTrs. These GTrs, in all, transfer sugars on to 52 different antibiotics. In GTrDB these 52 antibiotics are clustered further into 20 acceptor families based on similarity in the chemical structures of acceptor aglycon. Sequences of each of these 102 GTrs are used as query protein in BLAST search against NCBI non-redundant protein sequence database. BLAST searches were carried out with an E-value cut off of 10^{-6}. The accession numbers of matching GTrs for each of these queries were analyzed to check if they were linked to literature in PUBMED describing experimental characterization. The PUBMED IDs for the matches were compared with PUBMED IDs of 102 GTrs in GTrDB to eliminate matches already cataloged in GTrDB and only new entries were retained. This way 52 new GTrs involved in tailoring of 36 different new antibiotics (not listed in GTrDB) were identified. Chemical structures of the donor and acceptor moieties for each of these 52 new GTrs were obtained from the respective PUBMED articles. Detailed analysis of the chemical structures of acceptor groups indicated that, GTrs involved in tailoring of these 36 new antibiotics can be further grouped into 20 acceptor families. 7 of these 20 belong to acceptor families listed in GTrDB. 13 acceptors families were not listed in GTrDB and hence form new classes of acceptor families. Thus addition of these newly identified GTrs of known specificity to GTrDB knowledge-base resulted in a dataset of 154 GTr sequences belonging to 33 different acceptor families.

2.2.2 Building Hidden Markov Model (HMM) profiles for different acceptor families

Next task was to build acceptor family specific HMM profiles for glycosyltransferases belonging to each of the 33 acceptor families. It may be noted that, the clustering of GTrs into different acceptor families have been carried out based on chemical structure similarity of acceptor groups. Even though earlier studies suggest that, GTrs belonging to a given acceptor families in general show significant similarity in sequence, there might be cases when GTrs lacking significant sequence similarity would also be transferring donor groups to similar acceptor moieties. Such cases are likely to pose certain difficulties while building HMM profiles, because to build HMM profiles,

GTrs being considered from one group should have some minimum level of sequence identity and similarity. But this was not the case with certain acceptor groups particularly when GTrs are transferring sugars on same or similar acceptors but at different positions on acceptor. In such cases, either one group was divided into different sub-groups (e.g. vancomycin-I and II; macrolide I and II) or no HMM profiles are built for the respective acceptor group or sub-group if number of GTrs belonging to them was too small (e.g. Bislactone-II, Orthosomycin-II and III, Aminoglycoside-II etc). HMMbuild module of HMMER software version 3.0 was used for building HMM profiles(Eddy, 1998). Multiple sequence alignment (MSA) of the training set of sequences, which were used to build HMM profile for a given acceptor family, was obtained from the CLUSTALW software and the MSA was given as input to the HMMER package.

In some cases GTrs from one acceptor group showed significant sequence similarity with GTrs of other acceptor groups. In such cases again two approaches were taken. If number of GTrs in each of these acceptor families were large enough, separate profiles were built for each group. There were 14 such groups. On the other hand, if the different acceptor families showing significant similarity had less number of sequences in either group or both, they were merged to form a single group. There were four such groups. But in no case same GTr is repeated. Thus mixed group contains GTrs distinct in sequence to those, which form original acceptor group of the same class and incorporated in mixed group as they showed higher similarity to other acceptor classes compared to GTrs forming original acceptor group. For eight acceptor families no HMM profile could be built due to lack of adequate number of sequences and divergence even from the mixed groups.

2.2.3 Profile HMM searches for assigning acceptor families to experimentally uncharacterized GTrs

The HMM profiles for 18 different acceptor families were used to search the TrEMBL protein database to identify and cluster glycosyltransferases present among experimentally uncharacterized sequences of various genomes. HMMsearch and HMMpfam modules of HMMER package were used for this purpose and an E-value cut off of 10^{-25} was used classifying a match as statistically significant.

2.3 RESULTS

2.3.1 Identification of experimentally characterized GTrs and clustering them into different acceptor families

As discussed earlier in the methods section, the first step of our computational protocol involved enhancing the knowledge-base of GTrDB by identifying newly characterized GTrs. Instead of manual search of published literature, a semi-automated protocol involving BLAST search was employed. Sequences of well-characterized GTrs from GTrDB were used as query against NCBI non-redundant protein sequence database and matches having PUBMED links to manuscripts describing experimental characterization were identified. Manual curation of the corresponding publications helped in identification of donor and acceptor specificities of the respective newly characterized GTrs which were not cataloged in GTrDB. These newly identified GTrs were grouped into different acceptor families by visual comparison of the chemical structures of their acceptor substrates. **Table 1** gives list of 52 newly found GTrs along with 102 GTrs from GTrDB classified according to acceptor families. As can be seen from **Table 1**, out of these 52 newly identified GTrs, 26 GTrs belong to seven acceptor clusters already present in GTrDB, while the remaining 26 could be grouped into 13 new acceptor families (highlighted in blue color in **Table 1**). **Figure 2a, b** and **c** gives schematic diagrams of antibiotics or secondary metabolites for different acceptor families present in GTrDB, while **Figure 3a** and **b** shows the chemical structures of antibiotic or secondary metabolites belonging to new acceptor families found by the current BLAST search. As can be seen these molecules are much complex compared to synthetic molecules and many of them are known to be bioactive.

2.3.2 Building Hidden Markov Model (HMM) profiles for different acceptor families and their use in genome mining studies

As discussed earlier, the 102 GTrs present in GTrDB and the 52 newly identified GTrs of known donor and acceptor specificity were grouped into a total of 33 acceptor families based on similarity in the chemical structures of their acceptor moieties. In general GTrs belonging to same acceptor families show a high degree of sequence similarity than those belonging to different acceptor families. This enabled acceptor-based classification of GTrs in to different groups and also forming of meaningful HMMER profiles based on these acceptor groups. However, in some cases it is also

Table 1 gives list of 33 acceptor families arising from 88 acceptor moieties present in different antibiotics or secondary metabolites both. It distributes number of experimentally characterized GTrs in these different acceptor families. Columns 1 and 2 show the results from GTrs in GTrDB, while columns 3 and 4 show the newly identified GTrs by BLAST search and literature curation. New acceptor classes are labeled in blue.

GTrs from GTrDB		Newly found GTrs	
Antibiotic	No. of GTrs	Antibiotic	No. of GTrs
Group 1: Vancomycin Family (NRPS)			
A40296	2		
Chloroeremomycin	3		
Balhimycin	3		
Teicoplanin	3		
Vancomycin	4 (15)		
Group 2: Anthracycline (PKS)			
Aclacinomycin A	2	Aranciamycin	1
Adriamycin	1	Chalcomycin	2
Nogalamycin	3	Steffimycin	1 (4)
Daunomycin	1		
Daunorubicin	1		
Elloramycin	1 (9)		
Group 3: Polyene macrolide (PKS)			
Amphotericin B	1		
CE-108	1		
Candicidin-D	1		
Rimocidin	1		
Nystatin	1		
Pimaricin	1 (6)		
Group 4: 20 membered macrocyclic lactam core (PKS)			
Vicenistatin	1 (1)		
Group 5: 20 membered pentacyclic lactones (PKS)			
Avermectin	1 (1)		
Group 6: Orthosomycin group (PKS)			
Avilamycin	4		
Evernimicin	5 (9)		
Group 7: Hybrid (NRPS-PKS)			
Bleomycin	2 (2)	Tallysomycin	1 (1)
Group 8: Amino-glycoside			
Butirosin	1	Fortimicin A	1
Tobramycin	2	Kanamycin	2
Gentamicin	2	Ribostamycin	1
Streptomycin	1 (6)	Neomycin	1
		Kasugamycin	1 (6)
Group 9: Indocarbazole			
Rebeccamycin	1 (1)	K252a	1
		Staurosporine	1
		AT2433	2 (4)
Group 10: Polyether (modular type I PKS)			

Nanchangmycin	3	(3)			
Group 11: Enediyne (Iterative PKS)					
C1027	1		Maduropeptin	1	
Calicheamicin	4	(5)	Neocarzinostatin	1	(2)
Group 12: Aureolic acid (PKS)					
Chromomycin A3	4				
Mithramycin	4	(8)			
Group 13: Aminocoumarin					
Clorobioin	1				
Novobiocin	1				
Coumermycin A1	1	(3)			
Group 14: Angucycline					
Landomycin A	4		Sch-47554	2	(2)
Landomycin E	3				
Urdamycin A	4				
Simocyclinone	1	(12)			
Group 15: AHBA containing					
Mitomycin C	1	(1)			
Group 16: Macrolide (PKS)					
Erythromycin D	2		Concanamycin A	1	
Megalomicin A	3		Midecamycin	1	
Methymycin	1		Mycinamicin	2	
Pikromycin	1		Spiramycin	3	(7)
Oleandomycin	2				
Tylosin	3				
Spinosyn A	2	(14)			
Group 17: Gilvocarin					
Gilvocarin V	1	(1)			
Group 18: BIQ class (PKS)					
Granaticin B	1				
Medermycin	1	(2)			
Group 19: Pluramycin Iterative Type I PKS					
Hedamycin	2	(2)			
Group 20: Iterative Type II PKS					
Jadomycin B	1	(1) [102]			
Group 21: Aminonucleoside antibiotics					
			A201A	1	
			Hygromycin	1	(2) ·
Group 22: Bislactone (Type II PKS)					
			Chartreusin	2	(2)
Group 23: Spirotetronate Antibiotics					
			Chlorothricin	2	
			Tetrocarcin A	3	
			Kijanimicin	5	(10)
Group 24: Octaenoic acid derivative (Type I PKS)					
			ECO-0501	1	(1)
Group 25: Large linear Type I PKS product					
			ECO-02301	1	(1)

Group 26: Tetraterpenoid Derivative				
		KS-505a	1	(1)
Group 27: Lactonamycinone				
		Lactonamycin	1	
		Lactonamycin Z	1	(2)
Group 28: Acyclic Polyene Antibiotic (Polyenoyltetramic Acid Derivative)				
		α-Lipomycin	1	(1)
Group 29: Cyclopentane containing Antibiotic				
		Pactamycin	1	(1)
Group 30: Phenalinolactone Derivatives (Terpenes)				
		Phenalinolactone A	1	(1)
Group 31: Ansamycin Antibiotic				
		Rubradirin	1	(1)
Group 32: Aminocyclitol containing Antibiotic				
		Validamycin	1	(1)
Group 33: Benzonaphthacenequinone Antibiotic (PKS)				
		Pradimicin	2	(2) [52]

Figure 2 Schematic diagrams of molecules belonging to acceptor families already present in GTrDB. A representative structure of each acceptor class is shown. The serial numbers used for each class correspond to those given in **Table 1** for ease of comparison.

Figure 2a

9) Indocarbazole

10) Polyether

11) Enediyne

12) Aureolic Acid

13) Aminocoumarin

14) Angucycline

15) AHBA Derivative

16) Macrolide

17) Gilvocarin

18) BIQ Class

19) Pluramycin Iterative Type I PKS

20) Iterative Type II PKS

Figure 3 Schematic diagrams of molecules belonging to new acceptor families not listed in GTrDB. A representative structure of each acceptor class is shown. Again the serial numbers used for each class correspond to those given in **Table 1** for ease of comparison.
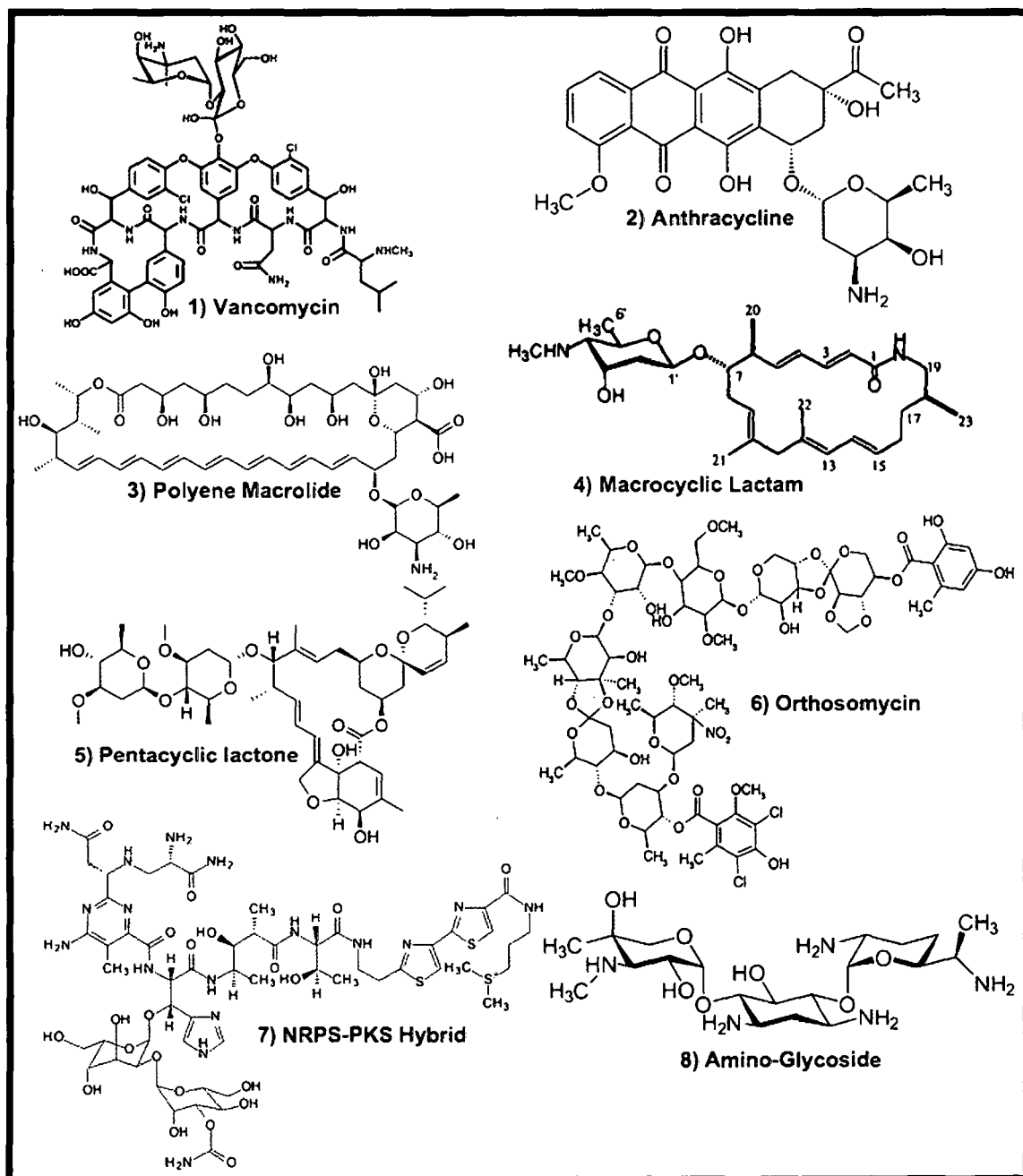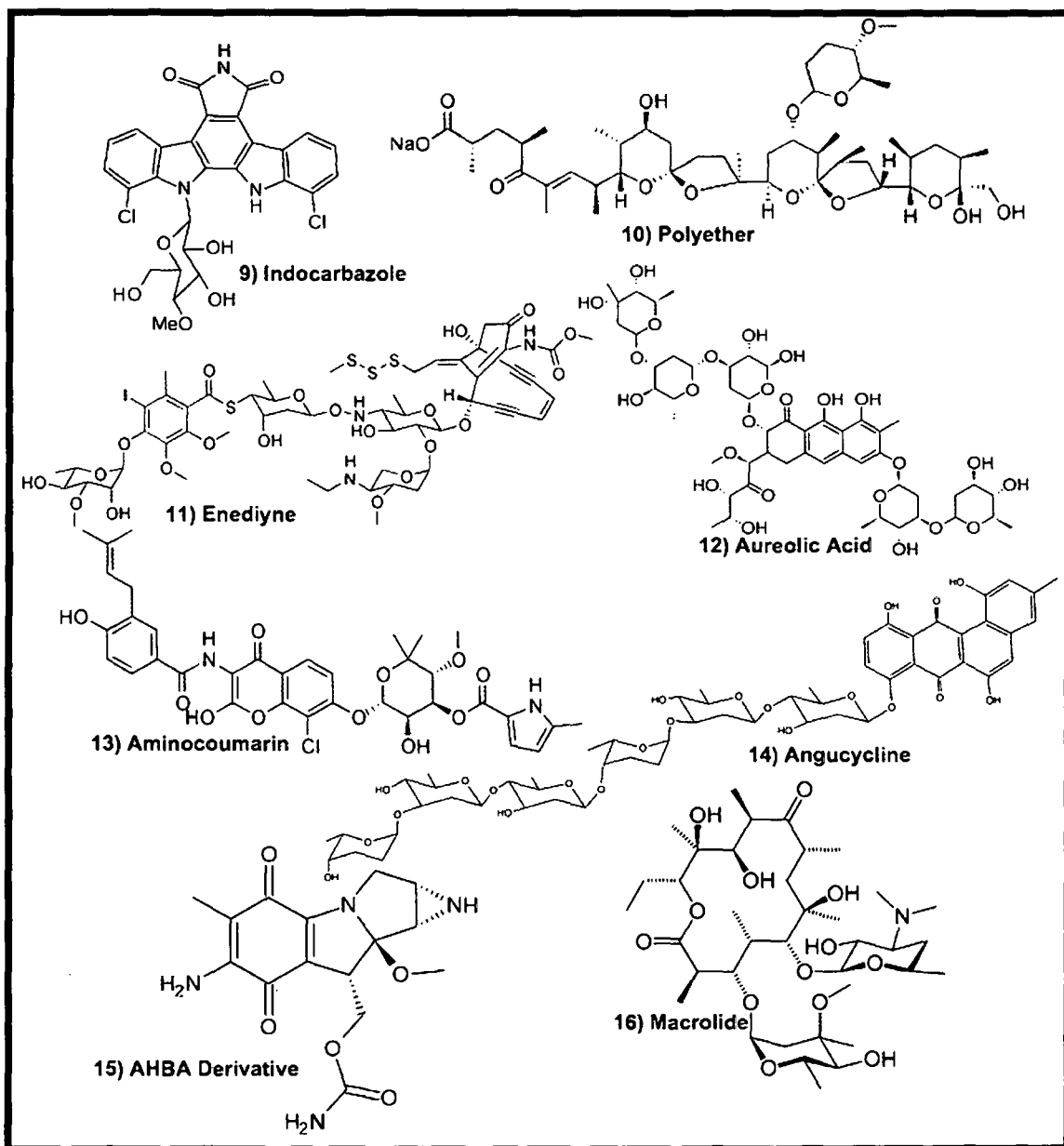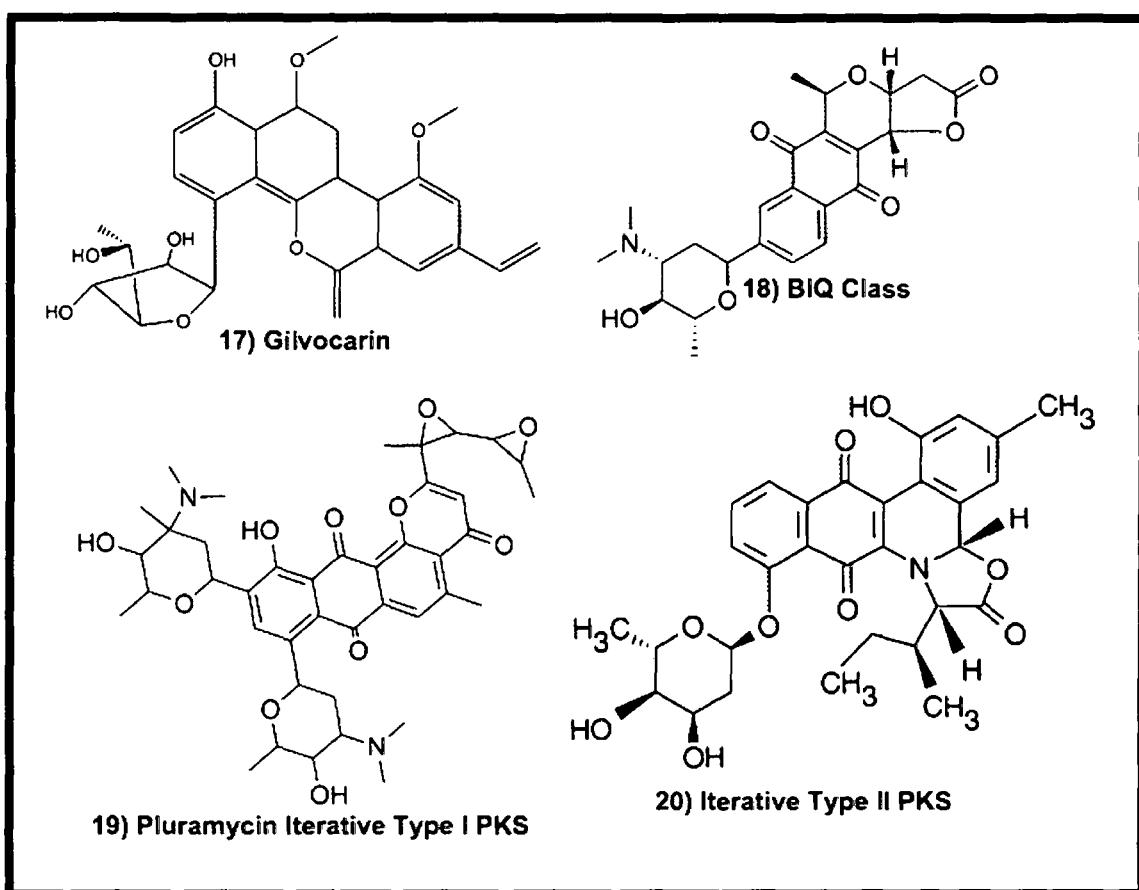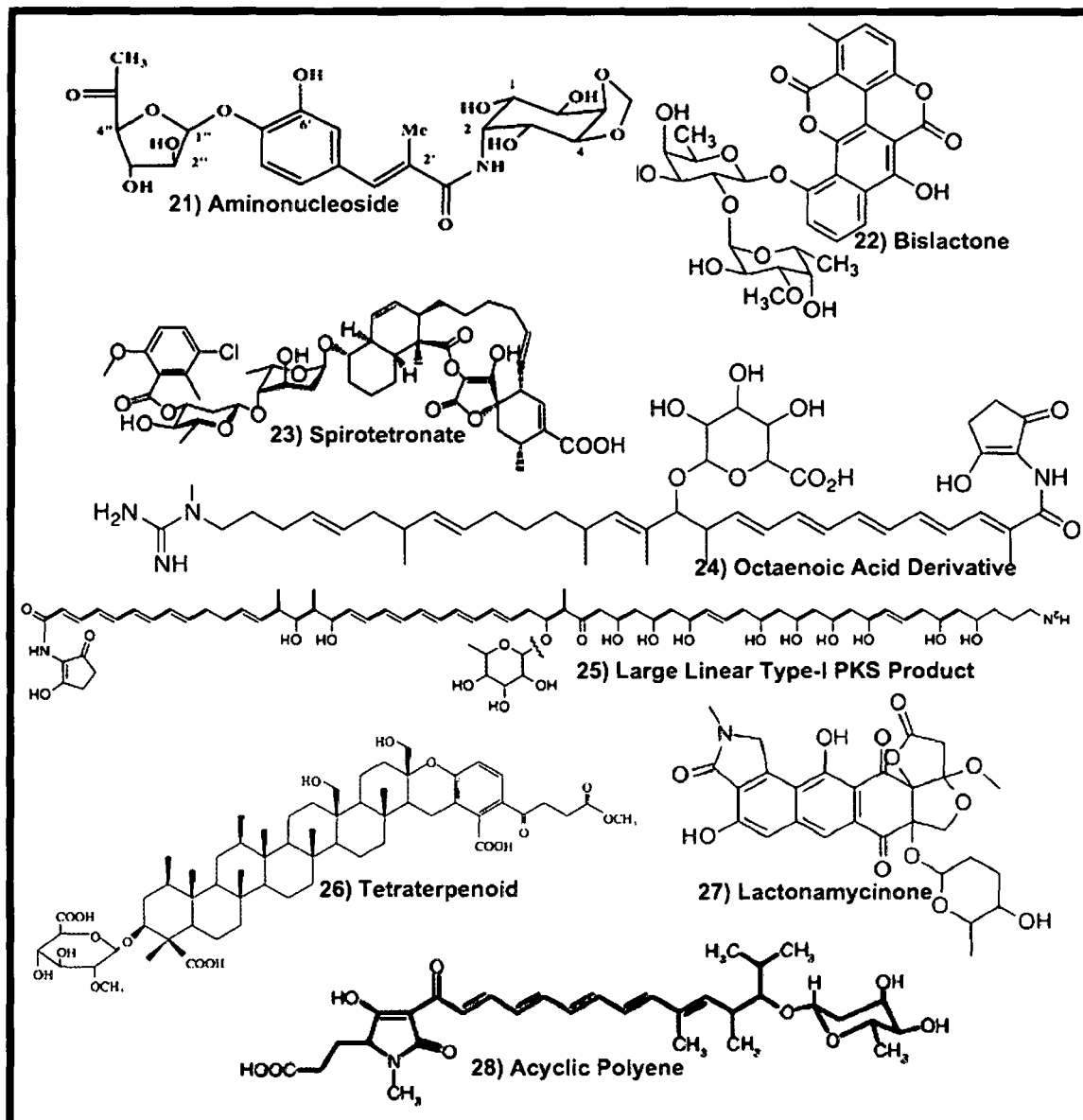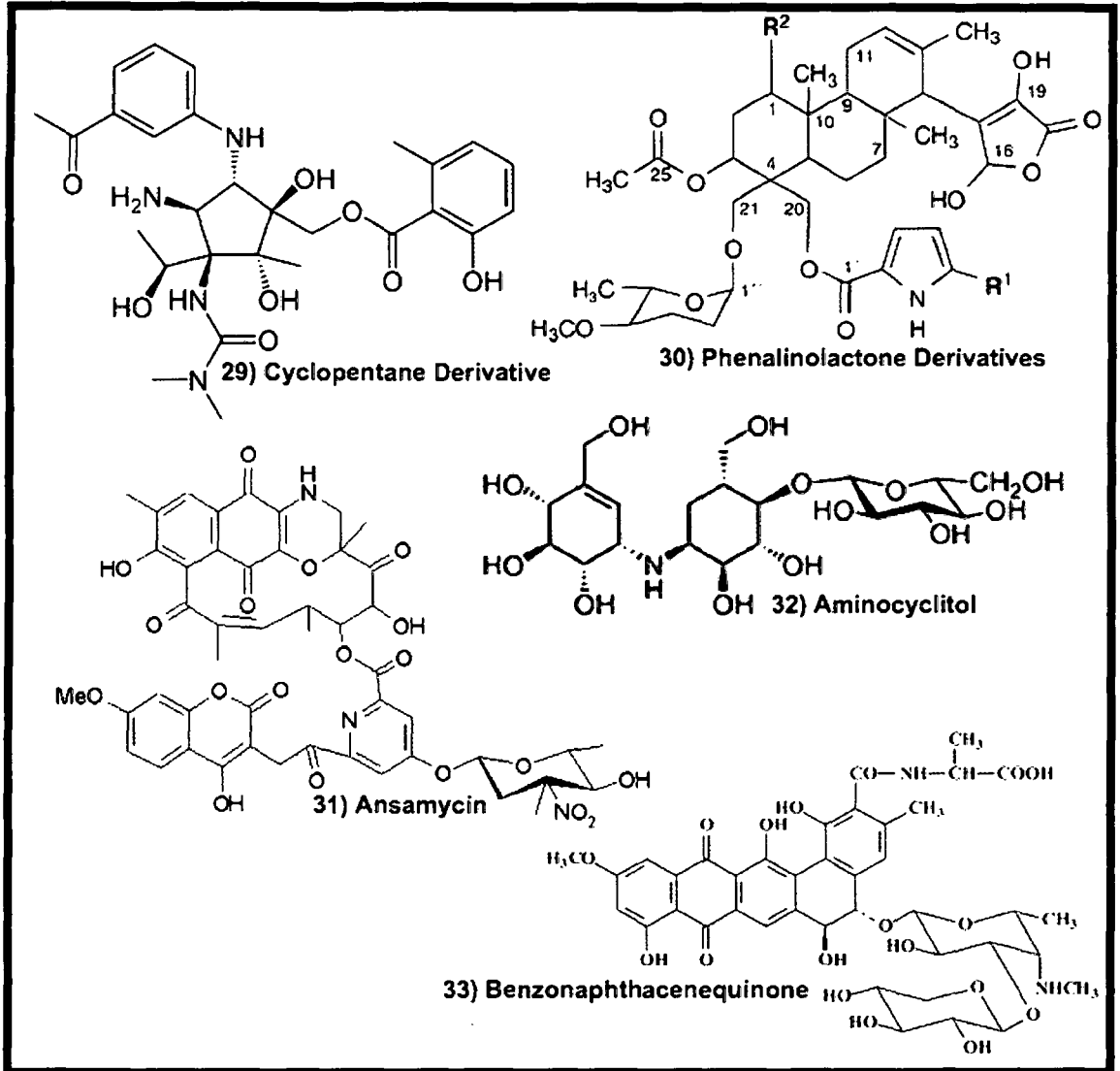
Figure 3a

Figure 3b

29) Cyclopentane Derivative

30) Phenalinolactone Derivatives

31) Ansamycin

32) Aminocyclitol

33) Benzonaphthacenequinone

possible that GTrs belonging to two different acceptor families can have significant similarity in sequence and conversely, sequentially divergent GTrs might also belong to the same acceptor group. As discussed earlier in the methods section, sequentially divergent GTrs belonging to the same acceptor families were divided into multiple sub-groups based on sequence similarity. Similarly, sequentially similar GTrs belonging to different acceptor families were merged into single groups. No HMM profiles were built when a group or sub-group lacked adequate number of sequences. Thus in total 18 HMMER profiles were built for different acceptor families. Each group has a lead GTr against which sequence identity and similarity of other GTrs of the group is compared. Phylogram (**Figure 4**) was built which included two GTrs from each group, one is the lead GTr and other GTr which is sequentially most divergent from the lead GTr. The branches of the phylogram have been labeled indicating the name of the antibiotic GTr and the acceptor group. As can be seen from **Figure 4**, the sequences have mostly clustered as per acceptor groups, though a few exceptions are present. This indicates that, HMM profiles cover significant space in acceptor diversity pool of GTrs and hence, searches made using these profiles will not only help in identifying new GTrs over diverse sequence range but also enable classification based on their putative acceptor substrates.

TrEMBL database was searched using HMMER profiles generated using GTrs from 18 different acceptor groups. Results are presented in five tables (**Table 2** to **Table 6**). Each table has three columns. First column gives the name of the acceptor group and serial number for this group used in table 1 and in figures 2a and 2b. Second column lists the lead GTr for the corresponding acceptor group, the other GTrs from the corresponding group used to build the HMM profile and their sequence similarities to the lead GTr. Third column is divided in four sub-columns, each of which gives number of GTrs found during searches against TrEMBL in different ranges of E-value cut off. For a given acceptor group, the first sub-column lists number of uncharacterized GTrs showing HMM profile matches having E-value less than 1e-100, second sub-column shows number of matches having E-values in the range 1e-100 to 1e-75. The third sub-column shows number of matches having E-values within 1e-75 to 1e-50, while the forth sub-column shows number of matches having E-values in the range of 1e-50 to 1e-25. As can be seen from the results reported in **Tables 2, 3, 4, 5** and **6**, HMM searches on TrEMBL database for identification of putative GTrs belonging to 18 different acceptor families revealed presence of more than 5000 GTrs at e-value cut-off of lower than 1e-25. It is

Figure 4 Phylogram of GTrs made from 2 GTrs of each of the 18 groups used for making HMMER profiles. Naming used to differentiate GTrs is as follows, first is name of the acceptor molecule then the arbitrary number assigned to each GTr of this acceptor molecule, then the name of the group to which it belongs.
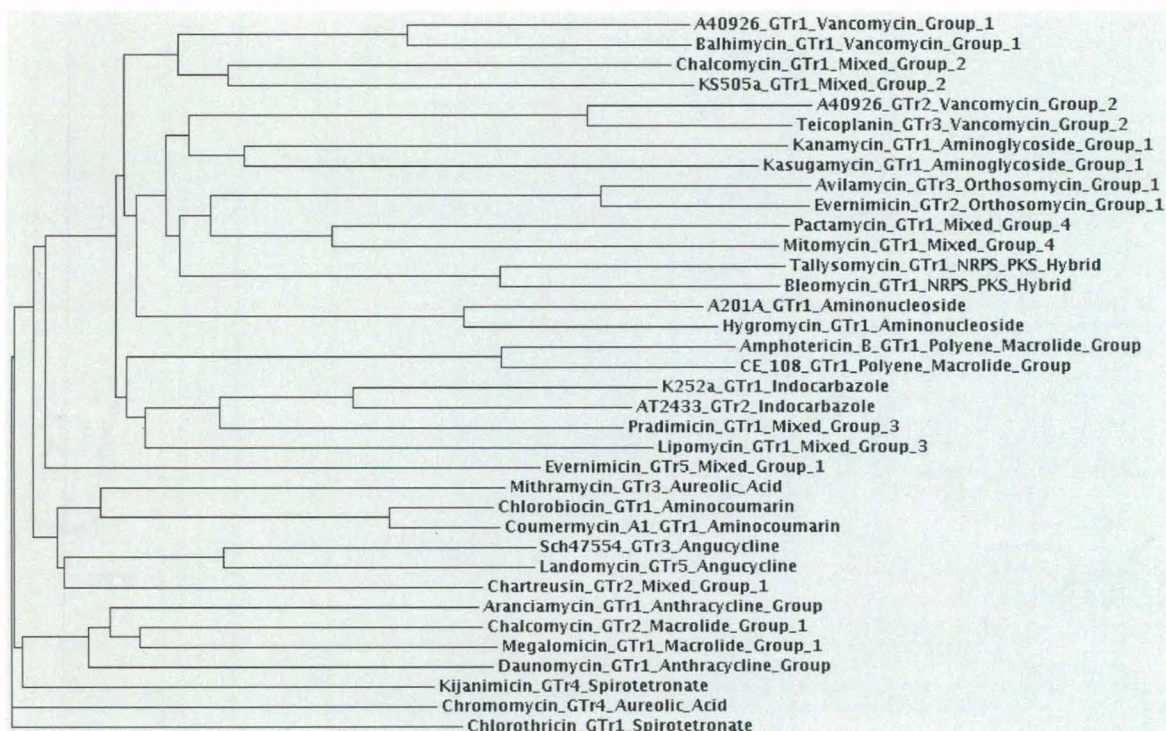

A40926_GTr1_Vancomycin_Group_1
Balhimycin_GTr1_Vancomycin_Group_1
Chalcomycin_GTr1_Mixed_Group_2
KS505a_GTr1_Mixed_Group_2
A40926_GTr2_Vancomycin_Group_2
Teicoplanin_GTr3_Vancomycin_Group_2
Kanamycin_GTr1_Aminoglycoside_Group_1
Kasugamycin_GTr1_Aminoglycoside_Group_1
Avilamycin_GTr3_Orthosomycin_Group_1
Evernimicin_GTr2_Orthosomycin_Group_1
Pactamycin_GTr1_Mixed_Group_4
Mitomycin_GTr1_Mixed_Group_4
Tallysomycin_GTr1_NRPS_PKS_Hybrid
Bleomycin_GTr1_NRPS_PKS_Hybrid
A201A_GTr1_Aminonucleoside
Hygromycin_GTr1_Aminonucleoside
Amphotericin_B_GTr1_Polyene_Macrolide_Group
CE_108_GTr1_Polyene_Macrolide_Group
K252a_GTr1_Indocarbazole
AT2433_GTr2_Indocarbazole
Pradimicin_GTr1_Mixed_Group_3
Lipomycin_GTr1_Mixed_Group_3
Evernimicin_GTr5_Mixed_Group_1
Mithramycin_GTr3_Aureolic_Acid
Chlorobiocin_GTr1_Aminocoumarin
Coumermycin_A1_GTr1_Aminocoumarin
Sch47554_GTr3_Angucycline
Landomycin_GTr5_Angucycline
Chartreusin_GTr2_Mixed_Group_1
Aranciamycin_GTr1_Anthracycline_Group
Chalcomycin_GTr2_Macrolide_Group_1
Megalomicin_GTr1_Macrolide_Group_1
Daunomycin_GTr1_Anthracycline_Group
Kijanimicin_GTr4_Spirotetronate
Chromomycin_GTr4_Aureolic_Acid
Chlorothricin_GTr1_Spirotetronate

Table 2 to Table 6

Results of searches made on TrEMBL database using HMMER profiles generated from GTrs of 18 different acceptor groups. Each table has three columns. First column gives the name of the acceptor group and serial number for this group used in **Table 1** and in **Figures 2** and **3**. Second column lists the lead GTr for the corresponding acceptor group, the other GTrs from the corresponding group used to build the HMM profile and their sequence similarities to the lead GTr.   Third column is divided in four sub-columns, each of which gives number of GTrs found during searches against TrEMBL in different ranges of E-value cut off. For a given acceptor group, the first sub-column lists number of uncharacterized GTrs showing HMM profile matches having E-value less than 1e-100, second sub-column shows number of matches having E-values in the range 1e-100 to 1e-75. The third sub-column shows number of matches having E-values within 1e-75 to 1e-50, while the forth sub-column shows number of matches having E-values in the range of   1e-50 to 1e-25

Table 2

Name of the Group	GTrs of the Group	No. of New GTrs by E-Value			
		A	B	C	D
Vancomycin Group 1 (Contains GTrs from group 1 of table 1)	A40926_GTr1 Tcicoplanin_GTr2 (74%) Chlorocremomycin_GTr2 (69%) Vancomycin_GTr2 (69%) Balhimycin_GTr2 (70%) Tcicoplanin_GTr1 (64%) Chlorocremomycin_GTr3 (62%) Balhimycin_GTr3 (63%) Chlorocremomycin_GTr1 (57%) Vancomycin_GTr1 (59%) Balhimycin_GTr1 (57%)	29	0	2	402
Vancomycin Group 2 (Contains GTrs from Group1 of table 1)	A40926_GTr2 Tcicoplanin_GTr3 (63%)	5	34	93	11
Aminoglycoside Group (Contains GTrs from group 8 of table 1)	Kanamycin_GTr1 Ribostamycin_GTr1 (60%) Ncomycin_GTr1 (63%) Gcntamicin_GTr2 (54%) Butirosin A_GTr1 (34%) Kanamycin_GTr2 (32%) Fortimicin_GTr1 (33%) Tobramycin_GTr1 (34%) Gcntamicin_GTr1 (30%) Tobramycin_GTr2 (34%) Kasugamycin_GTr1 (27%)	27	5	59	>500
Orthosomycin Group (Contains GTrs from group 6 of table 1)	Avilamycin_GTr3 Evcrnimicin_GTr2 (71%)	3	0	0	44

Table 3

Name of the Group	GTrs of the Group	No. of New GTrs by E-Value			
		A	**B**	**C**	**D**
NRPS-PKS Hybrid (Contains GTrs from group 7 of table 1)	Tallysomycin_GTr1 Bleomycin_GTr1 (62%)	4	0	0	0
Aminonucleoside Group (Contains GTrs from group 21 of table 1)	A201A_GTr1 Hygromycin_GTr1 (63%)	3	0	0	0
Polyene Macrolide Group (Contains GTrs from group 3 of table 1)	Amphotericin B_GTr1 Nystatin A1_GTr1 (85%) Pimaricin_GTr1 (66%) Candicidin_GTr1 (63%) Rimocidin_GTr1 (65%) CE 108_GTr1 (65%)	9	0	0	15
Indocarbazole Group (Contains GTrs from group 9 of table 1)	K252a_GTr1 Staurosporine_GTr1 (58%) Rebeccamycin_GTr1 (57%, 67%, 1e-109) AT2433_GTr2_dglucose (56%, 68%, 4e-104)	22	25	21	355
Aureolic acid Group (Contains GTrs from group 12 of table 1)	Chromomycin_GTr4 Mithramycin_GTr1 (51%) Chromomycin_GTr3 (43%) Chromomycin_GTr1 (40%) Mithramycin_GTr2 (38%) Mithramycin_GTr4 (37%) Chromomycin_GTr2 (33%) Mithramycin_GTr3 (30%)	25	117	111	66

Table 4

Name of the Group	GTrs of the Group	No. of New GTrs by E-Value			
		A	B	C	D
Aminocoumarin Group (Contains GTrs from group 13 of table 1)	Chlorobiocin_GTr1 Novobiocin_GTr1 (83%) Coumermycin_GTr1 (84%)	3	0	30	221
Angucycline Group (Contains GTrs from group 14 of table 1)	Sch47554_GTr3 Urdamycin_GTr1 (72%) Landomycin_GTr1 (58%) Simocyclinone_GTr1 (55%) Landomycin_GTr5 (60%)	15	3	14	208
Anthracycline Group (Contains GTrs from group 2 of table 1)	Aranciamycin_GTr1 Aclacinomycin_GTr2 (54%) Daunorubicin_GTr1 (51%) Steffimycin_GTr1 (52%) Nogalamycin_GTr2 (52%) Aclacinomycin_GTr1 (43%) Adriamycin_GTr1 (40%) Daunomycin_GTr1 (38%)	82	15	53	121
Macrolide Group (Contains GTrs from group 16 of table 1)	Chalcomycin-GTr2 Tylosin_GTr1 (62%) Midecamycin_GTr1 (59%) Methymycin_GTr1 (57%) Oleandomycin_GTr2 (55%) Spiramycin_GTr1 (53%) Mycinamicin_GTr1 (54%) Erythromycin_GTr2 (49%) Megalomicin_GTr3 (48%) Megalomicin_GTr2 (49%) Oleandomycin_GTr1 (49%) Erythromycin_GTr1 (47%) Megalomicin_GTr1 (47%)	80	21	49	120

Table 5

| Name of the Group	GTrs of the Group	No. of New GTrs by E-Value			
		A	**B**	**C**	**D**
Spirotetronate Group (Contains GTrs from group 23 of table 1)	Chlorothriciln-GTr1 Chlorothriciln_GTr2 (39%) TetrocarcinA_GTr1 (36%) Kijanimicin_GTr3 (37%) Kijanimicin_GTr2 (38%) TetrocarcinA_GTr3 (39%) Kijanimicin_GTr1 (36%) TetrocarcinA_GTr2 (35%) Kijanimicin_GTr5 (37%) Kijanimicin_GTr4 (36%)	21	89	124	47
Mixed Group-I (Contains GTrs from group 22, 25, 16, 14, 18, 19, 4, 6, 9, 2, 11, 30, 27, 17, 10, 20 of table 1)	Chartreusin_GTr2 ECO02301_GTr1 (36%) Spiramycin_GTr2 (41%) Sch47554_GTr1 (35%) Granaticin_GTr1 (34%) Hedamycin_GTr1 (38%) Urdamycin A_GTr3 (37%) Urdamycin A_GTr4 (36%) Landomycin_GTr2 (35%) Medermycin_GTr1 (37%) Landomycin_GTr6 (36%) Vicenistatin_GTr1 (31%) Urdamycin A_GTr2 (32%) Evernimicin_GTr1 (33%) AT2433_GTr1 (34%) Hedamycin_GTr2 (34%) Nogalamycin_GTr3 (33%) Spinosyn_GTr2 (29%) Landomycin_GTr7 (30%) Calicheamicin_GTr1 (34%) Elloramycin_GTr1 (33%) Phenalinolactone_GTr1 (31%) Spinosyn_GTr1 (31%) Lactonamycinz_GTr1 (29%) Calicheamicin_GTr3 (29%) Avilamycin A_GTr1 (30%) C1027_GTr1 (29%) Glivocarcin V_GTr1 (29%) Landomycin_GTr3 (29%) Lactonamycin_GTr1 (29%) Landomycin_GTr4 (31%) Nogalamycin_GTr1 (32%) Maduropeptin_GTr1 (26%) Nanchangmycin_GTr1 (28%) Jadomycin_GTr1 (31%) Evernimicin_GTr5 (27%)	130	87	46	203

Table 6

Name of the Group	GTrs of the Group	No. of New GTrs by E-Value			
		A	B	C	D
Mixed Group-II (Contains GTrs from group 2, 16, 33, 26 of table 1)	Chalcomycin-GTr1 Tylosin_GTr3 (67%) Mycinamicin_GTr2 (63%) ConcanamycinA_GTr1 (47%) Pradimicin_GTr2 (36%, 51%) KS505a_GTr1 (37%)	30	205	105	207
Mixed Group-III (Contains GTrs from group 28, 11, 5, 16, 33 of table 1)	Lipomycin_GTr1_digitoxose Calicheamicin_GTr4 (39%) Avermectin (39%) Calicheamicin_GTr2 (41%) Spiramycin_GTr3 (36%) Tylosin_GTr2 (37%) Pradimicin_GTr1 (29%)	21	54	264	261
Mixed Group-IV (Contains GTrs from group 29, 32, 15 of table 1)	Pactamycin_GTr1_ ValidamycinA_GTr1 (36%) Mitomycin_GTr1 (37%)	4	1	3	142

possible that, this number does not take in to account redundancy because certain GTr sequences are likely to show matches to multiple acceptor families. However, the number of such redundant sequences are expected to be low, because each HMM profile represents distinct regions of acceptor diversity pool as indicated in the phylogram shown in **Figure 4**. It may be noted that, distribution of these newly identified GTrs according to various acceptor groups is uneven. For example vancomycin group 1, aminoglycoside group, indocarbazole group, aureolic acid group, aminocoumarin group, angucycline group, anthracycline group, macrolide group, spirotetronate group all have more than 200 GTrs in total, indicating natural abundance of these molecules and hence the corresponding GTrs. On the other hand orthosomycin group, NRPS-PKS hybrid group, aminonucleoside group, polyene macrolide group have less than 50 total GTrs indicating restricted use of these GTrs in generation of secondary metabolite structural diversity. It is also possible that, genome sequences of the organisms which biosynthesize metabolites containing these acceptor moieties are under represented in TrEMBL database.

2.4 DISCUSSION

In this study an attempt has been made to identify novel antibiotic glycosyltransferases by *in silico* genome mining and decipher their acceptor specificities. The set of 102 known GTrs cataloged in GTrDB have been used as starting point and based on BLAST search and literature curation, we have identified new GTrs of known specificity. This analysis resulted in identification of 52 new GTrs involved in biosynthsis of 36 different antibiotics. The set of 102 GTrs present in GTrDB and the 52 newly identified GTrs of known donor and acceptor specificity were grouped into different acceptor families based on similarity in the chemical structures of their acceptor moieties. In general GTrs belonging to same acceptor families show a high degree of sequence similarity than those belonging to different acceptor families. Profile HMMs were built for each acceptor families and using these profiles nr database was searched and assigning acceptor family was assigned to experimentally uncharacterized GTrs. Thus in total 139 GTrs were divided into 18 different groups based on acceptor family to which they belong and sequence similarity. 14 of these groups are pure, meaning all their GTrs correspond to single acceptor family. Four of the groups have GTrs from different acceptor families put together due to sequence similarity. Studies like the one presented here can be considered as the first step towards truly understanding and exploiting

TH-2041 5

21
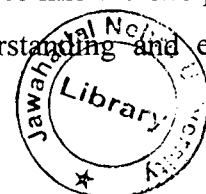
synthetic potential of nature's biosynthetic machinery. This study gives idea of expanse of biosynthetic pool of GTrs, an important class of tailoring enzymes. Expansions of such studies using in-silico and in-vitro approaches to fine-tune the specificity of each enzyme will be the next stage in evolution of methodologies to uncover untapped biosynthetic potential of this nature resource. Even though the current computational approach for identifying novel tailoring enzymes has been applied to the antibiotic glycosyltransferase family of enzymes, the approach is general enough for being applicable to other classes of tailoring enzymes like oxydoreductases, methyltransferases etc. Such computational methods for genome mining would be valuable resource for natural product based drug discovery.

Chapter 3

Understanding donor substrate specificity
of antibiotic glycosyltranaferases

3.1 INTRODUCTION

Bioactivity of secondary metabolite natural products (NPs) is known to depend on the glycosylation of the polyketide or nonribosomal peptide core by tailoring enzymes like glycosyltransferases (GTrs)(Weymouth-Wilson, 1997). Various studies have shown that number of sugar moieties, kind of sugars and their position of attachment affect biological activity of these secondary metabolites. Thus ability to selectively modulate tailoring GTrs will endow biochemists with the ability to produce sugar-based variants of NPs(Ahmed *et al.*, 2006). This is particularly important in case of GTrs involved in biosynthesis of vancomycin related antibiotics, because these molecules form last line of defense against gram positive bacterial infections and recently bacterial strains resistant to these antibiotics had been discovered. Hence, derivatives of vancomycin obtained by altered glycosylation patterns have the potential to show activity against these resistant strains(Losey *et al.*, 2002).

Vancomycin family of molecules constitutes an important class of clinically proven antibiotics. Microorganism *Amycolatopsis orientalis* is known to be involved in biosynthesis of two antibiotics of this class namely Vancomycin and Chloroeremomycin. The glycosyltransferase GtfD is involved in biosynthesis of Vancomycin, while biosynthesis of Chloroeremomycin involves glycosyltransferases, GtfA and GtfC. In the last stage of Vancomycin biosynthesis, GtfD transfers the sugar vancosamine from sugar donor TDP-vancosamine on to hepta-peptide acceptor Desvancosaminyl Vancomycin (DVV) to complete Vancomycin synthesis. The sugar is transferred onto the second hydroxyl of the glucose moiety of DVV (**Figure 1**). On the other hand, GtfA acts at the penultimate stage of the biosynthesis of chloroeremomycin. It transfers the sugar *epi*-vancosamine from TDP-*epi*-vancosamine to DVV, thus forming Chloroorientacin B (COB). The sugar *epi*-vancosamine is transferred on to side chain β-hydroxyl of the sixth residue of hepta-peptide acceptor DVV (**Figure 2**). Product of the enzyme GtfA i.e. COB is the acceptor substrate for the glycosyltransferase GtfC. GtfC transfers epi-vancosamine from the sugar donor TDP-epi-vancosamine onto the second hydroxyl of the glucose moiety on COB. This completes the biosynthesis of Chloroeremomycin (**Figure 3**)(Walsh *et al.*, 2003).

In view of the pharmaceutical importance of vancomycin family of antibiotics, deciphering the specificity determining code of vancomycin family of antibiotic GTrs have a topic of major interest to several research groups. Several experimental studies
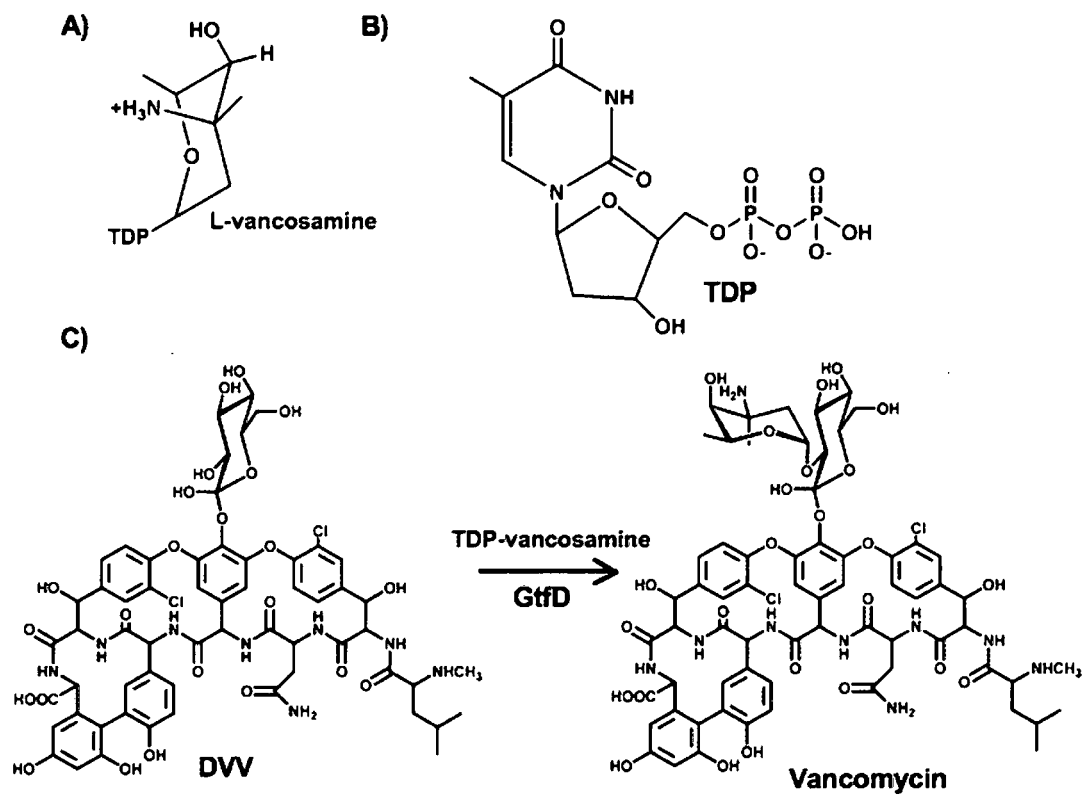
Figure 1: Structures of substrates and products of GtfD; A) Sugar donor TDP-vancosamine; B) Thymidine Diphosphate (TDP) donates sugar and acts as leaving group; C) Shows the reaction catalyzed by GtfD. DVV acts as acceptor substrate and Vancomycin is the final product.
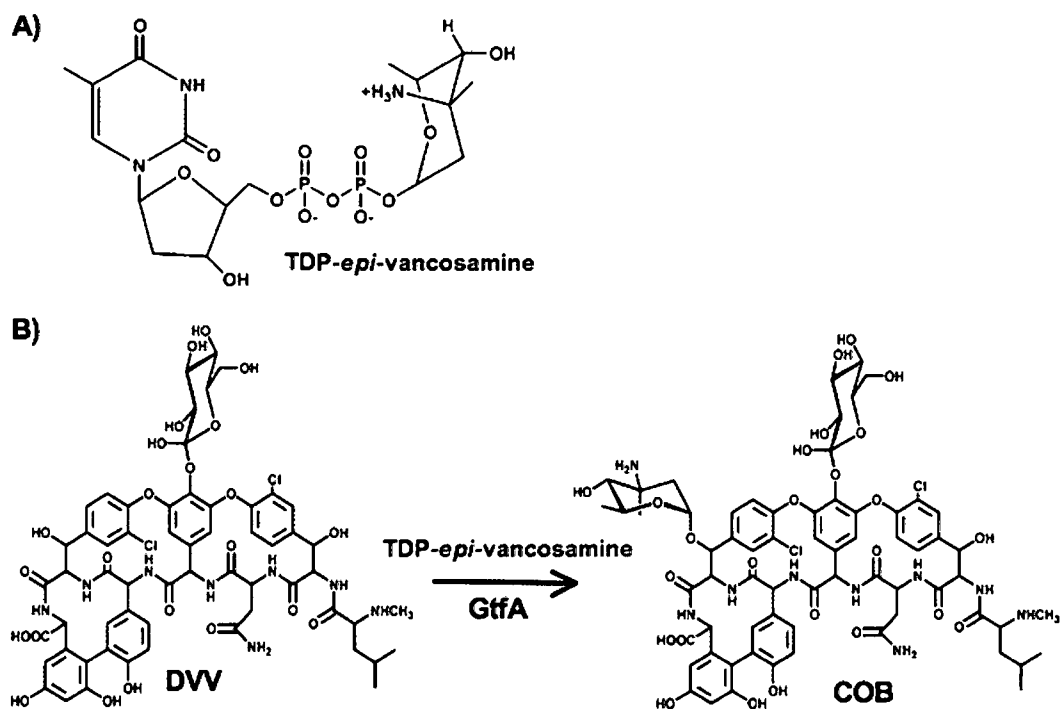
Figure 2: Schematic diagrams of the donor and acceptor substrates of the enzyme GtfA   A) Sugar donor TDP-*epi*-vancosamine; B) Sugar acceptor DVV and final product COB.
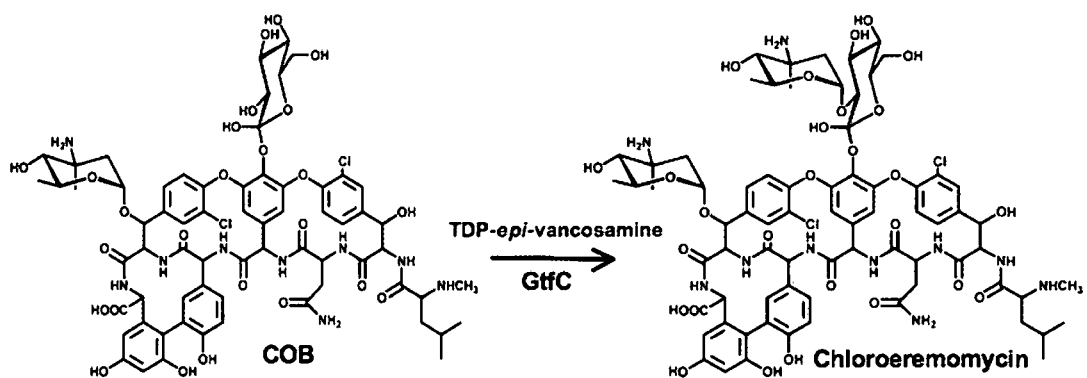
Figure 3: Depicts the reaction catalyzed by GtfC. It transfers epi-vancosamine from TDP-epi-vancosamine to COB to finish Chloroeremomycin biosynthesis.

involving biochemical analysis and enzymology have attempted to probe the donor and acceptor specificity of these GTrs(Baltz, 2002; Oberthur *et al.*, 2005). As a result detailed substrate specificity data is available for GTrs like GtfD, GtfA and GtfC. Detailed biochemical and enzymological studies by Oberthur *et al* have helped in characterization of *in vitro* donor substrate specificities of the three important vancomycin family of antibiotic GTrs, namely GtfA, GtfC and GtfD. **Table 1** summarizes the Kcat and Km values for the enzymatic reactions carried out by these three enzymes during transfer of their in vivo cognate as well as non-cognate donor sugar moieties to their cognate acceptors. As can be seen, all three GTrs are capable of differentiating between TDP-vancosamine and TDP *epi*-vancosamine. GtfD is the most versatile enzyme and can transfer both sugars at high catalytic rates. However, it transfers its cognate sugar TDP-vancosamine at much lower K_m than TDP-*epi*-vancosamine. Thus it can differentiate between both donor sugrars at lower concentrations. GtfA is the most stringent of the three as it is completely selective for TDP-*epi*-vancosamine. However, as can be seen from Kcat/Km values even for its cognate sugar TDP-*epi*-vancosamine, its activity is much lower than other two GTrs. GtfC shows residual activity for TDP-vancosamine, but again like GtfA, it is very selective for its natural substrate TDP-*epi*-vancosamine. Even though its rate of catalysis for transfer of epi-vancosamine is higher than GtfA, it is still significantly lower than GtfD. Hence, these in vitro substrate specificity data suggest that, even though vancosamine and *epi*-vancosamine differ only with respect to orientation of single hydroxyl group, the vancomycin family of GTrs have evolved to distinguish between them to generate structural and functional diversity in vancomycin class of antibiotics. GtfD and GtfC are versatile GTrs and have relaxed donor substrate specificities, while GtfA has a stringent specificity and is highly selective for epi-vancosamine.

Apart from biochemical data regarding substrate specificity a wealth of structural information is also available for antibiotic glycosyltransferases. Mulichak *et al* have reported the crystal structures of GtfA and GtfD both in presence and absence of the donor and acceptor substrates(Mulichak *et al.*, 2003; Mulichak *et al.*, 2004). In fact both the enzymes have very similar three dimensional fold, even though there are subtle differences which control regio-specificity of glycosylation. **Figure 4** shows the cartoon representation of the structure of GtfD as a representative of both GTrs. Both GTrs exhibit similar bi-domain architecture. N-terminal domain has a surface exposed acceptor binding site while cleft between two domains forms donor binding site and also the active

A)

HO—H

+H₃N

vancosamine

O

TDP

H

+H₃N =OH

epi-vancosamine

O

TDP

B)
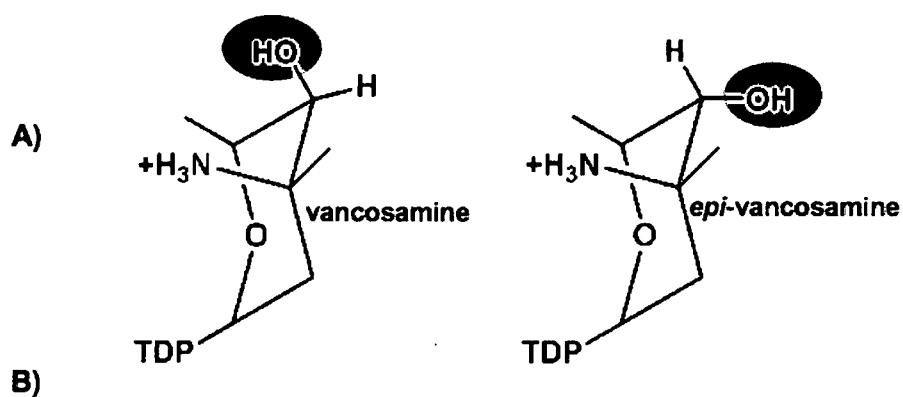
Sugar Donor	GtfD	GtfA	GtfC
TDP-vancosamine	++++ Kcat = 128 Km = 38	--- Kcat = 0	+ Kcat = 0.6 Km= 31
TDP-epi-vancosamine	+++ Kcat = 135 Km = 232	++ Kcat = 2.3 Km = 218	+++ Kcat = 41 Km = 199

Table 1: Detailed substrate specificities of GtfD, GtfA and GtfC based on experimental studies reported in literature. A) Schematic diagrams of sugar-donors TDP-vancosamine and TDP-epi-vancosamine, with hydroxyl at which they differ highlighted in grey; B) Table describes activity shown by each Glycosyltransferase towards aforementioned sugar donors in presence of natural acceptor. Activity scale is graded as Maximum (++++), high(+++), good(++), poor (+) and inactive (----) based on K_{cat}/K_m values. K_{cat} and K_m values are in s^{-1} and μM units respectively.

Enzyme 1	Enzyme 2	Similarity (%)	Identity (%)
GtfD	GtfC	81	71
GtfD	GtfA	64	53
GtfC	GtfA	69	58

Table 2: Sequence similarities between GtfD, GtfA and GtfC
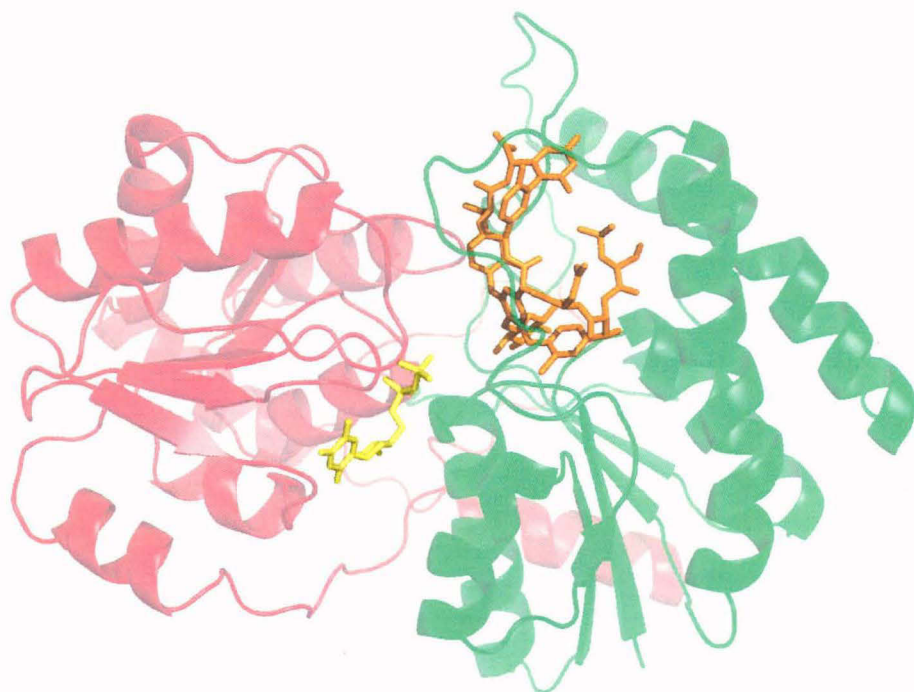
Figure 4: The crystal structure of GtfD in complex with TDP and DVV. N-terminal domain is shown in green, C terminal domain in red, donor TDP in yellow and acceptor DVV in orange.

site. Both N and C-terminal domains contain similar core structure of parallel β sheets connected by α helices commonly called Rossmann fold that is usually associated with dinucleotide binding. Both structures have completely resolved acceptor DVV bound (**Figure 4**) in acceptor binding site. However, in the donor binding site only TDP is bound (**Figure 4**) and no electron density could be located in the crystal structure for the sugar moiety in case of both GtfA and GtfD. As of today no crystal structure is available for GtfC, but given its high sequence similarity to GtfD it can be assumed that reliable model of GtfC can be built using GtfD structure as template.

Since a large amount of biochemical and structural data is available for these GTrs, they are ideal candidates for detailed structure based studies to identify the specificity determining residues (SDRs) of these tailoring enzymes and understand mechanistic details of their substrate selectivity. Such knowledge about mechanistic details of glycosylation by antibiotic GTrs and their specificity determining residues will provide valuable clues to the biochemists for designing GTrs with desired specificity. Ability to understand how biosynthetic tailoring enzymes achieve their unique substrate specificities will enable their systematic modulation to design enzymes with broader substrate specificities and higher catalytic rates.

In this study, a structure based computational method has been used to decipher the structural principles governing the donor substrate specificities of antibiotic glycosyltransferases using vancomycin GTrs as test case. Apart from evolutionary analysis of sequence and structural features of vancomycin GTrs, detailed docking and molecular dynamics simulations have been carried out on GtfA, GtfC and GtfD in complex with various cognate and non-cognate substrates to understand in general, why GtfD shows relaxed substrate specificities while GtfA and GtfC have stringent substrate specificities. This study has also attempted to answer the following specific questions. 1. Why GtfD can transfer both sugars vancosamine and *epi*-vancosamine at higher catalytic rate than GtfA and GtfC ? 2. Why GtfD has lower K_m for TDP-vancosamine compared to TDP-*epi*-vancosamine? 3. How and why GtfA and GtfC are selective for TDP-*epi*-vancosamine? 4. Why GtfA and GtfC have lower catalytic ability than GtfD ?

3.2 METHODS

In order to compare the binding energy of cognate and non-cognate donor sugar moieties to GtfA, GtfD and GtfC, it was necessary to obtain the donor and acceptor substrate bound structures for these three enzymes. The cognate and non-cognate donor

substrates were TDP-vancosamine or TDP-*epi*-vancosamine for these GTrs, while in each case the binding energy was computed in presence of the cognate acceptor which was either DVV or COB. **Table 3** lists all the six possible substrate bound GTr complexes used in this study. As mentioned earlier crystal structures were available for GtfD and GtfA in complex with TDP and DVV, but coordinates for vancosamine and *epi*-vancosamine moieties was missing in both these structures. Hence, sugar moieties were docked on to these TDP and DVV bound crystal structures of GtfD and GtfA to generate complexes 1 to 4 listed in **Table 3**. Since, no crystal structure was available for GtfC, the structure for GtfC was modeled using GtfD as structural template to which it showed 71% sequence identity and 81% sequence similarity (**Table 2**). The structure of GtfC in complex with TDP-vancosamine/TDP-epi-vancosamine and COB was generated by transforming the coordinates of TDP and DVV from the structural template and converting them to COB and TDP-vancosamine/TDP-epi-vancosamine by molecular modeling and docking.

3.2.1 Docking of TDP-vancosamine and TDP-*epi*-vancosamine in to GtfD and GtfA Active Sites

As mentioned earlier crystal structure are available for GtfD and GtfA in complex with TDP and DVV. But coordinates for the sugar moiety attached to the TDP was missing in these crystal structures. Therefore, in order to generate the donor/acceptor bound complexes 1, 2, 3 and 4 listed in **Table 3**, TDP-vancosamine and TDP-epi-vancosamine were docked in to empty donor binding site (after removal of bound TDP) of GtfD and GtfA using Autodock 4(Huey *et al.*, 2007). Coordinates of TDP were extracted from crystal structures of GtfD and GtfA and were converted to TDP-vancosamine and TDP-epi-vancosamine by molecular modeling technique using Pymol editor. TDP-sugars thus obtained were docked into empty donor substrate binding sites of GtfD and GtfA after removal of the bound TDP in the crystal structure. During docking, all bonds of TDP were kept rigid as its coordinates were taken from respective crystal structures and already were in optimized orientation. Only sugar part and the bond by which it is attached to TDP were kept flexible. During the docking of TDP-vancosamine and TDP-epi-vancosamine acceptor substrate DVV was already bound in the crystal structure. This indeed helped in mimicking the role that acceptor plays in determining proper orientation of donor substrate. During the Autodock run, the docking grid spanned the entire structures of the enzymes GtfD and GtfA and docking was carried out using the Lamarckian genetic algorithm (LGA) as conformational search method.

Serial No.	Glycosyltransferase	Donor Substrate	Acceptor Substrate
1)	**GtfD (1RRV)**	**TDP**-vancosamine	**DVV**
2)	**GtfD (1RRV)**	**TDP**-*epi*-vancosamine	**DVV**
3)	**GtfA (1PN3)**	**TDP**-vancosamine	**DVV**
4)	**GtfA (1PN3)**	**TDP**-*epi*-vancosamine	**DVV**
5)	GtfC (modeled using 1RRV)	TDP-vancosamine	COB
6)	GtfC (modeled using 1RRV)	TDP-*epi*-vancosamine	COB

Table 3: Catalogue of all glycosyltransferase structures required for proposed donor-substrate specificity analysis. The proteins or ligand groups highlighted in bold were from crystal structures, while the proteins or ligand fragments not shown in bold have been modeled by homology modeling or docking.

	Region 1	Region 2	Region 3	Region 4	Region 5
GtfD	T10,R11,G12,D13	P126,S127,P128,V129,Y130	G245,S246,S247,S248	H309,G310,S311,A312,G313,T314	N331,T332,D333,Q334
GtfC	S10,R11,G12,D13	Y127,S128,P129,N130,Y131	G246,S247,A248,S249	H310,G311,S312,A313,G314,T315	H332,T333,D334,Q335
GtfA	S10,R11,G12,D13	L124,S125,P126,D127,H128	G229,S230,S231,S232	H293,D294,S295,A296,G297,T298	V315,V316,D317-Q322

Table 4: List of five sugar-donor binding regions. Residues listed in the same column correspond to aligned positions in MSA obtained from ClustalW. Between D-317 to Q-322 of Region 5 of GtfA, there is four amino acid insertion not included in table which is as follows, N318,V319,V320,E321.

The docking parameters for LGA search method were 27,000 generations, 25,00,000 energy evaluations and 250 docking runs. The final set of 250 docked conformations of TDP-vancosamine or TDP-epi-vancosamine were scanned based on multiple criterions to identify correct binding pose. Criterions used were, binding pose should be one of the lowest energy binding pose, it should also be one of the highest frequency binding pose, the distance between C-1 atom of sugar through which it is bound to TDP, and the hydroxyl of acceptor onto which the sugar is transferred must be low so as to allow for transfer of the sugar moiety from TDP to DVV.

3.2.2 Modeling GtfC in complex with TDP-vancosamine/TDP-epi-vancosamine and COB.

The crystal structure of GtfD was used as template for building homology model of GtfC, because they shared a sequence similarity of 81% (**Table 2**). Modeller (version mod9v1) software was used for building the homology model for GtfC(Fiser & Sali, 2003). The ligands bound to the structural template GtfD i.e. TDP-vancosamine and DVV were transformed on to GtfC model during model building using options available in Modeller software. In order to generate the structure of GtfC in complex with its cognate acceptor COB, in the acceptor binding site of GtfC, the coordinates of DVV were converted to COB by modeling epi-vancosamine on β-hydroxyl of the sixth residue of heptapeptide DVV. Modeling of COB starting from the coordinates of DVV was carried out using molecule editor module of Pymol software. Using similar approach coordinates of TDP-vancosamine and TDP-epivancosamine were transformed from donor bound complexes of GtfD. This resulted in structure of GtfC in complex with COB acceptor and TDP-vacosamine as well as TDP-epi-vancosamine as donors (complexes 5 and 6 listed in **Table 3**).

3.2.3 Molecular Dynamics Simulations

The donor and acceptor substrate bound complexes of GtfD, GtfA and GtfC obtained from docking studies were further refined using molecular dynamics simulations in the explicit solvent environment. The refinement using MD simulations was required because, docking studies do not incorporate protein flexibility, flexibility of cyclic ring fragments in the ligands and they also do not take into account induced fit effect which are known to be involved in ligand recognition by enzymes. Explicit solvent MD simulations were carried out using AMBER 9 package(Case, 2006) on all the six

complexes listed in **Table 3**. Antechamber module of Amber was used to assign force field parameters for donor and acceptor substrates apart from the protein(Wang *et al.*, 2006). The protein-ligand complexes obtained from docking studies were energy minimized in vacuum to remove steric clashes if any and then the minimized complex was solvated in a water box which extended 9 Å from the outermost atom of the protein-ligand complex in X, Y and Z directions. Electrostatic interactions were computed using PME approach and a cut off of 8 Å was used for non-bonded interactions. The forcefield used was ff03(Duan *et al.*, 2003) for the protein and TIP3P water model for solvent. All the simulations were carried out for a period of 3 ns in NVT ensemble at 300K using a time step of 1 fs. SHAKE algorithm was used to constrain bonds involving hydrogens. The convergence of the simulations was monitored in terms of RMSD between the starting structure and the structures sampled during the simulations. After convergence the last 1ns of the trajectory was used to compute binding free energy and persistence of various interactions between the donor/acceptor substrates and the enzyme.

3.2.4 Analysis of sequence and structural features

Pair-wise sequence comparison between various GTr enzymes was carried out using the BLAST software from NCBI, while CLUSTALW was used for obtaining the multiple sequence alignments. Ptraj module of AMBER was used to monitor hydrogen bonds (H-bonds) between donor/acceptor substrates and amino acids lining ligand binding pockets of the various glycosyltransferases. Distance cut off of 3.5 Å between hydrogen and acceptor was used for monitoring H-bonds. Apart from analysis of hydrogen bonds in the starting structures and structures obtained at the end of entire 3 ns simulations, percentage occupancy of different hydrogen bonds was also computed from the last 1ns trajectory for each simulation. Pymol Viewer was used to for display and analysis of the structures sampled during the simulations.

3.2.5 Calculation of binding free energy

The MM_GB/SA module of AMBER was used to calculate the binding free energy between the ligand and the protein in various protein-ligand complexes. The structures sampled during last 1ns of the MD simulations were selected at an interval of 200ps and a set of 50 structures were used for MM_GB/SA calculations. Since binding free energy was evaluated by implicit solvent approach, the solvent molecules were excluded and only the coordinates of protein and ligand were used for calculation of

binding free energy. The computation of binding free energy in MM_GB/SA module of AMBER involves the following steps.

$$\Delta G_{total} = G_{complex} - G_{protein} - G_{ligand}$$

where ΔG_{total} is the binding free energy.

$$G = H_{gas} + G_{solvation} - TS\chi$$

$$H_{gas} = \text{Bonded Energies} + \text{Non Bonded Energies}$$

$$E_{bonded} = E_{bond\text{-}strech} + E_{angle\text{-}bend} + E_{rotate\text{-}along\text{-}bond}$$

$$E_{non\text{-}bonded} = E_{van\text{-}der\text{-}Waals} + E_{electrostatic}$$

E_{bonded} and $E_{non\text{-}bonded}$ i.e. molecular mechanical energy was calculated using *sander* program of AMBER with infinite cutoff for all interactions

$$G_{solvation} = \text{Polar Interactions} + \text{Non Polar Interactions}$$

Polar contribution to solvation free energy is calculated using generalized Born (GB) method(Cheatham *et al.*, 1998), while non-polar contribution was calculated using surface area dependant solvent accessibility term as implemented in GB/SA module of AMBER.

Apart from calculating the total binding free energy for TDP-vancosamine and TDP-epi-vancosamine, MM_GBSA module was also used for decomposition of calculated binding free energy into contributions from individual residues of the enzyme. All residues within 5 Å of vancosamine/*epi*-vancosamine sugar moiety were considered for evaluation of their contributions to the total binding free energy.

3.3   RESULTS

3.3.1   Architecture of the donor sugar binding site

The available crystal structures of substrate bound GtfD and GtfA contained only the coordinates of TDP and DVV. Since the structural model of GtfC was built using GtfD as template, only the coordinates of TDP and COB could be obtained from the template structure by coordinate transformation. Hence, for all the six complexes listed in **Table 3**, the binding site for donor sugar moiety was obtained by docking simulations. Since the major objective of the study was to understand the specificities of these GTrs for their cognate as well as non-cognate donor sugar groups, binding site for both vancosamine as well as epi-vancosamine moiety on each of these three GTrs was obtained from docking studies. As described in the methods section, instead of docking vancosamine or epi-vancosamine moiety alone , TDP-vancosamine and TDP-*epi*-

vancosamine were docked, but the TDP was kept rigid during docking as its coordinate has been taken from the crystal structure. During the docking of TDP-vancosamine and TDP-epi-vancosamine to various GTrs the cognate acceptor substrate was already bound in the orientation as obtained from crystallographic studies. In case of GtfC, cognate substrate COB was modeled based on the DDV bound template structure as mentioned in methods section.

Analysis of final docking poses of TDP-vancosamine and TDP-epi-vancosamine to all three GTrs indicated that, general architecture of the donor binding site was conserved across all three GTrs. In all the six complexes, TDP part of the donor substrate was bound to the same site as identified in the TDP bound crystal structures for GtfD and GtfA and C1 of the donor sugar was proximal to the attacking hydroxyl of the acceptor substrate. **Figure 5** shows the binding pose for TDP-vancosamine in the active site of GtfD. As can be seen, five distinct regions of the enzyme GtfD constitute the sugar-donor binding site. Thus amino acid residues of these regions in proximity of sugar moiety make up most likely candidates for specificity determining residues. Hence, these GTrs achieve their distinct sugar specificity through subtle variations in the sequence and structure of these five regions. **Figure 6** shows these five donor substrate binding regions highlighted in the multiple sequence alignment of GtfD, GtfA and GtfC obtained through ClustalW. For easy comparison of sequence variations in the donor binding site identified by the current docking studies, **Table 4** lists all the residues of each of these five regions for all the three GTrs. Analysis of these active site pocket residues in all the three GTrs indicate that, the residues identified by the current docking studies are consistent with the currently accepted catalytic mechanism of glycosyl transfer. One of the catalytic residues acts as a base and initiates catalysis by abstracting proton from acceptor hydroxyl i.e. 2nd hydroxyl of glucose of DVV in case of GtfD (**Figure 5**). This results in a negatively charged oxygen in the acceptor and it acts as a nucleophile and attacks relatively positive C-1 of sugar being transferred, i.e. vancosamine of TDP-vancosamine. TDP acts as a leaving group and vancosamine gets transferred on to attacking oxygen. As can be seen from **Table 4**, D 13 of region 1 is conserved in all three GTrs and clearly plays the role of catalytic base. The sugar moiety of TDP-sugar donor was found to be sandwiched between D 13 of Region 1 and another aspartic acid of Region 5, D 333 in case of GtfD. As can be seen from **Figure 5**, D 333 of region 5 forms a salt bridge with amino group of the vancosamine. Thus the results from current docking
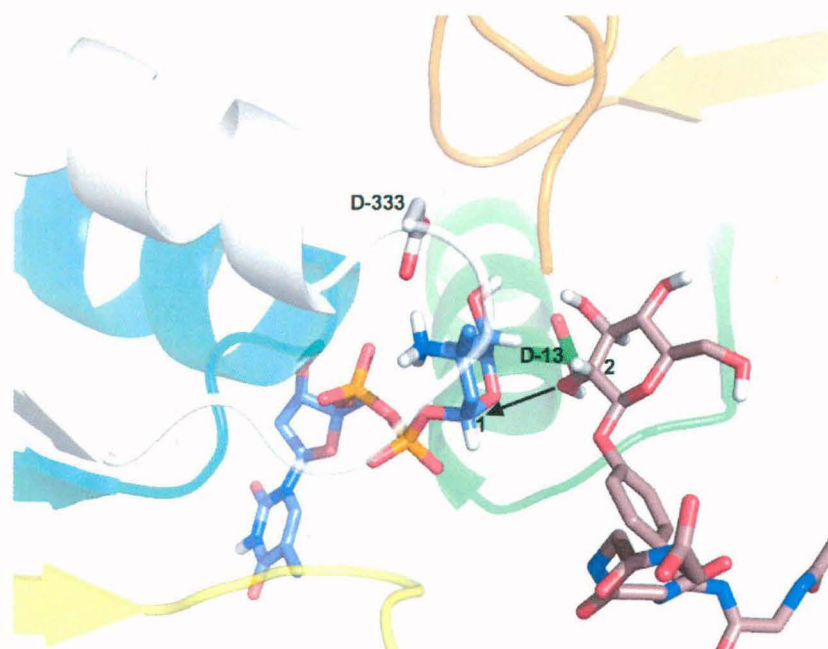
Figure 5: Shows bound conformation (obtained from docking calculations) of sugar donor TDP-vancosamine in the active site of GtfD. Five distinct regions (Regions 1 to 5) which constitute the donor binding site are shown in different colors. (Region 1: Green, Region 2: Orange, Region 3: Yellow, Region 4: Cyan, Region 5: Grey, TDP-vancosamine: Blue, Acceptor DVV: Brown). Second hydroxyl of glucose moiety in DVV onto which vancosamine is transferred after the reaction and C-1 of vancosamine which forms covalent bond with DVV are numbered.

```
                Region 1
GtfD    MRVLLSVCGTRGDVEIGVALADRLKALGVQTRMCAPPAAEERLAEVGVPHVPVGLPQH-M 59
GtfC    MRVLLSTAGSRGDVEPLVALAVRLQGLGVEARMCASPASAERLAEVGVPHVPVGLQLEGM 60
GtfA    MRVLITGCGSRGDTEPLVALAARLRELGADARMCLPPDYVERCAEVGVPMVPVGRAVRAG 60
        ****::  .*:***.*   **** **: **.::*** .*   ** ****** ****   .


GtfD    MLQEGMPPPPPEEEQRLAAMTVEMQFDAVPGAAEGCAAVVAVGDLAAATGVRSVAEKLGL 119
GtfC    LLQEGMPPPSPEEERRLAAKAIDMQFDEVPAAAEGCAAVVAAGELAAAAVRSVAEMLGI 120
GtfA    AREPGELPPGAAE---VVTEVVAEWFDKVPAAIEGCDAVVTTGLLPAAVAVRSMAEKLGI 117
        : *  ** . *      :.: .:   ** **.* *** ***:.* *.**..***:** **:
                Region 2
GtfD    PFFYSVPSPVYLASPHLPPAYDEPTTPGVTDIRVLWEERAARFADRYGPTLNRRRAEIGL 179
GtfC    PYFYAAYSPNYLPSPHHAPPEDERTTPGVTDNKVLWDERGQRFAKRYGDTLNSRRASVGL 180
GtfA    PYRYTVLSPDHLPSEQSQAERD-----------MYNQGADR---LFGDAVNSHRASIGL 162
        *: *:. ** :*.* :  . *            :::: . *    :* ::* :**.:**


GtfD    PPVEDVFGYGHGERPLLAADPVLAPLQP-DVDAVQTGAWLLSDERPLPPELEAFLAAGSP 238
GtfC    PPVEDVFGYGYSERPWLATDPILAPLPP-DFDAVQTGTWILPDERPLSAELEAFLAAGSP 239
GtfA    PPVEHLYDYGYTDQPWLAADPVLSPLRPTDLGTVQTGAWILPDERPLSAELEAFLAAGST 222
        ****.::.**: ::* **:**:*:** * *..:****:*:* ***** .. ********** .
                Region 3
GtfD    PVHIGFGSSSGRGIADAAKVAVEAIRAQGRRVILSRGWTELVLPDDRDDCFAIDEVNFQA 298
GtfC    PVYLGFGSASGPGIDDAARVAIEAIRAHGRRIVLLSGWADLVRPDDGADCFSVDEVNLQV 299
GtfA    PVYVGFGSSSRPATADAAKMAIKAVRASGRRIVLSRGWADLVLPDDGADCFVVGEVNLQE 282
        **::*****:*   . ***:**::*:*:** ***::* **::** *** *** :.***:*
                Region 4                    Region 5
GtfD    LFRRVAAVIHHGSAGTEHVATRAGVPQLVIPRNTD----QPYFAGRVAALGIGVAHDGPT 354
GtfC    LFSRAAAAIHHGSAGTEHLATLAGIPQIVIPRHTD----QPYYAERVADLGIGVALEGPV 355
GtfA    LFGRVAAAIHHDSAGTTLLAMRAGIPQIVVRRVVDNVVEQAYHADRVAELGVGVAVDGPV 342
        ** *.**.***.****  :*  **:**:*:*:: *  .*   *.*.* *** **:*** :**.


GtfD    PTFESLSAALTTVLAPETRARAEAVAGMVLTDGAAAAADLVLAAVGREKPAVPALEHHHH 414
GtfC    PTFDAMSAAVATALAPETRARATAVAGTIRTDGAAVAARLLLDAVSREKSAVLA------ 409
GtfA    PTIDSLSAALDTALAPEIRARATTVADTIRADGTTVAAQLLFDAVSLEKPTVPALEHHHH 402
        **:::.:***: *.**** **** *** ::  :.**** :***:: :*.:** **.:* *
```

Figure 6: Multiple Sequence alignment of GtfD, GtfA and GtfC. Five donor binding regions depicted in Figure 5 are highlighted and numbered.

studies suggest that D 333 might have a crucial role in proper orientation of the sugar moiety in the active site of these GTrs.

Thus the current docking studies helped in identifying the role of different donor binding residues in catalysis of glycosyl transfer. As can be seen from **Figure 5**, region 1 (green) contains catalytic base residue D 13, region 2 (Orange) forms the roof of active site and is in interacting distance with sugar amino and hydroxyl. Region 3 (yellow) is primarily involved in binding second phosphate, while region 4 (cyan) forms first phosphate binding region and also has a phosphate binding motif HHXXAGT. Region 5 (grey) contains conserved aspartic acid which forms salt bridge with sugar amino. Even though our docking studies defined the overall architecture of the donor binding site of vancomycin family of glycosyltransferases, based on these docking results alone it was not possible to provide a theoretical rationale for the experimentally observed substrate specificities of these three GTrs. Therefore, substrate bound complexes obtained from docking studies were further refined by explicit solvent molecular dynamics simulations and binding free energy values for various donor sugar groups were computed by MM-PB/SA approach.

3.3.2 Molecular dynamics simulations & MM_GBSA Analysis

Explicit solvent MD Simulations were performed on all the six substrate bound complexes obtained from docking studies. The convergence of the simulations were monitored by analyzing RMSD with respect to the starting structure for various conformations sampled in each of the six trajectories. **Figure 7** shows the variation of RMSD over the 3 ns trajectories for GtfD, GtfA and GtfC in complex with TDP-vancosamine as well as TDP-epi-vancosamine. As can be seen from **Figure 7**, both GtfD+TDP-vancosamine as well as GtfD+TDP-epi-vancosamine complexes remain close to the starting structure and show RMSD in the range of 1.0 to 1.5 Å. On the other hand, while GtfA+TDP-vancosamine complex shows RMSD in the range of 1.0 to 1.5 Å, GtfA+TDP-epi-vancosamine complex shows slightly higher RMSD and finally converges to a value of close to 1.9 Å. This indicates that, in contrast to GtfA+TDP-vancosamine complex which remains close to the conformation obtained from docking, GtfA+TDP-epi-vancosamine complex undergoes structural rearrangement during explicit solvent MD refinement. In case of GtfC, both the complexes involving TDP-vancosamine as well as TDP-epi-vancosamine show slightly higher RMSD values and finally converge to a value of close to 2.0 Å. Thus the results from MD simulations indicate that all the six
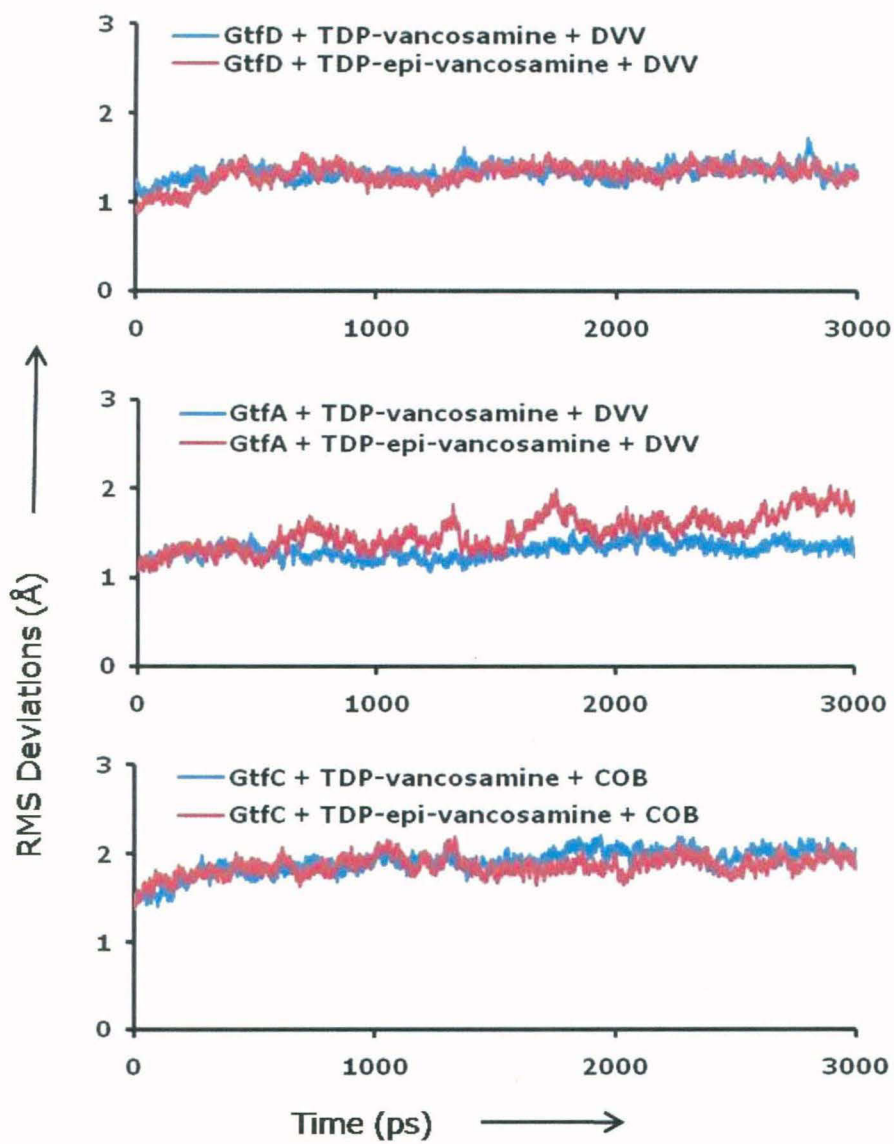
Figure 7: RMSDs (w.r.t. the starting structure) for various conformations sampled during 3 ns MD simulations on GtfD, GtfA and GtfC in complex with different donor sugar and cognate acceptors.

complexes obtained from docking studies remained stable during the explicit solvent MD simulations and the simulations had converged. Detailed analysis of various other structural parameters indicated that, even though the final structure obtained from MD simulations remained close to the structures obtained from docking in terms of overall RMSD, there had been subtle conformational rearrangement in terms of bound conformation of the donor as well as acceptor and also conformation of the substrate binding site. As will be discussed later many of these conformational rearrangements also had very interesting functional implications for substrate selection by these three enzymes.

The various conformations sampled during last 1ns in each of the six trajectories were used to compute the binding free energy between the TDP-vancosamine or TDP-epivancosamine and each of the three glycosyltransferases by MM-GB/SA approach as described in the methods section. **Figure 8** shows the MM-GB/SA binding free energy values for TDP-vancosamine and TDP-epi-vancosamine for GtfD, GtfA and GtfC. It is interesting to note that, GtfD has a binding free energy of approximately -100 kcal/mol for its cognate substrate TDP-vancosamine, while the binding free energy for TDP-epi-vancosamine is close to -90 kcal/mol. On the other hand, GtfA has a binding free energy of -80 kcal/mol for its cognate substrate TDP-epi-vancosamine, while the binding free energy for TDP-vancosamine is only -45 kcal/mole. Thus GtfD shows a difference of only about 10 kcal/mol in the binding free energy of vancosamine vs epi-vacosamine, while the corresponding value for GtfA is about 35 kcal/mole. Hence, the results from binding free energy calculations are consistent with the experimental observation which indicate that GtfA is completely selective for TDP-epi-vancosamine, while GtfD can transfer both vancosamine and epi-vancosamine to DVV. The MM-GB/SA binding free energy difference between TDP-vancosamine and TDP-epi-vancosamine was also found to be of the order of 15 kcal/mole. The experimental studies also suggest that GtfC can transfer both vancosamine as well as epi-vancosamine. Thus in case of GtfC the results from the current computational studies are also in agreement with experimental observations. However, it must be noted that the computational binding free energy is lower for the TDP-vancosamine, even though the cognate substrate for GtfC is TDP-epi-vancosamine.

Since the MM-GB/SA binding free energy values computed from MD trajectories for the various substrate bound complexes for GtfD, GtfA and GtfC were in qualitative agreement with experimental observations, each of the complexes were analyzed in
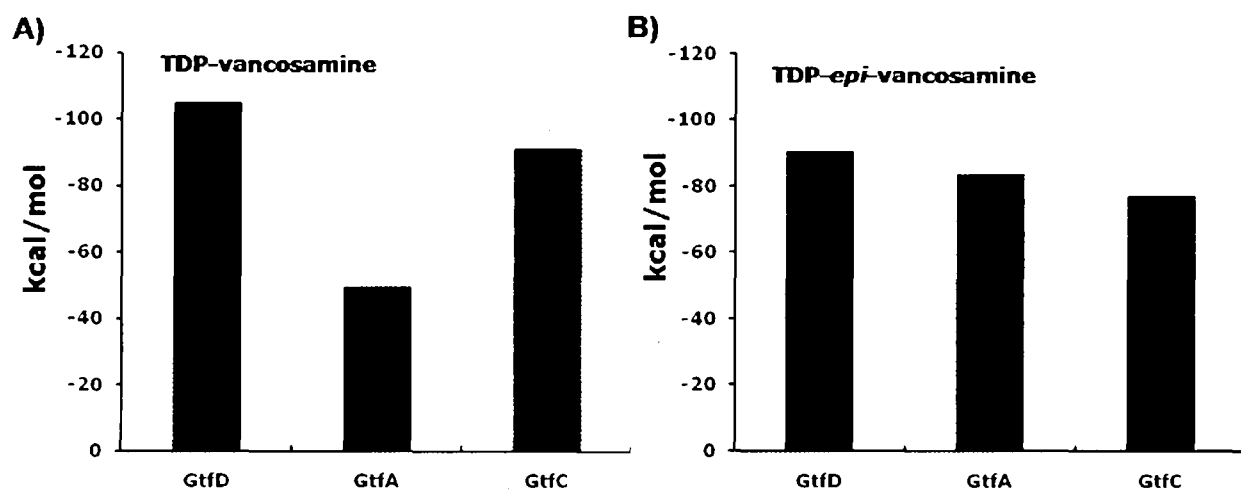
33

Figure 8: Results from MM_GBSA binding free energy calculations. A) Affinity of TDP-vancosamine for GtfD, GtfA and GtfC B) Affinity of TDP-*epi*-vancosamine for GtfD, GtfA and GtfC

details to identify crucial specificity determining residues based on contributions of various binding pocket residues to the binding free energy.

3.3.3 Binding of GtfD to TDP-vancosamine and TDP-*epi*-vancosamine

Figure 9 shows the binding pose of TDP-vancosamine in the active site of GtfD before MD simulation but after minimization. As can be seen, D 333 from region 5 and Y 130 from region 2 of the binding pocket are primarily involved in vancosamine binding. D 333 forms salt bridge with sugar amino while Y 130 is involved in hydrogen bonding with sugar hydroxyl. S 246 of region 3 and S 311 of region 4 are involved in phosphate binding. Figure 10 shows the bound conformation of TDP-vancosamine at the end of the 3 ns MD simulations to highlight the conformational rearrangements occurring during the simulations. As can be seen, TDP-vancosamine moves towards acceptor and distance between hydroxyl of acceptor and C1 of vancosamine reduces from 4.2Å to 3.4Å. Catalytic D 13 of region 1 is forms stable hydrogen bond with vancosamine hydroxyl, thus bringing donor and acceptor in close proximity. Before simulation vancosamine hydroxyl was away from region 1, and was in contact with Y 130 of region 2 as depicted in Figure 9. This movement of vancosamine is further assisted by T 10 of region1, which is now in contact with 2nd phosphate of TDP. But major conformational rearrangement occurs in region 3 where S 247 which was away from TDP-vancosamine before dynamics is found to form H-bond with 2nd phosphate of TDP, thus propelling donor close to acceptor. D 333 of region 5 maintains its salt bridge with vancosamine amino, though now hydroxyl of S 311 from region 4 competes with vancosamine amino for hydrogen bonding with D 333. This probably helps in release of vancosamine from TDP on to DVV after catalysis. It may be noted that, S 311 is conserved only among GTrs transferring amino sugar, and is replaced with Glycine in all other GTrs. Thus the current analysis shades light on probable role of this serine of region 4 during transfer of amino sugars. All the hydrogen bonds between TDP-vancosamine and binding pocket residues of GtfD were also monitored over the 3 ns trajectory and expressed as percentage of their occurrence. Figure 11 shows the percentage occurrence of different hydrogen bonds between the donor substrate and the binding pocket residues and also the contribution of different binding pocket residues to the MM-GB/SA binding free energy. As can be seen from Figure 11, the various binding pocket residues e.g. T 10, D 13, S 247 and D 333 etc which were involved in the conformational rearrangement in the
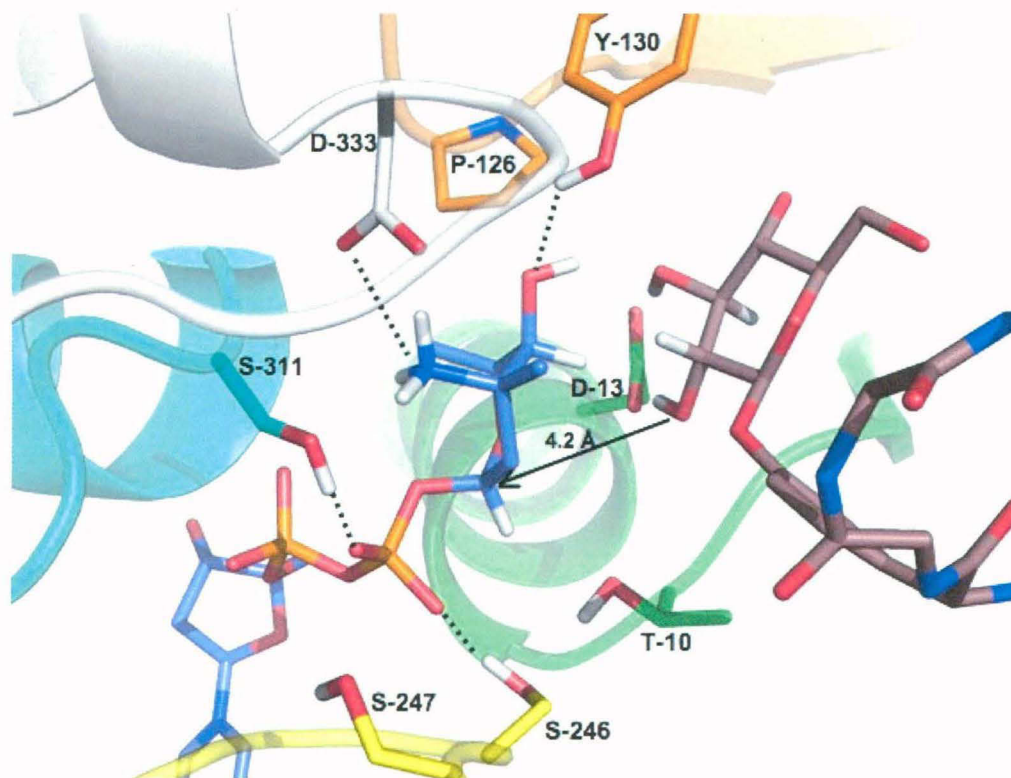
Figure 9: Active site of GtfD in complex with TDP-vancosamine and DVV before simulation. The five different regions constituting the substrate binding site are depicted in same colors as in Figure 5. The H-bonds between the enzyme and substrate are shown as dotted lines. Arrow shows distance between attacking hydroxyl of acceptor and C-1 of vancosamine.
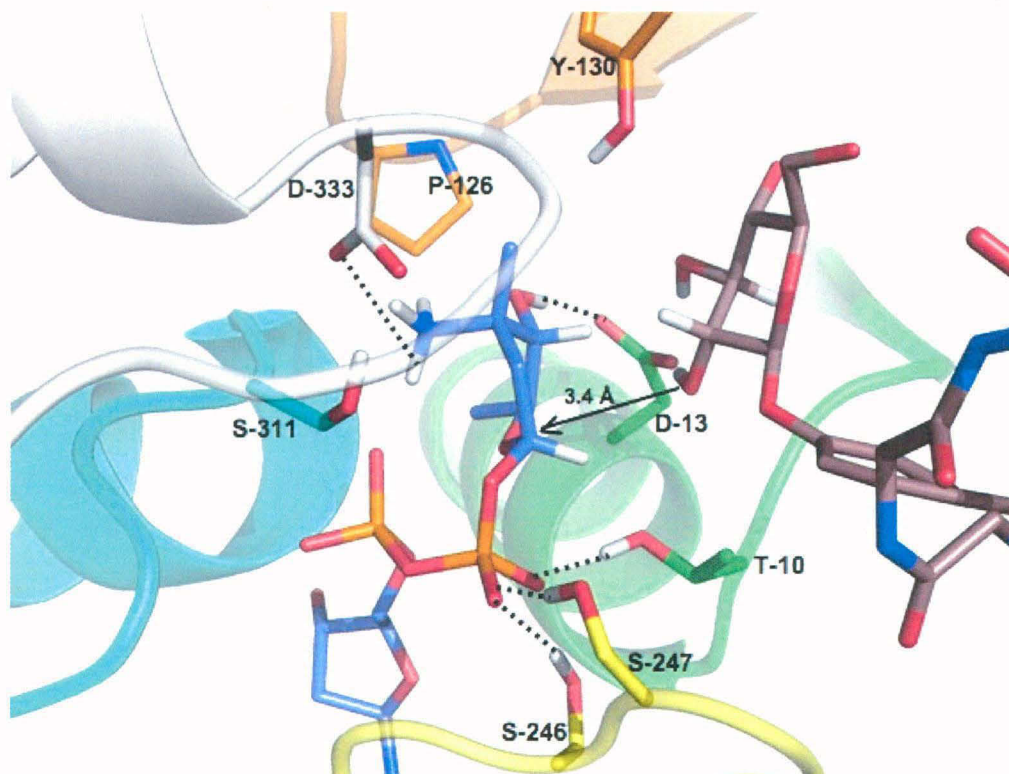
Figure 10: Active site of GtfD in complex with TDP-vancosamine and DVV after 3ns MD simulation. Dotted lines depict H-bonds which are stable over the 3 ns MD simulation (based on analysis by Ptraj module of AMBER).

A)

B)

Functional Group of TDP-vancosamine	GtfD Amino Acid Residue	Strength of H-bond (% of Trajectory)
vancosamine 3-Amino	D-333	93%
vancosamine 4-Hydroxyl	D-13	93%
2nd Phosphate	T-10	86%
	S-246	100%
	S-247	100%
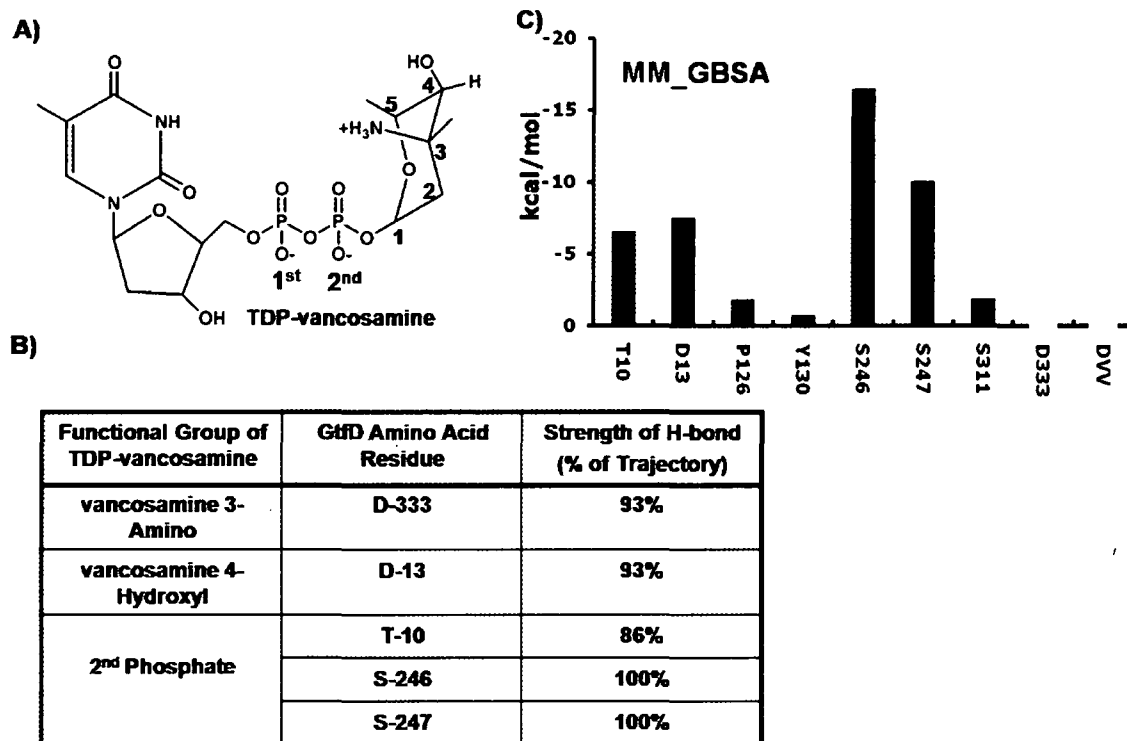
Figure 11: H-bond and MM_GBSA binding free energy analysis of the trajectory obtained from MD simulations on GtfD in complex with TDP-vancosamine and DVV. A) Schematic diagram depicting different functional groups of TDP-vancosamine which form H-bond with active site residues of GtfD, B) Result of H-bond Analysis C) Binding free energy contributions by different residues of GtfD.

binding site during dynamics also form stable hydrogen bonds and contribute significantly to the binding free energy.

TDP-epi-vancosamine bound complex of GtfD was also analyzed to understand how different orientation of single hydroxyl group affects binding. **Figure 12** shows the binding pose of TDP-*epi*-vancosamine in the active site of GtfD before MD simulations but after minimization. As opposed to TDP-vancosamine complex, hydroxyl of *epi*-vancosamine forms H-bond with catalytic D 13 of region 1 before MD simulation. Y 130 of region 2 also forms hydrogen bond with sugar hydroxyl. D 333 of region 5 forms salt bridge with sugar amino. S 246 and S 311 from regions 3 and 4 respectively bind phosphates of TDP. However, T 10 and S 247 are not involved in interaction with the substrate before simulation. **Figure 13** shows representation of TDP-epi-vancosamine bound active site of GtfD after MD simulation. Structure of active-site of GtfD in presence of TDP-*epi*-vancosamine as a sugar donor remains almost identical before and after simulation as shown in **Figure 13**. As can be seen, in contrast to TDP-vancosamine binding no rearrangement was observed after MD simulation in the TDP-epi-vancosamine:GtfD complex. Different orientation of sugar hydroxyl stabilizes TDP-*epi*-vancosamine in the conformation obtained from docking studies. Thus donor is not actively pushed closer to acceptor DVV due to lack of rearrangement. Hence, T 10 and S 247 are not involved in phosphate binding. However, as in the case of vancosamine binding D 13 and Y 130 form hydrogen bonds bonds with sugar hydroxyl, while S 311 and D 333 interact with the amino group of the donor sugar moiety. **Figure 14** shows the results of H-bond and MM_GBSA analysis. Particularly noticeable is the lack of contribution of T 10 and S 247 towards H-bonding and MM-GB/SA free energy and also higher contribution of D 333 to binding free energy compared to TDP-vancosamine.

Thus from these simulations on GtfD it can be reasoned that GtfD has higher Kcat/Km for its natural substrate TDP-vancosamine because of active site rearrangement is able to increase the proximity between donor and acceptor.

3.3.4 Binding of GtfA to TDP-vancosamine and TDP-epi-vancosamine

As mentioned earlier, GtfA transfers sugar on to β-hydroxyl group in the sixth residue of acceptor DVV. However, it has a highly stringent specificity for TDP-epi-vancosamine as sugar-donor, does not utilize TDP-vancosamine as a donor substrate. In order to understand the molecular basis of substrate selection by GtfA, hydrogen bonding and binding free energy analysis was also carried out for vancosamine and epi-
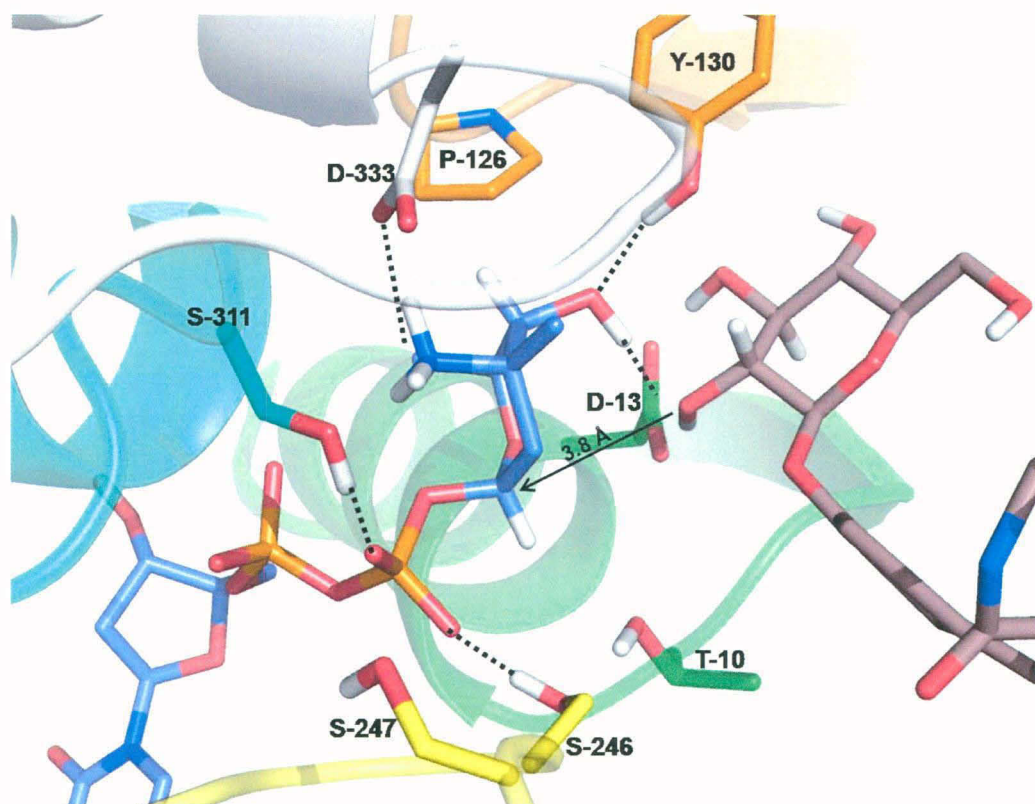
35

Figure 12: Active site of GtfD in complex with TDP-epi-vancosamine and DVV before simulation. The five different regions constituting the substrate binding site are depicted in same colors as in Figure 5. The H-bonds between the enzyme and substrate are shown as dotted lines. Arrow shows distance between attacking hydroxyl of acceptor and C-1 of epi-vancosamine.
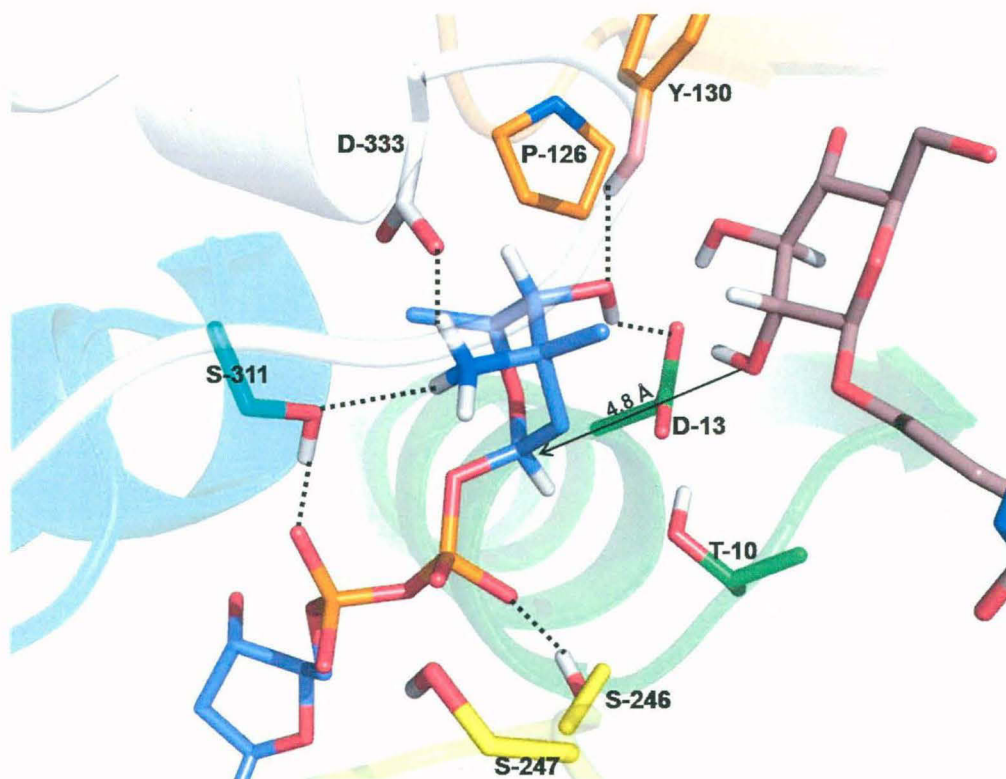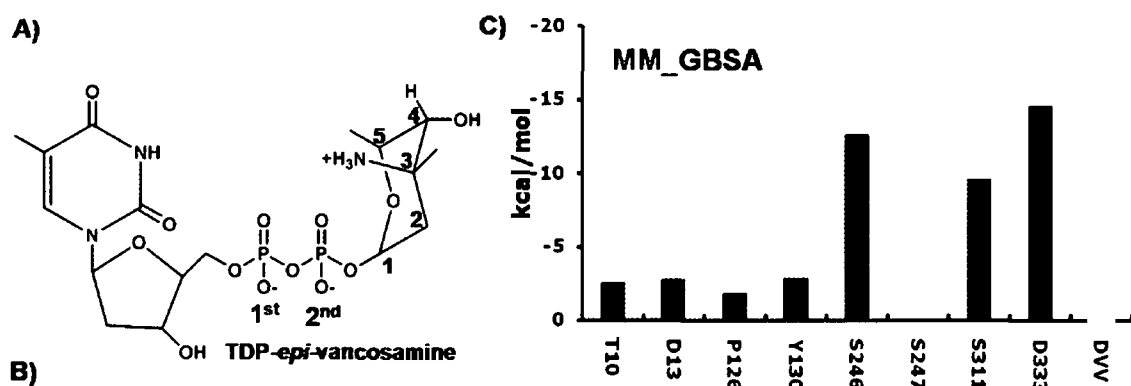
Figure 13: Active site of GtfD in complex with TDP-epi-vancosamine and DVV after 3ns MD simulation. Dotted lines depict H-bonds which are stable over the 3 ns MD simulation (based on analysis by Ptraj module of AMBER).

A)

B)

Functional Group of TDP-*epi*-vancosamine	GtfD Amino Acid Residue	Strength of H-bond (% of Trajectory)
epi-vancosamine 3-Amino	S-311	88%
	D-333	96%
epi-vancosamine 4-Hydroxyl	D-13	91%
	Y-130	98%
2ⁿᵈ Phosphate	S-246	100%
1ˢᵗ Phosphate	S-311	86%

C)

Figure 14: H-bond and MM_GBSA binding free energy analysis of the trajectory obtained from MD simulations on GtfD in complex with TDP-epi-vancosamine and DVV. A) Schematic diagram depicting different functional groups of TDP-epi-vancosamine which form H-bond with active site residues of GtfD, B) Result of H-bond Analysis C) Binding free energy contributions by different residues of GtfD.

vancosamine bound complexes of GtfA. **Figure 15** shows TDP-vancosamine bound active site of GtfA before MD simulation. GtfA donor binding regions were found to have multiple differences compared to GtfD donor binding regions. These differences and their effect are discussed below. For ease of comparison, **Figure 16** shows alignment of GtfD and GtfA with arrows pointing at unusual amino acids of GtfA compared to GtfD in sugar-donor binding regions. The conserved aspartic acid of region 5, D 317 of GtfA, is not available for making salt bridge with sugar amino as four amino acid insertion immediate next to it moves it away from active site. In addition R-313 (counterpart of I-329 and I-330 for GtfD and GtfC respectively) close to region 5, sequesters D 317 away from active site. On the other hand, unusual D 294 (equivalent of G 310 and G 311 for GtfD and GtfC respectively) of region 4 seems to be in a position to act as a replacement for D 317. However, it is also oriented away from active site due to interactions with R 314 in region 5. R 314 is in fact capable of destabilizing proper orientation of vancosamine because of repulsive interaction with sugar amino. Similarly, the conserved tyrosine of region 2 of GtfD and GtfC (Y 130 and Y 131 respectively), which forms hydrogen bonds with sugar hydroxyl in case of GtfD, is replaced by histidine in GtfA which falls short of hydrogen bonding capability with sugar hydroxyl. Thus before simulation, vancosamine moiety of TDP-vancosamine is unable to make any sustainable contact in case of GtfA and it is bound only through phosphate group of TDP.

Figure 17 shows active site of TDP-vancosamine bound GtfA after MD simulation. As can be seen, due to lack of stable contact and also possibly due to repulsive interaction with R 314 vancosamine moiety of TDP-vancosamine reverts and moves away from active site during simulation. Thus active site rearrangement seen in case of GtfD is absent during simulations on GtfA, because TDP-vancosamine fails to anchor in active site of GtfA. This explains the structural basis for the lower MM-GB/SA binding free energy and experimental observation of complete inactivity of GtfA when TDP-vancosamine is a sugar donor. Hydrogen bonding analysis and residue-wise decomposition of binding free energy shown in **Figure 18** also indicate that TDP-vancosamine interacts with GtfA through hydrogen bonds between phosphate groups and seine residues and sugar moiety contributes very little to the binding free energy.

The complex of TDP-epi-vancosamine with GtfA (**Figure 19**) was also analyzed to understand the high specificity of the enzyme towards epi-vancosamine. For the GtfA system, most notable difference when TDP-*epi*-vancosamine is a sugar donor instead of TDP-vancosamine is that opposite orientation of hydroxyl of *epi*-vancosamine enables
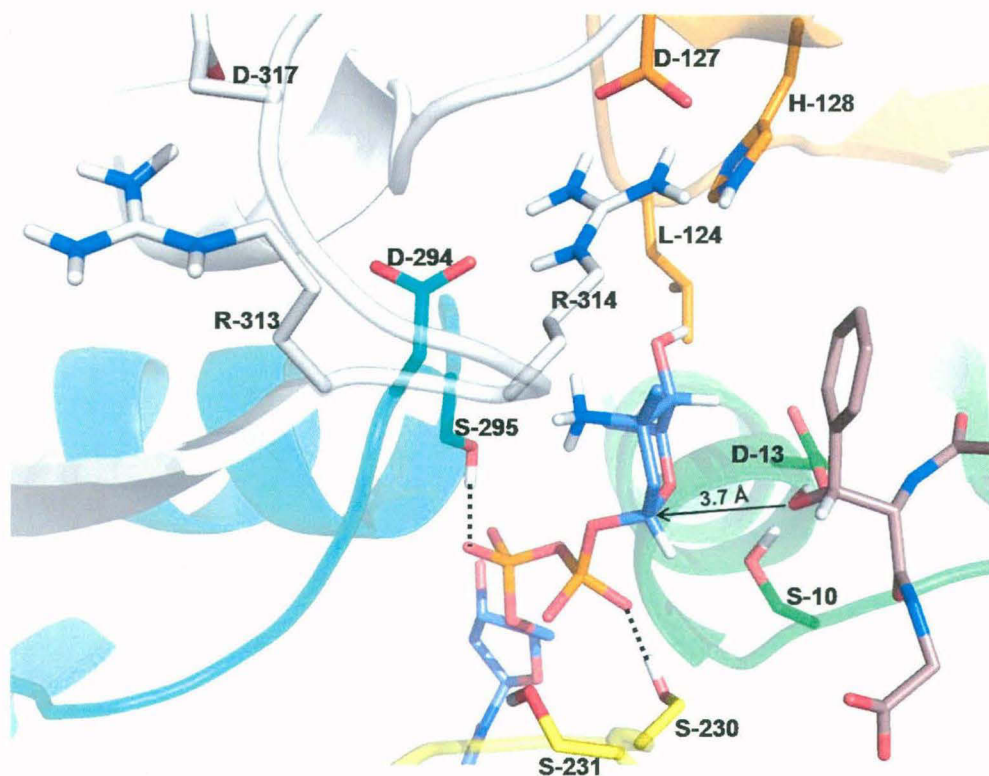
Figure 15: Active site of GtfA in complex with TDP-vancosamine and DVV before simulation. The five different regions constituting the substrate binding site are depicted in same colors as in Figure 5. The H-bonds between the enzyme and substrate are shown as dotted lines. Arrow shows distance between attacking hydroxyl of acceptor and C-1 of vancosamine.

```
GtfD  MRVLLSVCGTRGDVEIGVALADRLKALGVQTRMCAPPAAEERLAEVGVPHVPVGLPQHMM  60
GtfA  MRVLITGCGSRGDTEPLVALAARLRELGADARMCLPPDYVERCAEVGVPMVPVGRAVRAG  60
                Region 1

GtfD  LQEGMPPPPPEEEQRLAAMTVEMQFDAVPGAAEGCAAVVAVGDLAAATGVRSVAEKLGLP  120
GtfA  ARE--PGELPPGAAEVVTEVVAEWFDKVPAAIEGCDAVVTTGLLPAAVAVRSMAEKLGIP  118

GtfD  FFYSVPSPVYLASPHLPPAYDEPTTPGVTDIRVLWEERAARFADRYGPTLNRRRAEIGLP  180
GtfA  YRYTVLSPDHLPSEQS------------QAERDMYNQGADR---LFGDAVNSHRASIGLP  163
              Region 2

GtfD  PVEDVFGYGHGERPLLAADPVLAPLQP-DVDAVQTGAWLLSDERPLPPELEAFLAAGSPP  239
GtfA  PVEHLYDYGYTDQPWLAADPVLSPLRPTDLGTVQTGAWILPDERPLSAELEAFLAAGSTP  223

GtfD  VHIGFGSSSGRGIADAAKVAVEAIRAQGRRVILSRGWTELVLPDDRDDCFAIDEVNFQAL  299
GtfA  VYVGFGSSSRPATADAAKMAIKAVRASGRRIVLSRGWADLVLPDDGADCFVVGEVNLQEL  283
              Region 3

GtfD  FRRVAAVIHHGSAGTEHVATRAGVPQLVIPRNTD----QPYFAGRVAALGIGVAHDGPTP  355
GtfA  FGRVAAAIHHDSAGTTLLAMRAGIPQIVVRRVVDNVVEQAYHADRVAELGVGVAVDGPVP  343
              Region 4                    Region 5

GtfD  TFESLSAALTTVLAPETRARAEAVAGMVLTDGAAAAADLVLAAVGREKPAVPALEHHHHH  415
GtfA  TIDSLSAALDTALAPEIRARATTVADTIRADGTTVAAQLLFDAVSLEKPTVPALEHHHHH  403
```
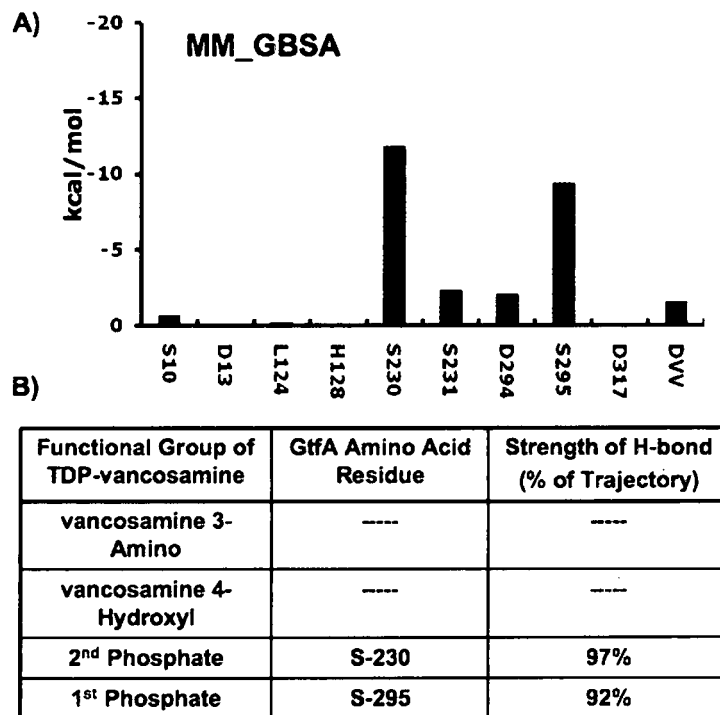
Figure 16: Sequence alignment of GtfD and GtfA. The five donor sugar binding regions are highlighted in blue. Arrows indicates critical differences between GtfA and GtfD in donor binding regions.
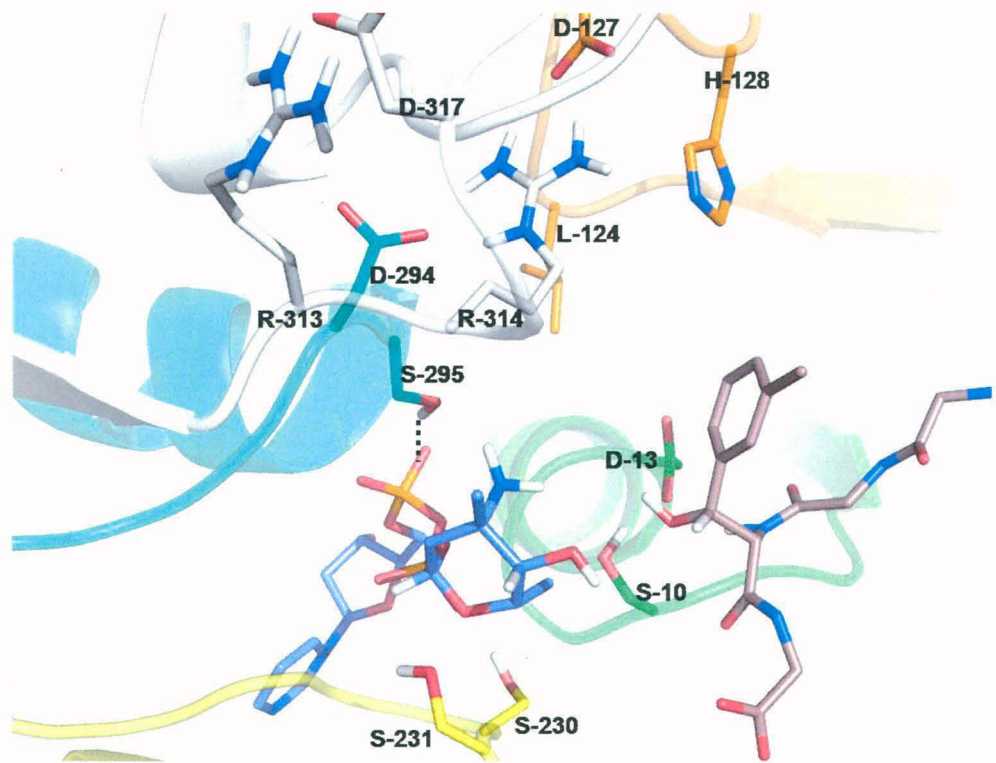
Figure 17: Active site of GtfA in complex with TDP-vancosamine and DVV after 3ns MD simulation.

A)

B)

Functional Group of TDP-vancosamine	GtfA Amino Acid Residue	Strength of H-bond (% of Trajectory)
vancosamine 3-Amino	----	----
vancosamine 4-Hydroxyl	----	----
2nd Phosphate	S-230	97%
1st Phosphate	S-295	92%

Figure 18: H-bond and MM_GBSA binding free energy analysis of the trajectory obtained from MD simulations on GtfA in complex with TDP-vancosamine and DVV. A) Binding free energy contributions by different residues of GtfD. B) Result of H-bond Analysis
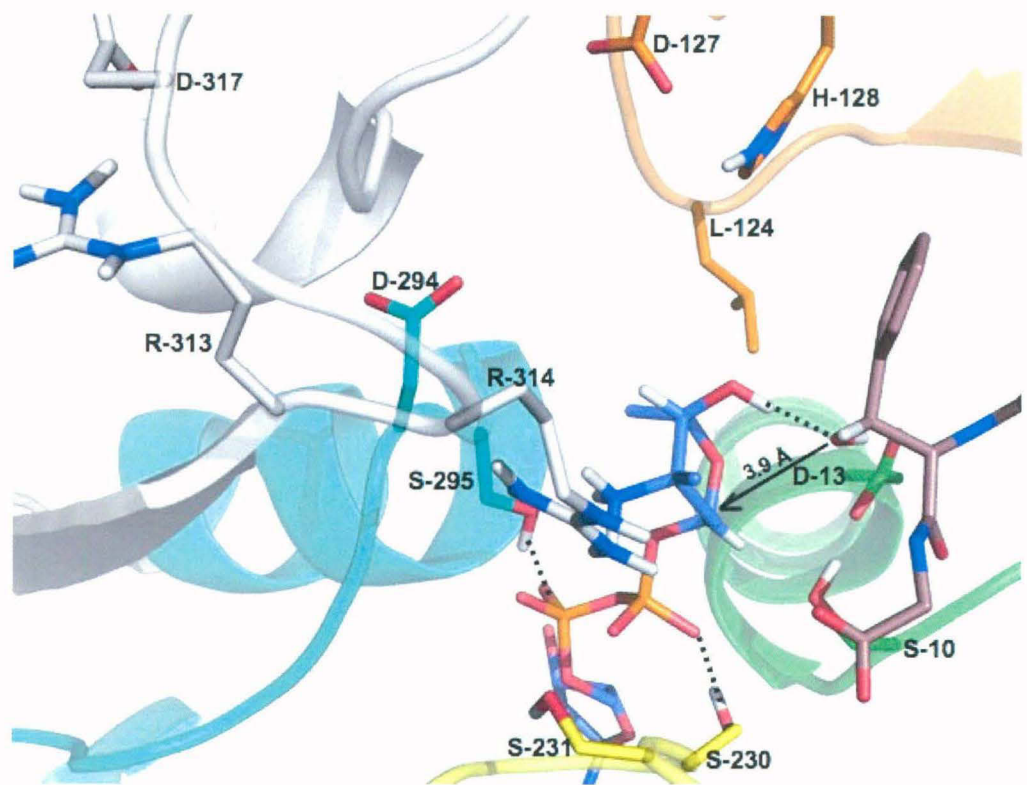
Figure 19: Active site of GtfA in complex with TDP-epi-vancosamine and DVV before simulation but after minimization. The five different regions constituting the substrate binding site are depicted in same colors as in Figure 5. The H-bonds between the enzyme and substrate are shown as dotted lines. Arrow shows distance between attacking hydroxyl of acceptor and C-1 of epi-vancosamine.

TDP-*epi*-vancosamine to make stable hydrogen bonds with-D13 of region 1 and also with hydroxyl of acceptor DVV as shown in **Figure 19**. These interactions seem to have stabilized proper orientation of *epi*-vancosamine in active site of GtfA. As opposite to the case of TDP-vancosamine, presence of stable TDP-epi-vancosamine destabilizes electrostatic interactions between R 314, D 127 and D 294. Thus R 314 moves away from active site even during minimization. This opens up the possibility of D 294 (equivalent of G 310 and G 311 for GtfD and GtfC respectively) of region 4 to act as replacement of D 317 of region 5 which has moved away from the active site due to the insertion. **Figure 20** shows the snapshot of TDP-epi-vancosamine bound active site of GtfA after MD simulations, while **Figure 21** shows the results from analysis of hydrogen bonding and residue-wise decomposition of binding free energy. As can be seen from **Figure 20**, during simulation sugar hydroxyl retains strong hydrogen bond with D 13 and hydroxyl of DVV. R 314 moves completely away from active site, as a result sugar amino forms stable salt bridge with D 294 of region 4. Thus this analysis explains how GtfA differentiates between TDP-vancosamine and TDP-*epi*-vancosamin based on difference in orientation of a single hydroxyl group. However, experimental studies indicate that catalytic ability of GtfA is much lower compared to GtfD even for its cognate substrate TDP-*epi*-vancosamine (**Table 1**). This reduced catalytic ability can be attributed to the fact that, D-294 of region 4 is a poor replacement for aspartic acid of region 5. In fact MM-GB/SA analysis (**Figure 21**) shows that D-294 contribution towards binding free energy for donor substrate is much lower than the contribution of D 333 in case of GtfD (**Figure 14**). Similarly H 128 of region 2 (counterpart of Y-130 of GtfD) is also a poor substitute for tyrosine as it is unable to form hydrogen bonds with sugar hydroxyl.

3.3.5 Binding of GtfC to TDP-vancosamine and TDP-epi-vancosamine

GtfC, like GtfA is selective for TDP-*epi*-vancosamine though it shows some residual activity for TDP-vancosamine. On the other hand it shows higher sequence similarity to GtfD compared to GtfA (**Table 2**) and the position on its acceptor substrate COB at which it transfers the donor sugar is structurally analogous to the sugar attachment site by GtfD on DVV. Thus GtfC shows a region-selectivity similar to GtfD, while its cognate substrate is same as that of GtfA. **Figure 22** shows the active site of GtfC in complex with TDP-vancosamine and acceptor substrate COB before MD simulations. **Figure 23** shows alignment of GtfC with GtfD and crucial differences in the
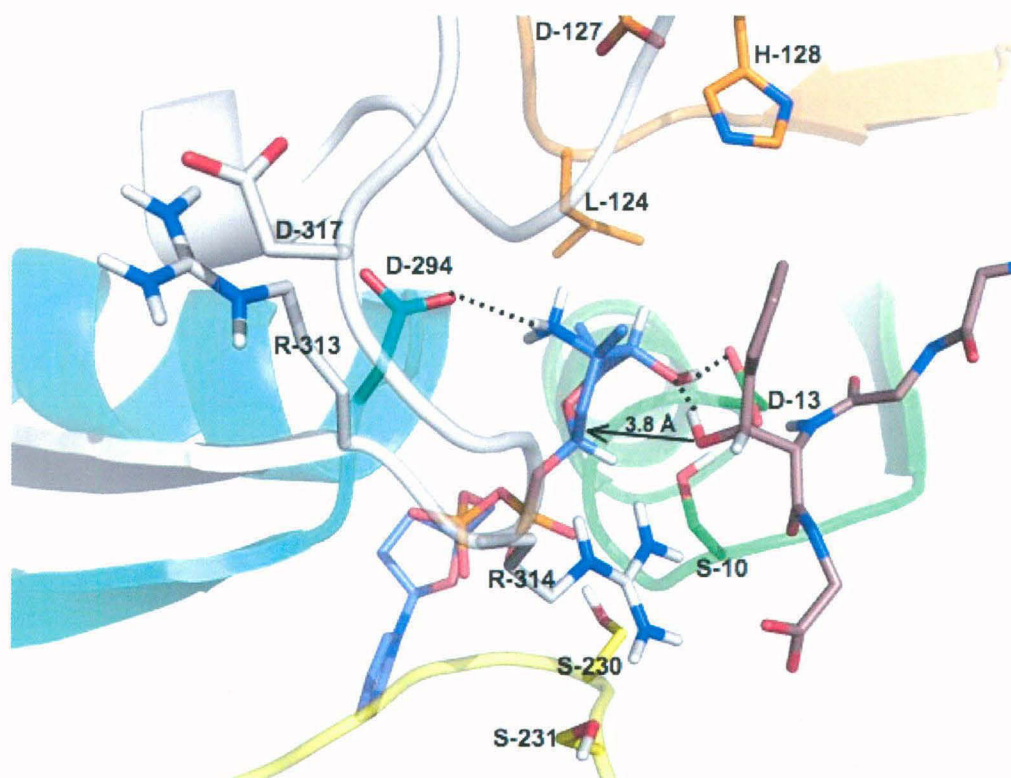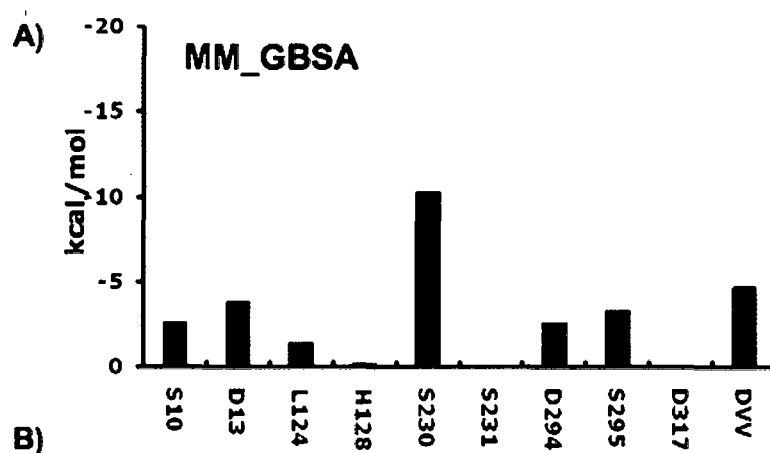
Figure 20: Active site of GtfA in complex with TDP-epi-vancosamine and DVV after 3ns MD simulation. Dotted lines depict H-bonds which are stable over the 3 ns MD simulation (based on analysis by Ptraj module of AMBER).

Functional Group of TDP-*epi*-vancosamine	GtfA Amino Acid Residue	Strength of H-bond (% of Trajectory)
epi-vancosamine 3-Amino	----	----
epi-vancosamine 4-Hydroxyl	DVV Reactive Hydroxyl (β-Hydroxyl of 6th Residue)	94%
	D-13	100%
2nd Phosphate	S-230	100%
1st Phosphate	S-295	50%

Figure 21: H-bond and MM_GBSA binding free energy analysis of the trajectory obtained from MD simulations on GtfA in complex with TDP-epi-vancosamine and DVV. A) Binding free energy contributions by different residues of GtfA. B) Result of H-bond Analysis

Figure 22: Active site of GtfC in complex with TDP-vancosamine and COB before MD simulation but after minimization. The H-bonds between the enzyme and substrate are shown as dotted lines. Arrow shows distance between attacking hydroxyl of acceptor and C-1 of vancosamine.

Figure 23: Multiple Sequence Alignment of GtfD and GtfC. Donor binding regions are highlighted in blue. Arrows indicate crucial differences in donor binding residues.

donor binding residues are indicated using arrows. As can be seen, five sugar donor binding regions of GtfC resemble more to GtfD than to GtfA. Before simulation, vancosamine does not form stable contact with catalytic D 13 of GtfC or acceptor hydroxyl. The orientation of vancosamine in the active site of GtfC is determined by the interaction of its amino group with D-334 of region 5 through salt-bridge interaction as in the case of GtfD. However, important difference compared to GtfD is in region 2, where Y 127 replaces innocuous P 126 of GtfD (**Figure 23**). This Y 127 can form hydrogen bonds with catalytic D 13, thus can act as a competitor of both donor and acceptor hydroxyls for D 13 and may negatively affect the catalytic efficiency of the enzyme. The other critical difference with respect to GtfD is in region 3 where counter part of S 247 of GtfD is A 248 for GtfC. It may be noted that, S 247 of GtfD plays crucial role in rearrangement or induce-fit effect which endows GtfD with high catalytic ability for TDP-vancosamine. Thus S 247 to A 248 mutation observed in GtfC might affect the structural rearrangement of active site which is crucial for recognition of TDP-vancosamine. MD simulations on GtfC-vanosamine complex indicate that there are no structural rearrangement in the active site because A248 in GtfC can not play the role of S 247 of GTfD. **Figure 24** shows binding of TDP-vancosamine to GtfC after MD simulation. Analysis of snapshots based on hydrogen bonding and binding free energy decomposition data (**Figure 25**) gives clear explanation for very low activity of GtfC for TDP-vancosamine. Thus vancosamine moiety of TDP-vancosamine fails to make H-bond with D 13 of region 1 or with acceptor hydroxyl due to absence of active site rearrangement. Hence, vancosamine is lifted away from acceptor COB as its sole stable contacts come from D 334 of region 5 and Y 131 of region 2. This manifests as increased distance between reactive hydroxyl of acceptor and C1 of vancosamine i.e. 5.6 Å (**Figure 24**).

Figure 26 shows TDP-epi-vancosamine bound active site of GtfC before simulations. Important difference compared to when TDP-vancosamine is sugar donor is that in case of TDP-*epi*-vancosamine, *epi*-vancosamine interacts with two different and opposite regions of binding cavity. Amino group of the sugar moiety forms salt-bridge with D 334 of region 5 and more importantly oppositely oriented hydroxyl of *epi*-vancosamine forms hydrogen bonds with D 13 of region 1 and acceptor hydroxyl. This binding mode brings donor and acceptor in close proximity and no structural rearrangement is required in the binding pocket for stable complex with epi-vancosamine. **Figure 27** shows the TDP-epi-vancosamine bound active site of GtfC after simulations.
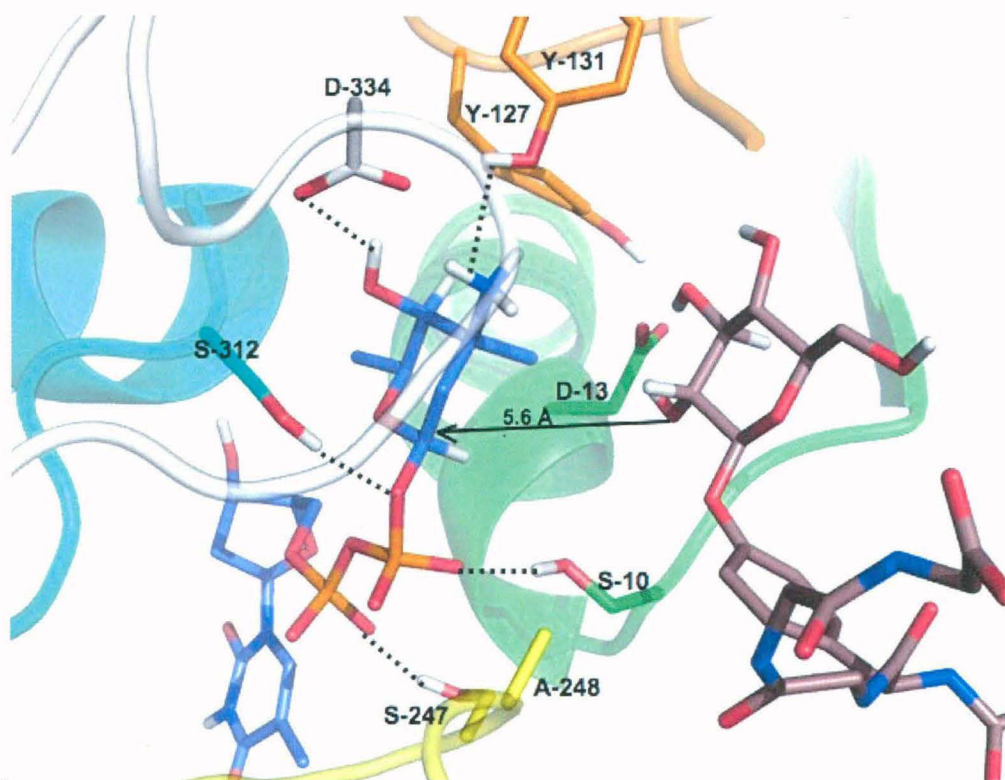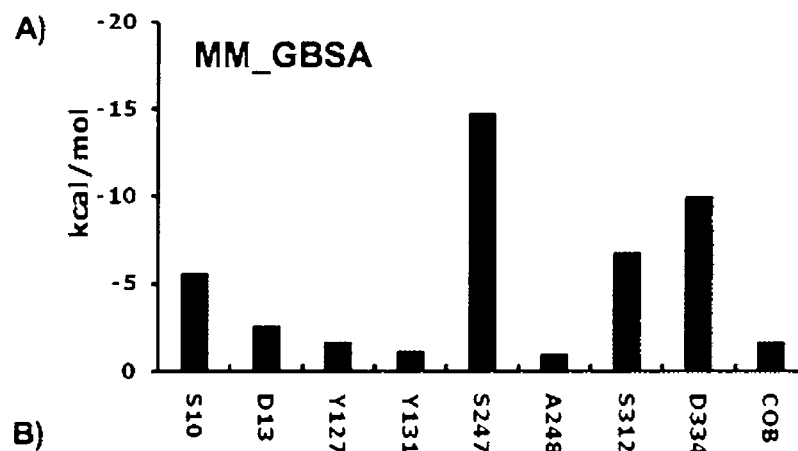
Figure 24: Active site of GtfC in complex with TDP-vancosamine and COB after 3ns MD simulation. Dotted lines depict H-bonds which are stable over the 3 ns MD simulation (based on analysis by Ptraj module of AMBER).

A)

MM_GBSA

y-axis: kcal/mol, labeled -20, -15, -10, -5, 0

x-axis categories: S10, D13, Y127, Y131, S247, A248, S312, D334, COB

B)

Functional Group of TDP-vancosamine	GtfC Amino Acid Residue	Strength of H-bond (% of Trajectory)
vancosamine 3-Amino	Y-131	61%
	D-334	66%
vancosamine 4-Hydroxyl	D-334	98%
2nd Phosphate	S-10	99%
	A-248	81%
	S-312	98%
1st Phosphate	S-247	99%

Figure 25: H-bond and MM_GBSA binding free energy analysis of the trajectory obtained from MD simulations on GtfC in complex with TDP-vancosamine and COB. A) Binding free energy contributions by different residues of GtfC. B) Result of H-bond Analysis.
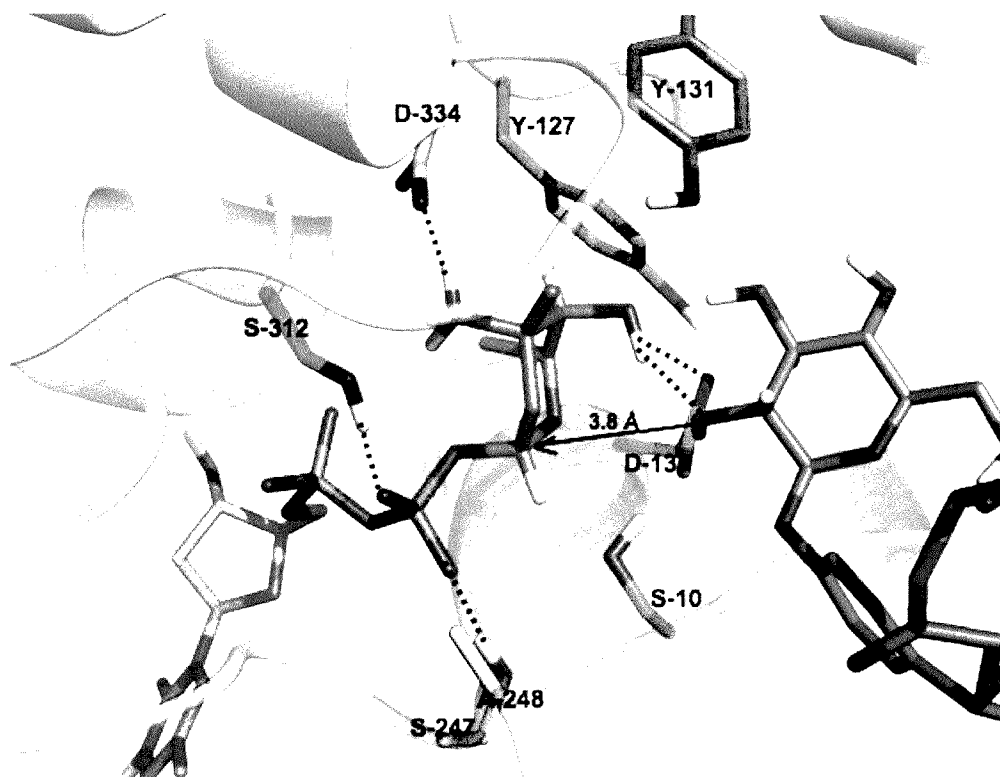
Figure 26: Active site of GtfC in complex with TDP-epi-vancosamine and COB before MD simulation. The H-bonds between the enzyme and substrate are shown as dotted lines. Arrow shows distance between attacking hydroxyl of acceptor and C-1 of vancosamine.
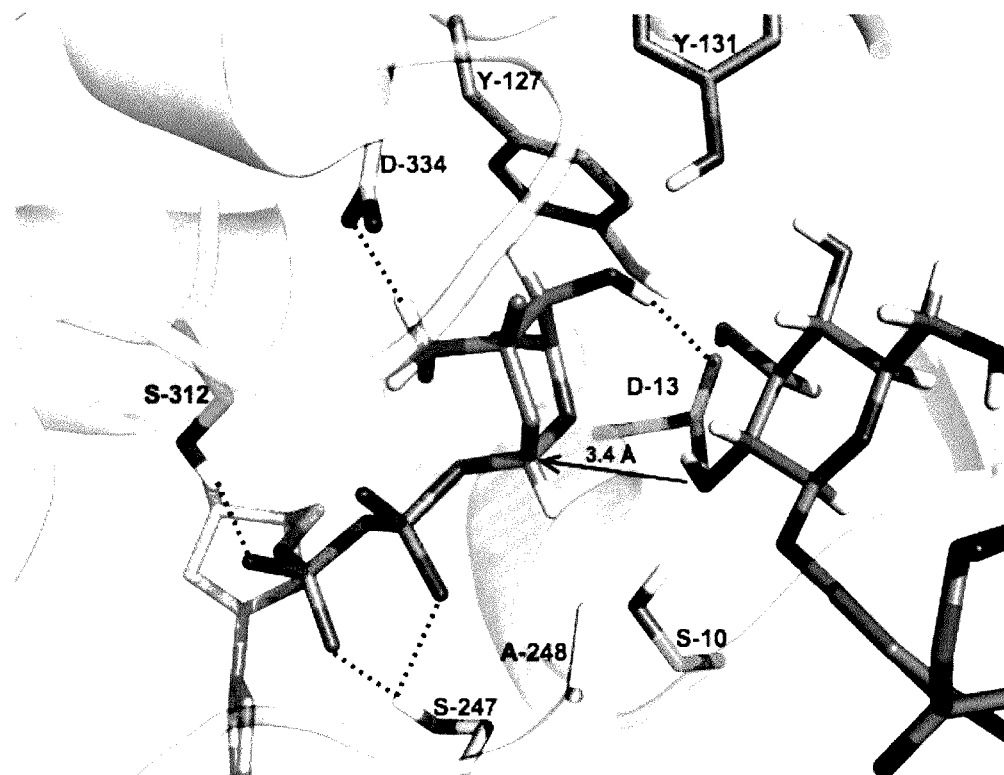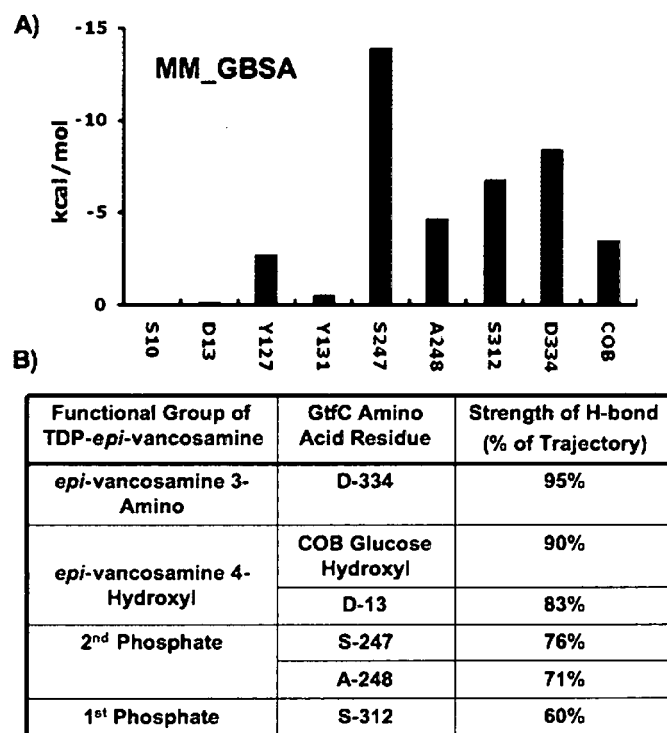
Figure 27: Active site of GtfC in complex with TDP-epi-vancosamine and COB after 3ns MD simulation. Dotted lines depict H-bonds which are stable over the 3 ns MD simulation (based on analysis by Ptraj module of AMBER).

A) MM_GBSA

kcal/mol

-15 -10 -5 0

S10 D13 Y127 Y131 S247 A248 S312 D334 COB

B)

Functional Group of TDP-*epi*-vancosamine	GtfC Amino Acid Residue	Strength of H-bond (% of Trajectory)
epi-vancosamine 3-Amino	D-334	95%
epi-vancosamine 4-Hydroxyl	COB Glucose Hydroxyl	90%
	D-13	83%
2nd Phosphate	S-247	76%
	A-248	71%
1st Phosphate	S-312	60%

Figure 28: H-bond and MM_GBSA binding free energy analysis of the trajectory obtained from MD simulations on GtfC in complex with TDP-epi-vancosamine and COB. A) Binding free energy contributions by different residues of GtfC. B) Result of H-bond Analysis.

As can be seen, there are no significant differences in the binding pose of TDP-epi-vancosamine before and after simulations. Sugar amino binds D-334 at the roof of active site cavity while sugar hydroxyl binds D-13 and acceptor hydroxyl forming bottom of the cavity. **Figure 28** shows the results from hydrogen bonding analysis in the active site pocket and residue-wise decomposition of binding free energy. Thus pattern of GtfC and TDP-*epi*-vancosamine interaction is exactly the same as pattern of GtfD-*epi*-vancosamine interaction, where no rearrangement in active site is involved in bringing donor and acceptor in proximity. Thus the results from the current simulations explain the reasons for higher selectivity of GtfC for TDP-epi-vancosamine compared to TDP-vancosamine.

3.4 DISCUSSION

GtfD, GtfA and GtfC are three important antibiotic GTrs belonging to vancomycin family and are known to exhibit differential specificity for sugar donors TDP-vancosamine and TDP-*epi*-vancosamine. However, the molecular basis of donor substrate selection by these GTrs was unknown. Synchronized use of biochemical data available from literature and docking and MD simulation carried out in the current study proved effective in giving novel insights into mechanisms by which these GTrs achieve their unique specificities. The MM-GB/SA binding free energy values obtained from the current computational analysis are in agreement with experimental observations. Detailed analyses of various substrate bound complexes have also revealed crucial specificity determining residues for each of these three GTrs. These studies suggest that GtfD has an active mechanism to transfer sugar from its natural sugar donor TDP-vancosamine involving active site rearrangement. It was also observed that GtfC has specific mutations compared to GtfD like Y 127 to P-126 and A 248 to S 247. These differences make GtfC more selective for TDP-*epi*-vancosamine. On the other hand GtfA was found to achieve its strict selectivity for TDP-epi-vancosamine by insertion near D 317 and compensatory introduction of D 294 which arises from glycine to aspartic acid mutation. Thus current studies provide novel clues for altering substrate specificity of GTrs by site directed mutagenesis studies. The current theoretical analysis also indicates that lesser activities shown by GtfA and GtfC for TDP-vancosamine and TDP-*epi*-vancosamine are result of evolutionary constraints to achieve selectivity for TDP-*epi*-vancosamine over TDP-vancosamine.

In summary, the systematic computational methodology used in this study relies on taking into account all available information about evolutionarily related enzymes and

39

then doing detailed structure based studies like docking and MD simulations. In view of the promising results obtained from antibiotic glycosyltransferases, this computational approach with combines bioinformatics and computational chemistry would be a valuable tool for natural product based drug discovery using biosynthetic engineering approach.

Chapter 4

Understanding substrate specificity of cytochrome P450 monooxygenase family of oxidoreductases

4.1 INTRODUCTION

Cytochrome P450 monooxygenase family of oxidoreductases are enzymes, which introduce polar oxygen containing groups in to its substrates and thus convert hydrophobic substance to hydrophilic. These enzymes use heme as a prosthetic group. Importance of oxidoreductases comes from their ability to catalyze oxidation of non-reactive C-H bonds(Li *et al.*, 2009a). Recently, crystal structure as well as detailed substrate specificity data from biochemical studies have became available for PikC, EryK and EryF, which are three important members of oxidoreductase family of tailoring enzymes involved in biosynthesis of natural products. Hence these enzymes can play a crucial role in generating diverse analogues of NPs. We describe below a detailed review of experimental information available on these three tailoring enzymes.

4.1.1 Biosynthetic Role of PikC

PikC is a versatile enzyme involved in biosynthesis of 12 and 14 membered ring macrolides produced by Streptomyces venezuelae (**Figure 1**)(Sherman *et al.*, 2006). It acts at the last stage of biosynthesis. Its 12 membered macrolide substrate is YC-17. It hydroxylates YC-17 at two different positions on macrolide ring i.e. at 10th and 12th carbons to give two major products methymycin and neomethymycin respectively in 1:1 ratio. PikC hydroxylates its 14 membered macrolide substrate narbomycin predominantly at a single position, C-12 on macrolide ring to give Pikromycin as a major product. Both substrates of PikC have a single sugar, desosamine attached to macrolide backbone. Thus substrates of PikC are essentially divided into two parts, macrolide part and sugar desosamine part, joined by oxygen linkage.

4.1.1.1 Structure of PikC and Its Specificity for Different Substrates

Multiple structures of PikC are available both in ligand free and ligand bound forms(Sherman *et al.*, 2006). Ligand free structures proved existence of two different conformations of PikC in absence of ligand, open and closed, as shown in **Figure 2**. Movable regions formed by BC-Helices and FG-Helices that line ligand-binding cavity bring about transitions between open and closed conformations. Heme forms the floor of ligand binding cavity. Function of such conformational plasticity was predicted to be for allowing substrate access to the active site and subsequent product release.

Structures of PikC in complex with its ligands YC-17 and Narbomycin showed presence of six distinct ligand binding regions (LBR) lining active site (**Figure 3**). As mentioned earlier, PikC substrates have two distinct parts, macrolide and desosamine. It was found that macrolide portion of substrates makes non-specific van der Waals
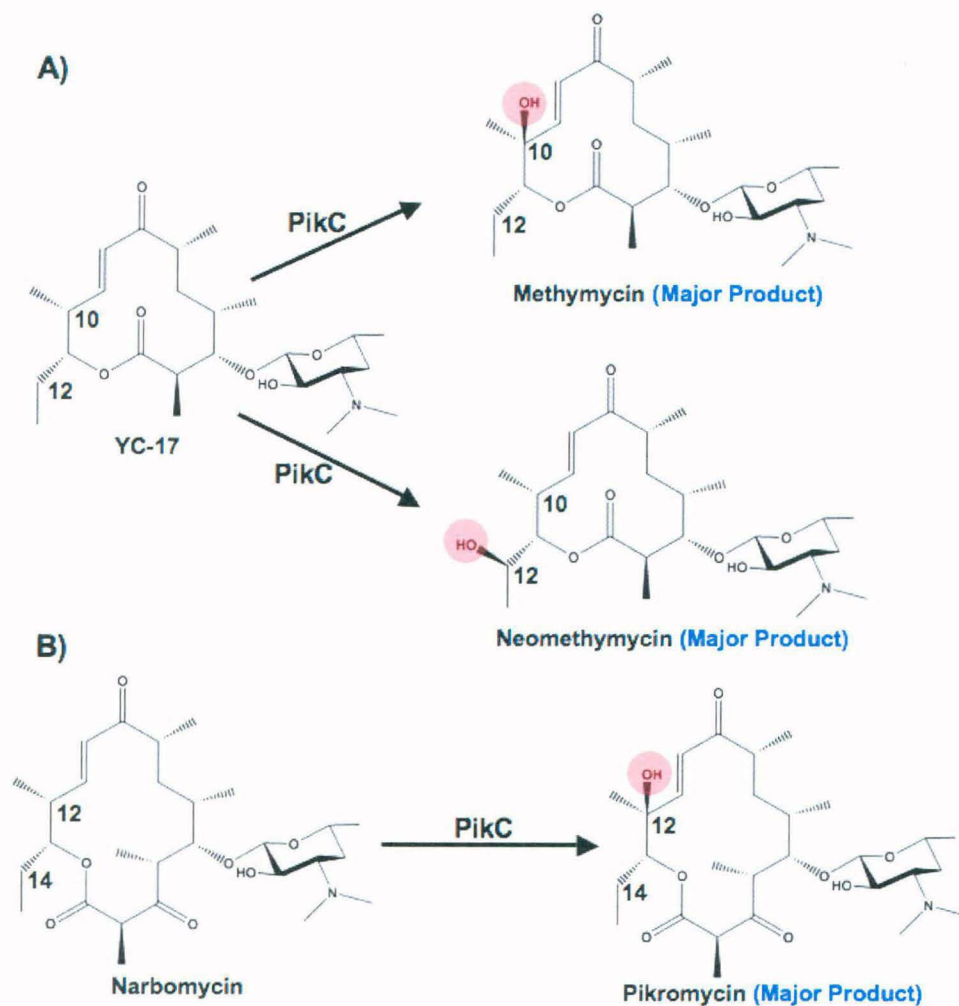
Figure 1: Substrates and major products of PikC. A) Methymycin and neomethymycin, two major products of PikC when YC-17 is the substrate. B) Pikromycin, only major product of PikC when Narbomycin is the substrate. Minor products include novamethymycin which is YC-17 hydroxylated at both C-10 and C-12 positions, neopikromycin which is Narbomycin hydroxylated at C-14 position and novapikromycin which is Narbomycin hydroxylated at both C-12 and C-14 positions. These minor products are not shown and considered here due to low yield compared to major products.
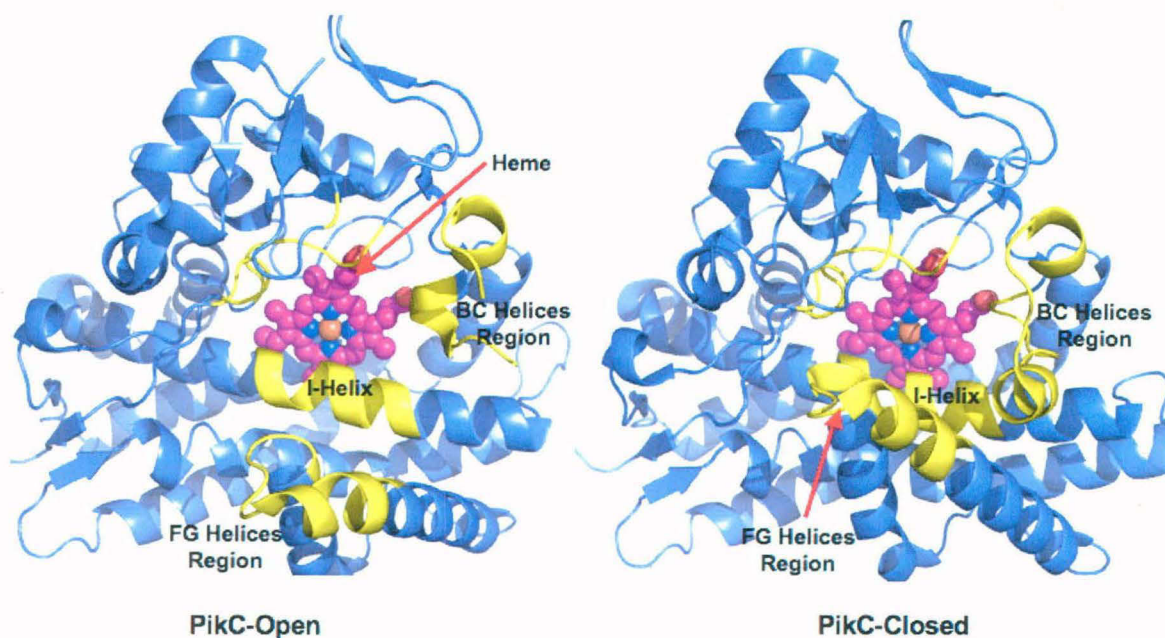
Figure 2: Open and closed structures of ligand free PikC. Images are created using Pymol from corresponding PDB. Color code: Protein-Purple, Ligand binding regions-Yellow, Heme-Magenta.
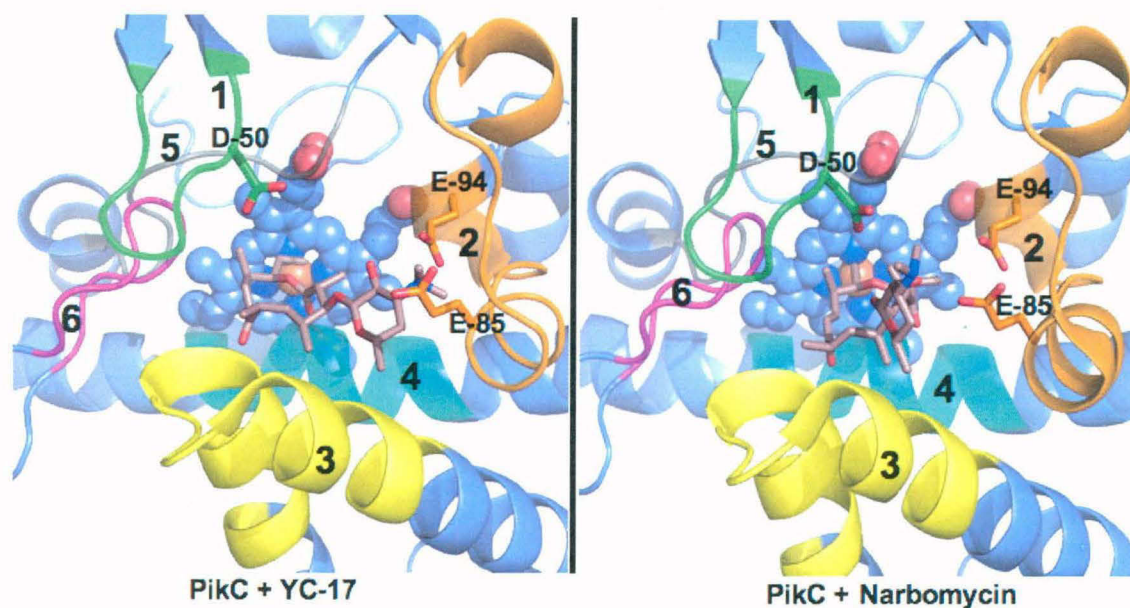
Figure 3: PikC in complex with its ligands YC-17 and Narbomycin. Images are generated from respective PDBs using Pymol. Six ligand-binding regions (LBR) are numbered and colored differently. Color code: Protein and heme-Purple, Ligands-Brown, LBR 1 (β-hairpin β1)-Green, LBR 2 (BC-Helices Region)-Orance, LBR 3(FG-Helices Region)-Yellow, LBR 4 (I-Helix)-Cyan, LBR 5 (K-Helix Terminal)-Grey, LBR 6 (β-hairpin β4)-Magenta. D-50, E-85 and E-94 are shown in stick representation and numbered.

interactions with ligand binding regions. Sugar desosamine on the other hand was found to make specific charged interactions with protein residues highlighting its importance. Desosamine of YC-17 is buried deeply between two negatively charged residues of BC-helices region E-85 and E-94. Thus Positive dimethyl-amino group of desosamine was stacked between two negative glutamic acids making ionic salt-bridge type interactions. Compared to YC-17, desosamine of Narbomycin was in surface exposed conformation, but once again sandwiched between two negatively charged residues namely D-50 of β-hairpin β1 region and E-85 of BC Helices region. Thus it was suggested that PikC binds two different ligands because of two different desosamine dependent binding modes. As desosamine and not macrolide part of both ligands was found making specific H-bond/ionic interactions with PikC, natural conclusion was that desosamine part of substrate acts as anchor and is essential for substrate recognition.

Thus based on these structures D-50, E-85 and E-94 residues of PikC were considered essential for substrate anchoring and for determining specificity. Site-directed mutagenesis studies have confirmed E-85 and E-94 residues to be essential for PikC activity by showing that E85Q and E94Q mutants of PikC are inactive or have highly reduced activity for YC-17 and Narbomycin. Thus negative charge of E-85 and E-94 was considered vital for anchoring positively charged desosamine. But contrary to the expectations, D50N mutant showed higher activity and higher affinity for both cognate ligands(Sherman *et al.*, 2006). **Table 1** lists binding affinity of PikC wild type (PikC$_{WT}$) and PikC$_{D50N}$ for YC-17 and Narbomycin.

From **Table 1** it becomes clear that YC-17 has 2.3 times higher binding affinity for PikC compared to Narbomycin. One more interesting observation is that for more active mutant PikC$_{D50N}$, binding affinity of YC-17 is increased four fold but it increases only slightly for Narbomycin i.e. 1.3 fold. Higher affinity of PikC$_{D50N}$ for YC-17 compared to Narbomycin was also expressed at the level of product yield. During 40 min of incubation time with enzymes it was observed that conversion of YC-17 to its products as a percentage of total was increased from 40% to 60% when PikC$_{D50N}$ replaces PikC$_{WT}$. But increase in Narbomycin conversion to its product was lesser i.e. from 35% to 44%(Li *et al.*, 2009b).

To solve this riddle of why PikC$_{D50N}$ is more active than PikC$_{WT}$, crystal structures were solved of PikC$_{D50N}$ in complex with YC-17 and Narbomycin. Structure of PikC$_{D50N}$ with YC-17 was very similar to the previously solved structure of PikC$_{WT}$ with

Substrate	Affinity of Enzyme $PikC_{WT}$ [K_D]	Affinity of Enzyme $PikC_{D50N}$ [K_D]
YC-17	98.9 μM	27.2 μM
Narbomycin	234.5 μM	171.9 μM

Table 1: Affinity of YC-17 and Narbomycin for PikC and $PikC_{D50N}$. Binding affinity is expressed in terms of dissociation constant K_D, thus smaller the value of K_D higher the affinity.

YC-17. In both the structures desosamine of YC-17 is buried between E-85 and E-94. But for Narbomycin, which had desosamine moiety bound in surface exposed pocket between D-50 and E-85 in PikC$_{WT}$, change were observed in binding to PikC$_{D50N}$. It was found in alternative binding mode similar to that of YC-17 in PikC$_{WT}$ and PikC$_{D50N}$, with its desosamine buried between E-85 and E-94. Based on these new structures it was proposed that buried pocket is the catalytic pocket and PikC$_{D50N}$ is more active because it facilitates passing of substrate from surface exposed pocket to buried pocket(Li *et al.*, 2009b).

All ligand bound structures showed desosamine moiety of substrate to be an anchor, which guides substrate into active site in proper orientation. Given ability of PikC to introduce stereo and regio selective hydroxylation on non-reactive sp^3 carbons of its substrates, researchers have attached desosamine to various organic molecules to see if PikC retains its activity of selective hydroxylation against non-natural substrates(Li *et al.*, 2009a). The objective of these studies was to develop an efficient enzyme with applications in synthetic organic chemistry. Among unnatural substrate tested were carbocyclic analogues like desosaminyl cyclododecane, also called carbolide-12 (mimic of YC-17), carbolide-13 and carbolide-14, mimic of Narbomycin (**Figure 4**). Enzyme used in this study was the self-sufficient and more active analogue of PikC i.e. PikC$_{D50N}$-RhFRED.

Binding affinities of carbolide-12, 13 and 14 for PikC were found out to be much lower than YC-17 and Narbomycin thus indicating that macrolide part plays possibly a vital role in substrate recognition by PikC(Li *et al.*, 2009a). As compared to YC-17, which gives 2 products, and Narbomycin, which gives single major product, carbolide-12, 13 and 14 gave 7, 6 and 6 different products respectively. Researchers have argued that this lack of selectivity by PikC against hydroxylation of unnatural substrates is likely due to higher degrees of freedom of these molecules compared to natural substrates and possibility of flipped binding in active site which will allow for hydroxylation on both sides due to substrate flipping in active site. In addition to this structures were solved for PikC$_{D50N}$ bound to carbolide-12 and carbolide-13. Significantly, no fixed orientation was observed for desosamine moiety of either substrates in the binding site. However, orientation of substrates was such that sugar residue was always pointing outward while cyclic hydrocarbon ring faced inner heme centre.

PikC performs hydroxylation of Narbomycin primarily at allylic C-12 carbon in case of Narbomycin. But in all solved structures of PikC/PikC$_{D50N}$ with Narbomycin C-12
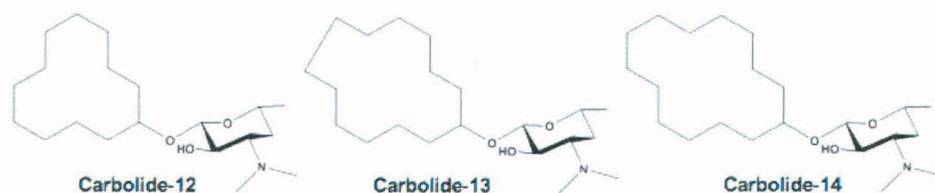
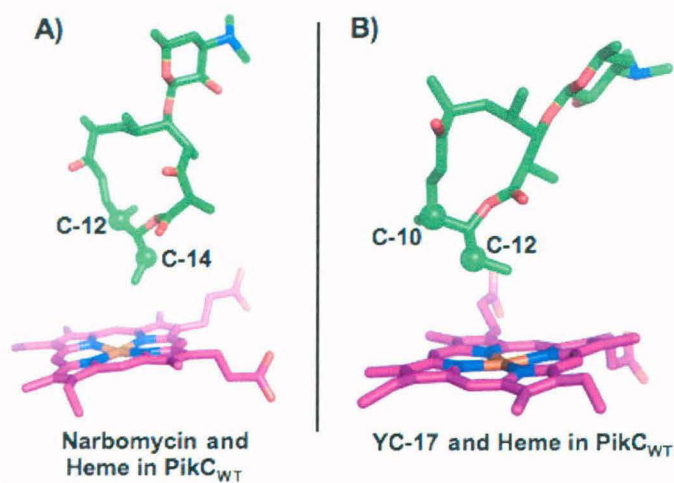**Figure 4**: Desosamine bound unnatural substrates of PikC.

Figure 5: Orientation of Narbomycin and YC-17 in PikCWT active site with respect to Heme. Images are made from PDB files of respective crystal structures. A) PikCWT in complex with Narbomycin, showing orientation of C-12 and C-14 of Narbomycin to oxidizing heme centre. B) PikCWT in complex with YC-17, showing relative distance of C-10 and C-12 of YC-17 from oxidizing heme centre.

is situated away from oxidizing heme centre at a distance of 7.1 Å (**Figure 5**). Instead methylene C-14 is situated more favorably at a distance of 5.3 Å. But ratio of C-12 to C-14 hydroxylation of Narbomycin is 40:1. Similarly, PikC primarily hydroxylates YC-17 at C-10 and C-12 in ratio 1:1. But here too allylic C-10 is situated at a distance of 7.5Å while methylene C-12 is situated at 5.3 Å (**Figure 5**).

4.1.2 Biosynthetic Role of EryK

Compared to PikC, EryK is a stringent enzyme (**Figure 6**). It acts at the second last stage of erythromycin biosynthesis by performing hydroxylation at the 12th position of its substrate Erythromycin D (ErD) to yield Erythromycin C (ErC)(Savino *et al.*, 2009). Substrate ErD has two sugar moieties attached to macrolide backbone, mycarose at C-3 and desosamine at C-5. Methyltransferase EryG then completes biosynthesis of erythromycin by transferring methyl on to mycarose hydroxyl to produce Erythromycin A (EryA). But EryG can competitively act on EryK substrate ErD to produce Erythromycin B (ErB). Interestingly EryK cannot recognize and hydroxylate ErB; this leads to significant accumulation of shunt product during ErA production. This fact highlights the specificity of EryK, as it is intolerant to addition of single methyl group on its large substrate ErD.

4.1.2.1 Structure and substrate specificity of EryK

Like PikC, in absence of ligands EryK exists in two different conformations, open and closed(Savino *et al.*, 2009). Again, like PikC, six distinct regions form ligand-binding cavity of EryK and heme forms bottom of cavity. **Figure 7** shows EryK open and closed structures. As in PikC, BC-Helices region and FG-Helices region are the movable parts of ligand-binding site and bring about open/close transitions of EryK. During crystallization studies on EryK, researchers have noted that under low salt concentration came open conformation is favoured, while at higher salt concentration closed form of EryK is more stable. At physiological salt concentration open to closed ratio was estimated to be 10:1.

Six ligand-binding regions of EryK are structural equivalent to those of PikC and are labeled accordingly as 1) β-Hairpin β1 Region, 2) BC Helices Region, 3) FG Helices Region, 4) I-Helix Region, 5) K Helix Terminal Region, 6) β-Hairpin β4 Region. BC Helices Region of EryK is much smaller than that of PikC, possibly to accommodate extra sugar on its substrate. Just like was the case for PikC, outward movement of BC and FG Helices Regions gives open conformation and inward movement gives closed conformation. During open to close transition in absence of ligand BC-Helices Region moves inward by 11Å. FG-Helices movement is not uniform as each part of FG-Helices
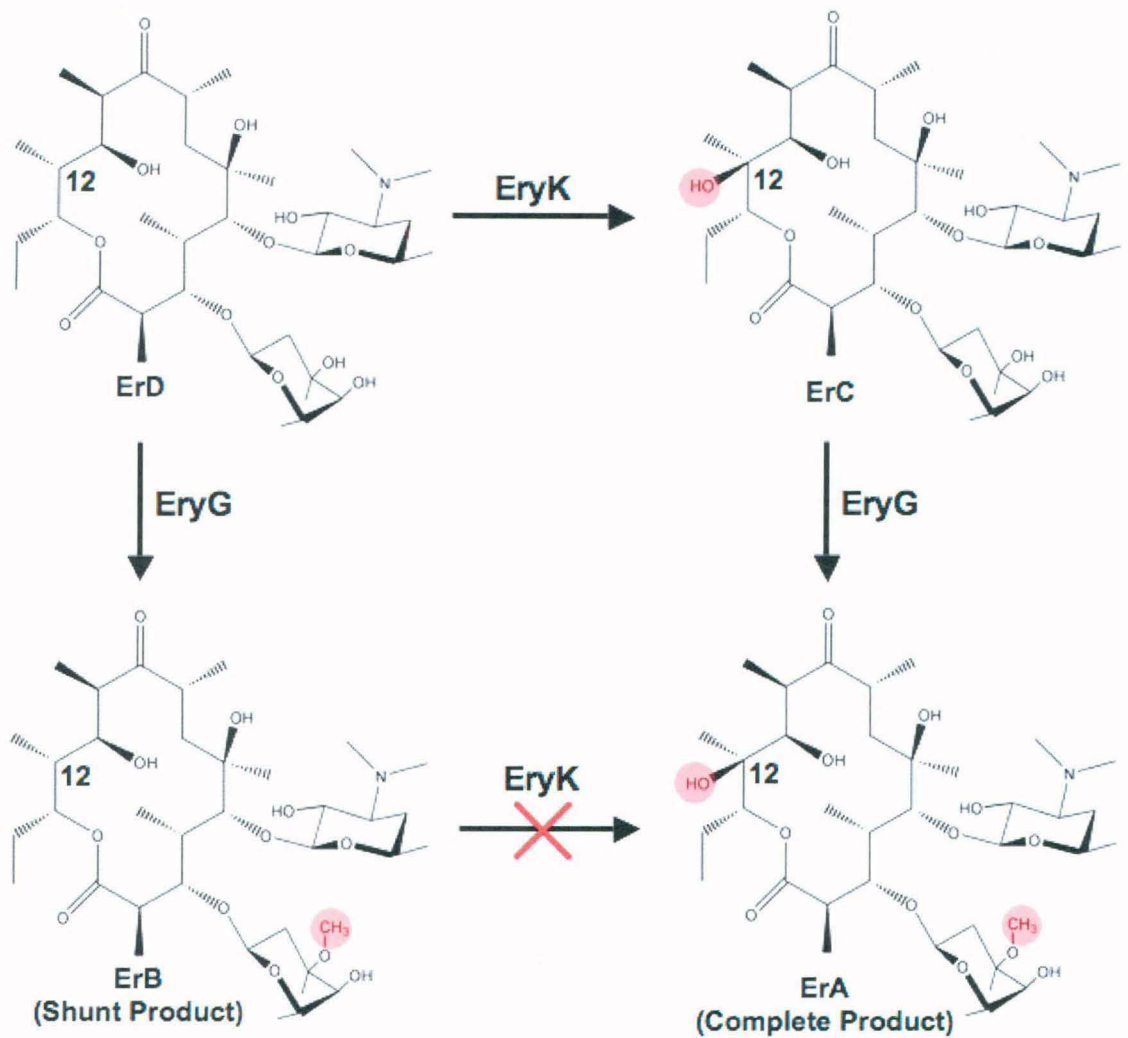
Figure 6: Reaction catalyzed by EryK. It converts ErD to ErC by C-12 hydroxylation. Diagram also shows function of methyltransferase EryG and the fact that if EryG acts before EryK it leads to accumulation of shunt product ErB as additional methyl group inhibits hydroxylation by EryK.
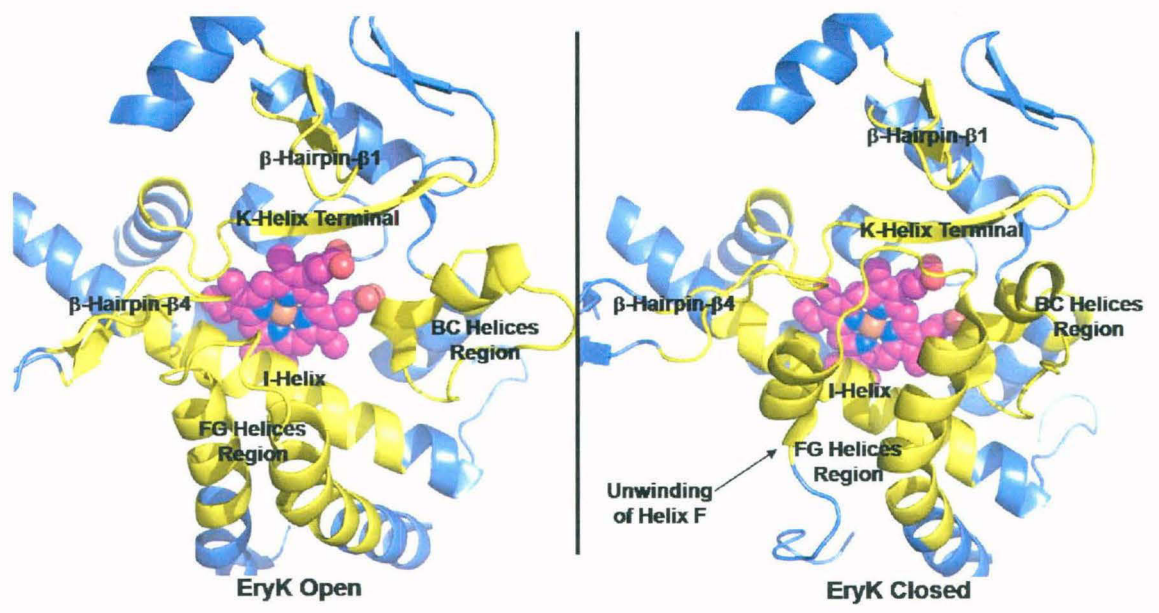
Figure 7: EryK open and closed structures. Color code: Protein-Purple, Ligand binding regions-Yellow, Heme-Magenta.

Region Responds differently. G-Helix N-terminal part moves inward by 10 Å, this movement is helped by kink at P-192 of G-Helix. This P-192 is also a part of HSWP network which stabilizes open conformation (**Figure 8**). This proline is not conserved across homologues P450s. This movement of Helix G pulls FG-loop and F-Helix inwards causing latter to unwind almost completely (**Figure 7, 8**).

4.1.2.2 HSWP Network in Open and Closed EryK in Absence of Ligand

Another difference between open and closed conformation is the presence of HSWP network in open conformation, consisting of H-243 (I-Helix), S-166 (F-Helix), W-165 (F-Helix) and P-192 (G-Helix). This network links I-Helix to G-Helix through F-Helix and stabilizes open conformation (**Figure 8**). In closed conformation without ligand, this HSWP network is broken along with unwinding of F-Helix. W-165 rotates and makes hydrogen bond with H-243. I-F-G Helices linkage is thus broken due to loss of H-S and W-P interactions during open to close transition in absence of ligand(Savino *et al.*, 2009).

4.1.2.3 EryK ErD Complex

ErD bound EryK assumed closed conformation but differs from ligand free closed state in two aspects, 1) HSWP network is preserved, 2) F-Helix is intact (**Figure 9**). Substrate ErD anchors to EryK via its sugars mycarose an desosamine (**Figure 9 A**). Mycarose forms hydrogen bonds with H-88 and E-89 of BC-Helices Region. Desosamine interacts with N-290 and N-292 of K-Helix Terminal Region. Macrolide ring of ErD makes Van der Waals interactions mainly with I-Helix but also with K-Helix Terminal and β-Hairpin regions(Savino *et al.*, 2009).

4.1.2.4 HSWP network in EryK in presence of ErD

Transformation from open to closed state after ligand binding involves movement of BC-Helices Region and FG-Helices Region. As depicted in diagram F-Helix maintains its conformation without unfolding during inward movement (**Figure 9 B**). This movement of F-Helix during open to close shift is achieved by its lifting by 3.7 Å and tilting by 13° as oppose to ligand free closed state where it unfolds. During this transition it maintains hydrogen bonds with neighboring regions of β-Hairpin and G-Helix. Helix-G movement is aided by kink at P-192. HSWP network is preserved during open to close transition in presence of substrate ErD. HSWP network is present in open conformation, it is absent in ligand free closed conformation but is present in closed conformation in presence of substrate ErD. Hence it was proposed that this HSWP network stabilizes and
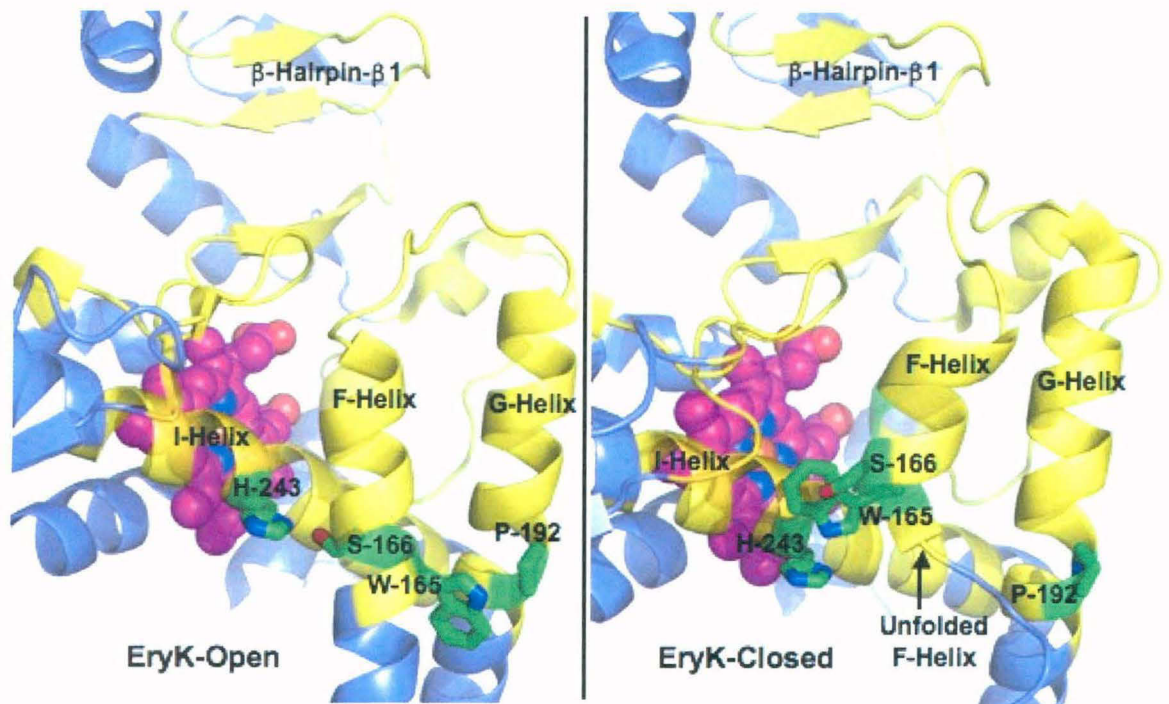
Figure 8: Side view of EryK open and closed conformations shown in **Figure** 7. This view shows HSWP linkage present in Open conformation and abscent in closed conformation due to unwinding of F-Helix and rotation of W-165.
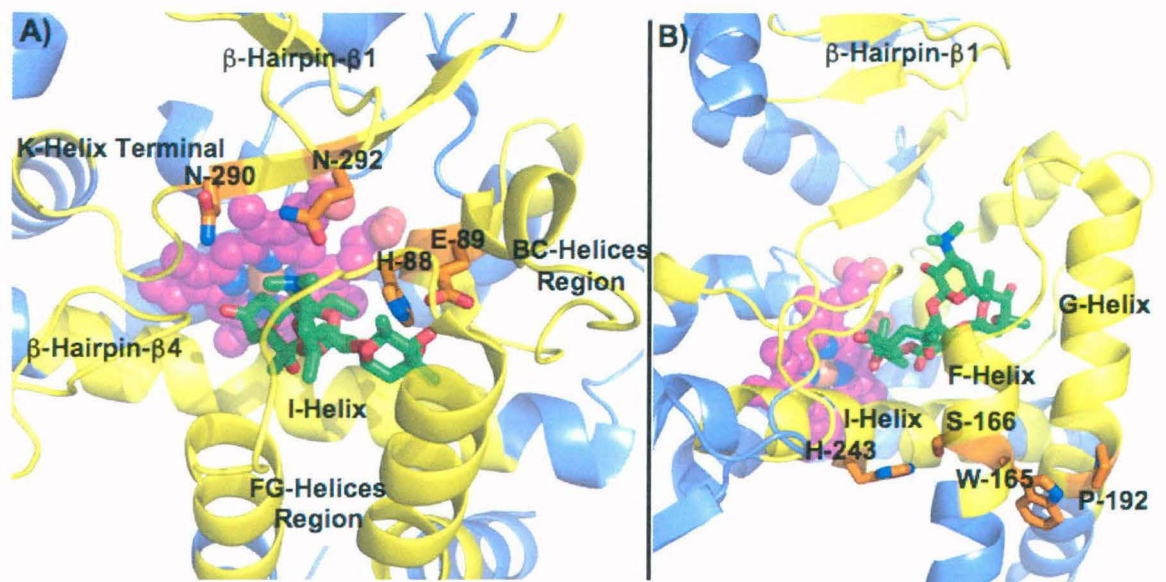

Figure 9: EryK with substrate ErD. Color Code: Protein-Purple, Ligand-Binding Regions-Yellow, Ligand-Green, Critical EryK Residues-Orange and Heme-Magenta. **A)** Top view of ligand binding site, showing ligand binding regions and position of important amino acids. **B)** Side view of EryK active site showing intact HSWP network.

locks closed conformation in presence of substrate ErD so as to allow catalysis, while due to lack of this interaction ligand free closed state is less stable hence leads to more open to close ratio in absence of ligand.

When ErD binds to EryK, its macrolide ring establishes extensive interactions with I-Helix, which induces pronounced bending in the helix. I-Helix then transfers this bending energy to FG-Helices Region as follows. I-Helix bending is accompanied by flipping and reorientation of its H-243 residue, which is now capable of sustaining its hydrogen bond with S-166 of F-Helix and induces movement of F-Helix. Due to stabilizing effect of this interaction, F-Helix retains its integrity and shifts forward without unwinding. W-165 of F-Helix exerts pressure on G-Helix through its interaction with P-192 of G-Helix. At this pressure G-Helix kinks at P-192, aiding its forward movement. Altogether, FG-Helices region moves inwards and acts as a lid to close ligand access channel. This repositioning of FG-Helices region allows it to make van der waals contacts with BC-Helices region. Movement of BC-Helices region to fortify these van der waals contacts leads to complete closure of active site channel(Savino *et al.*, 2009).

4.1.2.5 Structures of EryK with Inhibitors Clotrimazole and Ketoconazole

Clotrimazole (CLT) and Ketoconazole (KC) both inhibit EryK by coordinating heme iron through nitrogen moiety of azole ring. **Figure 10** shows schematic structures of both inhibitors. Structures of EryK with these inhibitors proved that CLT traps EryK in open conformation while KC traps EryK in closed conformation (**Figure 11**)(Montemiglio *et al.*). Stout and non-polar CLT binds EryK mostly through van der waals interactions while slender and long KC is capable of making some hydrogen bond/hydrophilic interactions with EryK. KC binds in nearly straight orientation along the ligand access channel so that nitrogen moieties of its piperazine ring can establish hydrogen bonds with K-Helix Terminal Region and β-Hairpin β4 Region. KC also makes hydrophobic interactions with some residues of FG-Helices Region. Proposed theory states that these inhibitors affect protein conformations primarily due to their differential interactions with flexible I-Helix Region and thus by affecting HSWP network. It was suggested that terminal acetyl moiety of KC interacts with FG-Helices Region of EryK and strain induced through these interactions pulls F-Helix up towards ligand access channel, thus disturbing HSWP network(Montemiglio *et al.*).
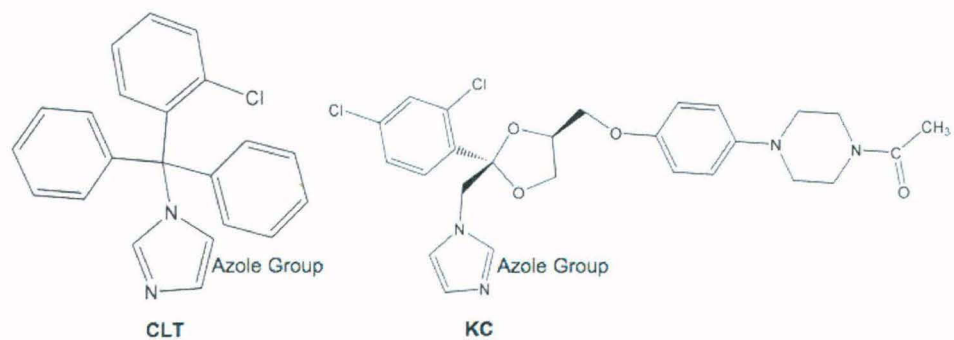
Figure 10: Structural representations of Clotrimazole and Ketoconazole, inhibitors of EryK. Both inhibitors coordinate heme iron of EryK through nitrogen moiety of azole ring to inhibit the enzyme.
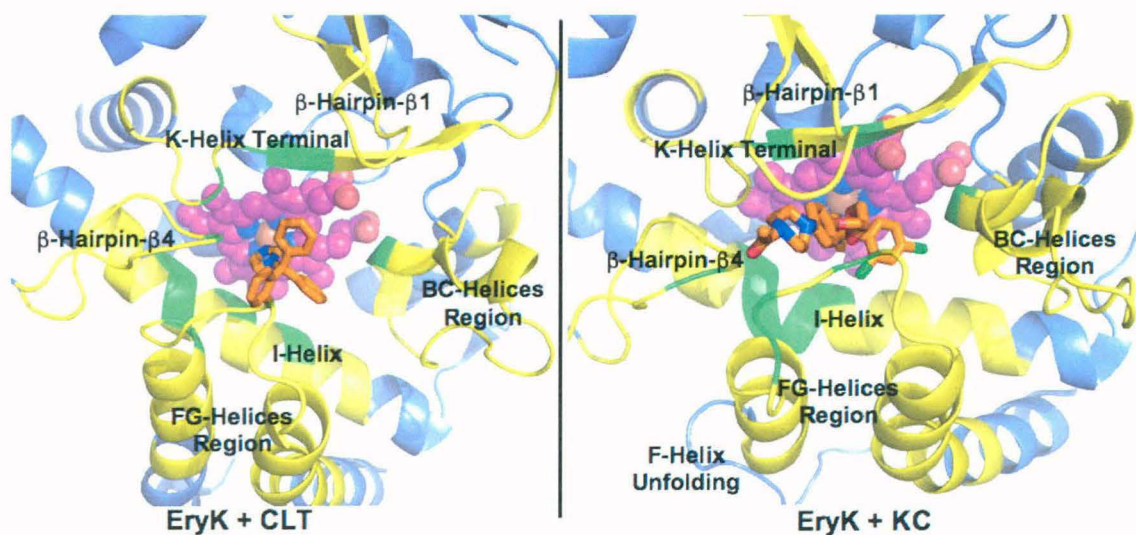

Figure 11: EryK in complex with CLT and KC. Color Code, Protein-Purple, Ligand-Binding Regions-Yellow, Position of residues in contact with ligand-Green, heme-Magenta, Ligands-Orange.
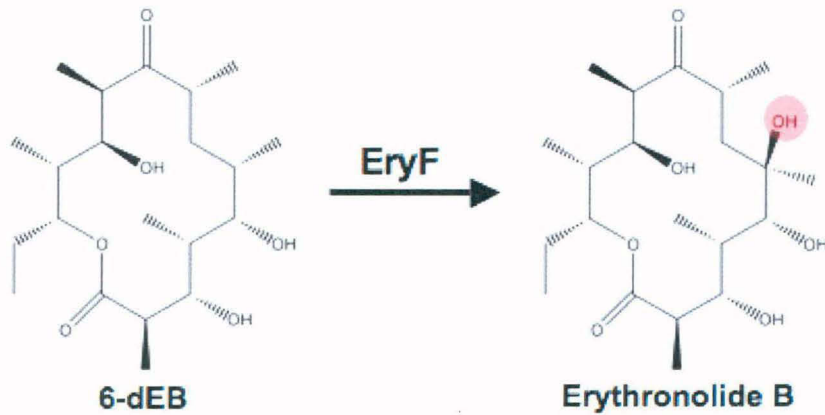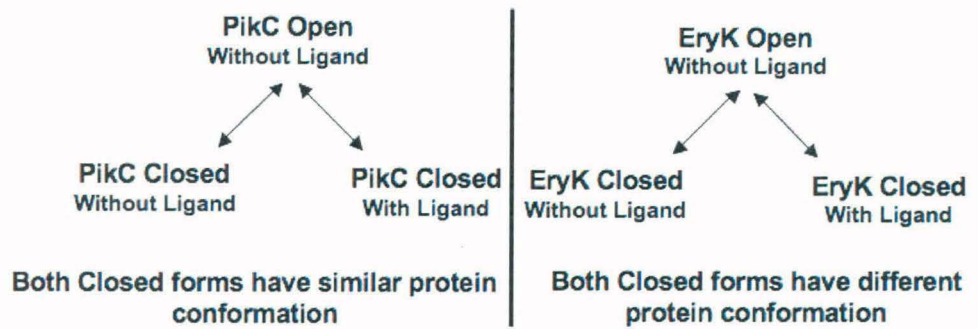
Figure 12: Schematic diagram showing conformational dynamics of PikC and EryK, and reaction catalyzed by EryF. As oppose to PikC and EryK, EryF substrate has no sugar attached.

4.1.2.6 Specificity of EryK for different substrates

As detailed earlier EryK has much stricter specificity compared to PikC as it fails to hydroxylate ErB, which differs from its large cognate substrate ErD by a single methyl group. Interestingly though EryK is known to hydroxylate substrates of PikC, at similar positions to PikC, albeit with 100 fold lesser yield than PikC(Lee *et al.*, 2004). **Table 2** gives binding affinity shown by EryK towards its substrate ErD and inhibitors CLT and KC(Montemiglio *et al.*). Compared to PikC, binding affinity showed by EryK for its substrate is much higher. Hydrophobic and stout inhibitor CLT binds EryK with very high affinity compared to ErD and another slender comparatively hydrophilic inhibitor KC.

4.1.3 Biosynthetic Role of EryF

EryF is the first tailoring enzyme involved in erythromycin biosynthesis. Thus it acts before glycosyltransferases and its substrate, 6-deoxyerythronolide B (6-dEB) has no sugar attached. It introduces hydroxyl group at the 6th position on macrolide backbone to produce erythronolide B (**Figure 12**)(Cupp-Vickery & Poulos, 1995).

4.1.3.1 Structure and Specificity of EryF

Structure of EryF in complex with its substrate 6-dEB once again shows conserved architecture of ligand-binding site in oxidoreductases (**Figure 13 A**). Ligand-binding cavity is formed by six regions, which are structural equivalents to those of PikC and EryK. Substrate 6-dEB is held in active site by non-polar residues from these regions through van der waals interactions. Both BC-Helices Region and FG-Helices Region are tapering and curved in. This architecture along with extended N-Terminal part of β-Hairpin β1 Region, which consists of disordered α-Helix, ensures tight closing of ligand binding cavity. It also makes ligand-binding cavity much smaller. Substrate 6-dEB is buried deeper in active site compared to substrates of PikC and EryK(Cupp-Vickery & Poulos, 1995). No ligand free structures of EryF are available, thus making it difficult to predict if it has different conformations in absence of ligand as was observed for PikC and EryK.

In EryF structure, A-245 is present at a position, which is occupied by conserved threonine in most other P450s. The hydroxyl of this threonine is implicated to act as a proton donor to heme bound dioxygen during reaction cycle. Instead C-5 hydroxyl of substrate 6-dEB (**Figure 13 B**) is at a structurally equivalent position to threonine

Enzyme	Affinity for Substrate ErD [K_D]	Affinity for Inhibitor Clotrimazole [K_D]	Affinity for Inhibitor Ketoconazole [K_D]
EryK	3.5 μM	Lesser than 0.01 μM	1.12 μM

Table 2: Binding affinity of EryK for its substrate ErD and inhibitors Clotrimazole and Ketoconazole. Affinities are expressed in terms of dissociation constant K_D, thus lower the values higher the affinity.
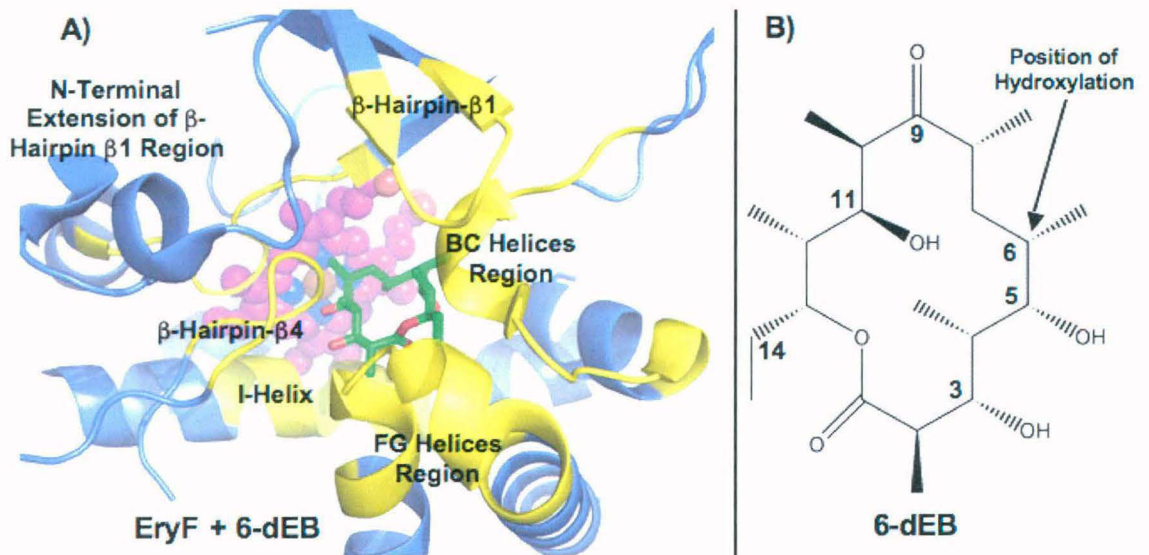

Figure 13: **A)** EryF in complex with substrate 6-dEB. Color Code, Protein-Purple, Ligand-binding region-Yellow, Substrate 6-dEB- Green, Heme-Magenta. **B)** Numbered schematic diagram of 6-dEB.

hydroxyl, as is postulated to take role of threonine hydroxyl. Analysis of substrate specificity of EryF revealed it to be the enzyme of strict specificity, as even slight modification of its substrate leads to significant loss of enzyme activity. Oxidation of C-5 hydroxyl (**Figure 13 B**) to carbonyl, reduction of C-9 carbonyl to hydroxyl and even replacement of methyl at C-14 with hydrogen all significantly affect activity of EryF for modified substrate(Andersen & Hutchinson, 1992).

4.2   RATIONALE

Structure and specificity data is available for three oxidoreductase tailoring enzymes PikC, EryK and EryF. However, a number of important questions relating to substrate selection by these tailoring enzymes are still unanswered. In this study an attempt has been made to apply structure-based methods involving docking and MD simulations on these enzymes for answering some of these questions. Since these enzymes are related to each other structurally, sequentially and functionally, comparative analysis of their sequence, structure and dynamics will give detailed insight into mechanisms employed by these enzymes for substrate recognition.

The specific questions the current study attempts to address are

1. Why hydroxylation of YC-17 by PikC gives two primary products while hydroxylation of Narbomycin by PikC gives only a single major product? Why does $PikC_{D50N}$ show higher catalytic activity for YC-17 compared to Narbomycin? Why cyclic hydrocarbons attached to desosamine have much reduced affinity for PikC compared to natural macrolide substrates? Why does PikC bind YC-17 and Narbomycin by desosamine anchoring while no desosamine anchoring is involved in recognition of carbolide-12 and carbolide-13?

2. Does desosamine of ErD play any anchoring role in ErD recognition by EryK? How does EryK discriminate between ErD and ErB which differ by a single methyl on mycarose hydroxyl? How do various factors like ligand binding, HSWP network, salt concentration etc control open/closed form conformational transition in EryK?

3. In contrast to PikC and EryK substrates which have attached sugar/sugars, EryF substrate 6-dEB is a 14-membered macrolide ring without attached sugars. How EryF achieves proper anchoring and orientation of its cognate substrate 6-dEB in the active site with high selectivity?

4.3   METHODS

4.3.1   Homology modeling and analysis of models as well as structural templates

The various crystal structures used in the current analysis were 2BVJ, 2CAO and 2C7X for PikC, 2JJN, 2WIO and 2JJO for EryK and 1Z8O for EryF. The open form structure of PikC obtained from PDB had several missing residues in flexible BC-helices region. The side chains of some residues in ligand binding site were also missing. Hence Modeller (Version mod9v1) software was used to build the complete open form structure of PikC taking fragments from other structures as template(Fiser & Sali, 2003). The various crystal structures of PikC, EryK and EryF along with structures obtained through docking and MD Simulations were analyzed using Pymol visualization software.   Other structural analysis were done though in-house scripts/programs.

4.3.2   Docking of substrates

To find possible alternate binding poses for PikC ligands YC-17 and Narbomycin, these ligands are docked in to PikC structure using Autodock 4(Huey et al., 2007). Open conformation of PikC is selected for docking studies as it has two advantages, 1) as it is non-liganded it will not have bias for one particular conformation of ligand for which side-chains of active site amino acids are already committed and 2) Open and larger active site will allow for inclusion of all possible binding modes in analysis.

4.3.3   Molecular dynamics simulations

To understand enzyme substrate dynamics explicit solvent MD Simulations were performed on the crystal structures of the enzyme-substrate complexes using AMBER 9 package(Case, 2006).

1) PikC + YC-17
2) PikC + Narbomycin
3) EryK + ErD
4) EryK + ErB

Antechamber module of Amber was used to assign force field parameters for donor and acceptor substrates apart from the protein (Wang et al., 2006). The protein-ligand complexes were energy minimized in vacuum to remove steric clashes if any and then the minimized complex was solvated in a water box which extended 9 Å from the outermost atom of the protein-ligand complex in X, Y and Z directions. Electrostatic interactions were computed using PME approach and a cut off of 8 Å was used for non-bonded interactions.   The forcefield used was ff03 (Duan et al., 2003) for the protein and TIP3P water model for solvent. Heme parameters are taken from published work. All the

simulations were carried out for a period of 5 ns in NVT ensemble at 300K using a time step of 1 fs. SHAKE algorithm was used to constrain bonds involving hydrogens. The convergence of the simulations was monitored in terms of RMSD between the starting structure and the structures sampled during the simulations.

4.3.4 Analysis of MD trajectories and evaluation of binding free energy

To extract desired data from simulation trajectories, MM-PB/SA module of AMBER was used for calculation of binding free energy between protein ligand complexes. This gives indication of strength of protein ligand complex. It also gives contribution of each energy component towards binding free energy i.e. electrostatic, van der Waals and solvation free energy etc. Ptraj tool of AMBER was used to monitor hydrogen bonds (H-bonds) between substrates and amino acid residues of enzymes within ligand binding region over entire 5 ns time scale. Cut off of 3.5 Å was used for monitoring H-bonds and strength of H-bonds is expressed as percentage of its presence over the course of trajectory.

4.4   RESULTS

4.4.1   Analysis of PikC

4.4.1.1 MD Simulations on PikC in Complex With YC-17 and Narbomycin

As detailed earlier in all $PikC_{WT}$ and $PikC_{D50N}$ structures in complex with Narbomycin, C-12 atom of Narmomycin, hydroxylation of which gives primary product is at more distance from oxidizing heme centre than C-14 atom, which gives secondary product. Similarly for structures of PikC with YC-17 ligand is bound in the active site such that in the entire cases C-12 atom of YC-17 which gives one of the primary product is closer to oxidizing heme centre than C-10 which gives another equal yield primary product. To see if protein flexibility and induced-fit effect can account for these discrepancies and to understand the dynamics of PikC-substrate interactions in more detail, MD simulations were carried out for PikC-substrate complexes. **Figure 15** shows RMSD plot showing smooth progress of trajectory along 5 ns time scale. It compares structures obtained over the course of trajectory to the starting structure. RMSD values are obtained through superimposing protein backbones.

Hydrogen bond analysis using Ptraj showed that throughout 5 ns of MD simulation, Narbomycin substrate showed no significant hydrogen bonds with PikC. Only recognizable hydrogen bond was between H-238 of I-Helix and macrolide carbonyl for 63 % of trajectory. Though positive desosamine sugar of Narbomycin remains suspended between different negatively charged amino acid residues i.e. D-50, E-85 and E-94
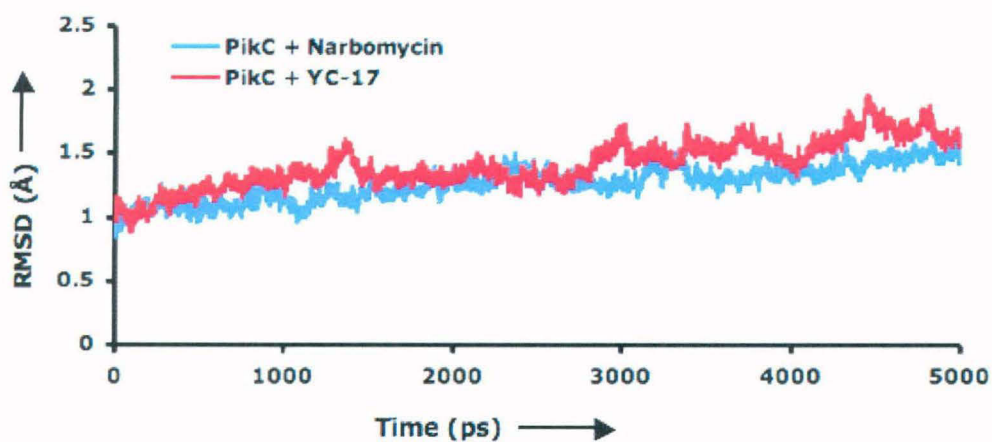
Figure 15: RMS Deviations of PikC in complex with alternate substrates YC-17 and Narbomycin to their initial structures over the course of 5 ns Simulations.
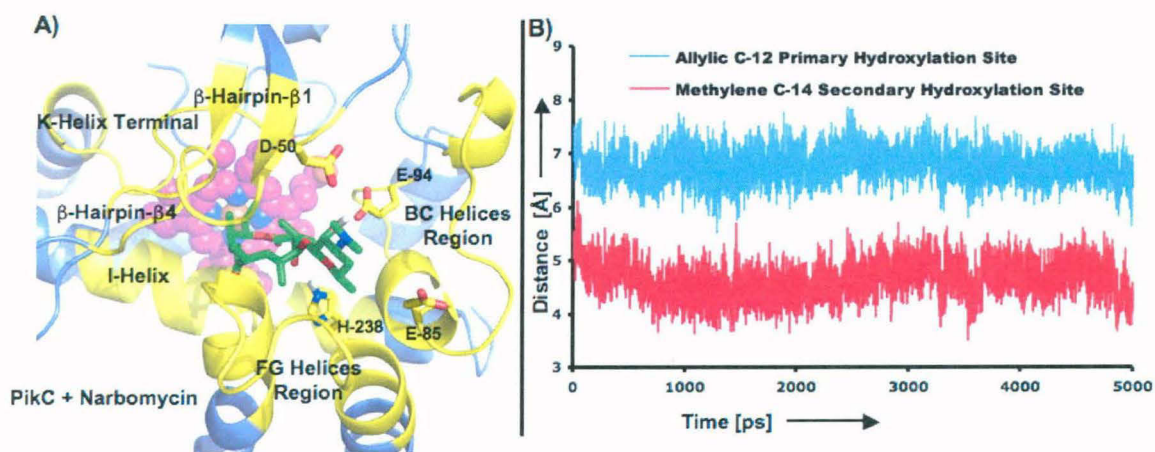

Figure 16: **A)** PikC image in complex with Narbomycin generated from the snapshot taken form final stages of simulation trajectory. **B)** Distances of C-12 and C-14 atoms of Narbomycin from heme centre over the course of 5 ns trajectory.

(**Figure 16 A**). It is also notable that hydroxyl of desosamine capable of acting as a donor and acceptor of H-bond also fails to make any significant hydrogen bonds. Plot in **Figure 16 B** shows distances of 12th and 14th carbon atoms of macrolide Narbomycin from oxidizing heme centre over the course of the trajectory. As can be seen C-12 atom, hydroxylation of which leads to primary product is consistently away from the heme compared to C-14. Hence there is possibility of another binding pose/mode for Narbomycin with C-12 atom near to the heme centre. Current binding pose might be responsible for very small fraction of the product resulting from C-14 hydroxylation and is likely to be more stable than binding pose which gives maximum yield, and hence available for structural studies.

YC-17 remains tightly bound in the buried pocket of PikC for the entire course of the trajectory (**Figure 17 A**). Positively charged dimethylamino group of its desosamine stays well pocketed between negatively charged E-85 and E-94 and is within salt-bridge distance from both. Desosamine part of YC-17 also makes two stable hydrogen bonds, hydroxyl of desosamine with E-94 and ring oxygen of desosamine with H-238. Perhaps because of this strong binding that during the course of dynamics it induces deformation in I-Helix to move even closer to oxidizing heme centre (**Figure 17 B**). This image is obtained by image in **Figure 17 A** to reveal the deformation in I-Helix. Also shown in the image are L-241 and H-245. Newly formed hydrogen bond between L-241 backbone carbonyl and H-245 side chain stabilizes this deformation. This observation highlights the inbuilt flexibility in I-Helix. This I-Helix region is known to be flexible in P450s and therefore may play role in substrate selection through induce-fit changes. **Figure 18** shows the plot of distances of 10th and 12th carbon atoms of macrolide ring of YC-17 from reactive heme centre of PikC. PikC performs hydroxylation at both positions in 1:1 ratio to give two primary products. But once again plot shows that only one of the two carbons consistently stays at a favorable distance form heme compared to another. Logical interpretation of this data means either this is a binding site for hydroxylation to occur at C-12 and C-10 hydroxylation occurs in another binding pose or this binding pose yields very less product but is found in crystal structure because of its stability over binding pose which yields actual products. Deformation of I-Helix which allows substrate to move closer to heme centre can also be noted with respect to C-10, C-12 carbon atoms of substrate in this plot between 900 to 1000 ps.
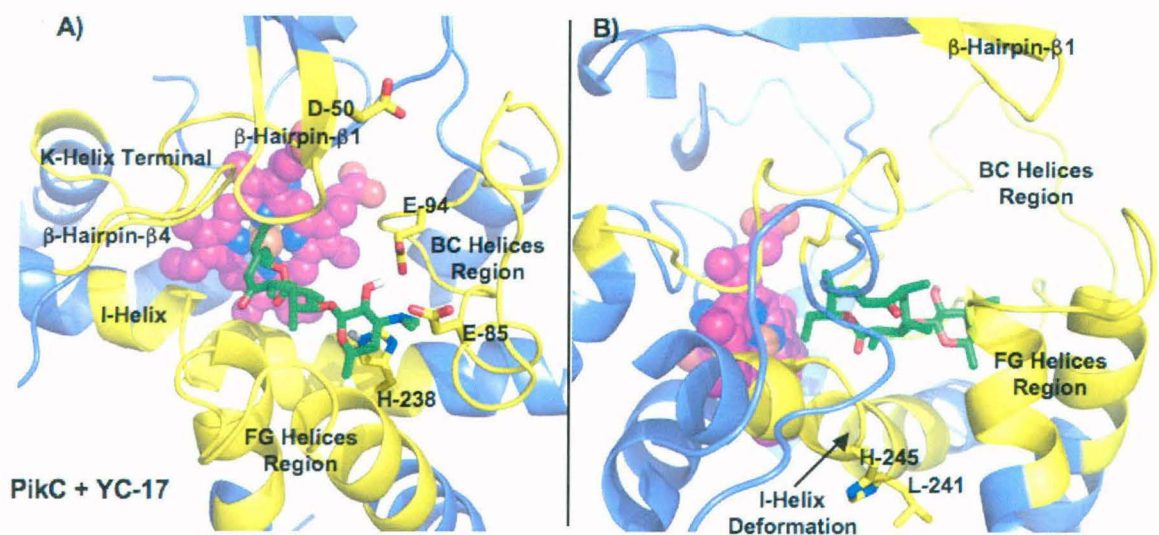
Figure 17: PikC in complex with YC-17. Images are generated from the snapshot taken from final stages of simulation trajectory. **A)** YC-17 binding in active site of PikC **B)** Side view of earlier image showing deformed I-Helix.
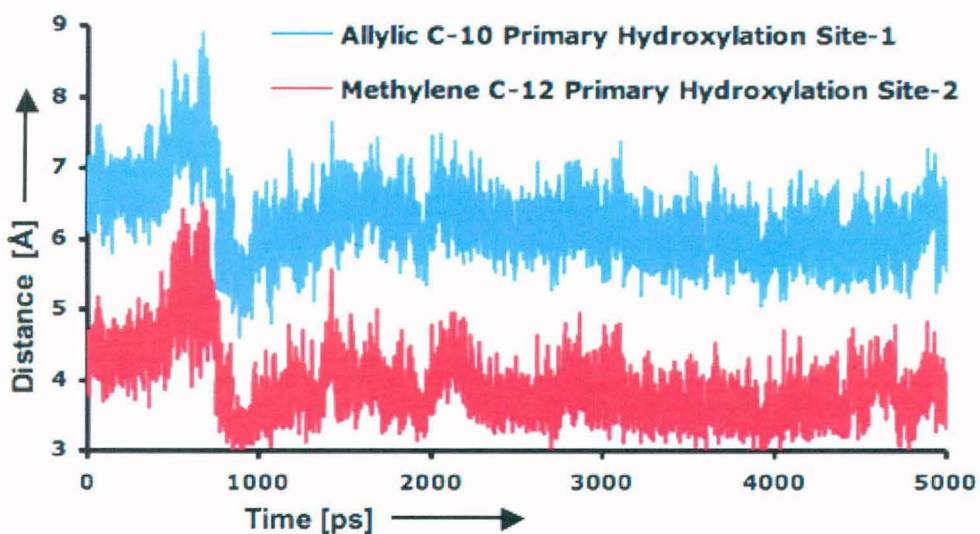

Figure 18: Graph showing distances of 10th and 12th carbon atoms of YC-17 from reactive heme centre of PikC over the course of 5 ns trajectory.

4.4.1.2 Detailed Analysis of Ligand Free Open and Closed Structures of PikC

Figure 19 A is generated form PikC open model built form PikC open crystal structure. Some part of flexible BC-Helices Region and side chains of some amino acids in ligand-binding site were missing in original PDB. Hence to get a complete idea of ligand binding site in open conformation modeler is used to create complete structural model of PikC form crystal structure. This model gives interesting insights in to mechanism of substrate identification by PikC. As can be seen in **Figure 19 A**, substrate binding regions of PikC are littered with negatively charged (acidic) residues i.e. aspartic and glutamic acids. These negatively charged residues are closely flanked by residues with partial or complete positive charge (Weak or Strongly Basic) like arginines, lysines, glutamines and asparagines. This architecture of active site completely complements its substrates, Narbomycin and YC-17 and is possibly used to trap substrates selectively in open confirmation (**Figure 20**).

Thus PikC substrates have multiple recognition points across active site of PikC. Positively charged dimethylamino group of desosamine of YC-17 and Narbomycin is free interact with multiple negatively charged aspartic and glutamic acid residues while weakly acidic ester group capable of acting as a hydrogen bond acceptor but not donor can be complemented by multiple basic/weakly basic residues capable of donating multiple hydrogen bonds. Hence both desosamine and macrolide parts of PikC substrates may have definite role to play in process of identification by PikC. This explains why carbolides (Cyclic hydrocarbons) with attached desosamine have much reduced affinity for PikC compared to its natural substrates.

Site directed mutational studies have proved residues E-85 and E-94 to be essential for PikC activity. Close analysis of PikC Closed structure reveals crucial role of these residues in determining integrity of BC-Helices Region and in bringing about open to close transition (**Figure 19 B**). E-94 and W-74 interaction determines shape and position of entire BC-Helices Regions while E-85 seems to have an essential role to play in open to close transition. Interaction of E-85 with Q-188 of G-Helix is likely to be essential for stabilizing the close structure. This interaction seals two movable regions i.e. BC-Helices Region and FG-Helices Region in closed conformation. This explains why E85Q and E94Q mutants have highly reduced activities. These mutations deform PikC structure and inhibit open to close transitions.

This analysis also gives definite clues about mechanism of substrate anchoring by PikC. It is likely that all the charged residues in the active site take random part in
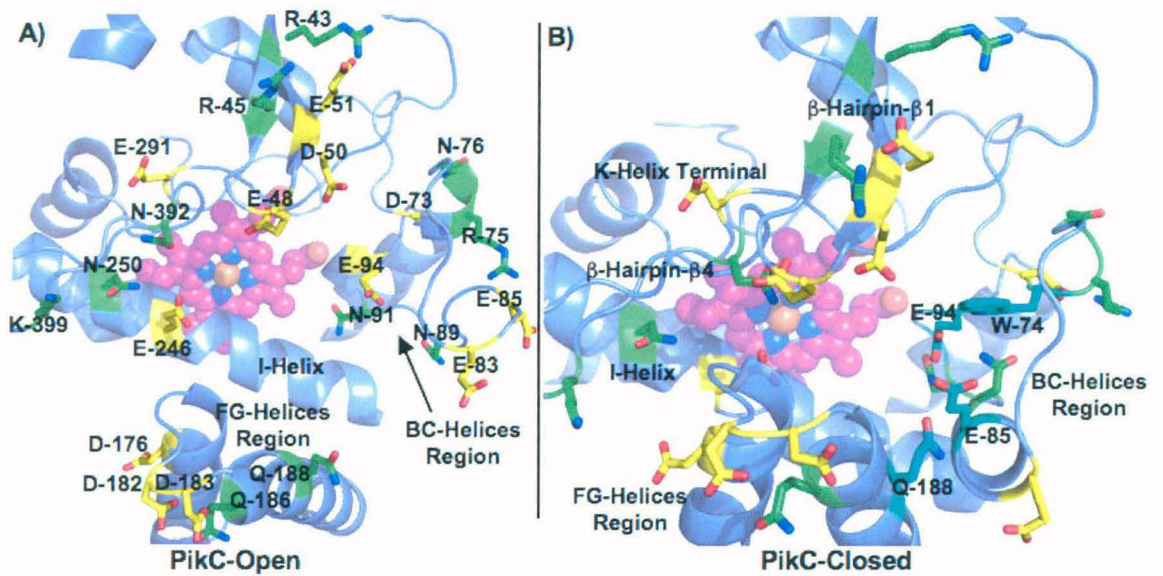
Figure 19: PikC open and closed conformation. **A)** PikC Open conformation showing ligand-binding site. Critical residues are showed in stick representation. Acidic residues are colored yellow and basic green. **B)** PikC Closed conformation highlighting residues critical for integrity of BC-Helices Region and for its movement during open to close transition. These residues are shown in stick representation and colored cyan (W-74, E-85, E-94, Q-188).
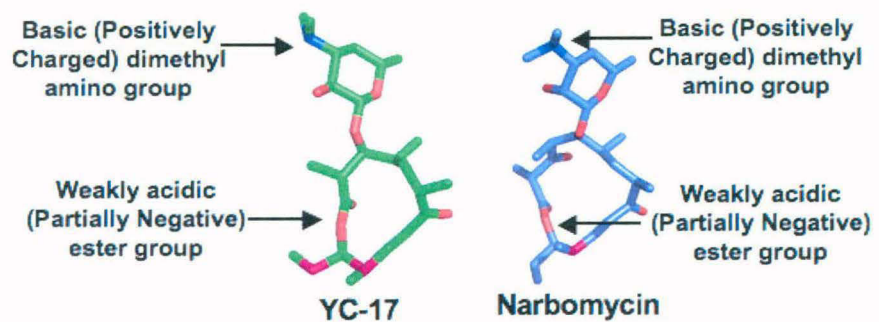

Figure 20: Polar architecture of PikC Substrates YC-17 and Narbomycin

influencing substrate binding and orientation. Most of the negatively charged (Asp/Glu) residues are crowded in top region of PikC mouth while positive (Asn/Gln/Arg/Lys) residues are either outside or inside the substrate binding cavity. This arrangement likely decides substrates orientation in active site, i.e. desosamine always at the top near the mouth of ligand binding site and macrolide below facing reactive heme centre. Though crystal structures have proved that PikC can exist in open and closed confirmation in absence of its ligands, it is also likely that substrate complementation of polar and charged binding site affects rate of conversion and stability of these open and closed stages. It is also a possibility that D50N mutation which increases overall catalytic activity of PikC, does so by affecting charge balance and hence opening and closing of PikC as it has no obvious structural role to play. But why D50N mutation has more effect on YC-17 binding and conversion compared to Narbomycin still remains a question.

4.4.1.3 Docking of YC-17 and Narbomycin in to PikC Active Site

Question still needs answering is, what are the possible binding poses of YC-17 in active site of PikC which allows C-10 and C-12 hydroxylation of macrolide in equal ratio, while prohibiting hydroxylations at other sites? Another question on the same line is, what is the binding pose for Narbomycin, which is responsible for its being hydroxylated primarily at a single position (C-12)? To get answers to these questions, YC-17 and Narbomycin are docked into PikC active site. Docking studies reveal a new binding pose of YC-17 in PikC active site (**Figure 21 A**). This pose corresponds to lowest energy and good frequency cluster from docking studies, thus proving its validity (**Figure 21 B**). As oppose to pose from crystal structure in which desosamine is anchored by Glu-95 and Glu-85, docking pose is stabilized by contact of positive dimethylamino of YC-17 with PikC Glu-246 very close to heme centre. In this new binding pose, desosamine moiety of YC-17 is stacked on to its macrolide moiety to form a U shape. This U shape allows YC-17 to get accommodated in this small binding pocket. This YC-17 shape is stabilized by intra-molecular hydrogen bond between desosamine hydroxyl and C-7 macrolide carbonyl (**Figure 22**). In this new pose no part of YC-17 is buried deeply in active site as oppose to the pose from crystal structure, this probably allows freedom for movement and hence hydroxylations at two different sites in this new found pose. Desosamine-macrolide interaction also restricts conformational flexibility of YC-17 as it prohibits rotation around oxygen-bridge through which they are attached. This rotational freedom was considered to be the primary factor for PikC being relatively non-selective towards hydrocarbon derivative of YC-17, Carbolide-12 (Gives 7 different products as oppose to
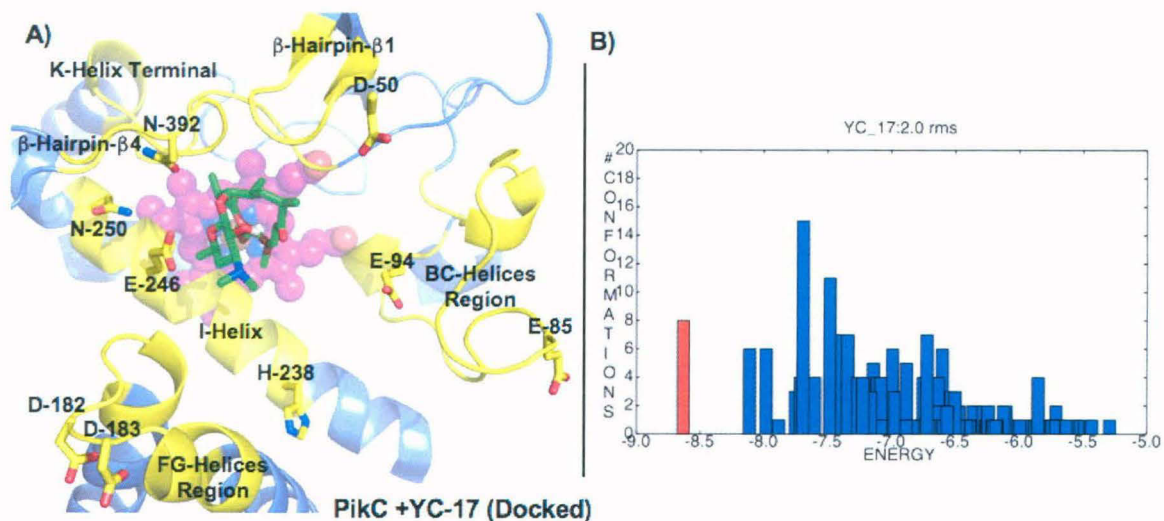
Figure 21: Docking of YC-17 in to PikC Open structure. **A)** Lowest energy and good frequency docking pose of YC-17 in to PikC active site. Color Code, Protein-Purple, Ligand-binding Regions-Yellow, Heme-Magenta, YC-17-Green. **B)** Docking result. Each binding pose is clustered according to its frequency and binding energy. One considered for further analysis is colored red. It's the lowest energy pose.
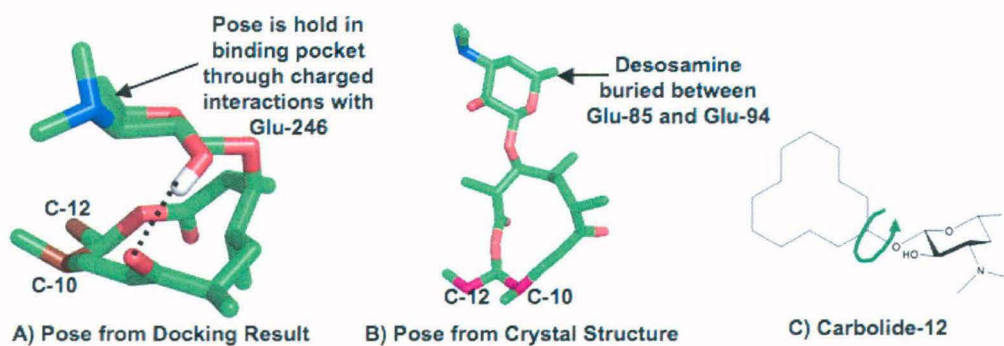

Figure 22: A) Binding pose of YC-17 obtained through docking B) Binding Pose of YC-17 found in crystal structure C) Rotational freedom of synthetic PikC substrate Carbolide-12

YC-17, which gives 2) as it permits flipping of macrolide with respect to desosamine in active site.

Figure 23 shows the binding pose obtained after docking of Narbomycin in to PikC Open active site. It's obtained as one of the lowest energy binding-pose during docking studies and chosen because in this pose Narbomycin is closest to Heme centre. Importantly in this pose C-12 of macrolide, which gives primary product Pikromycin is closer to Heme iron than C-14 giving secondary product. In this binding pose positively charged desosamine of Narbomycin is not interacting with any particular Asp/Glu residue but is suspended between many such residues. As oppose to YC-17, Narbomycin has a carbonyl group separated by just one carbon from the position of attachment of desosamine to macrolide. Hence desosamine hydroxyl can make H-bond with macrolide without bending of desosamine on macrolide, in erect orientation. This pose of Narbomycin was found in docking studies as shown in Figure 24. In this pose C-12 of macrolide lies most distant from the position of desosamine attachment while C-14 is pulled up towards desosamine by hydrogen bond between desosamine hydroxyl and macrolide C-3 carbonyl. This is likely to makes C-12 most vulnerable for oxidation by reactive Heme centre, by placing it most adjacent to it. Also, desosamine-macrolide intra-molecular hydrogen bond gives Narbomycin a rigid character which possibly prevents its flipping in active site thus making hydroxylation much more selective as it yields only one primary product as oppose to hydrocarbon counterpart Carbolide-14 which gives multiple products.

Thus, these docking studies have effectively answered questions regarding mechanisms by which PikC achieves its regio and stereo selectivity for its substrates YC-17 and Narbomycin.

4.4.1.4 MD Simulations on ligand free PikC and PikCD50N

Why PikCD50N mutant is more active than $PikC_{WT}$? Why $PikC_{D50N}$ mutation has more positive effect on YC-17 binding and resultant product yield compared to Narbomycin? These are the questions still unanswered. Thus to understand the effect of D50N mutation on PikC structure, MD simulations were performed on ligand free $PikC_{WT}$ and $PikC_{D50N}$ structures in closed conformations. To get the structure of $PikC_{D50N}$ for simulation studies D50N mutation is introduced on $PikC_{WT}$ Closed structure using Pymol Mutagenesis Wizard and thus modified structure is subjected to MD Simulation. Figure 25 shows RSM Deviations to stating structure over the 5 ns course of simulation
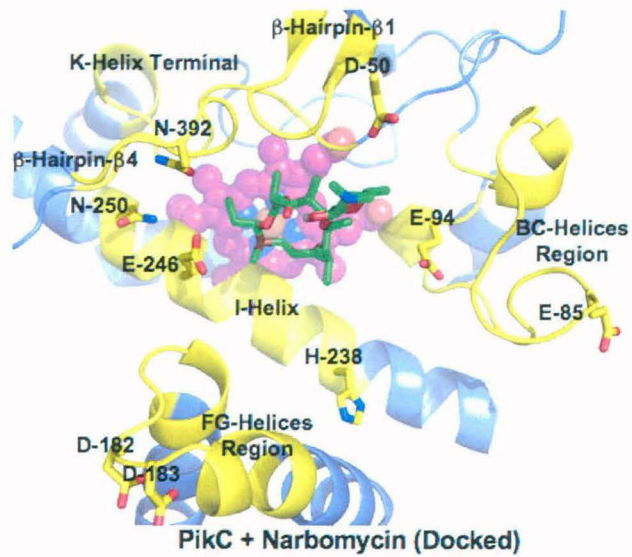
Figure 23: Binding Pose of Narbomycin in PikC active site, obtained through docking studies.

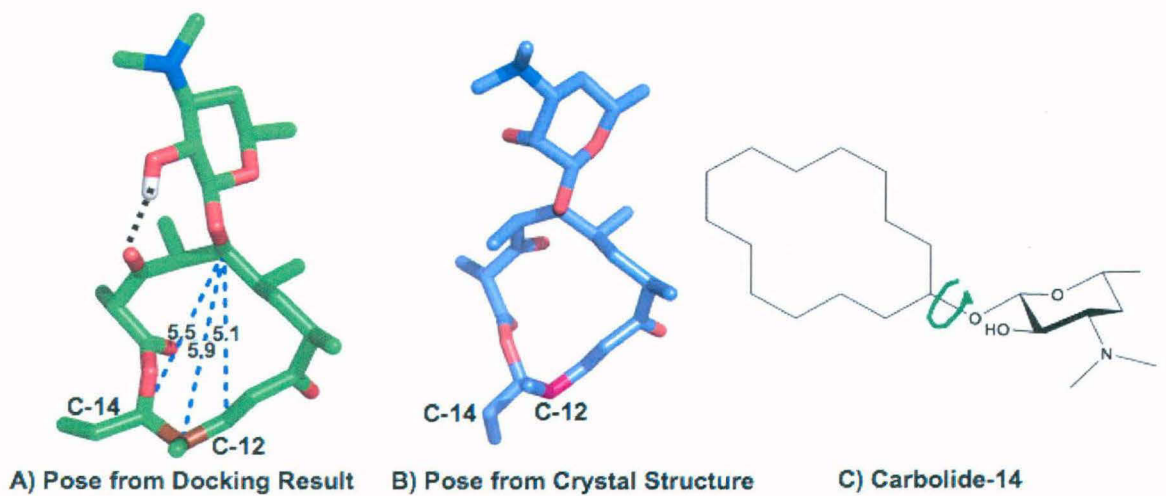**A) Pose from Docking Result** **B) Pose from Crystal Structure** **C) Carbolide-14**

Figure 24: A) Binding pose of Narbomycin obtained through docking B) Binding Pose of Narbomycin found in crystal structure C) Rotational freedom of synthetic PikC substrate Carbolide-14.
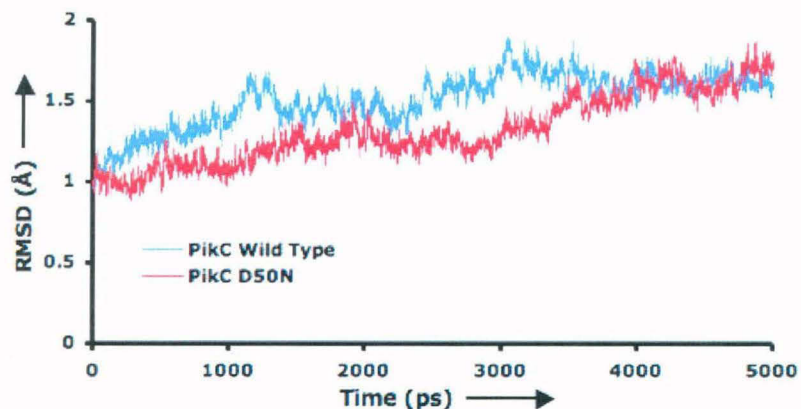
Figure 25: RMS Deviations of PikC$_{WT}$ and PikC$_{D50N}$ in Closed conformations to their initial structures over the course of 5 ns Simulations.
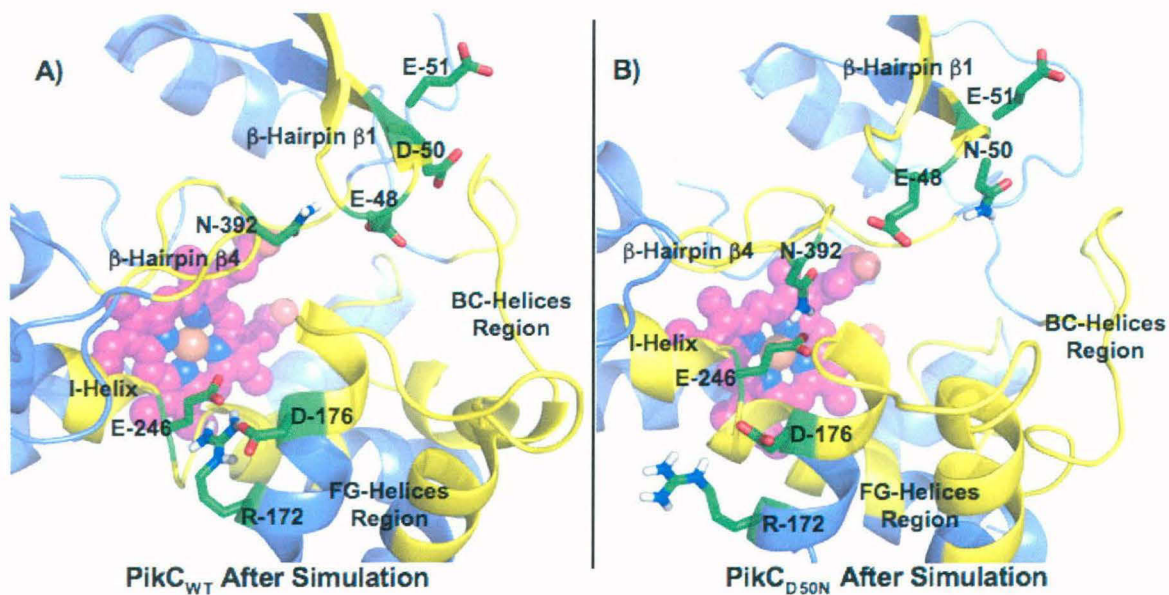

Figure 26: Images of PikC$_{WT}$ and PikC$_{D50N}$ taken after 5 ns of MD Simulations. **A)** PikC$_{WT}$ ligand-binding site showing sequestering of E-246 away from active site by R-172 and D-176. **B)** PikC$_{D50N}$ Ligand-binding site showing reorientation of E-246 towards active site centre due to E-246 and N-392 interactions.

trajectories of PikC$_{WT}$ and PikC$_{D50N}$. End structures and starting structures do not deviate much in terms of RMSDs, indicating smooth running of trajectories.

As clearly seen from PikC$_{WT}$ image (**Figure 26 A**), Glu-246, residue which was found important for YC-17 binding in docking studies, is sequestered by R-172 away from Heme and active site centre. This is because E-246, D-176 and R-172 are involved in salt-bridge type of interactions. It is highly probable that this interaction lessens availability of E-246 for interaction with desosamine of YC-17. On the opposite note in case of PikC$_{D50N}$ this E-246 is perfectly positioned for interaction with YC-17 (**Figure 26 B**). D50N mutation in PikC$_{D50N}$ makes N-392 of β-Hairpin β4 Region free of charged interactions with E-48, N-50, E-51 of β-Hairpin β4 Region. In PikC$_{WT}$ E-48, D-50, E-51 forms strong negative cluster, which keeps partially positive N-392 engaged in charged interactions. D50N mutation partially neutralizes and dispersed the negative charge. Thus freed N-392 bonds with E-246 pulling it up towards ligand binding region and increasing its availability for substrate interactions hence increasing catalysis. PikC$_{WT}$ and PikC$_{D50N}$ structures obtained after simulation are superimposed to highlight the difference in orientation of E-246 (**Figure 27**). This observation once again highlights I-Helix flexibility around Glu-246. This MD Simulation study also explains why D50N mutation has more effect on YC-17 catalysis than on Narbomycin catalysis. Glu-246, as proved by docking studies, is primarily involved in YC-17 binding.

4.4.1.5 Analysis of hydroxylation pattern of PikC substrates YC-17 and Narbomycin

Important fact to consider about PikC substrates is that Narbomycin is hydroxylated at single position (**Figure 28 A**) that is on the carbon farthest from point of desosamine sugar attachment supporting docking result in which sugar and macrolide part are forming a upright straight line, macrolide portion lying below and facing heme while desosamine lying on top and acting as a anchor. On the other hand YC-17 is hydroxylated at two positions (**Figure 28 B**), which are not farthest from point of sugar attachment, thus again supporting the docking result in which macrolide and desosamine parts are stacked horizontal to each other forming a U shape. Thus this observation of hydroxylation pattern supports the theory of different binding modes and sites for two PikC substrates and hence fortifies the work presented here.
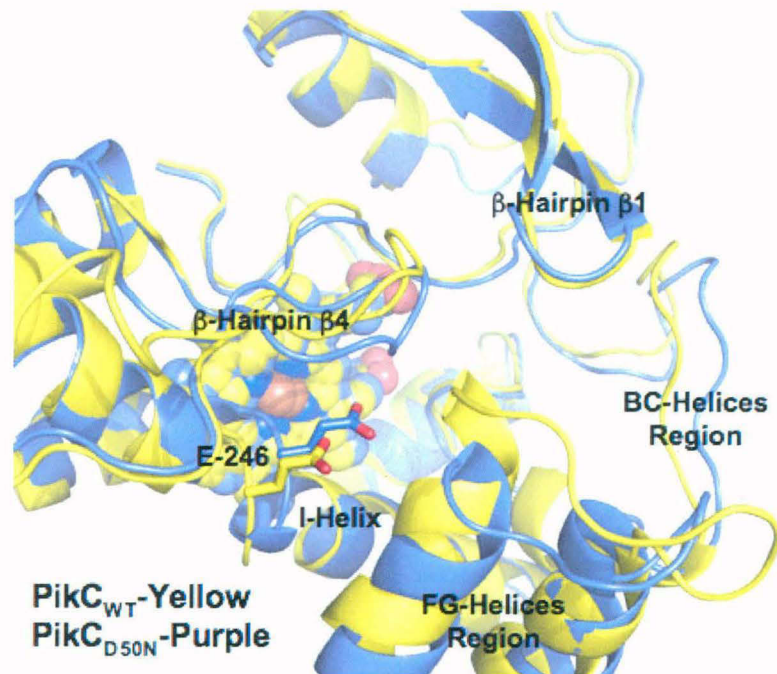
Figure 27: Superimposition of PikCWT and PikCD50N structures after 5 ns simulation highlighting difference between E-246 orientations.
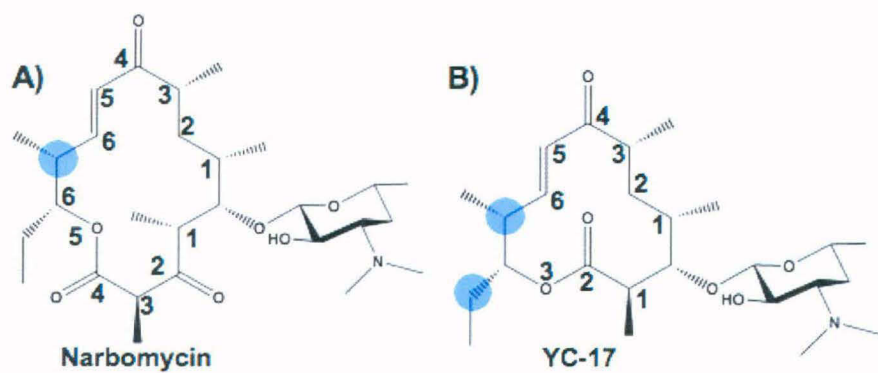

Figure 28: Schematic diagrams of A) Narbomycin and B) YC-17. Positions of hydroxylations are highlighted in blue. Numbering indicates distance of macrolide ring atoms from attached desosamine.

4.4.2 Analysis of EryK

4.4.2.1 Analysis of Desosamine Binding Site of EryK

EryK substrate ErD consists of 14-membered macrolide ring with two attached sugars namely desosamine and mycarose. Sugar desosamine is also present on PikC substrates Narbomycin and YC-17 and plays crucial role in substrate selection through slat-bridge type of interactions with negatively charged protein residues with its positive dimethylamino group. To see if desosamine of ErD plays such a role in substrate identification by EryK, residues surrounding desosamine binding-site are examined in structure of EryK-ErD complex (**Figure 29**).

This diagram clearly shows that desosamine binding-region is loaded with negatively and positively charged residues. This arrangement is similar to that of PikC and complements substrates polarity (dimethylamino of desosamine-positive and ester of macrolide-negative). Significantly charged residues are clustered mainly around desosamine binding region as compared to mycarose binding region. Inference that can be taken from this arrangement is, EryK identifies its substrates through desosamine moiety and the ester group of macrolide, very similar to PikC and mycarose sugar on ErD has little part to play in substrate identification process by EryK. It is also logical that EryK should identify its substrate preferably using desosamine as an anchor and not mycarose, because mycarose is the first sugar transferred during erythromycin biosynthesis, so anchoring through mycarose sugar will possibly lead to cross reactivity with mono-glycosylated substrates.

4.4.2.2 Analysis of open/closed form conformational transitions

According to present theory, open and closed transitions are driven/stabilized by I-Helix deformations and HSWP network. I-Helix deformations are caused by presence of ligand in active site. But this theory does not explain how these transitions occur in absence of ligand and how salt concentration can affect balance between open and close conformations. In PikC, these conformations are brought about and stabilized by charged nature of PikC ligand-binding site. As analysis of desosamine binding region exposed charged nature of EryK ligand-binding region, EryK structure is investigated further to see if charged interactions have any role to play in EryK opening and closing transitions. Study showed, as indicated in **Figure 30**, that open to close conformational change is driven by charged nature of movable FG Helix region. Its negative charge is complimented by positively charged residues present on β-Hairpin-β1 and K-Helix

Binding Site of Desosamine of ErD in EryK
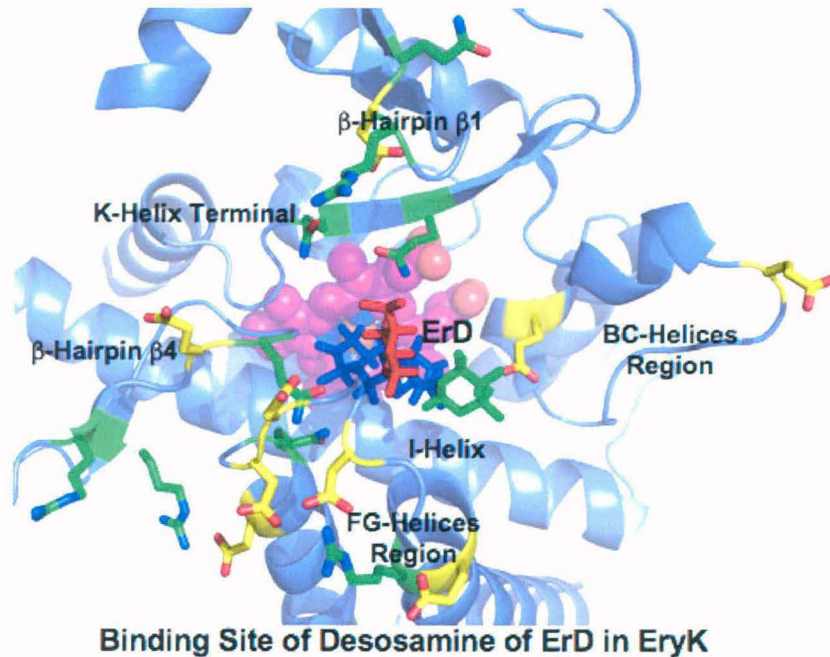
Figure 29: EryK-ErD Complex. Critical residues in region surrounding desosamine of substrate ErD are shown in stick representation. Acidic residues are colored yellow and basic green. ErD is shown in stick representation and its three parts are colored differently, Macrolide-Blue, Desosamine-Red, Mycarose-Green.

Figure 30: EryK Open Structure showing architecture of EryK mouth Region. Negatively charged Asp/Glu residues are colored yellow. Positively charged Arg/Asn/Gln residues are colored green.

Terminal Regions on opposite end of ligand binding region. It is very plausible that presence of ions disturbs local charged interactions of open conformation, thus powering it to attain a clo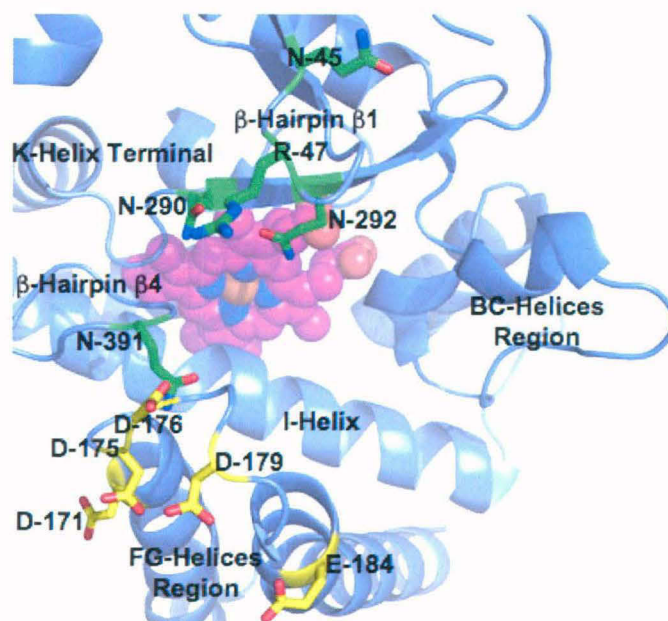sed conformation. It is also conceivable that charged ions mediate open to close conformational change by stabilizing intermediate conformations. This reasoning gives enough proof to assert that positively charged substrate ErD plays the same role that is played by ions in the active site of EryK. It destabilizes open conformation and stabilizes and mediates open to close transformation by bringing negative FG-Helices region and positive β-Hairpin-β1 and K-Helix Terminal Regions closer.

4.4.2.3 Analysis of EryK structures with inhibitors CLT and KC

CLT traps EryK in open conformation and KC in closed conformation as demonstrated by respective structures. This gives unique opportunity to analyze reason for different response of EryK for two inhibitors and to fortify the analysis presented above claiming charged nature of EryK as primary factor in open and closed conformational transitions. As shown in **Figure 31 A**, CLT is a stout hydrophobic molecule incapable of charged interactions. It binds to EryK in a way that no part of it is in a vicinity of charged residues forming top of ligand binding cavity (**Figure 31 B**). Form this analysis it can be concluded that CLT traps EryK in open conformation due to its inability to make charged salt-bridge type interactions with the charged residues that line top of EryK ligand-binding site, and bring about open to close transitions.

Compared to CLT, KC is a polar and much longer molecule (**Figure 32 A**). KC also has positively charged piperazine ring at one of its terminal. Analysis of EryK-KC structure shows this positive piperazine ring to be occupying similar position to that of positive desosamine of its natural substrate ErD (**Figure 32 B**). Thus as oppose to CLT, KC makes charged interactions with polar mouth region of EryK and this, very likely, brings about open to close transition. Based on this analysis and understanding, accepted reasoning that it's interaction between ligand and I-Helix and its after impact on HSWP network primarily leads to open and closed transitions can be effectively disproved. Instead it can be concluded that its charged nature of EryK ligand-binding region and ability of charged substrates or salt to affect this charged interaction network of EryK that is primarily responsible for open and closed transitions of EryK.

4.4.2.4 Analysis of Mycarose Binding Site of EryK

As described earlier EryK substrate ErD has two sugars, desosamine and mycarose, attached to macrolide backbone. Hence it is obvious that EryK has two

Figure 31: **A)** Schematic diagram of CLT, **B)** EryK-CLT complex. Ligand-binding regions of EryK are labeled. Positively charged residues are colored yellow and negatively charged are colored green.

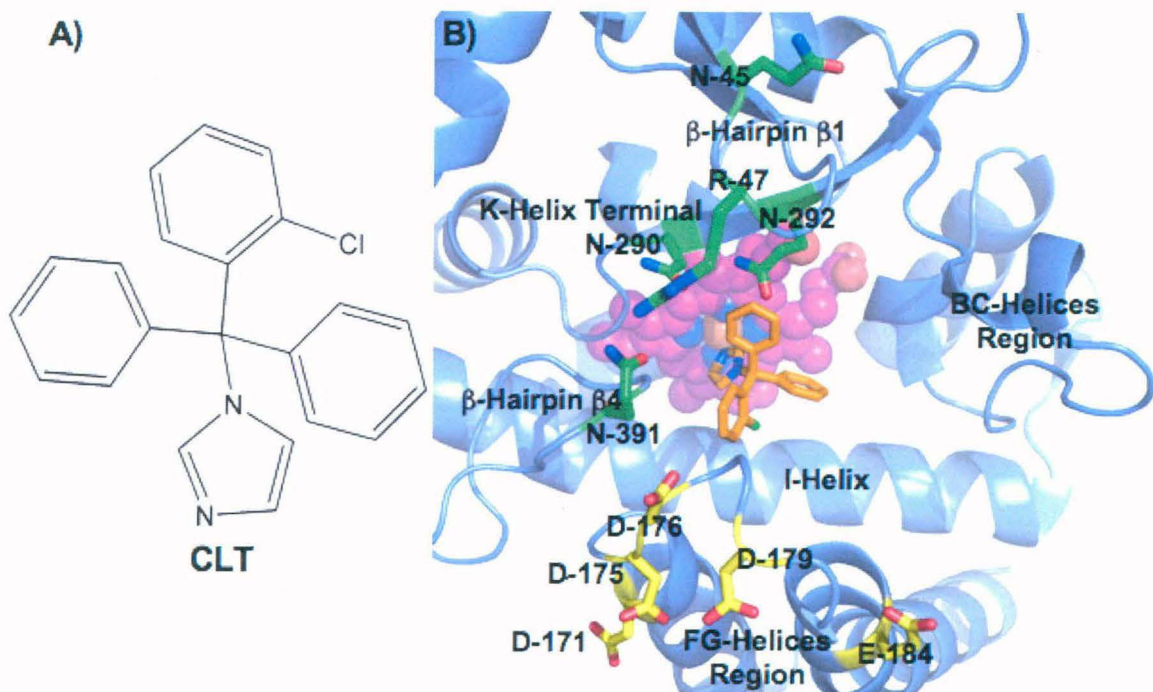**Figure 32**: **A)** Schematic diagram of KC, **B)** EryK-KC complex. Ligand-binding regions of EryK are labeled. Positively charged residues are colored yellow and negatively charged are colored green.
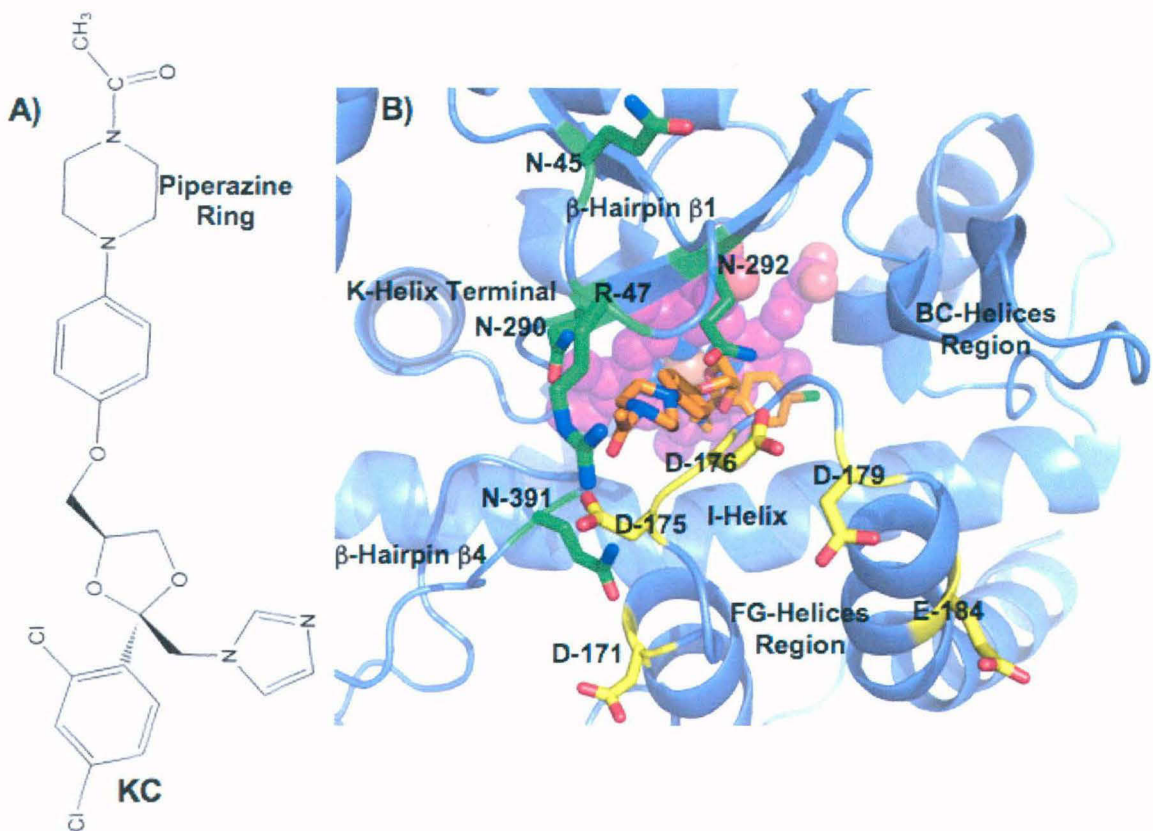
different sugar binding and recognizing regions for binding to its substrate ErD. Desosamine sugar is bulky and charged while mycarose is capable of making multiple H-bond interactions, acting as both donor and acceptor due to presence of two-hydroxyl groups easily accessible for protein. Analysis presented earlier proved that desosamine plays essential role in ErD recognition by EryK and in bringing about open to close transition, through its interactions with charged desosamine binding region. **Figure 33** shows EryK-ErD complex with emphasis on protein residues interacting with desosamine and mycarose. As can be seen from the diagram though desosamine directs ErD in to EryK active site, in actual binding site it is bound weakly to N-290 and N-292 of K-Helix Terminal Region through Hydrogen bonding. Desosamine has a positive charge and N-290 and N-292 residues have partial positive charge due to presence of amide nitrogen. Hence these interactions can be expected to be much weaker. On the other hand Mycarose binds strongly at the base of BC Helices region by multiple H-bonds with H-88 and E-89 through its two-hydroxyl groups.

Hence it can be surmised that ErD binding to EryK involves two distinct steps controlled by two different sugars on its substrate ErD. Sugar desosamine guides ErD to EryK active site and sugar mycarose holds ErD in active site in proper orientation. This arrangement is necessary to avoid flipping of a substrate, which will result in desosamine binding in mycarose binding region and vice versa, in which case regio selectivity during hydroxylation would have suffered. Thus to ensure the binding of the correct substrate ErD, EryK uses both sugars on its substrate to optimum, i.e. desosamine is essential for ErD recognition and closure of active site up on binding while mycarose sugar is essential for tight binding so that substrate holds its proper position and orientation during closing process so as to aid hydroxylation at desired position. This probably means that EryK cannot catalyze ErB hydroxylation, which has extra methyl on mycarose hydroxyl, compared to ErD because it interprets mycarose as desosamine due to reduced ability to H-bond and added bulkiness. This means ErD can hold its position in EryK active site while ErB will not during/ after closing of EryK because of the forces involved in closing of EryK. To verify this hypothesis MD simulations are carried out for EryK in complex with ErD and ErB to understand effect of extra methyl on ErB binding to EryK taking ErD as a standard.

4.4.2.5 MD Simulations and MM-PB/SA on EryK-ErD and EryK-ErB Complexes

To obtain the structure of EryK-ErB, methyl group was added on mycarose of ErD in already available EryK-ErD structure. Thus only difference between both
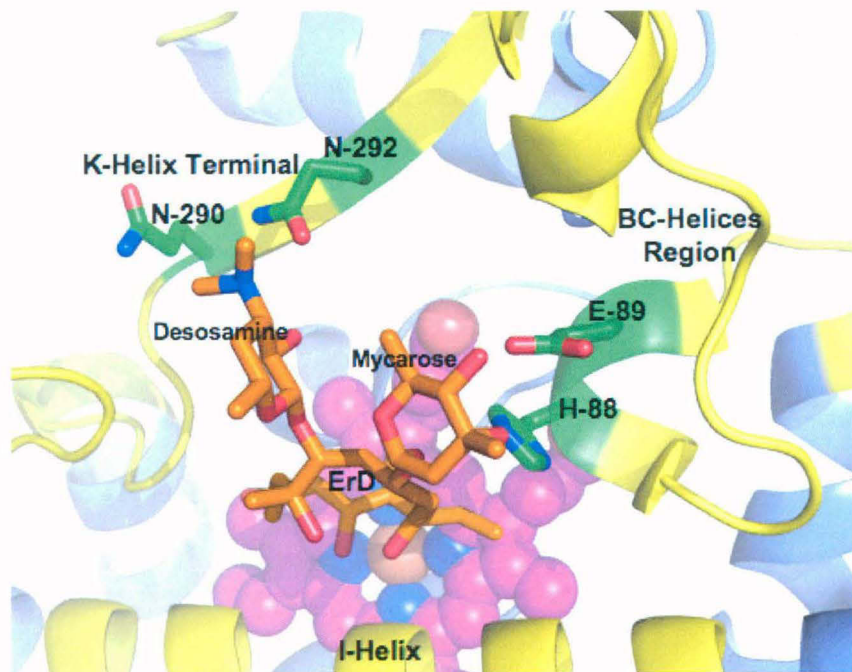
Figure 33: EryK-ErD Complex showing residues of EryK interacting with sugars desosamine and mycarose of ErD.

structures to begin with was the single added methyl group. Simulations are performed on both structures under explicit solvent conditions for duration of 5 ns. **Figure 34** shows backbone RMSD-plot comparing structures obtain during 5 ns simulation to starting structures. This plot indicates smooth running of simulation over the entire time course without major change in enzyme structures.

This analysis gives total binding free energy between enzyme substrate complex. Program also gives composition of total binding free energy, i.e. contribution of each component (Electrostatic, Van der waals, slovation) towards binding free energy. Analysis is performed on snapshots taken from last 1 ns trajectories of EryK + ErD and EryK + ErB systems. **Figure 35** shows result of this analysis. This analysis clearly shows that ErD binding has a very strong electrostatic component, even at the expense of higher, opposing solvation energy. On the other hand ErB has much lower electrostatic component, but is better solvated. This analysis possibly indicates that ErD binds tightly to protein through electrostatic interactions while ErB has a higher degrees of freedom, and its binding is influenced by all energy components. Under the light of this data, and data obtained through hydrogen bond monitoring during simulations, structures of EryK substrate complexes taken form final stages of simulation are examined.

Structure of EryK-ErD complex once again highlighted essential role of mycarose in ErD binding **(Figure 36 A)**. ErD was bound to EryK tightly primarily through mycarose. Mycarose made multiple hydrogen bonds through its hydroxyls with residues E-89 and H-88 of BC-Helices region **(Figure 36 B)**. E-89 came out as a single most important residue of EryK in ErD binding, making stable hydrogen bonds with both hydroxyls of mycarose that essentially ensured tight packing of ErD in EryK active site. Thus this structure and hydrogen bond analysis data supports strong electrostatic component observed in EryK-ErD binding free energy analysis. Due to these strong interactions of mycarose near the base of BC Helices region, ErD packs tightly against I-Helix.

After simulation structure of EryK-ErB showed reduced influence of methylated mycarose in ErB binding **(Figure 37 A)**. Due to methylation one of the hydroxyls of mycarose was no longer available for interaction with E-89. This led to reorientation of ErB in active site, so that other ligand-binding regions played more role in ErB binding. Thus desosamine was moved outward away from heme and was better placed between Q-290/Q-292 residues of K-Helix Terminal region making more stable hydrogen bond **(Figure 37 B)**. Macrolide ring of ErB was no longer tightly squeezed between heme and
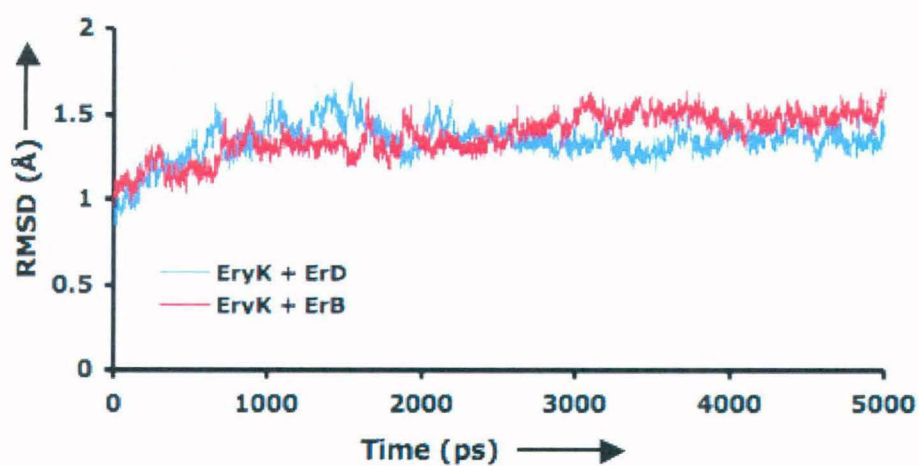
Figure 34: RMS Deviations of EryK + ErD and EryK +ErB complexes to their initial structures over the course of 5 ns MD Simulations.
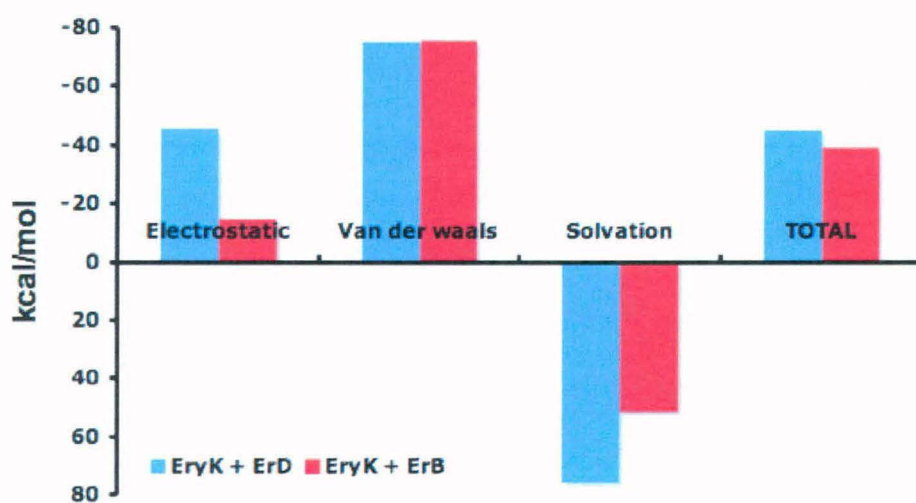
Figure 35: MM_PBSA binding free energy analysis of EryK + ErD and EryK + ErB Complexes

Figure 36: **A)** Structure of EryK-ErD Complex after MD Simulation showing major hydrogen bonds. **B)** Result of Hydrogen bond anlysis using Ptraj showing major hydrogen bonds between EryK and substrate ErD.

Sugar of ErD	EryK Amino Acid Residue	Strength of H-bond (% of Trajectory)
Mycarose	E-89	98 %
Mycarose	E-89	78 %
Mycarose	H-88	93 %
Desosamine	Q-292	59 %

Figure 37: **A)** Structure of EryK-ErB Complex after MD Simulation showing major hydrogen bonds **B)** Result of Hydrogen bond anlysis using Ptraj showing major hydrogen bonds between EryK and substrate ErB

Sugar of ErB	EryK Amino Acid Residue	Strength of H-bond (% of Trajectory)
Methylated-Mycarose	E-89	89 %
Methylated-Mycarose	H-88	98 %
Desosamine	H-88	71 %

I-Helix regions. Thus I-Helix now exhibited greater freedom. Interestingly it developed similar bend near H-243 that was observed with binding of inhibitors CLT/KC.

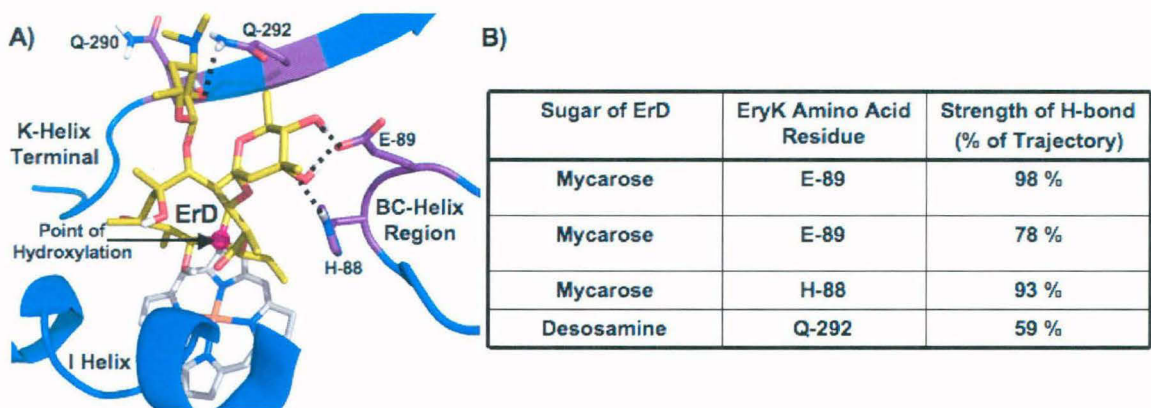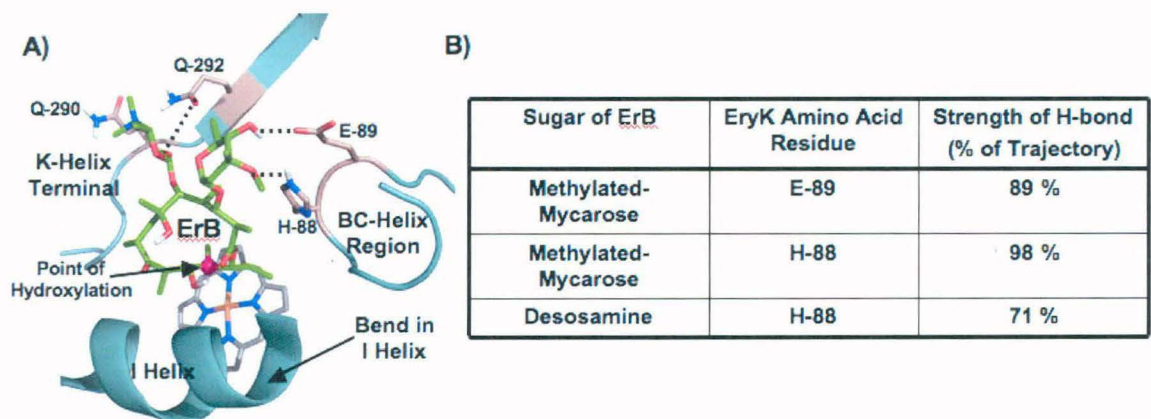After-simulation structures of EryK+ErD and EryK+ErB are superimposed to map the extent of movement and reorientation of ErB compared to ErD (**Figure 38 A**). This simplified diagram shows re-adjustment of ErB in active site of EryK. This happens primarily due to higher degree of freedom available for ErB due to weakened interactions of EryK BC Helices region with methyl-mycarose. Whole ErB molecule was found to be moving away from active site. To demonstrate this point further, distance between C-12 of macrolide (point of hydroxylation) and heme centre are plotted for ErD and ErB over entire 5 ns of trajectory (**Figure 38 B**). As can be seen, natural substrate ErD's reactive C-H clearly maintains lower distance from oxidizing Heme iron over 5 ns trajectory.

EryK analysis thus far reveals distinct binding features EryK has evolved to interact with two different sugars on its substrate ErD. It uses charged nature of desosamine sugar to identify its substrate and then in next stage of binding it uses strong hydrogen bonding capacity if mycarose to hold substrate in to active site in a way to allow for hydroxylation reaction at desired point on its substrate.

4.4.2.6 Effect of Simulations on F-Helix Integrity and HSWP network

EryK exists naturally in different conformations as proved by crystal structures. Its ligand free open form has an intact F-Helix and an HSWP network linking I-Helix to G-Helix through F-Helix. Its ligand free closed form has unfolded F-Helix and broken HSWP network. Its ligand bound (ErD bound) closed structure has intact F-Helix and intact HSWP network. Thus to see the effect of simulation on stability of F-Helix in EryK-ErD and EryK-ErB complexes, end structures after simulation are visualized to check integrity of F-Helix. As can be seen from **Figure 39 A**, during simulation F-Helix maintains its integrity in case of EryK-ErD complex, just as the case was in original crystal structure. On the contrary, F-Helix was found partially unfolded for EryK-ErB Complex (**Figure 39 B**), proving that effect of single methyl substitution on ErD has effect on over all binding of ErB to EryK and even on protein conformation. Based on this observation a new theory can be proposed for EryR–substrate binding. Binding of natural substrate leads to closing of EryK without unfolding of Helix F. Binding of unnatural substrate leads to closing of EryK but with unfolding of Helix F.

As expected both F-Helix and HSWP network were found to be intact in EryK-ErD Complex after simulation. For EryK-ErB Complex, though F-Helix was partially unfolded, HSWP network was still intact (**Figure 40 A**). As can be seen unfolding of F-
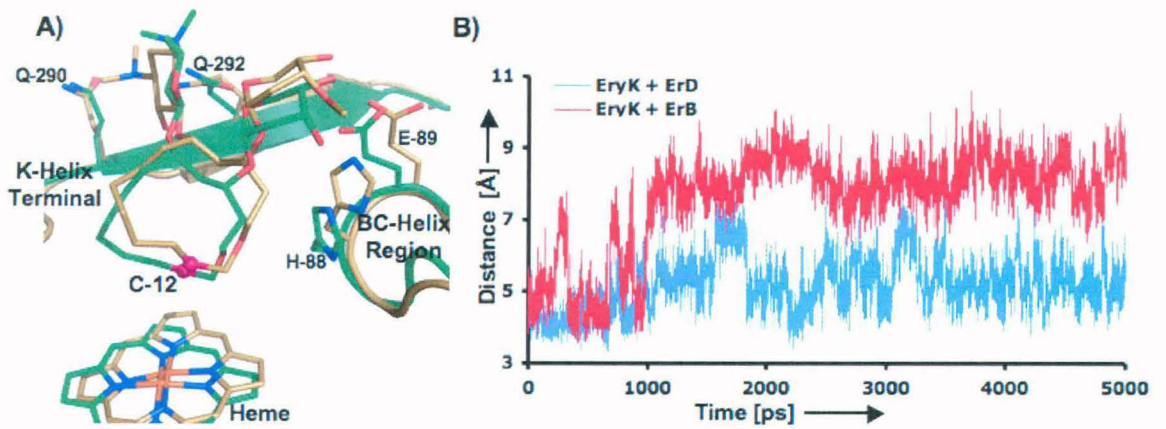
Figure 38: **A)** Simplified diagram made from superimposed structures of EryK-ErD and EryK-ErB showing relative positioning of ErB compared to ErD in EryK active site. Color Code: EryK-ErD Complex-Green Shading, EryK-ErB Complex-Brown Shading. **B)** Plot showing distance of C-12 of ErD and ErB from reactive heme centre over the course of 5 ns trajectory.
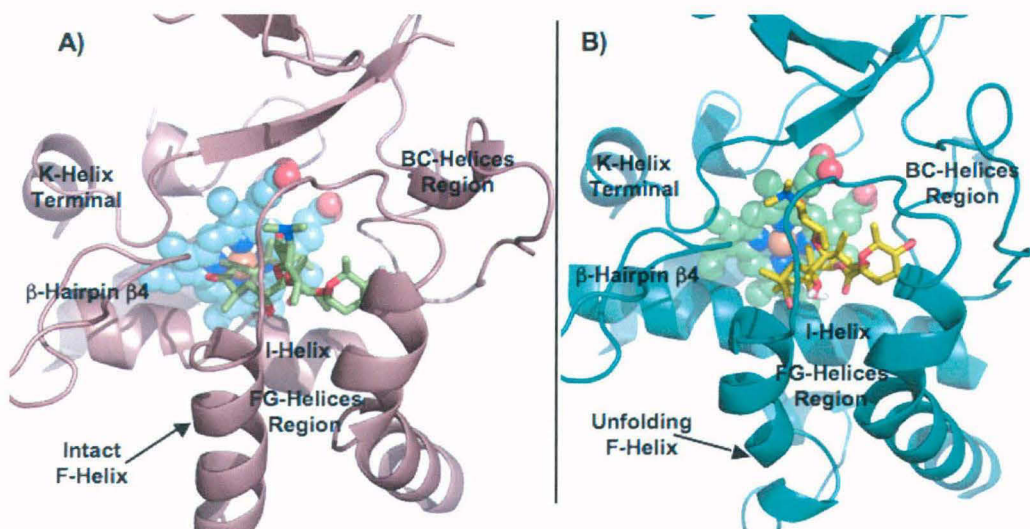
Figure 39: After simulation images of EryK-ErD and EryK-ErB Complexes. **A)** EryK-ErD complex showing intact F-Helix and **B)** EryK-ErB Complex showing unfolding F-Helix.

Figure 40: **A)** HSWP network in EryK+ErB structure after 5 ns MD simulation. **B)** Broken HSWP network in ligand free EryK closed structure. Image generated from original crystal structure.

Figure 41: HSWP network in EryK-KC complex

Helix has progressed till W-165. If F-helix has to unfold further as it happens in non-liganded closed form (**Figure 40 B**), movement of bulky W-165 from facing G-Helix to Facing I-Helix will be hindered by bulky ErB. Thus if F-Helix will unfold further, ErB will be ejected from the active site. Given that P-192 of G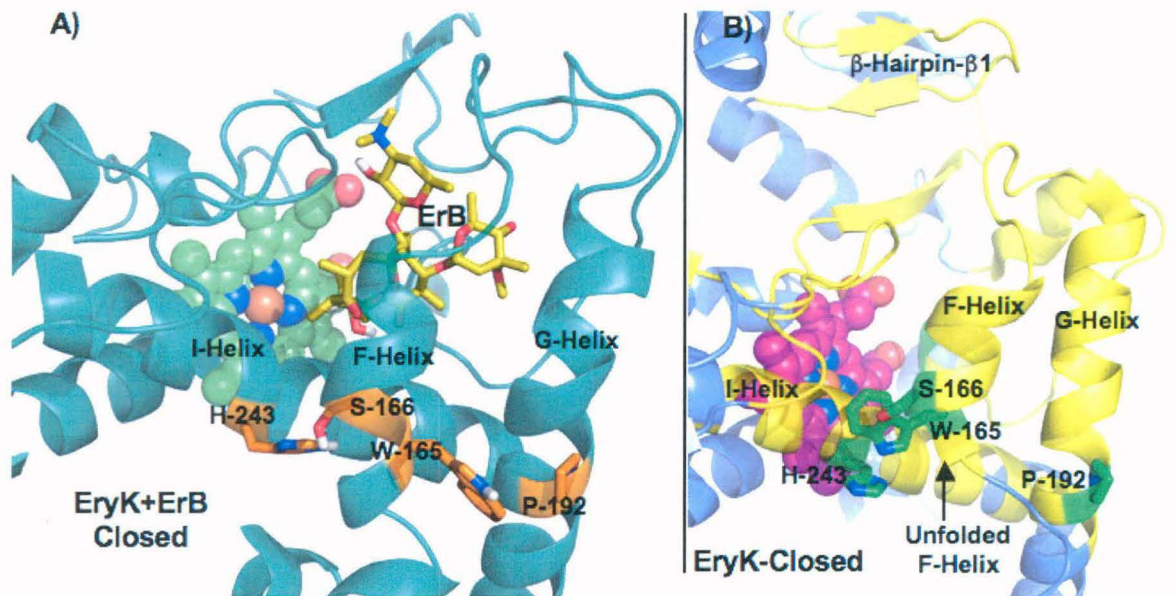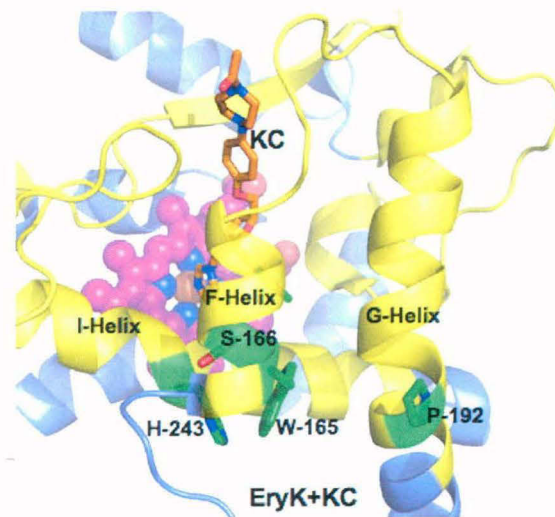-Helix of HSWP network is an unusual amino acid at this position, the usual being glutamic acid which will in that case form strong H-bond with W-165 of F-Helix so as to prevent unfolding and maintain HSWP network, it can be assumed that EryK deliberately uses HSWP mechanism and unfolding of F-Helix to prevent unwanted ligands from getting in to active site. In case where ErD, natural substrate, binds in flipped orientation, it will also be ejected out, as desosamine will not be able to match mycarose in strong H-bonding capacity.

Significantly, above reasoning also explains why EryK binds one of its inhibitor CLT much more strongly than the other inhibitor KC. CLT coordinates EryK in open conformation and makes mainly van der waals interactions. But KC on the other hand due to its charged tail, induces closing of EryK, but fails to make mycarose like interaction with EryK, thus getting ejected resulting in having higher dissociation constant. EryK-KC structure lends proof to this theory (**Figure 41**). Once again F-Helix is in partially unfolded state just as the case was for EryK-ErB complex. Thus far there are four cases presented here. In first, EryK is in closed conformation in absence of ligand (**Figure 40 B**). This closed conformation has maximum unfolding of F-Helix due to lack of hindrance from ligand and results in breakage of HSWP network. In second EryK is in complex with ErB (**Figure 40 A**). Here F-Helix unfolding stops at W-165 due to presence of bulky ligand ErB that does not allow space for further unfolding and results in intact HSWP network after 5 ns simulation. In third, EryK is in complex with KC. Again unfolding stops at W-165. This probably means that if F-Helix unfolds further, ligand would be expelled from active site, but coordination bond of imidazole of KC with Heme is strong enough to hold ligand in active site against forces that determine closed conformation in absence of HSWP network. In fourth case of EryK-ErD complex, natural substrate ErD binds tightly through mycarose at the base of BC-Helices region, pressing bulky ErD macrolide tightly against I-Helix. This completely inhibits F-Helix unfolding by not allowing space for it to occur (**Figure 39 A**).

4.4.3   Analysis of EryF

PikC and EryK both use sugars attached to their substrates to identify and bind their substrates. But then how EryF achieves its substrate specificity given no sugar is

attached to its substrate 6-dEB? Analysis of PikC and EryK presented above gives some clues around which EryF structure can be explored to find an answer to this question.

4.4.3.1 Analysis of binding pose of 6-dEB in EryF active site

EryF performs hydroxylation at 6th carbon of macrolide 6-dEB. As can be seen in **Figure 42**, 6-dEB has hydroxyl on immediate next 5th carbon of its macrolide ring. According to current knowledge this C-5 hydroxyl serves to replace the role of conserved threonine in related P450s as this threonine conserved across P450 families is replaced by alanine in EryF and plays a role of proton donor. This arrangement answers the issue of flip binding or binding in some other unintended orientation as substrate bound in flip or unintended orientation will lack the hydroxyl in vicinity to activate the reaction cascade, thus bringing regio-selectivity.

Figure 42 also reveals the rigid shape of macrolide ring held by C-1 ester carbonyl and C-11 hydroxyl. As can be seen, macrolide ring has three hydroxyl groups, a carbonyl group and a ester group, and thus resulting many other permutation and combination for H-bond network within this small macrolide ring, which will twist and distort its shape so that it will be perceived as completely different molecule by EryF. This explains the lack of activity of enzymes towards substrate with slight modification of 6-dEB. Indeed converting 9-carbonyl to hydroxyl will render much flexibility to 6-dEB, which will then bend on to it self. Also it seems, intently put methyl group at C-2, and C-14 helps existing H-bond between C-1 and C-11 and inhibit other possible H-bonds like that of between C-5 and C-1. This arrangement gives 6-dEB a rigid 3D shape to be recognized by enzyme and probably contributes vastly to its observed selectivity.

4.4.3.2 Substrate selectivity of EryF

Orientation of 6-dEB in EryF active site is such that its ester group faces upwards while 6th carbon at which hydroxylation occurs lies at its opposite end facing heme (**Figure 42**). Thus next question is how EryF selectively chooses 6-dEB from other ligands and manages this proper orientation in active site?

Close observation of substrate binding site of EryF reveals that it employs a sieve like mechanism to filter 6-dEB for other possible non-cognate molecules. It can be seen from **Figure 43 A** that the outer rim of substrate binding site consist of charged residues which form the mouth or entering passage of ligand binding cavity. These polar and charged residues form first filter to extract desired ligand 6-dEB as it has multiple hydroxyl residues, a carbonyl group and acidic charged ester group. These characters of 6-dEB are probably used to isolate it from many other candidates using this outer rim.
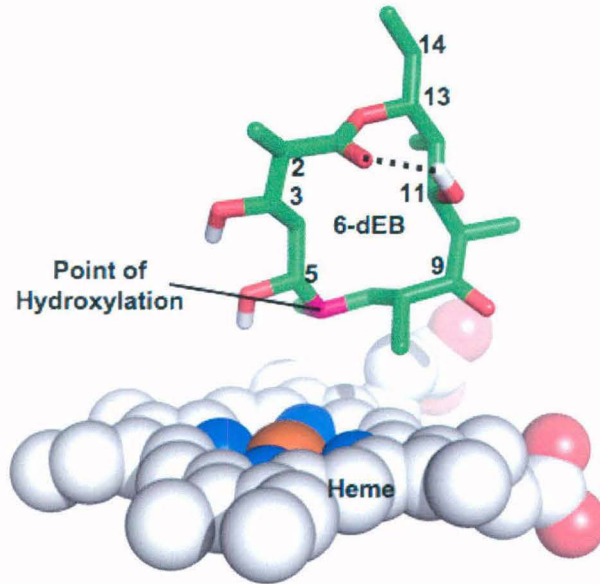
Figure 42: Binding pose of 6-dEB in EryF active site. Image is generated from structure of EryF in complex with6-dEB and shows relative orientation of reactive heme centre and substrate 6-dEB.
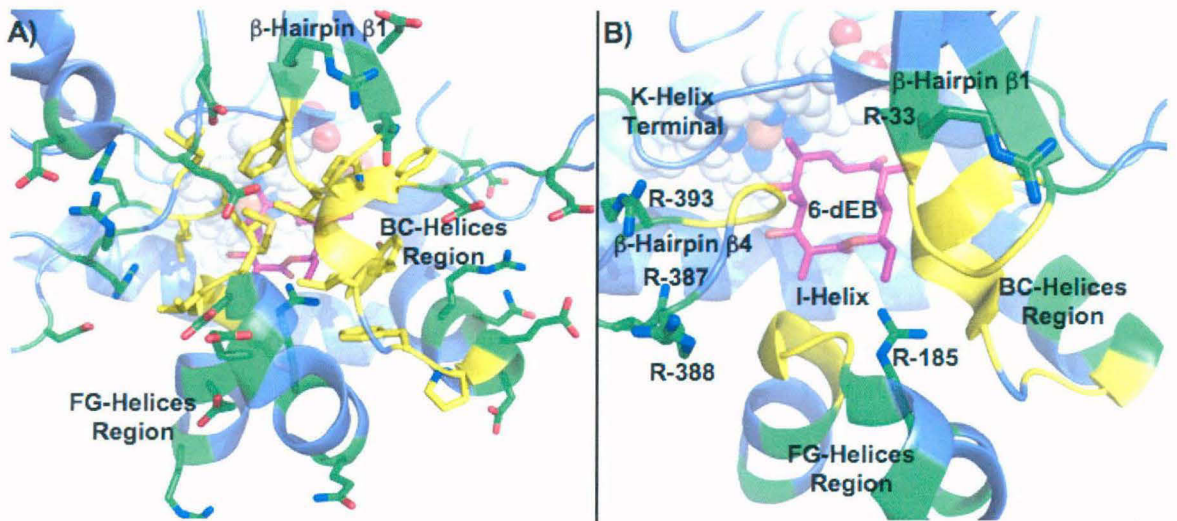
Figure 43: **A)** Substrate binding regions architecture in EryF. Polar residues forming outer rim showed in green, non-polar residues forming inner core shown in yellow, ligand 6-dEB is shown in magenta and heme in light grey. **B)** This image shown concentration of positive arginine residues near top region of EryF including R-185, which lies directly overhead of bound substrate 6-dEB.

The second sieve, actual binding site though is formed by non-polar residues and its size is reduced by much larger BC-Helices region compared to PikC and EryK, so that 6-dEB can pass through outer charged rim in to active site by virtue of non-polar nature of macrolide ring and smaller size, but other mono/di-glycosylated intermediates of ErA biosynthetic pathway will be blocked due to larger size, and higher polarity imparted by sugar group.

Only question still remains is how EryF decides proper 6-dEB orientation in active site? As can be seen in **Figure 43 B**, multiple basic Arg residues border the rim of active site of EryF, which directly compliments the acidic ester group of its substrate. Particularly so is Arg-185 which is absent in related P450s thus proving its importance of having a unique role. Its interaction with ester group decides orientation of 6-dEB in active site. Ester group faces upwards towards R-185, thus reactive 6th carbon faces downwards to Heme iron.

4.5 DISCUSSION

In order to understand the structural basis of substrate specificity of cytochrome P450 monooxygenase family of tailoring enzymes detailed structure based analysis involving docking and molecular dynamics simulations have been performed on EryK, PikC and EryF involved in stereo and regio specific hydroxylation of secondary metabolite natural products. The major conclusions and novel findings from the study are the following.

4.5.1 Substrate specificity of PikC

PikC ligand binding site has multiple negatively and positively charged residues. This complimentary arrangement of residues is likely to help PikC selectively sequester its cognate substrates YC-17 and Narbomycin for catalysis and filter out non-cognate substrate like non-glycosylated derivatives of its substrates 10-Deoxymethynolide and Narbonolide. Negatively charged PikC mouth region (E-48, D-50, E-51, E-83, E-85, D-176, D-182, D-183) facilitates anchoring of positively charged desosamine sugar. Small size of YC-17 macrolide ring allows it to bind PikC in active site pocket not accessible for Narbomycin. In this binding mode Glu-246 provides for electrostatic interaction necessary for substrate stabilization in active site. This observation explains higher catalytic activity of PikC for YC-17.

In both YC-17 and Narbomycin hydroxyl of desosamine makes H-bond with carbonyl of macrolide. This desosamine macrolide interaction is likely to be responsible for stereo and regio selectivity shown by PikC towards these two substrates as it gives

them rigid 3D shape and avoid their flipping in active site due to rotation around oxygen linkage between desosamine and macrolide.

YC-17 binds to PikC in U shape such that desosamine part is stacked on to macrolide part of YC-17. Resulting lack of grip on to YC-17 by PikC in this binding mode (As U shape makes inner part of desosamine and macrolide un-available for PikC to interact) probably results in C-10/C-12 hydroxylations in equal amount and hence two primary products. On the other hand for Narbomycin, macrolide and desosamine are stacked one above other forming a straight line instead of U shape. This in all probability makes macrolide carbon most distant from the position attachment of desosamine to macrolide, nearest to heme centre and hence its hydroxylation of this C-12 carbon that gives single primary product.

4.5.2 Substrate specificity of EryK

Open to closed conformational change in EryK is driven by charged interactions similar to that of PikC. EryK has an unusual P-192 residue compared to its related P450s. This residue is a part of HSWP network and plays a major role in regio selective hydroxylation of the natural substrate ErD.

Very strong binding of mycarose of ErD at the base of EryK allows EryK to close but without unfolding of Helix-F. ErD being large molecule, engages otherwise flexible I-Helix in to multiple van der waals interactions, thus forming a non-flexible wall against which F-Helix cannot unfold during closing of EryK. But small charged ligands like KC, can induce closing of EryK due to their charged nature, but fail to stop unfolding due to space available for F-Helix during closing of EryK, which results in ejection of ligand from active site of EryK in most case including the case of ErB, which has a large size but due to added methyl on its mycarose, it fails to hold its position during closing of EryK against unfolding forces of moving FG-Region

All in all, EryK uses characteristics of both sugars on its natural substrate to optimum. It uses desosamine on ErD to direct it to the active site, and induce closing of the active site region, and it uses sugar mycarose to hold it in to the active site during open to closed form conformational transition.

4.5.3 Substrate specificity of EryF

Active site of EryF consists of outer rim of charged residues and inner small cavity of non-polar residues. This arrangement utilizes both polar and non-polar aspects of macrolide ring to selectively choose it from other non-cognate substrates. Arg-185 interaction with acidic ester of macrolide determines proper orientation of macrolide ring

in active site. Regio-selectivity is further ensured by replacement of conserved Thr in most P450s with Ala-245. This change makes it essential for the carbon undergoing reaction to have properly oriented hydroxyl on a carbon before it to replace missing hydroxyl of threonine.

In summary, this analysis shows how substrate binding sites of the tailoring enzymes PikC, EryK and EryF have evolved to selectively use singular features of their respective ligands to recognize their cognate substrates and then orient them properly in an active site to achieve regio and stereo selectivity. These studies once again highlight the utility of comparative structural analysis of related enzymes to get valuable information regarding their catalytic mechanisms. Our computational studies demonstrate that such structure-based analysis of other cytochrome P450 monooxygenase family of tailoring enzymes will enable finer understanding of how this diverse class of enzymes in spite of having common structural fold interact with very wide range of substrates to generate metabolic diversity.

SUMMARY & CONCLUSIONS

Natural products have been prolific source of new drug candidates. Major advances in biosynthetic engineering during the last decade have demonstrated the feasibility of obtaining novel engineered natural products. Availability of complete genome sequences of increasing number of organisms has also opened up the possibility of discovering novel natural products by genome mining. In this thesis, we have attempted to develop and standardize powerful computational methods which can help in discovery of new NPs by genome mining and reprogramming of known biosynthetic pathways for producing novel NPs by rational manipulation of key tailoring enzymes.

Several pharmaceutically important secondary metabolite natural products biosynthesized by nonribosomal peptide synthetases (NRPS) and polyketide synthases (PKS) are often extensively modified by specialized tailoring enzymes for imparting additional structural diversity and biological activity through introduction of key functional groups. Glycosyltransferases and oxidoreductases are two important families of tailoring enzymes involved in altering structural and functional diversity of PKS/NRPS family of natural products. In this work, we have carried out genome mining studies for identification of novel glycosyltransferases and oxidoreductases (cytochrome P450 monooxigenase) involved in tailoring of secondary metabolite NPs. Detailed structural analysis involving docking and molecular dynamics (MD) simulations have been carried out for representative glycosyltransferases and cytochrome P450 monooxigenase family of oxidoreductases for understanding structural basis of their substrate preference and identifying crucial specificity determining residues.

For identification of newly characterized GTrs, database GTrDB which contains list of 102 GTrs of known specificity was used as a starting point. Searches have been performed with sequence information of these GTrs using protein BLAST against NCBI non-redundant protein database to identify new GTrs which have been experimentally characterized. Based on literature information, donor and acceptor substrate specificity data has been cataloged for these new GTrs. In this way 56 newly characterized GTrs have been identified belonging to 37 different NP biosynthetic gene clusters. Protein BLAST has been used to identify closest PDB homologue for each of these new GTrs for sequence structure comparison. The total of

158 secondary metabolite glycosyltransferases have been clustered into different groups as per their donor and acceptor specificities. HMM profiles have been built for each of these clusters. Searches on various genomes have been carried out using these HMM profiles of experimentally characterized GTrs to group experimentally uncharacterized GTrs into different specificity groups. Similar analysis has also been carried out for identifying different cytochrome P450 monooxigenase family of oxidoreductases present in secondary metabolite biosynthetic clusters of various genomes and grouping them into different functional subfamilies based on comparison with cytochrome P450 monooxigenases of known specificity.

Even though profile based analysis can group various glycosyltransfereases into broad functional families, it was necessary to carry out detailed structural analysis for understanding the structural basis of their donor and acceptor specificity and using this knowledge in design of altered secondary metabolites. Glycosyltransferases GtfA, GtfC and GtfD are involved in biosynthesis of vancomycin class of antibiotics were selected for this purpose because of the availability of crystal structures and specificity data. GtfD is involved in the final stage of vancomycin biosynthesis as it transfers sugar vancosamine from sugar donor TDP-vancosamine on to acceptor desvancosaminyl vancomycin (DVV) at second hydroxyl of glucose to complete vancomycin biosynthesis. GtfA and GtfC are involved in second last and last stage of chloroeremomycin biosynthesis respectively. Note that vancomycin and chloroeremomycin both contain glucosylated heptapeptide backbone DVV but differ with respect to other sugars and number and position of attachment of these sugars. GtfA transfers the sugar epi-vancosamine from TDP-epi-vancosamine on to β-hydroxyl of the sixth amino acid of heptapeptide DVV to form an intermediate chloroorientacin B (COB). GtfC then transfer another epi-vancosamine from TDP-epi-vancosamine to second hydroxyl of glucose of COB to complete chloroeremomycin biosynthesis. These enzymes have been biochemically characterized and the specificity data thus obtained reveals that GtfD is a versatile enzyme and can transfer both sugars vancosamine and epi-vancosamine on to its acceptor DVV. On the other hand, GtfA has been found to be totally selective for its natural substrate epi-vancosamine. GtfC transfers vancosamine at a much lower catalytic rate compared to its cognate substrate epi-vancosamine. In this work we have carried out docking and molecular dynamics simulations

to understand why GtfD is versatile, while GtfC and GtfA are stringent in selection of substrates. Crystal structures were available for both GtfA and GtfD in complex with TDP and DVV, only the sugar groups were modeled by docking simulations. Since no structures were available for GtfC, the structure of this enzyme was modeled based on crystal structure of GtfD with which it has a sequence similarity of 81%. Each of these enzymes was solvated in a large enough water box and MD simulations were carried out for 3ns using Amber 9 package. Analysis of the trajectories indicated that, throughout the 3ns simulations all the enzymes showed a RMSD of only around 2.5 Å from their starting structure and the substrates remained bound throughout the simulation. Binding free energy was calculated using GB/SA approach for structures extracted from this trajectory. Molecular dynamics and MM/GBSA energy calculations have provided novel insight to the structural basis of the specificity of donor substrate recognition by these enzymes. Interestingly free energy differences between the cognate and non-cognate substrates were in qualitative agreement with experimental studies. Based on the analysis of the contribution of various binding pocket residues to this free energy difference and visual analysis of different snapshots from MD trajectory, specificity-determining residues for vancosamine and *epi*-vancosamine were identified. Most notable of these residues is ASP 333 of GtfD, which is found to make salt bridge with amino group of sugar. Interestingly corresponding ASP 317 residue in GtfA is moved away from active site due to an insertion in the primary sequence. Instead a mutation of GLY to ASP at position 294 in GtfA is probably playing a role of ASP 333 in GtfD. L-vancosamine and L-*epi*-vancosamine differ only in orientation of single hydroxyl group. This hydroxyl group of TDP-L-*epi*-vancosamine is found to make crucial interactions with acceptor molecule and active site ASP 13 and SER 10, which may explain reason for specificity of GtfA for L-*epi*-vancosamine. Together these docking, molecular modeling and MD simulation studies provide novel insights in to the mechanism by which these GTrs achieve their unique specificities and demonstrate applicability of these computational tools in rational design of novel secondary metabolites.

Oxidoreductases (Cytochrome P450) are enzymes, which introduce polar oxygen containing groups in to its substrates and thus convert hydrophobic substance to hydrophilic. These enzymes use heme as a prosthetic group. Importance of oxidoreductases comes from their ability to catalyze oxidation of non-reactive C-H bonds. Hence these enzymes can play

a crucial role in generating diverse analogues of NPs. The tailoring enzymes EryK and PikC, which are involved in Erythromycin and Picromycin biosynthesis respectively, were selected for detailed analysis. Initial sequence based analysis of these enzymes revealed tendency of EryK homologues to bind larger substrates and that of PikC homologues to bind smaller substrates. Crystal structure was available for EryK in complex with its substrate Erythromycin D (ErD) and for PikC in complex with its substrates Narbomycin (intermediate in Pikromycin biosynthesis) and YC-17 (Intermediate in Methymycin biosynthesis). EryK has a stringent substrate preference and performs hydroxylation at 12th carbon of macrolide ErD. On the other hand, PikC is versatile and can attach hydroxyl groups at C12/C14 positions on Narbomycin and C10/C12 position on YC-17. In order to understand the structural basis of stringent specificity of EryK vis-à-vis relaxed specificity of PikC, molecular dynamics simulations have been performed on EryK and PikC in complex with their respective substrates. Forcefield parameters for ErD, Narbomycin and YC-17 were assigned using Antechamber module of amber and those for heme group were taken from published data. Simulations were carried out for 5 ns each in explicit solvent conditions. Calculation of binding energy by MM/GBSA analysis indicated that, a major component of the interactions of EryK with its cognate substrates was of electrostatic origin, while binding of PikC to Narbomycin and YC-17 was primarily mediated by van der Waals forces. Analysis of hydrogen bonds between these enzymes and their substrates also revealed presence of long lasting hydrogen bonds between EryK and mycarose sugar moiety, while the other sugar desosamine interacted mainly through van der Waals contacts. On the other hand in case of Narbomycin no significant hydrogen bonds were observed, while YC-17 was found to make hydrogen bonds with PikC through desosamine. In contrast to ErD which completely fills the active site of EryK and binds in a fixed orientation, PikC substrates Narbomycin and YC-17 are relatively smaller compared to the available volume in the active site and hence can bind in different orientations. In fact PikC can hydroxylate different substrates at different sites because of its ability to bind substrates in different orientations. Thus the results of our simulations are in agreement with the experimental observation about relaxed specificity of PikC *vs.* stringent specificity of EryK. Our computational studies demonstrate that such structure-based analysis of other cytochrome P450 monooxygenase family of tailoring enzymes will enable finer understanding of how this diverse class of

enzymes in spite of having common structural fold interact with very wide range of substrates to generate metabolic diversity.

In summary, in this work we have developed powerful profile based methods which enable rapid identification and broad categorization of key natural product biosynthetic enzymes by genome mining. This gives valuable clues about novel natural products encoded by secondary metabolite gene clusters in newly sequenced genomes. We have also demonstrated that sequence analysis, docking and molecular dynamic simulations on key tailoring enzymes can help in identifying crucial specificity determining residues which can be altered by genetic engineering techniques to generate altered natural products with desired chemical structures for obtaining diverse pool of natural products. Taken together, the computational methods developed and standardized in this thesis can help in drug design based on natural products.

Bibliography

Ahmed, A., Peters, N. R., Fitzgerald, M. K., Watson, J. A., Jr., Hoffmann, F. M., *et al.* (2006). Colchicine glycorandomization influences cytotoxicity and mechanism of action. *J Am Chem Soc, 128*(44), 14224-14225.

Andersen, J. F., & Hutchinson, C. R. (1992). Characterization of Saccharopolyspora erythraea cytochrome P-450 genes and enzymes, including 6-deoxyerythronolide B hydroxylase. *J Bacteriol, 174*(3), 725-735.

Andreeva, A., Howorth, D., Chandonia, J. M., Brenner, S. E., Hubbard, T. J., *et al.* (2007). Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*

Ansari, M. Z., Yadav, G., Gokhale, R. S., & Mohanty, D. (2004). NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. *Nucleic Acids Res, 32*(Web Server issue), W405-413.

Balamurugan, R., Dekker, F. J., & Waldmann, H. (2005). Design of compound libraries based on natural product scaffolds and protein structure similarity clustering (PSSC). *Mol Biosyst, 1*(1), 36-45.

Baltz, R. H. (2002). Combinatorial glycosylation of glycopeptide antibiotics. *Chem Biol, 9*(12), 1268-1270.

Baltz, R. H. (2006). Molecular engineering approaches to peptide, polyketide and other antibiotics. *Nat Biotechnol, 24*(12), 1533-1540.

Case, D. A., Darden, T.A., Cheatham, T.E., Simmerling, C.L., Wang, J., Duke R.E., Luo, R., Merz, K.M., Pearlman, D.A. (2006). AMBER9, University of California, San Francisco.

Cheatham, T. E., 3rd, Srinivasan, J., Case, D. A., & Kollman, P. A. (1998). Molecular dynamics and continuum solvent studies of the stability of polyG-polyC and polyA-polyT DNA duplexes in solution. *J Biomol Struct Dyn, 16*(2), 265-280.

Cupp-Vickery, J. R., & Poulos, T. L. (1995). Structure of cytochrome P450eryF involved in erythromycin biosynthesis. *Nat Struct Biol, 2*(2), 144-153.

Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., *et al.* (2003). A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem, 24*(16), 1999-2012.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics, 14*(9), 755-763.

Fiser, A., & Sali, A. (2003). Modeller: generation and refinement of homology-based protein structure models. *Methods Enzymol, 374*, 461-491.

Funa, N., Funabashi, M., Yoshimura, E., & Horinouchi, S. (2005). A novel quinone-
forming monooxygenase family involved in modification of aromatic polyketides.
J Biol Chem, 280(15), 14514-14523.

Gidrol, X., Fouque, B., Ghenim, L., Haguet, V., Picollet-D'hahan, N., *et al.* (2009). 2D
and 3D cell microarrays in pharmacology. *Curr Opin Pharmacol, 9*(5), 664-668.

Grabowaski, K., & Schneider, G. (2007). Properties and Architecture of Drugs and
Naturalm Products Revisited. *Current Chemical Biology, 1*, 115-127.

Gregory, M. A., Petkovic, H., Lill, R. E., Moss, S. J., Wilkinson, B., *et al.* (2005).
Mutasynthesis of rapamycin analogues through the manipulation of a gene
governing starter unit biosynthesis. *Angew Chem Int Ed Engl, 44*(30), 4757-4760.

Hermann, J. C., Ghanem, E., Li, Y., Raushel, F. M., Irwin, J. J., *et al.* (2006). Predicting
substrates by docking high-energy intermediates to enzyme structures. *J Am Chem
Soc, 128*(49), 15882-15891.

Hermann, J. C., Marti-Arbona, R., Fedorov, A. A., Fedorov, E., Almo, S. C., *et al.*
(2007). Structure-based activity prediction for an enzyme of unknown function.
Nature, 448(7155), 775-779.

Huey, R., Morris, G. M., Olson, A. J., & Goodsell, D. S. (2007). A semiempirical free
energy force field with charge-based desolvation. *J Comput Chem, 28*(6), 1145-
1152.

Jorgensen, W. L. (2004). The many roles of computation in drug discovery. *Science,
303*(5665), 1813-1818.

Kamra, P., Gokhale, R. S., & Mohanty, D. (2005). SEARCHGTr: a program for analysis
of glycosyltransferases involved in glycosylation of secondary metabolites.
Nucleic Acids Res, 33(Web Server issue), W220-225.

Keller, T. H., Pichota, A., & Yin, Z. (2006). A practical view of 'druggability'. *Curr Opin
Chem Biol, 10*(4), 357-361.

Koch, M. A., & Waldmann, H. (2005). Protein structure similarity clustering and natural
product structure as guiding principles in drug discovery. *Drug Discov Today,
10*(7), 471-483.

Koskiniemi, H., Metsa-Ketela, M., Dobritzsch, D., Kallio, P., Korhonen, H., *et al.* (2007).
Crystal structures of two aromatic hydroxylases involved in the early tailoring
steps of angucycline biosynthesis. *J Mol Biol, 372*(3), 633-648.

Lee, S. K., Basnet, D. B., Choi, C. Y., Sohng, J. K., Ahn, J. S., *et al.* (2004). The role of erythromycin C-12 hydroxylase, EryK, as a substitute for PikC hydroxylase in pikromycin biosynthesis. *Bioorg Chem, 32*(6), 549-559.

Li, S., Chaulagain, M. R., Knauff, A. R., Podust, L. M., Montgomery, J., *et al.* (2009a). Selective oxidation of carbolide C-H bonds by an engineered macrolide P450 mono-oxygenase. *Proc Natl Acad Sci U S A, 106*(44), 18463-18468.

Li, S., Ouellet, H., Sherman, D. H., & Podust, L. M. (2009b). Analysis of transient and catalytic desosamine-binding pockets in cytochrome P-450 PikC from Streptomyces venezuelae. *J Biol Chem, 284*(9), 5723-5730.

Losey, H. C., Jiang, J., Biggins, J. B., Oberthur, M., Ye, X. Y., *et al.* (2002). Incorporation of glucose analogs by GtfE and GtfD from the vancomycin biosynthetic pathway to generate variant glycopeptides. *Chem Biol, 9*(12), 1305-1314.

Mason, J. S., Good, A. C., & Martin, E. J. (2001). 3-D pharmacophores in drug discovery. *Curr Pharm Des, 7*(7), 567-597.

McArdle, B. M., & Quinn, R. J. (2007). Identification of protein fold topology shared between different folds inhibited by natural products. *Chembiochem, 8*(7), 788-798.

Montemiglio, L. C., Gianni, S., Vallone, B., & Savino, C. Azole drugs trap cytochrome P450 EryK in alternative conformational states. *Biochemistry, 49*(43), 9199-9206.

Mulichak, A. M., Losey, H. C., Lu, W., Wawrzak, Z., Walsh, C. T., *et al.* (2003). Structure of the TDP-epi-vancosaminyltransferase GtfA from the chloroeremomycin biosynthetic pathway. *Proc Natl Acad Sci U S A, 100*(16), 9238-9243.

Mulichak, A. M., Lu, W., Losey, H. C., Walsh, C. T., & Garavito, R. M. (2004). Crystal structure of vancosaminyltransferase GtfD from the vancomycin biosynthetic pathway: interactions with acceptor and nucleotide ligands. *Biochemistry, 43*(18), 5170-5180.

Newman, D. J., & Cragg, G. M. (2007). Natural products as sources of new drugs over the last 25 years. *J Nat Prod, 70*(3), 461-477.

Oberthur, M., Leimkuhler, C., Kruger, R. G., Lu, W., Walsh, C. T., *et al.* (2005). A systematic investigation of the synthetic utility of glycopeptide glycosyltransferases. *J Am Chem Soc, 127*(30), 10747-10752.

Rix, U., Fischer, C., Remsing, L. L., & Rohr, J. (2002). Modification of post-PKS tailoring steps through combinatorial biosynthesis. *Nat Prod Rep, 19*(5), 542-580.

Savino, C., Montemiglio, L. C., Sciara, G., Miele, A. E., Kendrew, S. G., *et al.* (2009). Investigating the structural plasticity of a cytochrome P450: three-dimensional structures of P450 EryK and binding to its physiological substrate. *J Biol Chem, 284*(42), 29170-29179.

Savino, C., Sciara, G., Miele, A. E., Kendrew, S. G., & Vallone, B. (2008). Cloning, expression, purification, crystallization and preliminary X-ray crystallographic analysis of C-12 hydroxylase EryK from Saccharopolyspora erythraea. *Protein Pept Lett, 15*(10), 1138-1141.

Schlunzen, F., Zarivach, R., Harms, J., Bashan, A., Tocilj, A., *et al.* (2001). Structural basis for the interaction of antibiotics with the peptidyl transferase centre in eubacteria. *Nature, 413*(6858), 814-821.

Sciara, G., Kendrew, S. G., Miele, A. E., Marsh, N. G., Federici, L., *et al.* (2003). The structure of ActVA-Orf6, a novel type of monooxygenase involved in actinorhodin biosynthesis. *Embo J, 22*(2), 205-215.

Shang, S., & Tan, D. S. (2005). Advancing chemistry and biology through diversity-oriented synthesis of natural product-like libraries. *Curr Opin Chem Biol, 9*(3), 248-258.

Sherman, D. H., Li, S., Yermalitskaya, L. V., Kim, Y., Smith, J. A., *et al.* (2006). The structural basis for substrate anchoring, active site selectivity, and product formation by P450 PikC from Streptomyces venezuelae. *J Biol Chem, 281*(36), 26289-26297.

Song, L., Kalyanaraman, C., Fedorov, A. A., Fedorov, E. V., Glasner, M. E., *et al.* (2007). Prediction and assignment of function for a divergent N-succinyl amino acid racemase. *Nat Chem Biol, 3*(8), 486-491.

Sousa, S. F., Fernandes, P. A., & Ramos, M. J. (2006). Protein-ligand docking: current status and future challenges. *Proteins, 65*(1), 15-26.

Stachelhaus, T., Mootz, H. D., & Marahiel, M. A. (1999). The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem Biol, 6*(8), 493-505.

Walsh, C. T., Losey, H. C., & Freel Meyers, C. L. (2003). Antibiotic glycosyltransferases. *Biochem Soc Trans, 31*(Pt 3), 487-492.

Wang, J., Wang, W., Kollman, P. A., & Case, D. A. (2006). Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model, 25*(2), 247-260.

Weymouth-Wilson, A. C. (1997). The role of carbohydrates in biologically active natural products. *Nat Prod Rep, 14*(2), 99-110.

Xue, Y., Wilson, D., Zhao, L., Liu, H., & Sherman, D. H. (1998). Hydroxylation of macrolactones YC-17 and narbomycin is mediated by the pikC-encoded cytochrome P450 in Streptomyces venezuelae. *Chem Biol, 5*(11), 661-667.

Zhang, M. Q., & Wilkinson, B. (2007a). Drug discovery beyond the 'rule-of-five'. *Curr Opin Biotechnol, 18*, 1-11.

Zhang, M. Q., & Wilkinson, B. (2007b). Drug discovery beyond the 'rule-of-five'. *Curr Opin Biotechnol, 18*(6), 478-488.

Appendix

Structural Studies on Human Neurotransmitter Transporters Based on Prokaryotic Template

INTRODUCTION

Neurotransmitter Transporters are proteins that span across plasma membranes and stop transmission of neural signals by pumping neurotransmitters back in pre-synaptic neurons and glial cells. This reuptake is driven by preexisting sodium gradient and neurotransmitter and two sodium ions are transferred in same direction thus earning these protein name Neurotransmitter Sodium Symporters (NSSs). These transporters are divided into different types based on neurotransmitters they transport, for example Dopamine Transporter (DAT), Norepinephrine Transporter (NET) and Serotonin Transporter (SERT)(Wang & Lewis).

As reuptake of neurotransmitters by these proteins stop transmission, inhibition of these pumps has stimulant effect on nervous system. Drugs, which block these transporters, are of pharmaceutical value in treatment of various nervous system related disorders like depression, anxiety, obsessive-compulsive disorder. Illicit addictive drugs like cocaine also act by non-specific inhibition of these NSS proteins. Thus to decrease addictive potential and maximize therapeutic potential, selective inhibitors of each type of transporters are developed like selective serotonin reuptake inhibitors (SSRIs), Selective Norepinephrine Reuptake Inhibitors (SNRIs) and Selective Dopamine Reuptake Inhibitors (SDRIs)(Serafini *et al.*, ; Westenberg, 1999).

Leucine Transporter

Leucine Transporter (LeuT) is a bacterial homologue of mammalian NSS proteins. In 2005 Yamashita A et al. solved structure of LeuT from *Aquifex aeolicus* along with its bound ligand Leucine and two sodium ions(Yamashita *et al.*, 2005). In 2007 Singh SK et al. and Zhou Z et al. separately solved crystal structures of LeuT along bound antidepressant (inhibitor)(Singh *et al.*, 2007; Zhou *et al.*, 2007). These structures revealed that inhibitor binds 11 Å above Leucine binding site on outer side of transport protein. As LeuT is the closest relative to human NSS proteins (hNSS) for which structure is available, attempts have been made to use LeuT as a template to understand functioning of hNSS proteins(Singh, 2008). Mutagenesis studies on hNSS have shown that this antidepressant binding-site and resulting inhibition of LeuT is probably preserved in hNSS proteins.

Figure 1 shows structure of LeuT in complex with antidepressant clomipramine. LeuT is formed by 12 trans-membrane (TM) Helices. First 10 of these helices form protein core. Leucine and sodium ions are bound within the protein core halfway across the membrane bi-layer in an occluded site devoid of water.

Clomipramine, antidepressant and inhibitor of LeuT, is bound 11Å above the substrate in an extra-cellular facing vestibule. This binding of clomipramine stabilizes extra-cellular gates of LeuT in closed conformation. As expected from information available through crystal structure it was shown that clomipramine non-competitively inhibits LeuT.

Availability of LeuT structure for comparative structural studies brought new wave of research in hNSS proteins. Zhou Z et al. solved structures of LeuT in complex with three selective serotonin inhibitors to suggest possible determinants of specificity for human SERT receptors(Zhou *et al.*, 2009). They have shown that LeuT has a halogen binding pocket (HBP) for antidepressant just like SERT has at antidepressant binding site and perhaps this pocket gives selective SERT inhibitor its specificity given it is these halogens that make known SERT inhibitors like Sertraline and Fluoxetine effective and selective. Though surprisingly only difference human SERT has to DAT and NET in this 7 residue HBP is alanine in DAT and NET compared to glycine in SERT (**Table 1**).

In other study, Indarte M et al. built a 3D model of DAT using LeuT as a template and docked dopamine and amphetamine in to substrate binding site to find out dopamine binding site and residues important for dopamine binding(Indarte *et al.*, 2008). In yet another study, Ravna AW et al. have shown that 3D model of SERT built using LeuT template is of sufficient accuracy to find out substrate binding residues using approaches like docking(Ravna *et al.*, 2006). They have proved that predictions from site-directed mutagenesis studies on SERT can be validated and understood at structural level using SERT model built from LeuT.

But Despite these advances, important questions still remain unanswered. Selective inhibitors exist for DAT, NET and SERT but how these molecules achieve their unique selectivity profile is still uncertain. Thus availability of LeuT structure gives unique opportunity to built 3D models of DAT, NET and SERT so that comparative structural analysis using methods like docking can be used to unearth reasons behind unique structural specificities of selective inhibitors. Predictions made using such approach will need further validation from biochemical mutagenesis studies in an
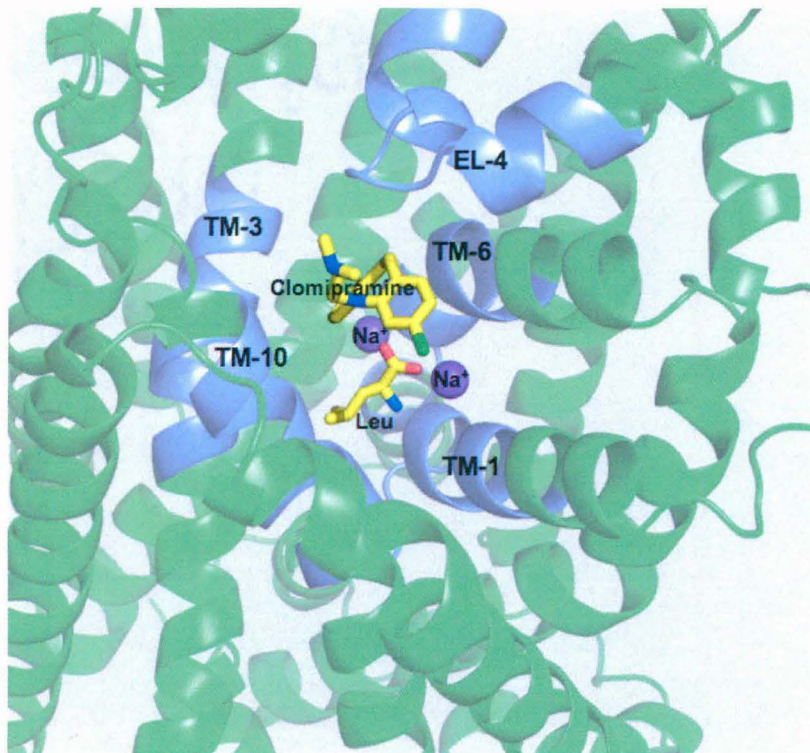
Figure 1 Leucine and Clomipramine binding sites on LeuT

LeuT	SERT	NET	DAT
Leu 25	Leu 99	Leu 76	Leu 80
Gly 26	Gly 100	Ala 77	Ala 81
Leu 29	Trp 103	Trp 80	Trp 84
Arg 30	Arg 104	Arg 81	Arg 85
Tyr 108	Tyr 176	Tyr 151	Tyr 156
Ile 111	Ile 179	Ile 155	Ile 159
Phe 253	Phe 335	Phe 317	Phe 320

Table 1 Sequence alignment at Halogen Binding Pocket between NSS proteins

iterative process so that each iteration will further the understanding of how inhibitors and ligands differentially interact with different hNSS proteins.

RATIONALE

Bupropion is a selective DAT inhibitor, Atomoxetine is a selective NET inhibitor and Fluoxetine is a selective SERT inhibitor(Bymaster *et al.*, 2002). **Table 2** gives specificity profile of each of above inhibitor in terms of Ki values.

Availability of LeuT crystal structure with conserved inhibitor binding site means that 3D models built from LeuT have potential use in understanding specificity determining residues of hNSS proteins and structural mechanisms working in them, which determines selectivity for inhibitors. Such understanding can be of help in developing selective inhibitors. **Table 3** gives sequence similarity data of LeuT to human transporters. As can be seen from table all hNSS proteins have sufficient similarity to LeuT to built reliable models for comparative structural study.

Thus plan here is to built models of SERT, DAT and NET from LeuT and dock Bupropion, Atomoxetine and Fluoxetine into inhibitor binding site of all these models in order to identify critical differences in DAT, NET and SERT inhibitor binding sites which gives them unique specificity profiles.

METHODOLOGY

Parameters used in doing docking using autodock and simulations using AMBER9 are similar to those which are followed in chapter 3 and 4 of the thesis and not discussed here. Given in this section is the step-wise procedure followed in order to extract desired information from LeuT, DAT, SERT and NET using structure based docking and simulation tools.

Step 1 Docking of Clomipramine into LeuT active site

Structures are available for LeuT both with bound inhibitor Clomipramine and without it. Crystal structure without bound inhibitor does not have side chains optimally oriented to bind inhibitor in inhibitor binding site. Thus to see if docking tool can correctly identify and put inhibitor in to proper orientation in LeuT inhibitor binding site, Clomipramine is docked into structure of LeuT obtained without bound inhibitor and the result is compared with available LeuT-Clomipramine crystal structure. As ultimate aim is to dock different inhibitors into structural models of human NSS proteins, this docking

Inhibitor	Ki for DAT (nM)	Ki for NET (nM)	Ki for SERT (nM)
Bupropion	562	>10000	>10000
Atomoxetine	1451	5	77
Fluoxetine	4752	1022	7

Table 2 Inhibition Constant (Ki) of Bupropion, Atomoxetine and Fluoxetine for hNSS proteins.

Query	Subject	% Identity	% Similarity	% Query Coverage
LeuT	DAT	24	43	89
LeuT	NET	29	47	92
LeuT	SERT	26	45	83

Table 3 Sequence similarity of LeuT to human NSS proteins calculated using
NCBI online protein-protein BLAST tool.

exercise will give idea about reliability of docking tool for Transporter-Inhibitor scenario. Docking tool used here is Autodock version 4.

Step 2 MD Simulations on LeuT

To understand dynamics of inhibitor-transporter interaction in Clomipramine-LeuT complex MD Simulations are performed on following LeuT complexes under explicit solvent conditions using Amber-9 Package.

1) LeuT + Leucine + 2Na$^+$
2) LeuT + Alanine + 2Na$^+$
3) LeuT + Leucine + Clomipramine + 2Na$^+$
4) LeuT + Leucine + Docked Clomipramine + 2Na$^+$

Alanine is another molecule that LeuT is known to transport other than Leucine. Both LeuT-Clomipramine crystal structure and LeuT-Clomipramine complex obtained by docking are used here in simulation study to see if docked complex behaves similar to the original crystal structure. All simulations are carried out for three nano-seconds. Trajectories obtained are analyzed using tools available with Amber-9 simulation package like MM_GBSA and Ptraj. MM_GBSA module is used to find out binding free energy between LeuT-Clomipramine complex. MM_GBSA module is also used in identifying critical residues for Clomipramine binding by residue-wise splitting of binding free energy. Ptraj is used for monitoring distance between Arg-30 and Asp-404, residues that form salt-bridge gate at the base of inhibitor binding site. Purpose was to see effect of inhibitor on this gate by comparing simulations in absence and presence of inhibitor.

Step 3 Building models of DAT, NET and SERT and docking of inhibitors

Modeller (Version 9v1) is used for building models of DAT, NET and SERT based on LeuT template. Autodock is then used to dock inhibitors Bupropion, Atomoxetine and Fluoxetine into these models. Autodock Tool (ADT) is then used to find out best binding pose. Inhibitor-Model complex thus obtained is then subjected to manual structural analysis using Pymol.

RESULTS AND DISCUSSION

1) Docking of Clomipramine into LeuT inhibitor binding site

As mentioned earlier crystal structures of LeuT are available both without inhibitor clomipramine and in presence of clomipramine. This created an opportunity to check efficiency of docking procedure. LeuT crystallized with clomipramine has its amino acid side-chains optimally oriented for binding clomipramine. But interest here was to see if docking could correctly identify Clomipramine binding site in LeuT structure crystallized in absence of Clomipramine. With this intention Clomipramine was docked into LeuT structure. The lowest energy and highest affinity pose of Clomipramine in LeuT obtained by docking was comparable to LeuT-Clomipramine co-crystal structure. **Figure 2** shows images of LeuT-Clomipramine co-crystal structure and LeuT-Docked Clomipramine structure for visual comparison.

As can be seen Clomipramine is docked into same position and with similar orientation to that of original crystal structure, thus validating efficiency and accuracy of docking tools.

2) MD Simulations on LeuT

As mentioned earlier, inhibitor binding-site is on extra-cellular facing surface of LeuT. This site forms part of TM passage through which ligands Leucine sodium ions are transported. Amino acid residues Arg-30 and Asp-404 are among those, which form gates for selective passage for desired ligands. These residues form salt-bridge across the passage to close the gate in absence of proper ligand. This salt-bridge forms bottom of inhibitor binding site. Along with Leucine, LeuT is also known to transfer Alanine from extra-cellular space to intra-cellular space. Thus to see effect of inhibitor on Arg-30 and Asp-404 salt-bridge and understand LeuT-Clomipramine dynamics, MD simulations are performed on following protein ligand systems as detailed in methodology section.

1) LeuT + Leucine + 2Na$^+$
2) LeuT + Alanine + 2Na$^+$
3) LeuT + Leucine + Clomipramine + 2Na$^+$
4) LeuT + Leucine + Docked Clomipramine + 2Na$^+$

Another reason behind this dynamics is to see if docked Clomipramine has similar effect on LeuT to that of Clomipramine co-crystallized with LeuT.
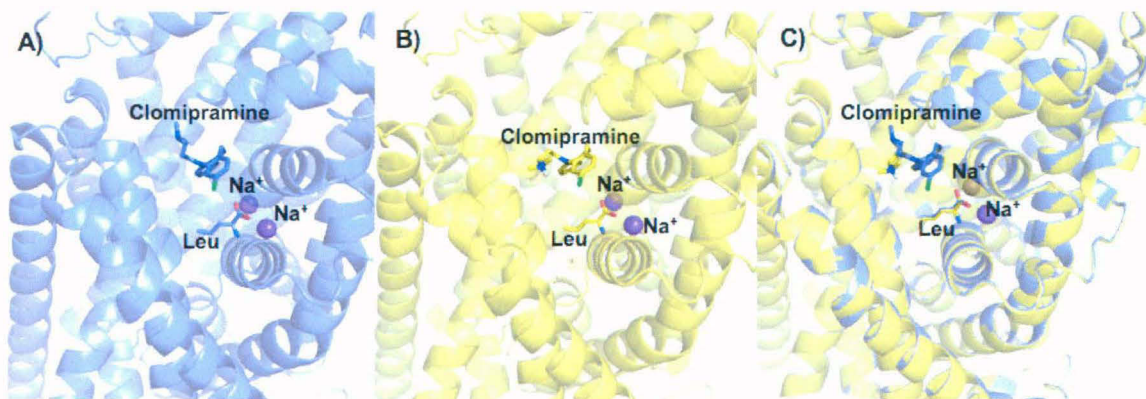
Figure 2 A) Structure of LeuT + Clomipramine co-crystal structure (Purple) **B)** Structure of LeuT + Docked Clomipramine (Yellow) **C)** Superimposed structures of A) and B)

2.1 RMSD Analysis

Starting structure is compared with all snapshots collected during dynamics for each system to ensure all systems have held stable during simulations (**Figure 3**).

2.2 Critical Residues at Inhibitor Binding Site of LeuT

Next step was to identify amino acid residues of LeuT important in clomipramine binding and to check their conservation profile across hNSS proteins. For this MM_GBSA analysis is performed on snapshots taken form last 1 ns trajectory of LeuT + Clomipramine (co-crystal structure) system. Total binding free energy between LeuT and Clomipramine was found out to be -46.18 Kcal/Mol. Residue-wise splitting of binding free energy helped identifying LeuT residues contributing maximally in Clomipramine-LeuT binding (**Table 4**).

Figure 4-A shows conservation profile of these five residues across hNSS proteins. **Figure 4-B** shows schematic diagram of Clomipramine molecule.

Arg-30 and Asp-404 of LeuT form salt-bridge at the base of inhibitor binding-site (**Figure 5**). This salt-bridge forms one of the gates along the TM passage. As expected this gate is conserved across all human NSS proteins. Only minor difference being SERT having Glu instead of Asp. But Gln-34 of LeuT, which makes polar interactions with Chlorine of Clomipramine, is replaced by hydrophobic Leu/Ile residues in human NSS proteins. Phe-320 another hydrophobic aromatic residue of LeuT is replaced by hydrophobic non-aromatic residues in hNSS proteins (Pro/Ala). But most critical difference occurs at Asp-401 of LeuT. This negatively charged residue forms salt-bridge type interaction with tertiary amine of Clomipramine and is absent in human transporters, replaced by Thr in DAT/NET and Leu in SERT.

2.3 Effect of Clomipramine on Salt-Bridge Gate

To see what effect inhibitor has on salt-bridge gate, distance between Arg-30 and Asp-404 is monitored over 3 ns time scale in all four simulations. **Figure 6A** shows comparison between LeuT + Leucine and LeuT + Alanine simulations. As told earlier LeuT can transport both Leucine and Alanine as substrates and in both simulations no inhibitor was present in inhibitor binding site. As can be seen from graph distance between Arg and Asp salt-bridge varies widely between 3Å to 7Å in both cases indicating that in absence of inhibitor salt-bridge gate varies between open and closed
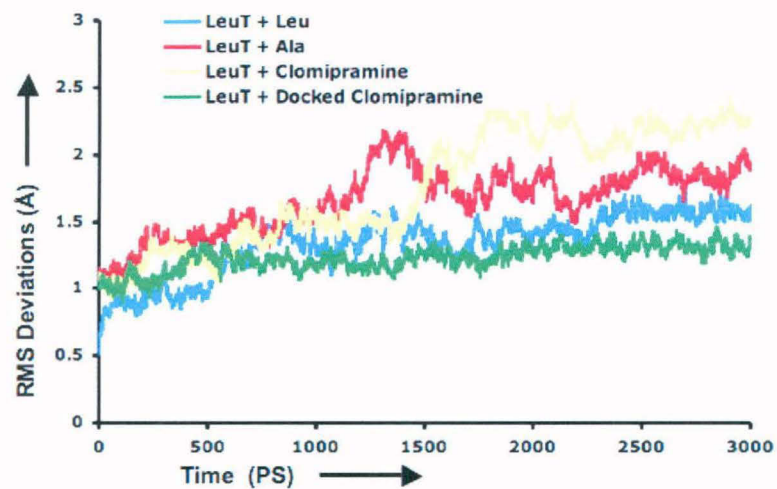
Figure 3 Shows RMSDs to starting structure for MD simulations on LeuT

LeuT Residue	Contribution in Clomipramine Binding (Kcal/Mol)
Arg-30	-9.84
Gln-34	-2.08
Phe-320	-4.69
Asp-401	-7.61
Asp-404	-2.48

Table 4 Residues of LeuT contributing maximally in Clomipramine binding.

A)

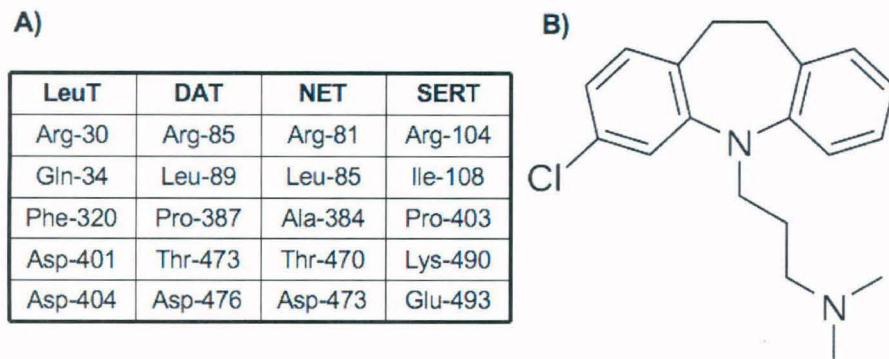| LeuT	DAT	NET	SERT
Arg-30	Arg-85	Arg-81	Arg-104
Gln-34	Leu-89	Leu-85	Ile-108
Phe-320	Pro-387	Ala-384	Pro-403
Asp-401	Thr-473	Thr-470	Lys-490
Asp-404	Asp-476	Asp-473	Glu-493

B)

Figure 4 A) Conservation of key residues of LeuT in Clomipramine binding across hNSS proteins. **B)** Schematic diagram of Clomipramine
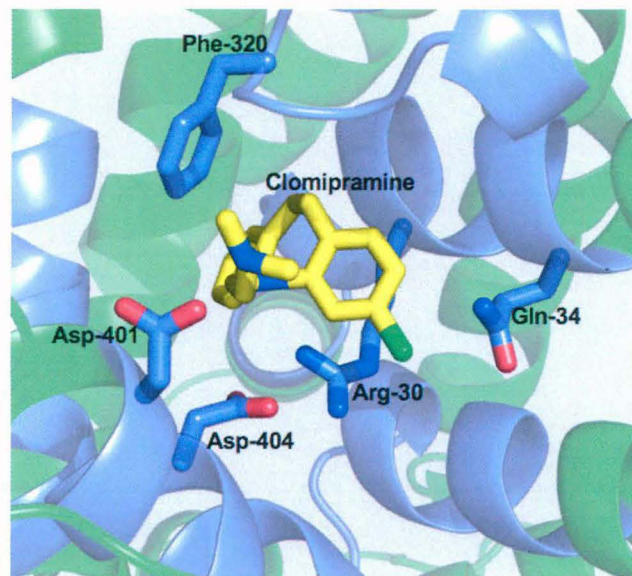

Figure 5 Clomipramine binding in LeuT, showing five residues contributing most in binding free energy of the complex as per MM GBSA analysis of simulation trajectory.

conformations. Graph in Figure **6B compares** salt-bridge distance in LeuT + Leucine simulation with LeuT + Leucine + Clomipramine (co-crystal structure) simulation. As can be seen Clomipramine binding shuts the gate in close conformation by stabilizing the salt-bridge that runs across the TM channel. **Figure 6C** compares gate distance in LeuT + Leucine Simulation with LeuT + Leucine + Clomipramine (docked Clomipramine) Simulation. In this case Clomipramine was originally absent in PDB and was docked as described in methodology. As can be seen, docked Clomipramine has similar stabilizing effect on salt-bridge gate to that of original LeuT + Clomipramine complex taken from co-crystal structure. This data fortifies docking results.

2.4 Docking of Selective Inhibitors into human NSS Protein Models

As detailed in methodology section, Modeller is used to build models of DAT, NET and SERT proteins from LeuT. Each model is docked with three small molecule ligands, Bupropion the selective DAT inhibitor, Atomoxetine the selective NET inhibitor and Fluoxetine the selective SERT inhibitor. **Figure 7** shows schematic diagrams of natural ligands and selective inhibitors of DAT, NET and SERT. As can be noticed, both ligands and inhibitors have similar general architecture, a primary or secondary amine attached to aromatic group/groups through alkyl chain. To understand how differences around this conserved architecture influence inhibitor selectivity for these transporters all inhibitors are docked into binding site on these models, which correspond, to inhibitor binding site in LeuT. Purpose is to understand how these selective inhibitors achieve their unique specificity against human NSS proteins (**Table 2**).

Analysis of docking results showed all dockings to be successful, i.e. in all cases at least one of the clusters belonging to lowest binding free energy was in supposed inhibitor binding site. **Table 5** lists binding free energy of each selected NSS-Inhibitor complex. It also gives contribution of electrostatic and van der waals forces towards observed binding free energy.

Each docking pose listed in table is selected based on two criteria. First, it should be one of the lowest energy binding-pose and second, inhibitor should be in supposed inhibitor binding site. As can be seen from the table, all NSS-Inhibitor complexes have similar binding energies and it is impossible to predict selective inhibitory properties of Bupropion, Atomoxetine and Fluoxetine based on it. Though it can be noted that DAT and NET both have higher electrostatic component for interaction with inhibitors compared to SERT.
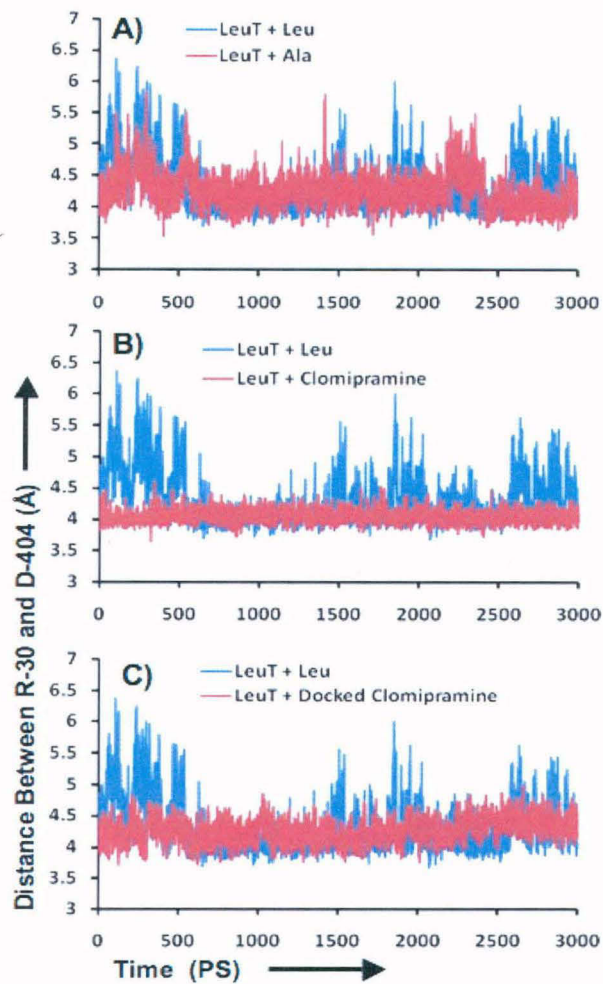
Figure 6 Comparison of distance between Arg-30 and Asp-404 slat-bridge over the course of 3 ns simulations for different LeuT systems. **A)** LeuT + Leucine and LeuT + Alanine **B)** LeuT + Leucine and LeuT + Clomipramine **C)** LeuT + Leucine and LeuT + Docked Clomipramine
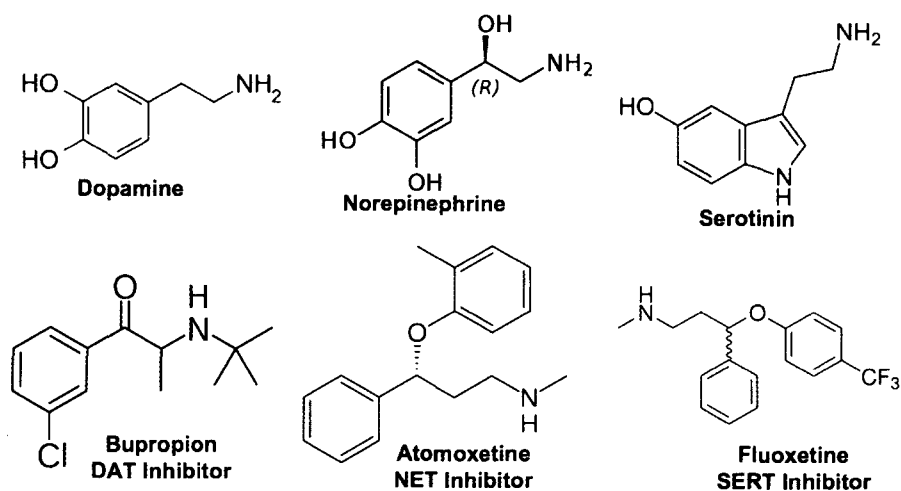
Figure 7 Natural ligands of DAT, NET and SERT i.e. Dopamine, Norepinephrine and Serotonin respectively and their selective inhibitors.

Transporter	Inhibitor	Binding Free Energy (Kcal/mol)	Electrostatic Component (Kcal/mol)	Van der Waals Component (Kcal/mol)
DAT	Bupropion	-8.16	-1.46	-7.7
NET	Bupropion	-7.04	-0.82	-7.07
SERT	Bupropion	-7.64	-0.18	-7.44
DAT	Atomoxetine	-8.75	-1.67	-8.22
NET	Atomoxetine	-7.89	-1-19	-7.74
SERT	Atomoxetine	-7.42	-0.35	-7.98
DAT	Fluoxetine	-9.65	-1.91	-9.41
NET	Fluoxetine	-7.61	-1.21	-7.74
SERT	Fluoxetine	-7.45	-0.32	-8.43

Table 5 Binding free energy of chosen NSS-Inhibitor complexes obtained from docking studies. Table also shows division of total binding energy in to electrostatic and van der waal components.

2.5 Sequence Comparison of Inhibitor Binding Site of LeuT, DAT, NET and SERT

Sequences of LeuT and human NSS proteins are aligned using ClustalW. Residues of LeuT interacting with Clomipramine are extracted from LeuT-Clomipramine crystal structure. Similarly, residues of DAT, NET and SERT interacting with Bupropion, Atomoxetine and Fluoxetine are extracted from respective docked models and these residues are highlighted on sequence alignment to point out inhibitor binding regions of these transporters (**Figure 8**).

As can be seen from alignment, five separate regions form inhibitor binding-site in all the cases. And these regions superimpose in sequence alignment between these transporters. **Region-1** is formed by C-terminal half of TM Helix-1 and contains Arginine, essential component of salt-bridge gate discussed earlier. In case of LeuT, it also contains Gln-34 (**Figure 5**), which makes polar interactions with chlorine of Clomipramine, but this Gln is absent in human transporters. **Region-2** is formed by mid-region of TM Helix-3 and contains mostly aromatic and non-polar residues with notable exception in SERT. In SERT this region is shifted downwards on sequence alignment compared to other regions but superimposes well in structural alignment. This variation in SERT compared to others is likely due to six-residue deletion downstream 20 amino acids unique to SERT among human NSS proteins. Due to this Region-2 of SERT contains polar Asp and Gln Residues. **Region-3** is formed by amino terminal end of TM Helix-6 and contains mainly a conserved phenylalanine residue, which seems to be forming another gate below Arg-Asp Salt-Bridge along the TM channel. **Region-4** is formed by extra-cellular loop-4 (EL-4). In case of LeuT, this region comprises non-polar residues, but for human NSS proteins, this region has negatively charged residue i.e. Asp for DAT and SERT and Glu for NET. Interestingly, SERT has one amino acid insertion (Ala) after its Asp compared to DAT and NET. **Region-5** is formed by N-terminal part of TM Helix-10. In case of all proteins it contains counterpart of positively charged Arg of Region-1 i.e. negatively charged Asp/Glu residue. These oppositely charged residues form salt-bridge across TM channel as detailed earlier. But in addition to this conserved Asp (Asp-404), Region-5 of LeuT contains another Asp (Asp-401), which interacts with tertiary amine of inhibitor Clomipramine in LeuT-Clomipramine co-crystal structure. Importantly, this Asp is absent in human NSS proteins. In DAT and NET it is replaced by Thr and in SERT it is replaced by a residue carrying opposite charge to that of Asp i.e. Lys. To understand how differences in these five regions accounts for inhibition by

```
DopamineT    --MSKSKCSVGLMSSVVAPAKEPNAVGPKEVELILVKEQNGVQLT-SSTLTNPRQSP--- 54
NoradT       MLLARMNPQVQPENNGADTGPEQPLRARKTAELLVVKERNGVQC-----LLAPRDG---- 51
SeroT        METTPLNSQKQLSACEDGEDCQENGVLQKVVPTPGDKVESGQISNGYSAVPSPGAGDDTR 60
LeuT         ------------------------------------------------------------
```

Region 1

```
DopamineT    -------------VEAQDRETWGKKIDFLLSVIGFAVDLANVWRFPYLCYKNGGGAFLVP 101
NoradT       -------------DAQPRETWGKKIDFLLSVVGFAVDLANVWRFPYLCYKNGGGAFLIP 97
SeroT        HSIPATTTTLVAELHQGERETWGKKVDFLLSVIGYAVDLGNVWRFPYICYQNGGGAFLLP 120
LeuT         -------------MEVKREHWATRLGLILAMAGNAVGLGNFLRFPVQAAENGGGAFMIP 46
                             ** *..::.::*:: * **.*.*. ***   . :******::*
```

```
DopamineT    YLLFMVIAGMPLFYMELALGQFNREGAAGVW-KICPILKGVGFTVIL------ISLYVGF 154
NoradT       YTLFLIIAGMPLFYMELALGQYNREGAATVW-KICPFFKGVGYAVIL------IALYVGF 150
SeroT        YTIMAIFGGIPLFYMELALGQYHRNGCISIWRKICPIFKGIGYAICI------IAFYIAS 174
LeuT         YIIAFLLVGIPLMWIEWAMGRYGGAQGHGTTPAIFYLLWRNRFAKILGVFGLWIPLVVAI 106
               * :  :: *:**::* *:*::         *   ::      :: :      *.: :.
```

Region 2

```
DopamineT    FYNVIIAWALHYLFSSFTTELPWIHCNNSWNSPNCSDAHPGD-SSGDSSGLNDTFGTTPA 213
NoradT       YYNVIIAWSLYYLFSSFTLNLPWTDCGHTWNSPNCTDPKLLNGSVLGNHTKYSKYKFTPA 210
SeroT        YYNTIMAWALYYLISSFTDQLPWTSCKNSWNTGNCTNYFSED------NITWTLHSTSPA 228
LeuT         YYVYIESWTLGFAIKFLVGLVPEP-------PPNATDPDSIL--------------RPF 144
             :*   * :*:* : :.. :.  :*         . *.::                   *
```

```
DopamineT    AEYFERGVLHLHQSHGIDDLGPPRWQLTACLVLVIVLLYFS-LWKGVKTSGKVVWITATM 272
NoradT       AEFYERGVLHLHESSGIHDIGLPQWQLLLCLMVVVIVLYFS-LWKGVKTSGKVVWITATL 269
SeroT        EEFYTRHVLQIHRSKGLQDLGGISWQLALCIMLIFTVIYFS-IWKGVKTSGKVVWVTATF 287
LeuT         KEFLYSYIGVPKGDEPILKPSLFAYIVFLITMFINVSILIRGISKGIERFAKIAMPTLFI 204
               *:      :     :    : . .     : :     :.:   : :  : **::  .*:. *  :
```

Region 3

```
DopamineT    PYVVLTALLLRGVTLPG-AIDGIRAYLSVDFYRLCEASVWIDAATQVCFSLGVGFGVLIA 331
NoradT       PYFVLFVLLVHGVTLPG-ASNGINAYLHIDFYRLKEATVWIDAATQIFFSLGAGFGVLIA 328
SeroT        PYIILSVLLVRGATLPG-AWRGVLFYLKPNWQKLLETGVWIDAAAQIFFSLGPGFGVLLA 346
LeuT         LAVFLVIRVFLLETPNGTAADGLNFLWTPDFEKLKDPGVWIAAVGQIFFTLSLGFGAIIT 264
               ..*     :.     *   * *  *:          :: :* :. *** *. *: **.:::
```

Region 4

```
DopamineT    FSSYNKFTNNCYRDAIVTTSINSLTSFSSGFVVFSFLGYMAQKHSVPIGDVAKD-GPGLI 390
NoradT       FASYNKFDNNCYRDALLTSSINCITSFVSGFAIFSILGYMAHEHKVNIEDVATE-GAGLV 387
SeroT        FASYNKFNNNCYQDALVTSVVNCMTSFVSGFVIFTVLGYMAEMRNEDVSEVAKDAGPSLL 406
LeuT         YASYVRKDQDIVLSGLTAATLNEKAEVILG-GSISIPAAVAFFGVANAVAIAKAGAFNLG 323
             ::**   : ::     ..: :: :*  :..  *    ::.. :*        :*.  . .*
```

```
DopamineT    FIIYPEAIATLPLSSAWAVVFFIMLLTLGIDSAMGGMESVITGLIDEF-QLLHRHRELFT 449
NoradT       FILYPEAISTLSGSTFWAVVFFVMLLALGLDSSMGGMEAVITGLADDF-QVLKRHRKLFT 446
SeroT        FITYAEAIANMPASTFFAIIFFLMLITLGLDSTFAGLEGVITAVLDEFPHVWAKRRERFV 466
LeuT         FITLPAIFSQTAGGTFLGFLWFFLLFFAGLTSSIAIMQPMIAFLEDEL-KLSRKHAVLWT 382
             **    . ::   . .: ..::*.:*: *: *::. :: :*: : *:: :: :: :.
```

Region 5

```
DopamineT    LFIVLATFLLSLFCVTNGGIYVFTLLDHFAAGTSILFGVLIEAIGVAWFYGVGQFSDDIQ 509
NoradT       FGVTFSTFLLALFCITKGGIYVLTLLDTFAAGTSILFAVLMEAIGVSWFYGVDRFSNDIQ 506
SeroT        LAVVITCFFGSLVTLTFGGAYVVKLLEEYATGPAVLTVALIEAVAVSWFYGITQFCRDVK 526
LeuT         AAIVFFSAHLVMFLNKS-----LDEMDFWAGTIGVVFFGLTELIIFFWIFGADKAWEEIN 437
             :.:        :.  .       .  :: :*  .:: * * :. *::* :    :::
```

```
DopamineT    QMTGQRPSLYWRLCWKLVSPCFLLFVVVVSIVTFRPPHYGAYIFPDWANALGWVIATSSM 569
NoradT       QMMGFRPGLYWRLCWKFVSPAFLLFVVVVSIINFKPLTYDDYIFPPWANWVGWGIALSSM 566
SeroT        EMLGFSPGWFWRICWVAISPLFLLFIICSFLMSPPQLRLFQYNYPYWSIILGYCIGTSSF 586
LeuT         RGGIIKVPRIYYYVMRYITPAFLAVLLVVWAREYIPKIMEETHWTVWITRF-------YI 490
             .            :    ::* ** .::              :. *   .         :
```

```
DopamineT    AMVPIYAAYKFCSLPGSFREKLAYAIAPEKDRELVDRGEVRQFTLRHWLKV 620
NoradT       VLVPIYVIYKFLSTQGSLWERLAYGITPENEHHLVAQRDIRQFQLQHWLAI 617
SeroT        ICIPTYIAYRLIITPGTFKERIIKSITPETPTEIP-CGDIRLNAV------ 630
LeuT         IGLFLFLTFLVFLAERRRNHESAGTLVPR--------------------- 519
              :  :  :.           ..     :.*.
```

Figure 8 Sequence alignment of LeuT with DAT, NET and SERT showing regions of these transporters forming inhibitor binding-site.

different selective inhibitors, docked complexes of DAT, NET and SERT proteins with inhibitors are examined further using Pymol.

2.6 Bupropion Binding to DAT

Bupropion is a selective DAT inhibitor (Table 2). To understand how Bupropion binds to DAT, detailed visual analysis is performed on DAT-Bupropion complex. **Figure-9** shows binding pose of Bupropion inside DAT inhibitor binding site.

Bupropion binds on the top of salt-bridge formed by Arg-85 of Region-1 and Asp-476 of Region-5. Region-2 and Regon-3 consist mainly of aromatic or non-polar residues. But as can be seen Asp-385 of Region-4 makes salt-bridge type of interactions with secondary amino group of Bupropion. In case of LeuT and its interaction with Clomipramine, this part is played by Asp-401 in Region-5, which is absent in human transporters. But Asp of Region-4 is conserved across human transporters except minor changes i.e. Glu instead of Asp in NET and single amino acid insertion in SERT (Ala) after Asp. Thus according to this analysis, human transporters DAT, NET and SERT have Asp/Glu residue in Region-4, which corresponds to Asp-401 in Region-4 of LeuT. Carbonyl moiety of Bupropion can form H-bond with Tyr-88 of Region-1. Chloro-Benzene moiety of Bupropion is tucked into polar backbone of extra-cellular loop of Region-4. Thus relatively small Bupropion makes specific interactions primarily with two regions of DAT, Region-1 and Region-4.

2.7 Bupropion Binding to NET

As can be seen from table 2 Bupropion has much lesser affinity for NET compared to DAT. Though, as can be seen from **Figure 10**, pattern of its interaction with NET is essentially similar to that of DAT.

DAT has Asp-385 in Region-4, which is replaced by E-382 in NET. Stout Asp and EL-4 loop of Region-4 form perfect hook for small Bupropion in DAT. But as NET has longer Glu instead of Asp, chlorobenzene moiety is not hooked inside EL-4 backbone as Glu pulls secondary amine of Bupropion in opposite direction. This possibly explains more affinity of DAT for Bupropion compared to NET.

2.8 Bupropion Binding to SERT

As can be seen from **Figure 11**, Bupropion interacts differently with SERT compared to DAT and NET.
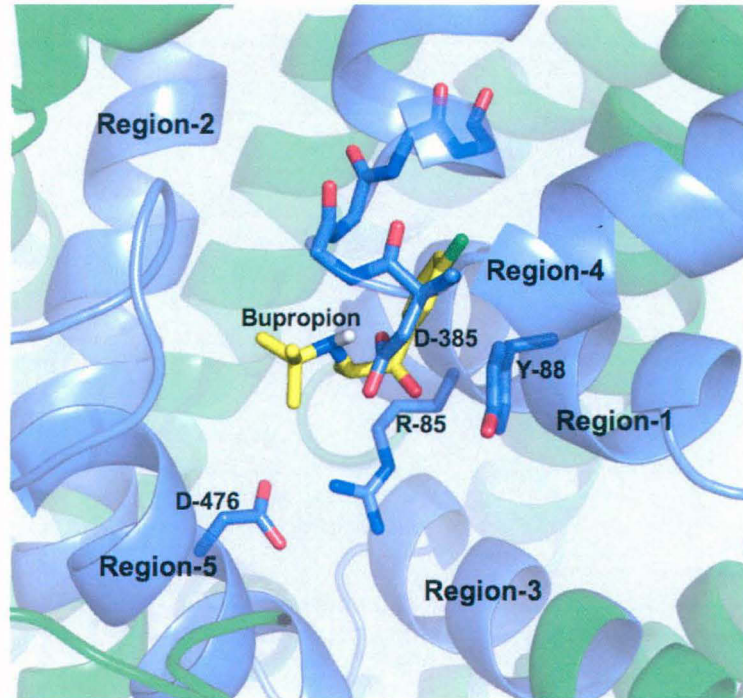
Figure 9 Binding pose of Bupropion in DAT inhibitor binding site as obtained by docking. Color Code:
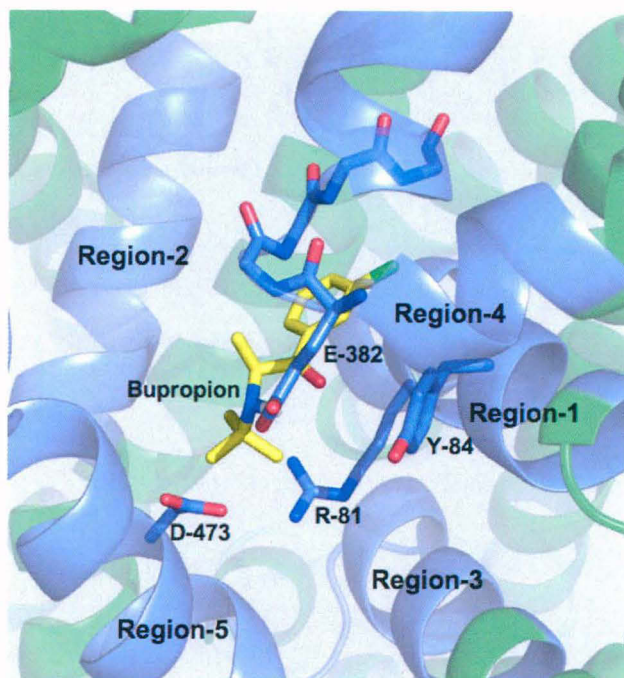
Figure 10 Binding pose of Bupropion in NET inhibitor binding site as obtained by docking. Color Code: NET-Green, Inhibitor Binding Regions of NET-Purple, Bupropion-Yellow.
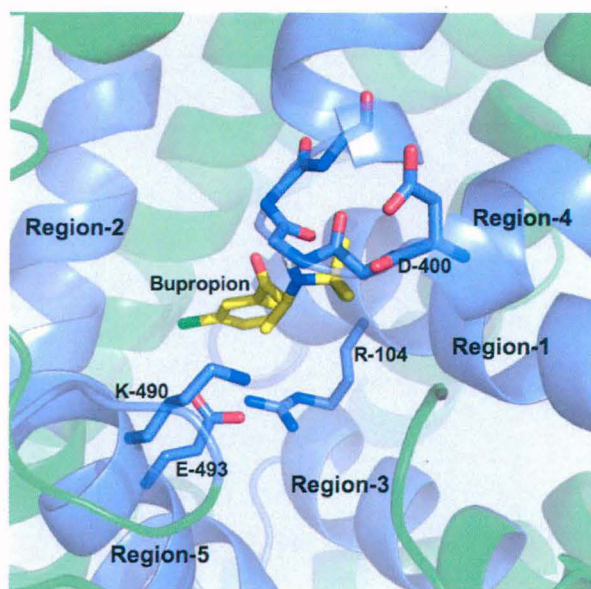

Figure 11 Binding pose of Bupropion in SERT inhibitor binding site as obtained by docking. Color Code: SERT-Green, Inhibitor Binding Regions of SERT-Purple, Bupropion-Yellow.

Secondary amino group of Bupropion cannot bind Asp-400 of Region-4 of SERT. This is due to one amino acid insertion after D-400 in SERT compared to respective negatively charged residue in DAT and NET. This insertion pushes Asp of Region-4 of SERT away from inhibitor binding site. Thus there is no salt-bridge type of interaction between Bupropion and SERT. Thus Bupropion binds further down the TM channel compared to DAT and NET with its chloro-benzene group pointing inside TM channel as oppose to DAT and NET where it's either inserted in on near EL-4 loop.

This explains why Bupropion has much lesser affinity for SERT compared to DAT. Asp of Region-4 is not available for binding amino group of Bupropion due to single insertion. As if to compliment this mutation positively charged Lys-490 of Region-5 protrudes in TM channel, thus further destabilizing binding of inhibitor in a position, which will enable it to bind Asp of Region-4. This is because positive amine of Lys-490 will repel positive secondary amine of inhibitor and Lys-490 protrudes in same position where inhibitor amine lies, in case of DAT and NET.

2.9 Atomoxetine Binding to NET

Atomoxetine is a selective NET inhibitor as shown by its inhibition constant (Ki) values for NET, DAT and SERT. Though it is selective for NET, it also has some activity against SERT but is least effective against DAT. In docked Atomoxetine-NET complex once again positive secondary amine of Atiomoxetine anchors it to NET through interaction with negative Glu-382 of Region-4 (**Figure 12**).

Atomoxetine is a non-polar aromatic molecule larger than Bupropion, and longer side-chain of Glu-382 compared to Asp in SERT/DAT apparently helps proper accommodation of Atomoxetine in inhibitor binding site. Atomoxetine has single oxygen atom which links propanamine chain to toluene moiety. This oxygen is in a position to form hydrogen bond with Y-84 of Region-1. Again Atomoxetine binding position lies right on top of salt-bridge gate formed by Arg-81 and Asp-473. Except for above-mentioned interactions, Atomoxetine binds NET mainly through non-specific van der waal interactions.

2.10 Atomoxetine Binding to SERT

As was the case with Bupropion, Asp-400 of Region-4 is not available for Atomoxetine binding (**Figure 13**). But this possibly gives greater freedom to Atomoxetine inside inhibitor binding site. Atomoxetine is oriented in such a way that its
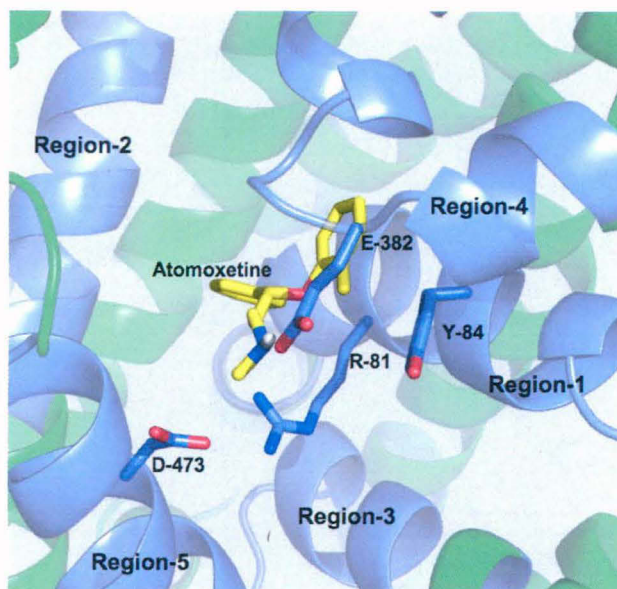
Figure 12 Binding pose of Atomoxetine in NET inhibitor binding site as obtained by docking. Color Code: NET-Green, Inhibitor Binding Regions of NET-Purple, Atomoxetine-Yellow.
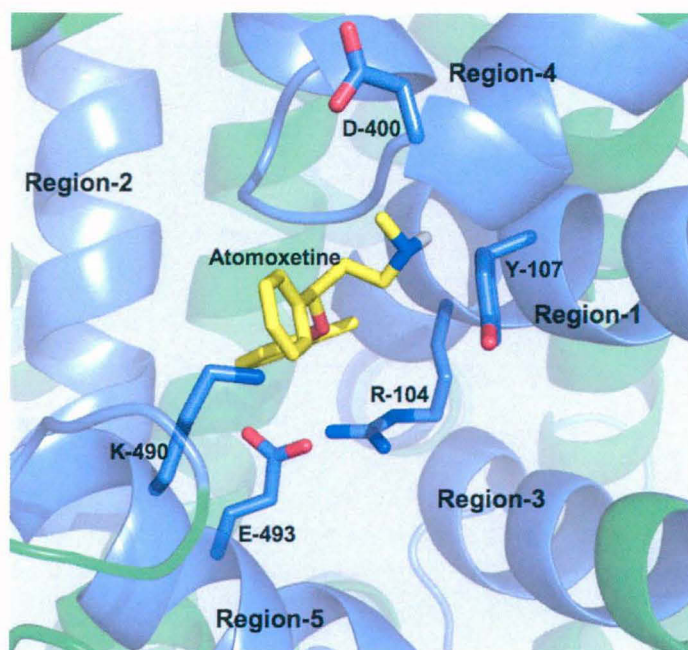

Figure 13 Binding pose of Atomoxetine in SERT inhibitor binding site as obtained by docking. Color Code: SERT-Green, Inhibitor Binding Regions of SERT-Purple, Atomoxetine-Yellow.

secondary amine group can interact with backbone atoms of loop forming Region-4. In real situation it might even get access to Asp-400 through this EL-4 loop. This possibly accounts for activity shown by Atomoxetine against SERT. Only oxygen in Atomoxetine is also in a position to form hydrogen bond with Tyr-107 of Region-1. Again, Lys-490 of Region-5 possibly ensures different orientation of Atomoxetine in SERT compared to NET/DAT by its ability to make repulsive interactions with secondary amine of Atomoxetine.

2.11 Atomoxetine Binding to DAT

Atomoxetine binds to DAT differently than it does to NET, primarily due to D-385 of Region-4 (**Figure 14**). Its counterpart in NET is Glu-382. Shorter length of Asp-385 probably gives less freedom for Atomoxetine in DAT binding site. Thus Atomoxetine is bound to DAT in such a way that its aromatic rings are flipped opposite compared to its binding pose in NET. In this orientation, oxygen of Atomoxetine is not available for binding Tyr-84 of Region-1. Together these features are likely to be responsible for low activity of Atomoxetine against DAT.

2.12 Fluoxetine Binding to SERT

Fluoxetine is a selective SERT inhibitor as can be seen in Table 2. It is highly selective for SERT compared to NET and DAT. Visual analysis of docked Fluoxetine-SERT structure once again highlights unique binding by SERT to its inhibitors compared to DAT and NET (**Figure 15**).

As discussed earlier, two factors make SERT different. First is unavailability of Asp-400 of Region-4 due to single amino acid insertion after Asp (Ala-401) and second repulsive interaction by Lys-490 of Region-5. Due to these differences, positive secondary amine of Fluoxetine cannot bind in a supposed position, instead it is inserted in loop of Region-4, where it is in a position to form hydrogen bond with Ser-404. Interestingly amine of Fluoxetine is also in a position to bond with Asp-400 as its side-chain along with side-chain of Ser-404 forms roof of Region-4 loop. Ser-404 is unique to SERT, highlighting its importance (Replaced by Gly in NET and DAT).

Trifluoromethyl group of Fluoxetine is inserted further down the TM channel and interestingly this strongly electronegative group is in contact with charged/polar residues of SERT i.e. Asp-193 and Gln-194 of Region-2. Again this polar nature of Region-2 is
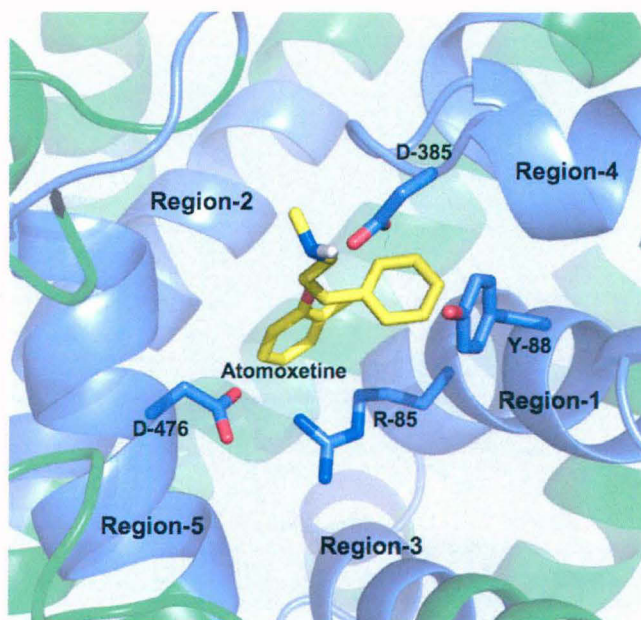
Figure 14 Binding pose of Atomoxetine in DAT inhibitor binding site as obtained by docking. Color Code: DAT-Green, Inhibitor Binding Regions of DAT-Purple, Atomoxetine-Yellow.
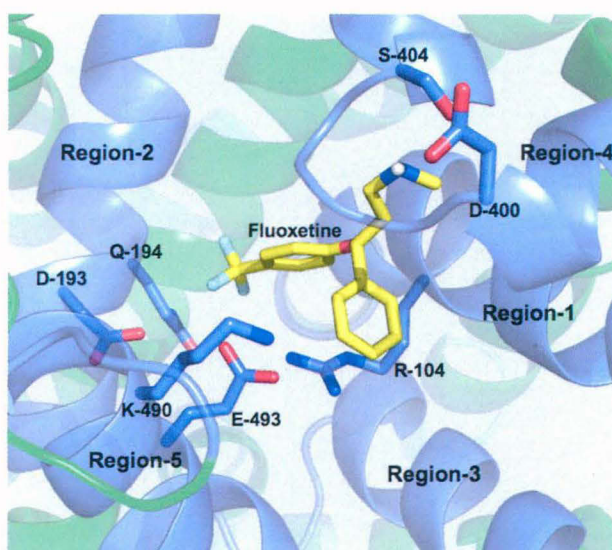

Figure 15 Binding pose of Fluoxetine in SERT inhibitor binding site as obtained by docking. Color Code: SERT-Green, Inhibitor Binding Regions of SERT-Purple, Fluoxetine-Yellow.

found only in SERT as six amino acid deletion further down the TM Helix of Region-2 compared to DAT and NET means that SERT has charged residues in Region-2 of inhibitor binding site. These residues perhaps form the halogen binding-site (HBS) and give selective SERT inhibitors their unique property. Zhou Z et al pointed out presence of halogen binding site in transporters but missed the correct residues as their study is based on SSRI-LeuT structures and thus failed to take into account deletion in SERT which ultimately defines its Region-2 and makes it unique in comparison to DAT and NET, which lack this deletion.

2.13 Fluoxetine Binding to NET

As mentioned earlier, Fluoxetine is a selective SERT inhibitor and Atomoxetine is a selective NET inhibitor. Interestingly, both molecules are very similar except one has – CH3 at ortho position of a phenyl ring that is attached to secondary propanamine through oxygen i.e. Atomoxetine and other has –CF3 at para position of this phenyl i.e. Fluoxetine (Figure 7). As –CH3 is a small non-polar group, it can be said it's the addition of –CF3 at the para position of oxy-phenyl, which converts NET inhibitor into SERT inhibitor.

As expected, secondary amine of Fluoxetine forms salt-bridge with E-382 Region-4 of NET (**Figure 16**). Trifluoromethyl group of Fluoxetine is extended in TM channel just as the case was with SERT, but contrary to SERT, no polar amino acid is present in NET to interact with –CF3 in what has been termed halogen binding site. This factor most likely discriminates Fluoxetine as a selective SERT inhibitor. Instead of Asp-193 and Gln-194 of SERT, NET has aromatic W-158 pointing inside TM channel.

2.14 Fluoxetine Binding to DAT

As can be seen from **Figure 17**, Fluoxetine binding to DAT is very similar to that of its binding to NET. Only difference is DAT having D-385 in Region-4 instead of Glu-382 of NET.

CONCLUSION

Selective inhibitors of DAT, NET and SERT are docked into structural models of these transporters generated based on LeuT template in an attempt to identify differences in inhibitor binding site which selective inhibitors manage to exploit. These studies showed how minor differences in five structurally distinct regions of DAT, NET and
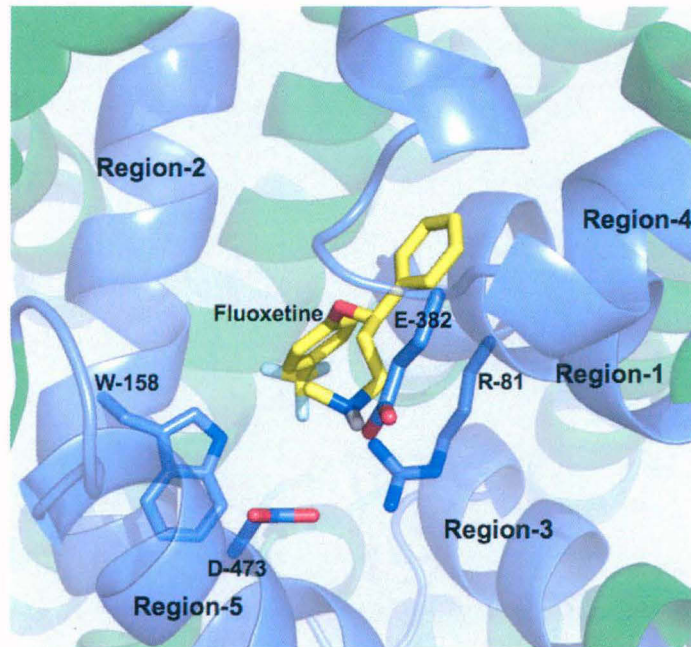
Figure 16 Binding pose of Fluoxetine in NET inhibitor binding site as obtained by docking. Color Code: NET-Green, Inhibitor Binding Regions of NET-Purple, Fluoxetine-Yellow.
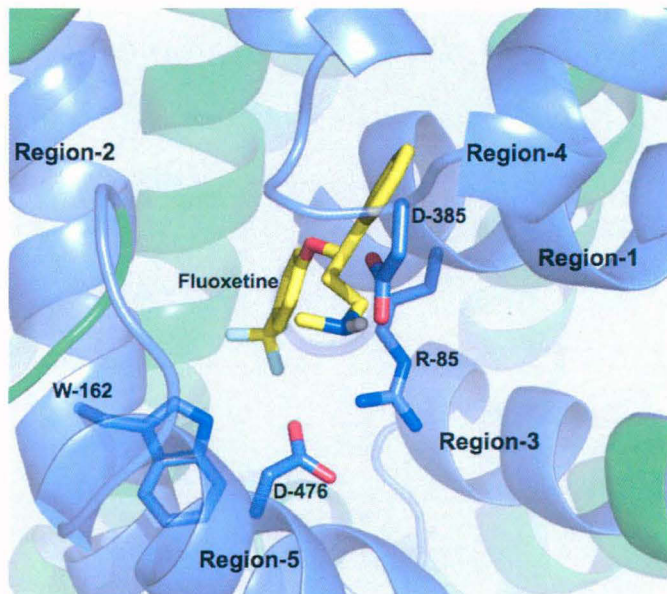

Figure 17 Binding pose of Fluoxetine in DAT inhibitor binding site as obtained by docking. Color Code: DAT-Green, Inhibitor Binding Regions of DAT-Purple, Fluoxetine-Yellow.

SERT, which make up inhibitor binding site, are perceived by respective selective inhibitors Bupropion, Atomoxetine and Fluoxetine. Inferences made here based on docking studies will need further verification from biochemical studies and have potential to be of help in developing new selective inhibitors of human neurotransmitter transporters.

Bymaster, F. P., Katner, J. S., Nelson, D. L., Hemrick-Luecke, S. K., Threlkeld, P. G., *et al.* (2002). Atomoxetine increases extracellular levels of norepinephrine and dopamine in prefrontal cortex of rat: a potential mechanism for efficacy in attention deficit/hyperactivity disorder. *Neuropsychopharmacology, 27*(5), 699-711.

Indarte, M., Madura, J. D., & Surratt, C. K. (2008). Dopamine transporter comparative molecular modeling and binding site prediction using the LeuT(Aa) leucine transporter as a template. *Proteins, 70*(3), 1033-1046.

Ravna, A. W., Jaronczyk, M., & Sylte, I. (2006). A homology model of SERT based on the LeuT(Aa) template. *Bioorg Med Chem Lett, 16*(21), 5594-5597.

Serafini, G., Pompili, M., Fusar-Poli, P., Porfiri, G., Giordano, G., *et al.* Bupropion and panic disorder: case report and review of the literature. *J Neuropsychiatry Clin Neurosci, 23*(2), E47-50.

Singh, S. K. (2008). LeuT: a prokaryotic stepping stone on the way to a eukaryotic neurotransmitter transporter structure. *Channels (Austin), 2*(5), 380-389.

Singh, S. K., Yamashita, A., & Gouaux, E. (2007). Antidepressant binding site in a bacterial homologue of neurotransmitter transporters. *Nature, 448*(7156), 952-956.

Wang, C. I., & Lewis, R. J. Emerging structure-function relationships defining monoamine NSS transporter substrate and ligand affinity. *Biochem Pharmacol, 79*(8), 1083-1091.

Westenberg, H. G. (1999). Pharmacology of antidepressants: selectivity or multiplicity? *J Clin Psychiatry, 60 Suppl 17*, 4-8; discussion 46-48.

Yamashita, A., Singh, S. K., Kawate, T., Jin, Y., & Gouaux, E. (2005). Crystal structure of a bacterial homologue of Na+/Cl--dependent neurotransmitter transporters. *Nature, 437*(7056), 215-223.

Zhou, Z., Zhen, J., Karpowich, N. K., Goetz, R. M., Law, C. J., *et al.* (2007). LeuT-desipramine structure reveals how antidepressants block neurotransmitter reuptake. *Science, 317*(5843), 1390-1393.

Zhou, Z., Zhen, J., Karpowich, N. K., Law, C. J., Reith, M. E., *et al.* (2009). Antidepressant specificity of serotonin transporter suggested by three LeuT-SSRI structures. *Nat Struct Mol Biol, 16*(6), 652-657.