

***IN SILICO* STRUCTURE-FUNCTION,
SPECIFICITY AND STABILITY STUDIES
OF N-TERMINAL NUCLEOPHILE
HYDROLASE ENZYMES**

THESIS SUBMITTED TO

SAVITRIBAI PHULE PUNE UNIVERSITY

FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY IN

BIOTECHNOLOGY

BY

PRIYABRATA PANIGRAHI

RESEARCH GUIDE
Dr. C.G. SURESH

DIVISION OF BIOCHEMICAL SCIENCES
CSIR-NATIONAL CHEMICAL LABORATORY
DR HOMI BHABHA ROAD, PUNE 411008
MAHARASHTRA, INDIA

JUNE 2015



सीएसआयआर-राष्ट्रीय रासायनिक प्रयोगशाला

(वैज्ञानिक तथा औद्योगिक अनुसंधान परिषद)

डॉ. होमी भाभा मार्ग, पुणे - 411 008. भारत

CSIR-NATIONAL CHEMICAL LABORATORY

(Council of Scientific & Industrial Research)

Dr. Homi Bhabha Road, Pune - 411008. India



CERTIFICATE

This is to certify that the work incorporated in the thesis "*In silico* structure-function, specificity and stability studies of N-terminal nucleophile hydrolase enzymes" submitted by Mr. Priyabrata Panigrahi was carried out by the candidate under my supervision/guidance. Such material as has been obtained from other sources has been duly acknowledged in the thesis.

Research Supervisor
Dr. C. G. Suresh
Chief Scientist
Division of Biochemical Sciences
CSIR-National Chemical Laboratory

Date: June 2015



Communications Channels
NCL Level DID : 2590
NCL Board No. : +91-20-25902000
Four PRI Lines : +91-20-25902000

FAX

Director's Office : +91-20-25902601
COA's Office : +91-20-25902660
SPO's Office : +91 20 25902664

WEBSITE

www.ncl-india.org

DECLARATION BY THE CANDIDATE

I hereby declare that the thesis entitled “*In silico* structure-function, specificity and stability studies of N-terminal nucleophile hydrolase enzymes” submitted by me for the degree of Doctor of Philosophy is the record of work carried out by me during the period from July 2011 to June 2015 under the guidance of **Dr. C.G. Suresh, Chief Scientist** and has not formed the basis for the award of any degree, diploma, associateship, fellowship, titles in this or any other University or other institution of higher learning.

I further declare that the material obtained from other sources has been duly acknowledged in the thesis.

Priyabrata Panigrahi

Division of Biochemical Sciences

CSIR-National Chemical Laboratory

Pune - 411 008

June 2015

Acknowledgement

This thesis is the end of my journey towards obtaining Ph.D. It would not have been possible to accomplish this without the help, support and encouragement from many people including my well wishers, my friends and colleagues. I would like to express my thanks to everyone who have made this thesis possible and gave me an unforgettable experience. First and foremost, I would like to thank beloved **BapDada** and *Brahmakumaris World Spiritual University* for giving me inner strength and willpower to pursue this journey and also their staunch support during tough times.

I am extremely thankful to **Dr. C. G. Suresh**, my research guide, for giving me complete freedom to execute my work. This work would not have been possible without his guidance, support and encouragement. I learnt not only the way he approaches to solve a problem but also influenced by his disciplined life and time management skills. I appreciate all his contributions of time, ideas to make my Ph.D. experience productive and stimulating. I would also like to thank **Dr. Sureshkumar Ramasamy**, for his critical suggestions in the work and constant encouragement in a very friendly manner. I am especially grateful to him for the cloning of *PdPGA* gene which was one of the turning points of my thesis where I step forward towards wetlab. Being from a pure computational background, entering to experimentation in the 3rd year of Ph.D. was one of the most-challenging periods for me. However due to the constant supports of my lab members, specifically **Deepak** and **Ruchira**, to whom I lovingly call ATH (Any Time Help) friends. Deepak was there throughout the wet lab experiments, without even bothering day or night, as well as his own work. Also the special thanks go to **Ruby** and **Tulika**, for their help

in CD and Fluorescence experiments. I am especially grateful to Tejashree who helped me in maintaining the *PdPGA* clone.

Coming to computational work, I enjoyed a lot working with **Manas** and **Ranu**. I thank Manas for his useful contributions in shaping all manuscripts and this thesis. As a part of this thesis, I had developed iRDP web server. With his useful discussions and brain-storming sessions, I could see constant progress in iRDP web server development. **Avinash Ghanate** has an important contribution towards the development of the front end of the iRDP web server. I would like to thank Ranu for giving me moral support, encouragement and valuable discussion.

I am grateful to my other lab members **Urvashi, Nishant, Payal, Manu, Ameya, Deepanjan, Shridhar, Teju, Yashpal, Vijay and Aditi** for providing a great working environment in the laboratory. Special thanks are due for my dearest friends **Pandurang, Supriya, Rekha and Surabhi** who gave a lot of moral support during tough times. I would like to acknowledge **Dr. Asmita Prabhune, Dr. Sushama Gaikwad and Dr. Vaijayanti Kumar**, for allowing me to use various instruments in their lab. I'm grateful to **Dr. Vidya Gupta**, Head, Division of Biochemical Sciences, CSIR-NCL, Pune and **Dr. J.K. Pal**, University of Pune, for being members of my evaluation committee. Their inputs have been invaluable to the progress of this work. I would like to thank Director, CSIR-National Chemical Laboratory (CSIR-NCL, Pune) for giving me the opportunity to work in this great institution. I would like to thank Council of Scientific & Industrial Research (CSIR) for the fellowship and CoESC for funding. Finally I would like to thank my parents and sister for their support and faith in me.

Priyabrata Panigrahi

Table of Contents

ABBREVIATIONS.....	i
LIST OF TABLES.....	iv
LIST OF FIGURES.....	vi
ABSTRACT.....	1

Chapter 1

1.1 Enzyme-substrate interactions	6
1.2 Substrate specificity of enzymes.....	7
1.3 Enzyme classification	7
1.4 Enzyme kinetics.....	8
1.5 Factors effecting enzyme activity	9
1.6 Sequence, structure and substrate specificity relationship in enzymes.....	11
1.7 Sequence, structure and stability relationship in enzymes.....	13
1.8 Ntn-hydrolase enzyme superfamily	18
1.8.1 Self-processing cysteine-dependent Ntn-hydrolase enzyme superfamily (NtCn-hydrolases)	26
1.8.1.1 Family C59.....	26
1.8.1.2 Family C44.....	28
1.8.1.3 Family C69.....	29
1.8.1.4 Family C89.....	29
1.8.1.5 Family C95.....	30
1.8.1.6 Family C45.....	30
1.8.2 Self-processing serine-dependent Ntn-hydrolase enzyme superfamily (NtSn-hydrolases)	31
1.8.3 Self-processing threonine-dependent Ntn-hydrolase enzyme superfamily (NtTn-hydrolases)	35
1.8.3.1 Family T1.....	35
1.8.3.2 Family T2.....	36
1.8.3.3 Family T3.....	37
1.8.3.4 Family T5.....	37
1.9 Scope of the thesis	38
1.9.1 Study of sequence-structure & specificity relationship in CGH family.....	38
1.9.1.1 Bile Salt Hydrolases, their physiological role and clinical importance	39
1.9.1.2 Penicillin V acylases, their physiological role and pharmaceutical importance	43

1.9.1.3 BSH and PVA: sequence and structural homology and difficulty in their distinction.....	47
1.9.2 Study of sequence-structure & stability relationship through development of iRDP web server..	48
1.9.3 Study of sequence-structure & stability relationship in PGA family.....	49
1.10 Tools and techniques used in this study.....	53

Chapter 2

2.1 Introduction.....	55
2.1.1 Difficulty in annotation of CGH enzymes as BSH/PVA and the need for an improved annotation method.	56
2.2 Materials and Methods.....	58
2.2.1 Retrieval of CGH family members and phylogenetic analysis.....	58
2.2.2. Structure retrieval and preprocessing.....	59
2.2.3 Prediction of substrate binding modes using docking analysis.....	59
2.2.4 Molecular dynamics Simulations.....	59
2.2.5 Estimation of Binding site similarity (BSS) scores for all CGH sequences	60
2.3 Results and Discussions.....	61
2.3.1 Dataset generation and phylogenetic analysis of the BSH/PVA sequences	61
2.3.2 Analysis of the substrate specificity and binding site properties of CGH enzymes	63
2.3.3 Mode of substrate binding among CGH enzymes	66
2.3.3.1 Modes of GCA binding.....	66
2.3.3.2 Polar complementarity: probable basis for GCA specificity	68
2.3.3.3 Modes of penicillin V binding among CGH enzymes	70
2.3.3.4 Aromatic interactions in the active site might influence penicillin V binding.....	73
2.3.4 Substrate specificity annotation of family members.....	74
2.3.5 Physiological role of BSH/PVA enzymes	75
2.3.6. Evolutionary basis for the divergence of CGH family members into two clusters.....	78
2.4. Summary.....	91

Chapter 3

3.1 Introduction.....	94
3.1.1 Molecular determinants of protein thermostability.....	94
3.1.2 Development of iCAPS (<i>in silico</i> Comparative Analysis of Protein Structures) module.....	96
3.1.3 Identification of potential sites for structural stabilization.....	97

3.1.4 Development of iStability (<i>in silico</i> Analysis of Stability Change in Protein Structures) module	98
3.1.5 Evaluating potential thermostabilization sites using molecular interactions	99
3.1.6 Development of iMutants (<i>in silico</i> Comparative Analysis of Interactions in Protein Mutants) module.	99
3.2 Materials and Methods	101
3.2.1 <i>In silico</i> Comparative Analysis of Protein Structures (iCAPS)	103
3.2.2 <i>In silico</i> Analysis of Stability Change in Protein Structures (iStability)	105
3.2.3 <i>In silico</i> Comparative Analysis of Interactions in Protein Mutants (iMutants)	106
3.3 Results and Discussion	107
3.3.1 Analysis of structure stabilization mechanisms using iCAPS	107
3.3.2 Demonstration of applicability of iCAPS module.	113
3.3.3 Identification of potential stabilizing mutations in a protein using iStability module	118
3.3.4 Evaluating mutations through interaction framework and evolutionary residue conservation at mutation sites using iMutants	124
3.3.5 Demonstration of the utility of iMutants module	125
3.3.6 iATMs (<i>in silico</i> Analysis of Thermally stable Mutants): An information resource.	129
3.4 Summary	130

Chapter 4

4.1 Introduction	132
4.2 Materials and Methods	134
4.2.1 Computational screening strategy for obtain of ptPGAs (putative thermostable PGAs)	134
4.2.2 Removal of signal and spacer peptide	134
4.2.3 Homology modeling	134
4.2.4 Molecular dynamics simulations	135
4.2.5 Estimation of non-bonded interactions	135
4.3 Results and Discussions	137
4.3.1. Putative Thermostable PGAs (ptPGAs): Screening	138
4.3.2 Removal of signal and spacer peptide	138
4.3.3 Homology modeling and model validation	140
4.3.4 Sequence-based consensus approach for thermostability analysis	142
4.3.5 Structure-based approach of thermostability analysis	147
4.3.5.1 Stabilization by Disulfide Bridges	149

4.3.5.2 Preference of Arg over Lys: A mechanism for thermostabilization	149
4.3.5.3 Presence of higher ion-pair networks: A stabilizing mechanism	153
4.3.5.4 Loop stabilization by proline residues	154
4.3.5.5 Decreased content of thermolabile residues	157
4.3.5.6 Contribution from Hydrogen bonds	157
4.3.5.7 Contribution from interactions involving aromatic residues.....	159
4.4 Summary	160

Chapter 5

5.1 Introduction.....	162
5.1.1 <i>Paracoccus denitrificans</i>	162
5.2 Materials and Methods.....	164
5.2.1 Cloning of PdPGA gene	164
5.2.2 Preparation of competent cell	164
5.2.3 Transformation of plasmid	165
5.2.4 Cryopreservation of Bacterial culture	165
5.2.5 Expression of PdPGA	165
5.2.6 Cell lysis.....	166
5.2.7 Purification of PdPGA by Immobilized Metal Ion Affinity chromatography (IMAC).....	166
5.2.8 Removal of imidazole by desalting.....	167
5.2.9 Purification by size exclusion chromatography	167
5.2.10 SDS - Polyacrylamide Gel Electrophoresis (SDS-PAGE).....	167
5.2.11 Western blot.....	168
5.2.12 Protein estimation	169
5.2.13 Penicillin G Acylase activity.....	169
5.2.14 Thermal unfolding measured using steady state fluorescence	170
5.2.15 Estimation of thermal unfolding using CD spectroscopy	171
5.2.16 Hydrophobic dye binding.....	172
5.3 Results and Discussion	173
5.3.1 Confirmation of PdPGA gene clone	173
5.3.2 Purification of PdPGA.....	176
5.3.3 Enzyme kinetics	178
5.3.4 Alkaline stability of PdPGA	178

5.3.5 Temperature stability profile of <i>Pd</i> PGA.....	179
5.3.6 Probing structural changes using fluorescence.....	181
5.3.7 Circular Dichroism study.....	181
5.4 Summary.....	183

Chapter 6

Chapter 6.....	186
Bibliography.....	190
List of Publications.....	

ABBREVIATIONS

Abbreviation	Long form
Ntn	N-terminal nucleophile
NtCn	N-terminal cysteine nucleophile
NtSn	N-terminal serine nucleophile
NtTn	N-terminal threonine nucleophile
PDB	Protein Data Bank
RMSD	Root Mean Square Deviation
bp	Base pair
aa	Amino acids
HMM	Hidden Markov Model
PSSM	Position Specific Scoring Matrices
Chapter 1	
GAT	Glutamine amidotransferase domain
GPATase	Glutamine phosphoribosylpyrophosphate amidotransferase
GFAT	Glucosamine-fructose-6-phosphate aminotransferase
AS	Asparagine synthetase
AC	Acid ceramidase
NAAA	N-acylethanolamine hydrolyzing acid amidase
GGT	Gamma-glutamyltransferase
Chapter 2	
CGH	Cholyglycine hydrolase family
BSH	Bile Salt Hydrolase
PVA	Penicillin V Acylase
<i>B</i> BSH	<i>Bifidobacterium longum</i> BSH
<i>Cp</i> BSH	<i>Clostridium perfringens</i> BSH
<i>Bt</i> BSH	<i>Bacteroides thetaiotaomicron</i> BSH
<i>Bsp</i> PVA	<i>Bacillus sphaericus</i> PVA
<i>Bsu</i> PVA	<i>Bacillus subtilis</i> PVA
<i>Pa</i> PVA	<i>Pectobacterium atrosepticum</i> PVA
BSS	Binding site similarity

PenV	Penicillin V
BS	Bile Salts
GCA	Glycocholic acid
PAA	Phenoxy acetic acid
6-APA	6-aminopenicillanic acid
Indel	Insertion-deletion

Chapter 3

iRDP	<i>in silico</i> Rational Design of Proteins web server
iCAPS	<i>in silico</i> Comparative Analysis of Protein Structures module
iStability	<i>in silico</i> Analysis of Stability Change in Protein Structures module
iMutants	<i>in silico</i> Comparative Analysis of Interactions in Protein Mutants module
iATMs	<i>in silico</i> Analysis of Thermally stable Mutants information resource
ASA	Accessible surface area
SS	Secondary structure elements
IP	Ion-pairs
AAI	Aromatic-aromatic interactions
ASI	Aromatic-sulphur interactions
CPI	Cation- π interactions
DB	Disulfide bridges
HB	Hydrogen bonds
HP	Hydrophobic interactions
TS	Thermophilic proteins
MS	Mesophilic proteins
CScore	Evolutionary conservation score

Chapter 4

PGA	Penicillin G acylase
<i>Ec</i> PGA	<i>Escherichia coli</i> PGA
<i>Af</i> PGA	<i>Alcaligenes faecalis</i> PGA
<i>Ax</i> PGA	<i>Achromobacter xylosoxidans</i> PGA
<i>Sw</i> PGA	<i>Sphingomonas wittichii</i> PGA
<i>Pd</i> PGA	<i>Paracoccus denitrificans</i> PGA

<i>Ao</i> PGA	<i>Acinetobacter oleivorans</i> PGA
<i>Kc</i> PGA	<i>Kluyvera cryocrescens</i> PGA
ptPGAs	putative thermostable PGAs (<i>Sw</i> PGA, <i>Pd</i> PGA and <i>Ao</i> PGA)
IPs	Total number of ion-pairs
IPnets	Percentage of ion-pairs that are involved in network formation
Bt2P	Proline residues at 2 nd position of β -turns
CNHB	Charged-neutral hydrogen bond

Chapter 5

OD	Optical density
LB	Luria-Bertani media
Ni-NTA	Ni ²⁺ -nickel-nitrilotriacetic acid
SDS-PAGE	Sodium dodecyl sulfate Polyacrylamide Gel Electrophoresis
CD	Circular dichroism
MRE	Mean residue ellipticity
ANS	8-Anilino-1-naphthalene sulfonic acid
PenG	Penicillin G
DTT	Dithiothreitol

LIST OF TABLES

Table	Description	Page
Chapter 1		
1.1	List of Ntn-hydrolase enzymes, their taxonomic distribution, domain architecture and function.	24
1.2	Sequence and structural similarity of BSH and PVA enzymes.	47
Chapter 2		
2.1	List of experimentally characterized BSH and PVA enzymes considered in the analysis.	57
2.2	SiteMap quantitative estimation of binding site properties of CGH enzymes.	65
2.3	Summary of free energy of binding (GlideScore) of all predicted protein-ligand complex structures.	68
2.4	Quantitative estimation of polar complementarities for the three hydroxyl groups of the GCA molecule.	70
2.5	Quantitative estimation of interface area between individual subunits of <i>B</i> /BSH and <i>Bt</i> BSH in their quaternary structures.	79
2.6	Functional annotation of CGH family members into BSH or PVA based on binding site similarity (BSS) based scoring system.	81
2.7	BSS based functional annotations for those members for which experimental evidence of their BSH or PVA activity is known.	88
2.8	Binding site similarity (BSS) based annotation of the Gram-positive members that were previously annotated by Lambert <i>et al.</i> , 2008.	89
Chapter 3		
3.1	List of currently available protein structural analysis tools, their usefulness and need for further improvement.	95
3.2	List of available stability prediction tools, their usefulness, limitations and the need for further improvements.	100
3.3	List of tools used by iRDP web server for estimation of few structural parameters.	101
3.4	List of various quantitative parameters generated by the iCAPS module.	108

3.5	Comparative analysis of various thermostability factors among 16 thermophilic-mesophilic pairs of protein.	114
3.6	Details of proteins considered in iCAPS validation.	115
3.7	Validation of iStability using the four protein engineering strategies.	121
3.8	The iMutants analysis on 51 mutations in Arc Repressor protein of Bacteriophage P22.	127
Chapter 4		
4.1	List of Cys residue containing PGAs from S45.001 family of MEROPS database.	137
4.2	The residue re-numbering after the removal of signal and spacer peptide among the PGA enzymes under study.	139
4.3	Evaluation statistics for molecular models of <i>Ax</i> PGA, <i>Sw</i> PGA, <i>Pd</i> PGA and <i>Ao</i> PGA sequences by various model validation tools.	139
4.4	List of 11 experimentally characterized mutations known to enhance thermostability in <i>Ec</i> PGA.	142
4.5	List of 24 sites identified as thermostabilization sites by sequence-based consensus approach.	144
4.6	The average and standard deviations of the RMSD (Root Mean Square Deviations) of C α atoms of PGA enzyme conformations during the production phase of molecular dynamics simulations (330, 400 and 500K).	147
4.7	Amino acid composition of six PGA enzymes under study.	150
4.8	List of thermolabile Asn-Gly bond positions in <i>Ec</i> PGA and the corresponding amino acids in other PGAs.	156
Chapter 5		
5.1	Percentages of various secondary structure elements of <i>Pd</i> PGA estimated at various temperatures using CDPro software.	182

LIST OF FIGURES

Figure	Description	Page
Chapter 1		
1.1	Folding of linear polypeptide chain.	5
1.2	Free energy diagram depicting progress of an enzyme catalyzed reaction.	9
1.3	Effect of temperature, pH, and enzyme, substrate concentration on enzyme activity.	10
1.4	Ntn-hydrolase structural fold.	19
1.5	Reaction mechanism of Ntn-hydrolase enzymes.	20
1.6	Auto-catalytic cleavage events amongst enzymes of Ntn-hydrolase superfamily.	22
1.7	Distribution of Ntn-hydrolase enzymes across different taxonomic group.	23
1.8	Cartoon representation of enzymes of NtCn-hydrolase enzymes.	27
1.9	Cartoon representation of enzymes of NtSn-hydrolase enzymes.	33
1.10	Cartoon representation of enzymes of NtTn-hydrolase enzymes.	34
1.11	Structures of bile acids and glycine or taurine conjugated bile salts.	40
1.12	Cholesterol homeostasis inside body.	41
1.13	Structures of various β -lactams.	44
1.14	Penicillin catalyzed reaction leads to 6-APA production.	45
Chapter 2		
2.1	Structures of Glycocholic acid and Penicillin V.	55
2.2	Workflow of BSS-based annotation of CGH family members.	58
2.3	The methodology of the Binding site similarity (BSS) based scoring and annotation system.	60
2.4	Dendrogram prepared based on the phylogenetic analysis of the sequences of CGH family.	62
2.5	Three-dimensional structure of <i>Cp</i> BSH, geometrical rearrangement of its catalytic residues and structural superposition of the four substrate binding site loops.	64
2.6	Mode of GCA binding in <i>B</i> /BSH and penicillin V binding in <i>Bsu</i> PVA.	66

2.7	Modes of GCA binding in <i>B</i> /BSH, <i>Cp</i> BSH, <i>Bt</i> BSH, <i>Bsp</i> PVA and <i>Bsu</i> PVA.	67
2.8	Radial distribution of receptor polar atoms around three hydroxyl groups of GCA in <i>B</i> /BSH, <i>Cp</i> BSH, <i>Bt</i> BSH, <i>Bsp</i> PVA and <i>Bsu</i> PVA.	69
2.9	Modes of penicillin V binding in <i>B</i> /BSH, <i>Cp</i> BSH, <i>Bt</i> BSH, <i>Bsp</i> PVA and <i>Bsu</i> PVA.	71
2.10	Geometrical arrangements of the aromatic planes of the residues in the vicinity of phenyl ring of substrate penicillin V among all five CGH enzymes during molecular dynamics simulation.	72
2.11	Box plot depicting distribution of angle between the phenyl ring planes of substrate penicillin V and aromatic residues in its vicinity.	72
2.12	Binding Site Similarity (BSS) based annotation of CGH family members into BSH or PVA enzymes.	74
2.13	Illustration of tetramer assembly motif in <i>B</i> /BSH and <i>Bt</i> BSH enzymes.	77
Chapter 3		
3.1	Workflow of rational protein engineering experiments.	93
3.2	List of sequence- and structure-based features that can be analyzed using iCAPS module.	96
3.3	Identification of potential sites of thermostabilization in the input protein structure through iStability module.	98
3.4	Analysis of change in local interactions due to mutation near the mutation site through iMutants module.	100
3.5	Workflow of the working modules implemented in the iRDP web server.	102
3.6	The user interface of iCAPS module.	103
3.7	The user interface of iStability module.	105
3.8	The user interface of iMutants module.	107
3.9	Sample output of iStability module.	118
3.10	Sample output of iMutants module.	125
Chapter 4		
4.1	Workflow towards identification of potential thermostable PGA enzymes.	136
4.2	Sequence alignment showing 34 residue regions involved in disulfide bond	138

	formation in PGA enzymes.	
4.3	Results of model validation programs for <i>Pd</i> PGA homology model.	141
4.4	Location of 24 sites of thermostabilization on three-dimensional structure of <i>Ec</i> PGA.	145
4.5	Illustrates the residues corresponding to the disulfide bridge forming Cys residue positions of <i>Af</i> PGA among all PGA enzymes.	148
4.6	<i>Af</i> PGA structure is shown to highlight the N-terminal nucleophilic β Ser1 residue and disulfide bond.	148
4.7	Line plots depicting average percentages of ion-pairs and ion-pair networks of <i>Af</i> PGA, <i>Ax</i> PGA, <i>Sw</i> PGA, <i>Pd</i> PGA and <i>Ao</i> PGA relative to <i>Ec</i> PGA.	151
4.8	Time evolution of number of short-range ion-pairs amongst six PGAs at 330K, 400K and 500K molecular dynamics simulations.	152
4.9	Three-dimensional structures of <i>Ec</i> PGA and <i>Pd</i> PGA showing the ionic interactions between the acidic and basic residues.	153
4.10	Bar plot depicting the average percentages of Bt2P residues among all PGAs relative to <i>Ec</i> PGA.	155
4.11	The acid-base and β -aspartyl shift mechanism of protein deamidation.	156
4.12	The average percentage of total hydrogen bonds and charged-neutral hydrogen bonds among PGA enzymes relative to <i>Ec</i> PGA.	158
4.13	Line plots depicting various statistics of interactions involving aromatic residues among the six PGA enzymes during molecular dynamics simulations.	159
Chapter 5		
5.1	Illustrates <i>Paracoccus denitrificans</i> .	162
5.2	The vector map of pETKat vector containing the <i>Pd</i> PGA gene.	173
5.3	Illustrates approximately 900 bases of <i>Pd</i> PGA gene sequenced by T7 forward primer.	174
5.4	Illustrates approximately 900 bases of <i>Pd</i> PGA gene sequenced by T7 reverse primer.	175
5.5	Elution profile of Gel filtration chromatography.	177

5.6	12% SDS-PAGE of various fractions obtained during 3-step <i>Pd</i> PGA purification protocol and Western blot.	177
5.7	The Michaelis-Menten plot depicting the rate of <i>Pd</i> PGA enzyme activity at different substrate concentrations.	178
5.8	The optimum temperature, optimum pH, temperature stability and pH stability profile of <i>Pd</i> PGA.	179
5.9	Trp fluorescence spectra of <i>Pd</i> PGA subjected to various temperatures.	180
5.10	Far UV CD spectra of <i>Pd</i> PGA subjected to various heat treatments.	182

Abstract

N-terminal nucleophile hydrolases or **Ntn-hydrolases** form a superfamily of hydrolytic enzymes which are functionally amidases. These enzymes are found in a variety of organisms ranging from microbes to higher organisms such as mammals. The N-terminal amino acid residue of these enzymes acts as both nucleophile and base during enzymatic action. Based on the residues at the N-terminal, **Cys**, **Ser** or **Thr**, the superfamily can be classified into N-terminal cysteine nucleophile (NtCn-hydrolase), N-terminal serine nucleophile (NtSn-hydrolase) and N-terminal threonine nucleophile (NtTn-hydrolase), respectively. On the basis of the substrate specificity, members of these superfamilies can further be categorized into different families and their subfamilies.

All the enzymes of Ntn-hydrolase superfamily show a similar fold of their catalytic domain (Ntn-hydrolase fold) and the active site topology. The Ntn-hydrolase fold comprises of a four-layer sandwich of α helices and β sheets ($\alpha\beta\beta\alpha$ core) which is shared by all Ntn-hydrolase enzymes. These enzymes are also mechanistically related. However, owing to differences in size, shape and properties of the substrate binding sites, a wide variation of substrate specificity is generally observed amongst the Ntn-hydrolase enzymes.

Cholylglycine hydrolase (CGH) family, belonging to the NtCn-hydrolase superfamily, consists of enzymes of immense pharmaceutical importance such as *Bile Salt Hydrolases* (BSH) and *Penicillin V Acylases* (PVA) which have been shown to play a vital role in cholesterol metabolism and semi-synthesis of β -lactam antibiotics, respectively. Due to a significant degree of homology between these two enzymes, their annotation based on substrate specificity remains a challenging problem. Owing to the medical importance associated with the functions of these enzymes, a high resolution sequence based annotation is highly desirable. Since the function of an enzyme is determined by its overall structure which in turn depends on the sequence, we have studied the *sequence-structure substrate specificity relationship* in order to develop a method for the differentiation of these two types of enzymes in terms of their function. By incorporating phylogenetic, binding site and substrate specificity information, an improved method based on binding site similarity (BSS) was developed using which all CGH family members were accurately annotated as BSH/PVA enzymes. Through docking and molecular dynamics simulations the substrate binding modes among CGH enzymes were explored and the probable

basis for their variations in substrate specificity were analyzed. Evolution of family members with respect to the antibiotics selection pressure theory was studied and the physiological roles of enzymes were discussed. This work has been described in **Chapter 2** of the present thesis.

Penicillin G Acylase (PGA) family, belonging to NtSn-hydrolase superfamily, comprises of members which are widely used in industry for the manufacture of many semi-synthetic antibiotics which show higher efficiency compared to natural antibiotics. Since the rate of an enzymatic reaction is expected to increase with temperature, the role of PGA enzymes as biocatalysts tends to be more attractive if their stability at higher temperatures can be improved. Thermal stability of an enzyme depends on its three-dimensional structure as well as its sequence. If the molecular determinants contributing to the thermostability of an enzyme are studied carefully, they can not only be used to screen novel sources of thermostable enzymes but also to improve the stability of a lesser stable enzyme. With this in mind a set of computational tools were developed to explore various mechanisms adopted by nature for protein thermostabilization. This work has been developed in the form of **iRDP web server**, available at <http://irdp.ncl.res.in>. The server provides three separate modules namely **iCAPS**, **iStability** and **iMutants**. iCAPS focuses on the comparative analysis of large number of protein structures for factors contributing to their structural stability, iStability uniquely offers *in silico* implementation of known thermostabilization strategies in proteins for identification and stability prediction of potential stabilizing mutation sites. iMutants aims to evaluate any mutations based on change in local interaction framework and degree of residue conservation at the mutation sites. In addition to these three modules, iRDP introduces **iATMs** information resource which provides detailed information about the local structural and interaction changes that occur near the mutation sites for all known experimentally validated mutations listed in the ProTherm database. Thus iATMs provide a better understanding of correlation between experimental observations with the interaction rearrangements due to mutations, leading to better application of derived knowledge towards efficient protein engineering. The iRDP server was built on a Linux platform using R, Perl, HTML and PHP. The development and implementation of this server has been described in **Chapter 3**.

PGA enzymes have huge applications in antibiotics industry. With the objective of identifying novel sources of thermostable PGA enzymes, a computational approach based on sequence and structure analysis of several PGA family members was followed. Presence of disulfide bridges was given the highest priority. Various other factors such as high arginine to lysine ratio, less content of thermolabile amino acids, presence of proline in β -turns, more number of ion-pair and other non-bonded interactions were considered. These parameters were estimated using iRDP web server. We have also designed a modified sequence based consensus approach that considers stabilizing residue positions by site-specific comparison between mesostable and thermostable PGAs. Based on these approaches we have selected candidate PGAs with unknown stabilities. A most likely thermostable enzyme identified from the analysis was **PGA from *Paracoccus denitrificans* (PdPGA)**. This was cloned, expressed and checked for thermostability using biochemical and biophysical experiments. The computational approach of selection of PdPGA is described in **Chapter4** while its experimental characterization is described in **Chapter 5**.

Overall, the thesis is organized into 6 chapters, **Chapter1** introducing the Ntn-hydrolase enzyme superfamily and describes the objectives of the present work, followed by **Chapter2** describing the sequence based substrate specificity annotation of CGH family members. **Chapter3** describes iRDP web server, an integrated rational protein engineering platform while **Chapter4** describes the computational screening of PGA family members towards identification of putative thermostable PGA enzymes. **Chapter5** describes the purification and characterization of PdPGA enzyme. **Chapter6** summarizes some of the generalized conclusions derived from the work and highlights the importance and future directions.

Chapter 1

*An introduction to
characteristics of enzymes belonging to
Ntn-hydrolase superfamily*

All living cells perform numerous biochemical reactions for their functioning. A great majority of these reactions are non-spontaneous in nature and thus occur at a very slow rate. Catalysis is the process of accelerating a chemical reaction using substances which themselves do not undergo any permanent chemical modifications. Enzymes are such biological molecules that catalyze biochemical reactions inside living cells without themselves undergoing any chemical change.

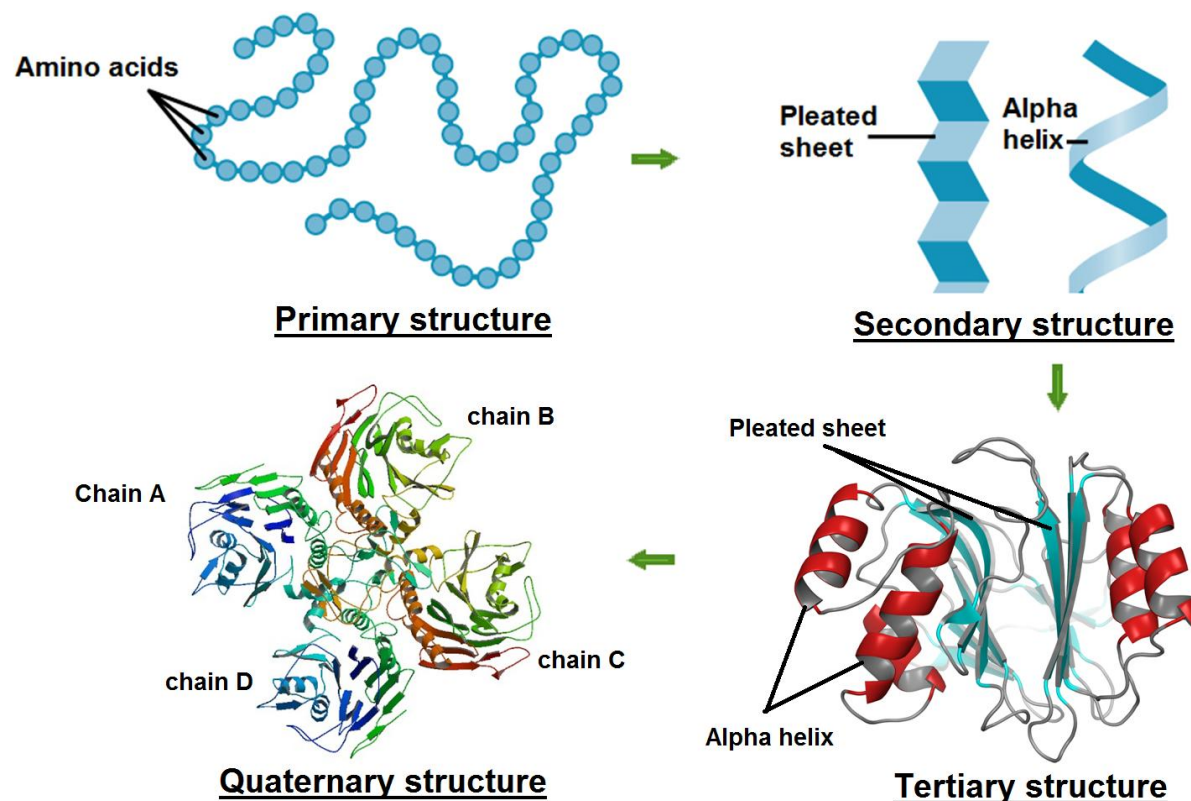


Figure 1.1: Illustrates the folding of linear polypeptide chain (primary structure) to secondary, tertiary and quaternary structure.

Enzymes are comparatively high molecular weight compounds, mainly proteins. They are made up of amino acids linked together by peptide bonds. Inside cells, there are 20 different kinds of natural amino acids whose serial order determines the *primary sequence* of an enzyme. The linear poly peptide chain of an enzyme folds either during or after the completion of translation to form regular periodic *secondary structural* elements such as helices, strands or turns (Fig. 1.1). The secondary structure folds further in three-dimension to form the *tertiary structure* which ultimately determines the enzyme function. Sometimes enzymes are made up of

multiple polypeptide chains which assemble in the form of *quaternary structure*. Many enzymes often require other non-protein *cofactors* for their functioning. The cofactors can either be coenzymes or prosthetic groups or metal-ion activators. Enzymes without any cofactors are termed as *apoenzymes* while enzymes in presence of cofactors are termed as *holoenzymes*. A *coenzyme* is an organic substance which is loosely attached to the enzyme via non-covalent interactions and thus can be dialyzable and separable from protein. However a *prosthetic group* is an organic substance covalently attached to the enzyme and therefore difficult to separate. *Metal activators* usually include ions like K^+ , Cu^{2+} , Fe^{2+} , Fe^{3+} , Zn^{2+} , Co^{2+} , Mn^{2+} , Mg^{2+} , Ca^{2+} , and Mo^{3+} . In 1947, James B. Sumner of Cornell University was awarded Nobel Prize for his breakthrough work of isolation and crystallization of enzyme *urease* from jack bean. He shared the Nobel Prize with John H. Northrop and Wendell M. Stanley for their discovery of a complex protocol of pepsin isolation. From then onwards many enzymes as well as functional proteins have been discovered and characterized using biochemical and biophysical techniques.

1.1 Enzyme-substrate interactions

The region of enzyme structure where the substrate molecule binds is defined as the *substrate binding site* while the region where the actual catalysis occurs is defined as *active site* or *catalytic site*. The binding site provides right shape and orientation of functional groups that facilitates binding of substrate molecules. Two popular hypotheses for enzyme-substrate interaction are proposed namely the *Lock & Key hypothesis* of Emil Fischer and *Induced fit hypothesis* of Daniel Koshland. According to the *Lock & Key hypothesis*, substrates perfectly fit in the active site just as a key fits a lock. Like a lock which can only be opened by its perfectly complementary key, a strict shape and interaction complementarity has to be followed in order for a substrate to bind to enzyme's active site. So, in this hypothesis both enzyme and substrate are assumed to be rigid molecules. On the contrary, the *Induced fit hypothesis* assumes flexible enzyme and flexible substrate model where substrate induces conformational changes in the enzyme's active site upon binding. The conformational flexibility of enzymes is mainly due to side chain and loop movements. However, sometimes substrate binding induces domain motion (Gutteridge & Thornton, 2004; Herschlag, 1988). The degree of complementarity between substrate and the binding site influences the binding affinity. An enzyme can bind to different substrates with different binding specificity and affinity. The active site contains key residues

important for catalysis known as *catalytic residues*. Substrate molecules interact with these residues via non-covalent interactions such as electrostatic, hydrogen bonding, van der Waals and hydrophobic interactions. The step by step sequence of reaction by which substrate is converted to product is termed *reaction mechanism*.

1.2 Substrate specificity of enzymes

One of the unique and important properties of an enzyme is its specificity towards the substrate molecules. The enzyme can have four different specificities (Bennett & Frieden, 1969) such as

1. **Absolute specificity** in which an enzyme is specific to only one substrate molecule.
2. **Group specificity** where an enzyme is specific to substrates that share specific functional groups such as phosphate, methyl and amino groups.
3. **Linkage specificity** in which an enzyme is specific to substrates having particular type of chemical bonds.
4. **Stereo specificity** where an enzyme is specific to particular type of stereo or optical isomer.

1.3 Enzyme classification

The first enzyme commission of International Union of Biochemistry (I.U.B.) in 1961, created a system for enzyme classification in which enzymes could be assigned unique code numbers based on the type of chemical reactions they catalyze. The code number, so called the *Enzyme Commission number* (EC number) has four numbers separated by dots.

1. First number correspond to one of the six main enzyme classes namely:
 - a. **Oxidoreductases**: These enzymes catalyze oxidation-reduction reactions. The systematic names assigned to these enzymes are *donor:acceptor oxidoreductase* while common names can be *dehydrogenase, reductase, catalase* and *oxidase*.
 - b. **Transferases**: These enzymes catalyze reactions involving transfer of functional groups such as methyl, acetyl, phosphate, glycosyl groups etc., from donor to acceptor molecules. The systematic names assigned to these enzymes are *donor:acceptor grouptransferase* while common names can be *acceptor*

- grouptransferase* or *donor grouptransferase* (e.g. Methyl transferase, Hydroxymethyl transferase, Acetyl transferase, Protein kinase and Formyl transferase).
- c. **Hydrolases:** These enzymes catalyze the hydrolytic cleavage of chemical bonds like C-O, C-N, C-C and other bonds. The systematic name always includes *hydrolases* while common name can be formed by *substrate name*, with suffix *-ase* (e.g. Protease, Exonuclease, Endonuclease and Phosphatase).
 - d. **Lyases:** These enzymes catalyze elimination reaction involving cleavage of C-C, C-O, C-N, and other bonds, leaving double bonds or rings. Conversely they also catalyze addition of functional groups to double bonds. The systematic name includes *substrate-lyase* while common name includes expressions like *dehydratase*, *decarboxylase*, *aldolase* etc.
 - e. **Isomerases:** These enzymes catalyze reactions involving geometrical or structural changes within one molecule. Based on isomerism types, they can be *racemases*, *mutases*, *tautomerases*, *epimerases*, *rotamase*, *isomerases* and *cis-trans-isomerases*.
 - f. **Ligases:** These enzymes catalyze the reactions involving joining of two molecules at the expense of hydrolysis of a diphosphate bond of ATP or other triphosphate molecules. The systematic name includes *X-Y ligase* while *synthetase* is usually used as common names (e.g. DNA ligase, RNA ligase and Aminoacyl-tRNA synthetase).
2. Second number represents the sub-class.
 3. Third number represents the sub-subclass.
 4. Fourth number is the serial number of enzyme in its sub-subclass.

For example the EC number of *catalase* is 1.11.1.6 where first digit (1) suggests that catalase performs oxidoreductase reaction while subsequent numbers represent its sub-class and sub-subclass within oxidoreductase class. The present study in this thesis involves hydrolase class of enzymes belonging to **Ntn-hydrolase enzyme superfamily** (EC 3.x.x.x).

1.4 Enzyme kinetics

Every chemical reaction requires some amount of energy to proceed. The energy barrier that prevents the reaction from proceeding is sometimes referred to as *activation energy* (Fig. 1.2). The magnitude of this energy barrier decides the rate of a reaction. Enzymes accelerate the rate of a reaction by lowering this activation energy and thus within a given period of time more

substrate can be converted to products (Berg *et al.*, 2002). The equation below represents basic enzyme catalyzed reaction in which the substrate S react with an enzyme E to form an Enzyme-Substrate complex (ES). The ES complex then breaks down to form product P while releasing the enzyme E in a chemically unmodified form.



Like most chemical reactions, enzyme-catalyzed reactions also maintain steady-state conditions or chemical equilibrium in which the rate of forward reaction is same as the rate of backward reaction. The equation below represents the basic equation on which most enzyme kinetics studies are based. The kinetics characteristics of an enzyme are described by V_{\max} , maximum velocity of enzyme catalyzed reaction and K_m , the substrate concentration at which rate of enzyme catalyzed reaction is half its maximum velocity. These two kinetic parameters are measured using the Michaelis-Menten equations.

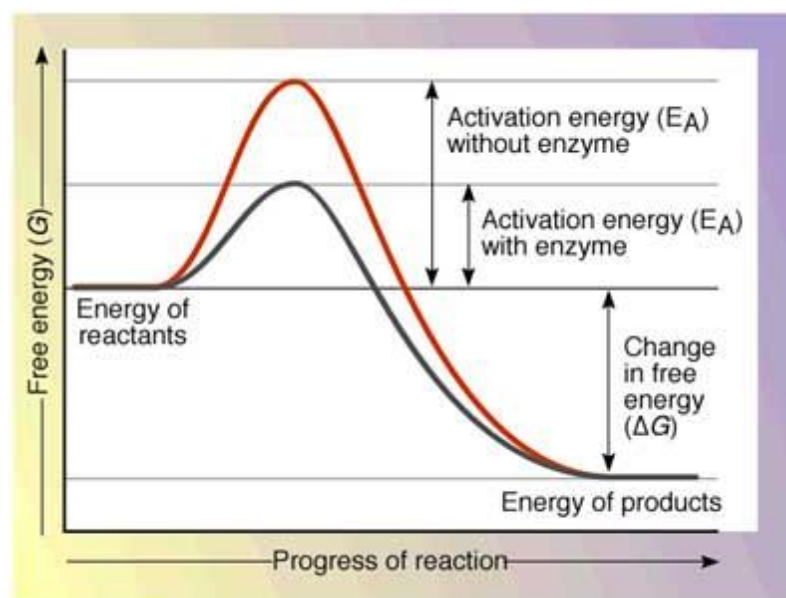


Figure 1.2: Free energy diagram depicting progress of an enzyme catalyzed reaction: The x-axes correspond to the progress of a reaction while y-axes correspond to free energy. It can be observed that the energy of products is comparatively less than energy of substrates. The activation energy without enzymes is higher compared to required activation energy in presence of enzyme. Image adapted from www.pathwayz.org.

1.5 Factors effecting enzyme activity

Several factors such as temperature, pH, concentrations of the enzyme and the substrates, presence of inhibitors and activators, influence rate of an enzymatic reaction (Fig. 1.3). As the **temperature** rises, kinetic energy of substrate and enzyme molecules increases and thus chances

of a perfect collision also increase resulting in enhancement of enzyme activity. The temperature at which maximum activity of enzyme can be obtained is referred as *optimum temperature*. At temperatures above and below optimum value, enzyme activity starts decreasing. Like optimal activity, enzyme also possesses a temperature stability profile i.e. region of *optimal temperature stability*. Like temperature, enzyme works well within certain range of **pH**. The pH at which the enzyme shows highest activity is known as *optimum pH*. Extreme acidic and basic pH usually denatures enzymes thus resulting in complete loss of enzyme activity. Like optimal pH, enzyme also shows a region of *optimal pH stability*.

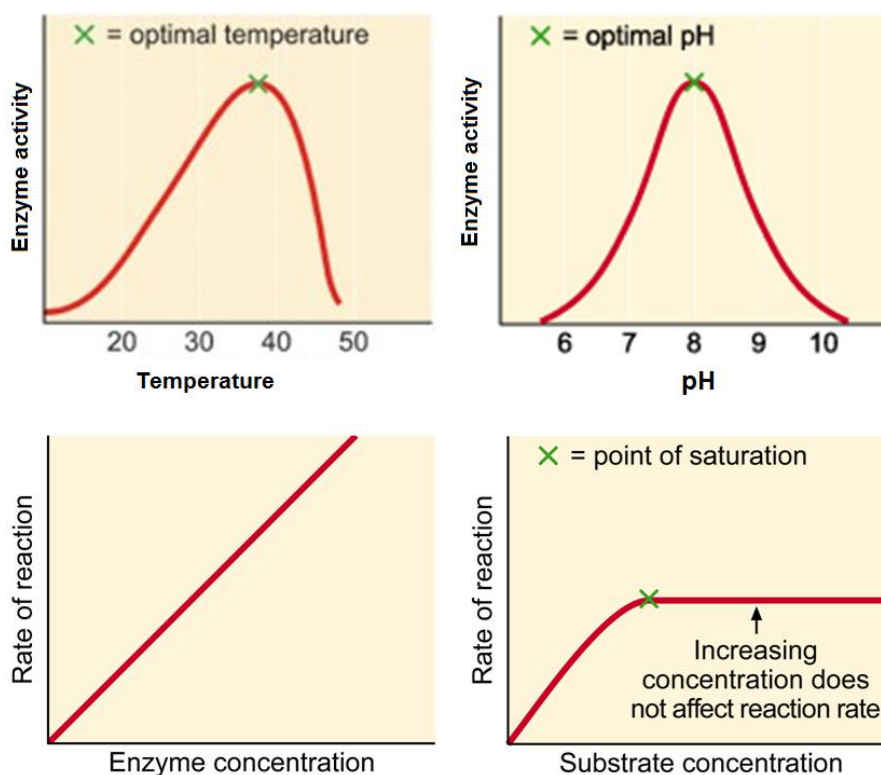


Figure 1.3: The effect of temperature, pH, and enzyme, substrate concentration on enzyme activity. Image adapted from <http://www.rsc.org/>.

The effect of **concentration of enzyme** on its activity can only be studied if substrates are present in sufficient quantity, such that the rate of product formation will depend on enzyme concentration alone. Under such circumstances, reaction follows zero-order kinetics where reaction rate is independent of substrate concentration and hence increases as enzyme concentration increases. Like enzyme concentration, when **substrate concentration** increases, enzyme activity also increases. However the reaction rate increases up to a point (point of

saturation) above which no further increase in reaction rate is observed upon increase in substrate concentration. This is because at this point all the enzyme active sites are saturated with the substrates.

Inhibitors are compounds which inhibit enzyme catalyzed reactions. The inhibition can either be reversible or irreversible. In case of *competitive inhibition*, the inhibitor binds at the same site as that of the substrate, leading to an increase of K_m values but V_{max} remains unchanged. In case of *uncompetitive inhibition*, inhibitor binds only to an enzyme-substrate complex in which case both V_{max} and K_m decreases. In case of *noncompetitive inhibition*, inhibitor binds to a site other than substrate binding site resulting in a decrease in V_{max} values but K_m remains unchanged. In case of *mixed inhibitions*, the inhibitor can bind both to free enzyme as well as enzyme-substrate complex resulting in decrease of V_{max} and increase in K_m values.

1.6 Sequence, structure and substrate specificity relationship in enzymes

The function of an enzyme is determined by its three-dimensional structure which in turn depends on its amino acid sequence. Thus, enzymes having similar sequences fold into similar structural folds and perform similar function. This assumption of *similar sequence* folding to *similar structure* and performing *similar function* forms the basis of most protein functional annotation and prediction tools (Sadowski & Jones, 2009). In case of enzymes, the substrate specificity and binding affinity depends on the catalytic framework near the active site and the active site chemistry, which in turn is decided by how primary structure folds into a well defined conformation. Enzymes sharing similar active site topology and chemistry usually prefer similar substrates although with varied degree of binding affinity.

A collection of related enzymes performing similar functions together form a **family** of enzymes. The members of an enzyme family are often evolutionarily related and are described as functional *homologs* of each other (Berg *et al.*, 2002). Homologs from different species performing similar function are more specifically termed as *Orthologs* while homologs within one species performing different functions are termed as *Paralogs*. As the sequence similarity or homology among enzymes decreases, the structural and functional diversity increases (Wilson *et al.*, 2000). Enzymes acting on different but chemically and/or structurally related substrates often form *distant homologs* of each other. Such distant homologs form a **superfamily** of enzymes. As

we move from family to superfamily level, the sequence homology generally decreases and consequently the functional diversity increases.

There are many public domain databases in existence such as Pfam (Punta *et al.*, 2012), MEROPS (Rawlings *et al.*, 2012), SUPERFAMILY (Gough *et al.*, 2001), SCOP (Murzin *et al.*, 1995) and CDD (Marchler-Bauer *et al.*, 2013), which classify proteins into families and superfamilies based either purely on sequence information or by also considering structural information. **Pfam** is a database which classifies protein sequences into various families, each represented by multiple sequence alignments and hidden Markov models (HMMs). Related families are grouped together to form clans. Each Pfam family represents a functional domain in a protein. Different combinations of such functional domains give rise to vast range of proteins found in nature. The Pfam database, version 27.0, includes a total of 14831 Pfam families. Unlike Pfam, **SCOP** is a hierarchical protein structural classification database in which proteins are manually classified into various families based on their structures. Each family provides comprehensive information about structural and evolutionary relationship among proteins. In SCOP database, there are 4 hierarchical levels of classification namely class, architecture, folds and families. Proteins clustered in a SCOP family are evolutionarily related and defined by having a common structural fold. The current version 1.75 includes a total of 1195 structural folds and 3902 protein families. **SUPERFAMILY** is another database built upon the SCOP database, provides structural and functional annotations for all proteins and genomes. It consists of large collections of hidden Markov models, each representing one structural domain in SCOP database. Those domains that are evolutionary related are grouped together as superfamily. In the current version of this database (version 1.75) the information is generated by scanning all protein sequences of over 2478 completely sequenced genomes against the HMM libraries. Conserved Domain Database (**CDD**) is a protein annotation resource in which proteins are represented as one or more conserved domains. Each conserved domain is represented by position-specific scoring matrices (PSSM) and the RPS-BLAST is used as domain identification tool. CDD uses 3D-structural information to define explicitly the domain boundaries and thus provides accurate insight into the sequence-structure-function relationship. **MEROPS** is another useful information resource, exclusively for peptidases, which uses a hierarchical structure-based classification of peptidases into various families. Each family includes members that are

homologs and therefore share significant degree of similarity in terms of their sequence and structure. Related families are grouped together into various Clans.

Among the many protein superfamilies available, **N-terminal nucleophile hydrolase superfamily** is one of the most diverse superfamily of hydrolytic enzymes. This superfamily includes many families of enzymes having physiological, clinical and pharmaceutical importance of which the **cholyglycine hydrolase (CGH) family** has been studied in this thesis to understand the *sequence-structure and substrate specificity relationship* among the family members. Despite the presence of significant degree of sequence and structural similarity, a wide variation in substrate specificity is observed among the CGH family members.

1.7 Sequence, structure and stability relationship in enzymes

Stability of an enzyme under varied environmental condition such as temperature, pH, organic solvents and ionic strengths is another important property of an enzyme which determines its applicability under these conditions. Similar to the sequence, structure and substrate specificity relationship, the stability of enzymes under varied conditions are also dependent on enzyme's 3D structure and the underlying primary sequence. The native conformation seems to be the most stable conformation of folded protein although the energy difference between the unfolded and folded state is small. The process by which a polypeptide chain folds to find its correct conformation is usually fast. Theoretically there are many possible conformations, even for a small protein, which makes it difficult to predict the folding pathway. Some proteins fold directly without any intermediate stages while for others, folding involves intermediates such as molten globules (Matthews, 1993). In cells, chaperones are the proteins which aid in the folding of misfolded proteins. In certain cases such as insulin, protein can only be folded properly as a precursor protein. Once the folding is complete, protein undergo post-translational processing in which enzyme gets activated by removal of polypeptide segments. The possible role of such segment could be to remove kinetic barrier in the protein folding pathway (Matthews, 1993).

The stability of a folded protein can be due to several factors, including intra molecular factors as well as external factors. **Hydrophobic effect** is considered as the principal driving force facilitating protein folding and is thus believed to have a significant role in protein stability

(Dill, 1990). Hydrophobicity results in the burial of hydrophobic residues in the protein core while the polar residues are exposed towards the solvent. Decrease in hydrophobic surface area is thought to be one of the mechanisms of thermostabilization among proteins (Knapp *et al.*, 1999).

It has long been believed that thermostability of a protein can be correlated to its *amino acid composition*. The intrinsic properties of amino acids have always been known to be of prime importance in providing thermostability to a protein. Statistical comparison among a set of mesophilic and thermophilic proteins showed several trends of residue preference as a mechanism of protein thermo-stabilization (Vieille & Zeikus, 2001). Some of the widely observed trends were higher preference of Arg compared to Lys residues, lower content of uncharged polar residues (at the expense of higher polar and charged residues). Other minor trends showed importance of aromatic residue content. However it is important to note that these trends cannot be applied universally to all proteins. Facchiano *et al.*, 1998, showed that *secondary structure stability* also plays an important role towards protein stability. They observed higher stability of helices of thermophilic proteins compared to their mesophilic counter part. Lower preference of branched side chain residues such as Val, Ile and Thr in helices of thermophilic proteins were observed (Facchiano *et al.*, 1998). Among many hyperthermophilic proteins, two mechanisms of loop stabilization have been observed: *shortening of loop* and *loop anchoring* (Auerbach *et al.*, 1997; Auerbach *et al.*, 1998). Shortening of loop regions are due to increase in periodic secondary structure content such as helices and strands, while loop anchoring is achieved by hydrogen bonding, ionic and hydrophobic interactions. In some cases *anchor of N- and C-terminal region* of protein is also observed to provide thermostability in a similar way to stabilization by loop anchoring (Hennig *et al.*, 1995).

Many intra molecular interactions were also shown to influence protein stability such as disulfide bridges, ionic interactions and hydrogen bonds. *Disulfide bridges* are thought to provide stability to protein through entropic effect (Matsumura *et al.*, 1989). The entropic effect is theoretically predicted to increase in proportion to the logarithm of the size of the loop connecting the two Cys partners. Reduction of disulfide bridges have been experimentally shown to reduce stability. Experimental evidence suggests that conformational environment and solvent

accessibility are important determinants of protection of disulfide bridges against thermal destruction (Vieille & Zeikus, 2001). Introduction of disulfide bridges through protein engineering is one of the most employed protein engineering strategy (Matsumura *et al.*, 1989).

Electrostatic interactions play significant role in maintaining conformation of a folded protein (Perutz, 1978). A single isolated ion-pair has been calculated to contribute approximately 3 to 5 kcal/mol of energy towards stabilization of T4 Lysozyme (Anderson *et al.*, 1990). Many thermophilic proteins were observed to maintain higher number of ion-pairs compared to their mesophilic counterpart (Yip *et al.*, 1995). Ion-pair networks are energetically more favorable compared to isolated ion-pairs, thus large ion-pair networks were also observed amongst thermophilic proteins. Since the protonated state of acidic and basic residues are determined by pH of the environment, it is expected that pH optimum and pH stability profile of enzymes could also be affected by ionic interactions (Vieille & Zeikus, 2001).

Site-directed mutagenesis experiments carried out on RNase T1 showed contributions of approximately 110 kcal/mol energy by **hydrogen bonds** towards its stability (Shirley *et al.*, 1992). Tanner *et al.*, 1996, showed a positive correlation between GADPH thermostability and percentage of charged-neutral hydrogen bonds. A possible reason could be that less de-solvation penalty would be required to bury charged-neutral hydrogen bonds compared to charged-charged hydrogen bonds (ion-pairs). Similarly due to charge-dipole interactions, charged-neutral hydrogen bonds have higher enthalpic rewards than neutral-neutral hydrogen bonds (Tanner *et al.*, 1996). This trend of higher percentage of charged-neutral hydrogen bonds is also observed in *T. maritime* ferredoxin thermostability (Macedo-Ribeiro *et al.*, 1996).

Although insufficient experimental evidence exists to describe the role of interactions involving aromatic residues towards protein stability, these interactions exist among proteins (Vieille & Zeikus, 2001). **Aromatic-aromatic interactions** are formed when the aromatic ring centroids are within the range of 4.5 to 7 Å. Thermitase from *Thermoactinomyces vulgaris* contains 16 aromatic residues participating in aromatic-aromatic interactions while its mesophilic counterpart, *Bacillus amyloliquefaciens* subtilisin BPN' possesses only six aromatic pairs (Teplyakov *et al.*, 1990). Single and double mutants carried out on Tyr13-Tyr17 pair in *B. amyloliquefaciens* RNase showed contribution of -1.3 kcal/mol towards protein thermostabilization (Serrano *et al.*, 1991).

Aromatic-sulphur interactions have been observed to occur more commonly in protein structures. Computational analysis suggested that configurations having sulphur atoms over the aromatic rings are important towards protein stability and folding (Ringer *et al.*, 2007). The interacting sulphur atoms were observed to show affinity towards aromatic ring edges rather than the region above the π -electrons of the ring (Reid *et al.*, 1985).

Like aromatic-sulphur interactions, *cation- π interaction* is another form of electrostatic interaction involving aromatic and positively charged centers. In this interaction, positively charged side chains such as that of Arg or Lys and metal cations interact with aromatic π -electron centers. The stabilization energy decreases with distance 'r' as a function of $1/r$. Low de-solvation energies associated with the burial of aromatic residues in hydrophobic environment makes these interactions a potential stabilizing mechanism in proteins (Dougherty, 1996).

Apart from molecular interactions, other factors were also observed which influence protein thermostability such as entropic stabilization by proline residues, helix dipole stabilization, conformational strain release, reduction of deamidation damage and oligomerization. **Proline residues** being conformationally most rigid provide *entropic stabilization* to protein by reducing the entropy of protein's unfolded state. Glycine on the other hand has highest conformational entropy. Thus Gly \rightarrow X and X \rightarrow Pro mutations have been considered as engineering strategy for enhancement of protein thermostability. Thermophilic proteins have also been observed to maintain *higher proline content* than their mesophilic counterpart (Watanabe *et al.*, 1997). Through site-directed mutagenesis it has been shown that proline residues if introduced at **second position of beta-turns** or at **N-cap position of helices**, would result in an enhancement of thermostability (Suzuki *et al.*, 1987; Watanabe *et al.*, 1994).

Residues with left-handed helical conformation are shown to be less stable than residues with right-handed helical conformation by 0.5 to 2 kcal/mol. The short distance between the carbonyl oxygen and beta-carbon of residues in left-handed helical conformation causes local conformational strain. Releasing this strain has been shown to enhance thermostability. This mechanism of *conformational strain release* has been considered as one of the engineering strategies in case of *B. subtilis* DNA binding protein HU where the conformational strain on Glu15 residue was removed by substitutions with Gly resulting in significant increase in

enzyme's thermostability (Kawamura *et al.*, 1996). Similarly Lys95→Gly mutation enhanced thermostability of *E. coli* RNase H1 through conformational strain release (Kimura *et al.*, 1992).

Helix dipole stabilization is considered as another mechanism of protein thermostabilization in which negatively charged residues are observed near N-terminal end while positively charged residues are observed near C-terminal end of helices. Nichol森 *et al.*, 1988, have estimated roughly 0.8 kcal/mol contribution by N-cap stabilization to enzyme's ΔG_{stab} (Nicholson *et al.*, 1988). Though the stabilization due to helix dipole neutralization seems marginal, observation in thermophilic *B. stearrowthermophilus* (16 Ncaps and 13 Ccaps) and *T. maritima* PGKs (17 Ncaps and 14 Ccaps) shows higher occurrence of dipole stabilized helices compared to the mesophilic pig (9 Ncaps and 12 Ccaps) and yeast (10 Ncaps and 9 Ccaps) PGKs (Auerbach *et al.*, 1997).

Presence of *reduced number of thermo-labile bonds* has been observed to be another thermostabilization strategy observed amongst thermophilic proteins. Asn-Gly bond is considered as the most thermolabile bond since Asn residues undergo deamidation at higher temperature by β -aspartyl shift mechanism (Robinson, 2002). This deamidation results in conversion of Asn into Asp and isoAsp residues which disturbs the protein backbone and thus reduces protein stability. Ser and Ala are also residues which promote deamidation, although at a much slower rate when compared to Gly. Thermophilic proteins either substitute Asn or Gly/Ser/Ala with bulkier residues to prevent the deamidation. In another proposed acid-base mechanism of deamidation, Ser and Thr are the residues which promote the deamidation of Asn/Gln. Thus many thermophilic proteins were observed to reduce their uncharged polar residue (Asn+Gln+Ser+Thr) content (Vieille & Zeikus, 2001). It is well known that **metals** also play critical role in enzyme activation and stability (Smith *et al.*, 1999).

Although several mechanisms of structural stabilization have been listed above, there seems to be not a single universal mechanism which can be considered when dealing with the problem of protein thermostabilization. Sometimes it may not be the global structural features which determine overall enzyme stability but a small number of highly selected mutations which when taken together are enough to enhance protein thermostability. The *sequence-structure-stability* relationship is a complex relationship and its understanding requires both computational and experimental approach. In this thesis we have attempted to understand this relationship by

studying enzymes of **penicillin G acylase (PGA)** family, a family belonging to N-terminal nucleophile hydrolase superfamily. The PGA enzymes are widely used in pharmaceutical industry for the manufacture of semi-synthetic antibiotics where thermostability plays a critical role in determining their usability. Exploring the above mentioned thermostability parameters among large number of protein structures could be time consuming and error-prone. Therefore a set of computational tools were also developed to automate large scale protein structural analysis.

The following section (section 1.8) describes the characteristics of Ntn-hydrolase enzymes along with brief description of each enzyme family classified under the superfamily. The next section (section 1.9) explores the scope of the work which has been carried out in this thesis highlighting the cholyglycine hydrolase and penicillin G acylase family. Finally the last section (section 1.10) describes the tools and techniques used for the analysis.

1.8 Ntn-hydrolase enzyme superfamily

N-terminal nucleophile (Ntn) hydrolases are a recently discovered superfamily of enzymes; which functionally belong to hydrolase class of enzymes but more specifically they are amidases (Artymiuk, 1995; Brannigan *et al.*, 1995). The N-terminal amino acid residue of these enzymes acts as *nucleophile* during catalysis. The nucleophilic residue can either be **Cys** in N-terminal cysteine nucleophile (NtCn-hydrolase) or **Ser** in N-terminal serine nucleophile (NtSn-hydrolase) or **Thr** in N-terminal threonine nucleophile (NtTn-hydrolase) superfamily, where the side chain sulfhydryl (-SH) or hydroxyl (-OH) group act as nucleophile. The free α -amino group (-NH₂) of the same N-terminal residue serves as *base* in order to generate the nucleophilic atom. The catalytic domain of all Ntn-hydrolase enzymes share a characteristic four layered $\alpha\beta\beta\alpha$ core structural fold, known as the **Ntn-hydrolase fold** (Fig. 1.4), in which two central anti-parallel β -sheets is sandwiched between layers of α -helices on either side (Brannigan *et al.*, 1995; Oinonen & Rouvinen, 2000). The peculiar feature about this $\alpha\beta\beta\alpha$ sandwich is that the active site usually lies in a narrow pocket between edges of the two β -sheets and one of the β -strands contains the N-terminal nucleophile residue. All Ntn-hydrolase members show spatially conserved active site topology and chemistry due to which the enzymes have similar reaction mechanisms (Duggleby *et al.*, 1995). However, due to variation in size and shape of the substrate binding pockets, enzymes display variation in their substrate specificity.

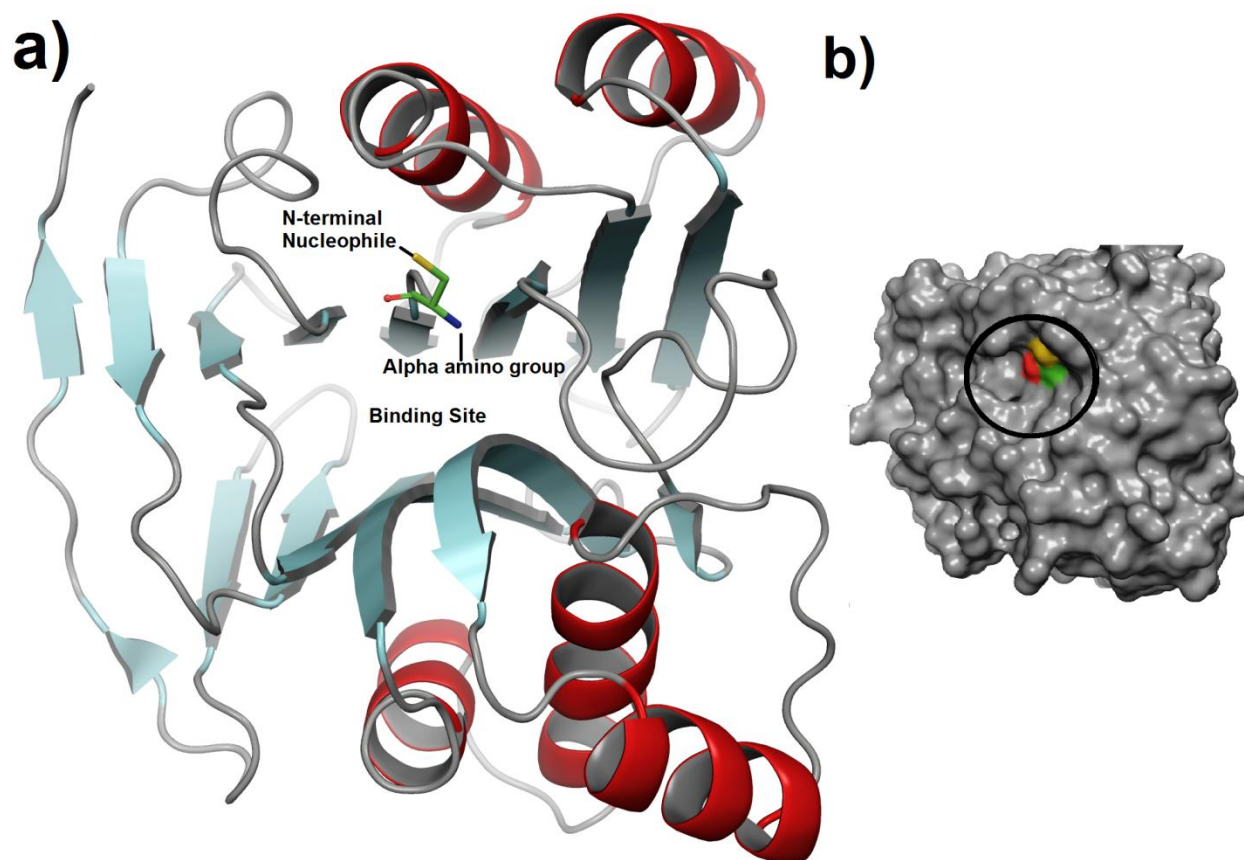


Figure 1.4: (a) Illustrates the $\alpha\beta\alpha$ Ntn-hydrolase structural fold containing the N-terminal Cys nucleophile residue located at the edge of one of the β -strands. Helices are colored red while strands are colored cyan. The binding site cleft between the two anti-parallel β -sheets is marked. The free alpha-amino group of Ntn-Cys which acts as base and the sulfhydryl group that acts as nucleophile during catalysis are also marked. This fold corresponds to glutaminase domain of Glucosamine 6-phosphate synthase enzyme (PDB ID: 1XFF). (b) The surface view showing the binding site cleft (encircled) that is located deep inside between the β -sheets near the N-terminal nucleophile residue (colored surface).

The reaction mechanism usually involves initial generation of nucleophile atom of N-terminal Cys/Ser/Thr residue by its own α -amino group. This is followed by the nucleophilic attack of the activated nucleophile on the carbonyl carbon of the amide bond of substrate. This nucleophilic attack results in formation of a tetrahedral intermediate which is stabilized in an oxyanion hole. Next step involves collapse of intermediate to release the leaving group and formation of acyl-enzyme adduct. Last step involves the formation of products by deacylation of enzyme-adduct through general acid-base mechanism (Duggleby *et al.*, 1995). **Bile Salt Hydrolase** enzyme from NtCn-hydrolase superfamily has been used to depict the mechanism of

hydrolytic reaction (Fig. 1.5). This enzyme catalyzes the hydrolysis of amide bond of substrate taurocholic acid. The reaction begins with the generation of nucleophile by proton transfer from sulfhydryl (-SH) group to α -amino (-NH₂) group of N-terminal Cys residue. Next step involves nucleophilic attack by Ntn-Cys residue on the carbonyl carbon of taurocholate generating negatively charged tetrahedral intermediate which is stabilized by oxyanion hole forming residues. The next step involves protonation of amide nitrogen of taurocholate by α -amino group of Ntn-Cys resulting in the release of leaving group (taurine) and formation of acyl-enzyme adduct. The final step involves cleavage of acyl-enzyme adduct by the attack of a nucleophilic water molecule (Lodola *et al.*, 2012). The members of Ntn-hydrolase superfamily not only differ with respect to the nucleophile residue (Cys/Ser/Thr) but also with respect to oxyanion hole forming residues. The reaction mechanisms of Ntn-hydrolases are usually compared to that of serine proteases involving the S-H-D triad, although the catalytic residues are different.

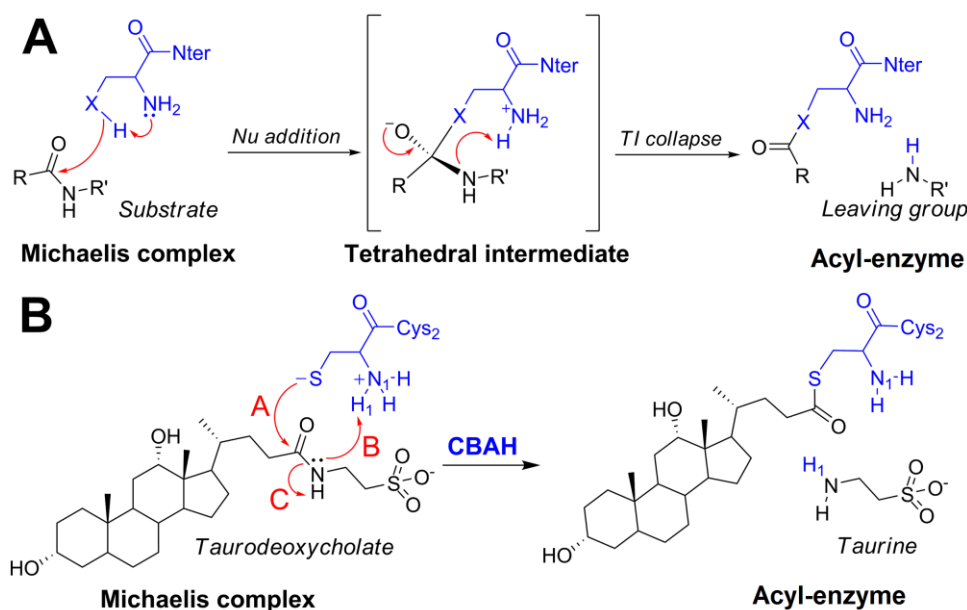


Figure 1.5: Reaction mechanism of Ntn-hydrolase enzymes is depicted taking bile salt hydrolase enzymes as example. In the image, X represents the nucleophile atom (sulphur). Adapted from Lodola *et al.*, 2012.

Many of the enzymes of Ntn-hydrolase superfamily play important function in metabolic pathways such as purine, amino acid and amino sugar biosynthesis. Some enzymes are widely used in pharmaceutical industry for β -lactam antibiotics synthesis. Some enzymes show clinical significance since their malfunctioning leads to lethal disorders. Some enzymes are potential drug target (Table 1.1). These enzymes show wide distribution in animals to microbes, and even

reported presence in viruses. The most remarkable feature of the members of Ntn-hydrolase superfamily is their evolutionary history. It is indeed interesting to understand how nature has utilized the Ntn-hydrolase domain amongst enzymes performing diverse functions. Many enzymes possess multiple domain for their functioning of which one of the catalytic domain is observed to be Ntn-hydrolase domain (Table 1.1). Some enzymes function as monomer while others as oligomers. The enzymes are so divergent that homology cannot be detected at their sequence level; the homology only exists at the 3D structural level.

One of the unique features of most Ntn-hydrolase enzymes is that they are produced as inactive pro-enzymes. The precursor pro-enzyme activates itself by a post-translational **intramolecular autocatalytic peptide bond cleavage** that exposes its N-terminal nucleophile residue responsible for catalysis (Guan *et al.*, 1996; Seemuller *et al.*, 1996; Tikkanen *et al.*, 1996; Xu *et al.*, 1999; Zwickl *et al.*, 1994). This autocatalytic hydrolytic activity justifies the classification of all Ntn-hydrolase pro-enzymes as peptidases and their inclusion in MEROPS database, an information resource of peptidases (Rawlings *et al.*, 2012). However, once the enzymes are activated, the mature form of enzymes does not possess any further peptidase activity. Exceptions include the Ntn-hydrolases such as proteasomes (NtTn-hydrolase) and aminopeptidase DmpA (NtSn-hydrolase) where the mature enzymes indeed are peptidases. Ntn-hydrolase enzymes can follow one of the four types of cleavage event. In some enzymes such as glycosylasparaginases and acid ceramidases, autocatalytic cleavage event cleaves a peptide bond in the polypeptide chain to produce two chain active enzymes (Fig. 1.6a). This cleavage event results in liberation of catalytically active nucleophile residue (Ntn residue) at N-terminal of one of the chain. In enzymes such as penicillin G acylases and cephalosporin acylases, the primary intra-chain cleavage of pro-enzyme is followed by a secondary cleavage event resulting in the removal of the spacer peptide (Fig. 1.6b). In enzymes like penicillin V acylases, proteasomes and glutamine-PRPP aminotransferases, enzyme maturation involves autocatalytic removal of N-terminal pre-peptide segment prior to the Ntn Cys/Ser/Thr residue (Fig. 1.6c). Finally, in case of enzymes like bile salt hydrolases and Ntn-glutamine aminotransferases, maturation involves removal of initiator methionine residue by methionyl aminopeptidase enzymes (Fig. 1.6d).

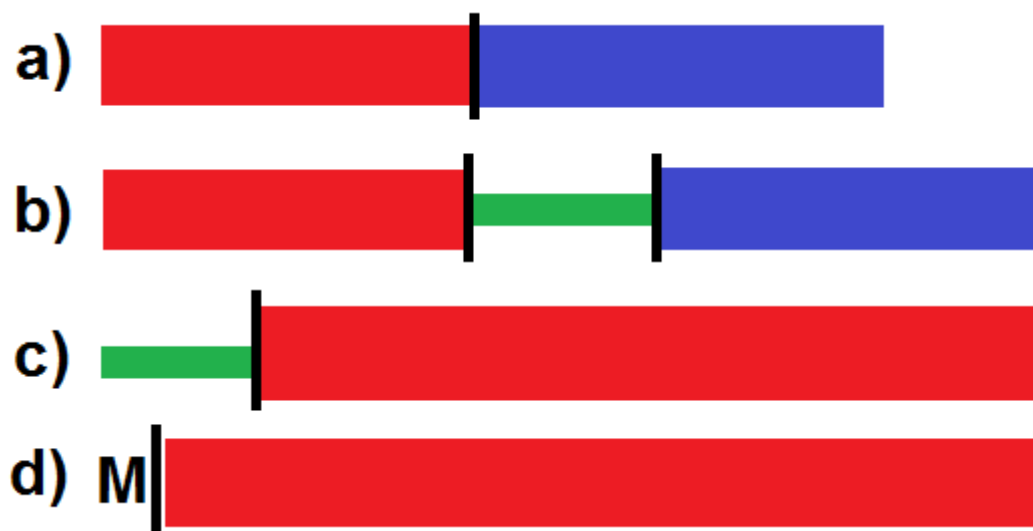


Figure 1.6: Schematic representation of various auto-catalytic cleavage events amongst enzymes of Ntn-hydrolase superfamily. (a) Cleavage occurs inside the polypeptide chain leading to split of single polypeptide chain into heterodimers. No removal of any peptide segment occurs. (b) Two cleavage event, primary and secondary, resulting removal of spacer peptide (green). (c) Removal of pre-peptide (green) by single auto-catalytic cleavage event at the N-terminal of a polypeptide chain. (d) No autocatalytic cleavage event. Instead the initiator Met residue is removed by methionyl aminopeptidases. The black line represents the cleavage site.

The **MEROPS** database, an information resource for all peptidases, has classified Ntn-hydrolase enzymes under Clan PB (Rawlings *et al.*, 2013). This clan has been classified into subclan PB(C), PB(S) and PB(T), based on the type of Ntn residue, Cys in case of PB(C), Ser in PB(S) while Thr in PB(T). Each of these subclans further includes many families of enzymes having diverse function (Table 1.1). Apart from the MEROPS database, information about Ntn-hydrolase superfamily is also available among other public domain databases. In SCOP database, Ntn-hydrolase enzymes are classified under **Class:** alpha and beta proteins ($\alpha+\beta$), **Fold:** Ntn hydrolase-like and **Superfamily:** N-terminal nucleophile aminohydrolases. The current version 1.75 of SCOP database includes seven families under Ntn-hydrolase superfamily. Figure 1.7 shows distribution of Ntn-hydrolase domains across different taxonomy groups present in SUPERFAMILY database. In Pfam database all Ntn-hydrolase enzymes are classified under clan NTN (Pfam ID: CL0052). The current statistics shows 14 families under this clan including a total of 47927 sequences belonging to 5468 species having 255 unique domain architectures.

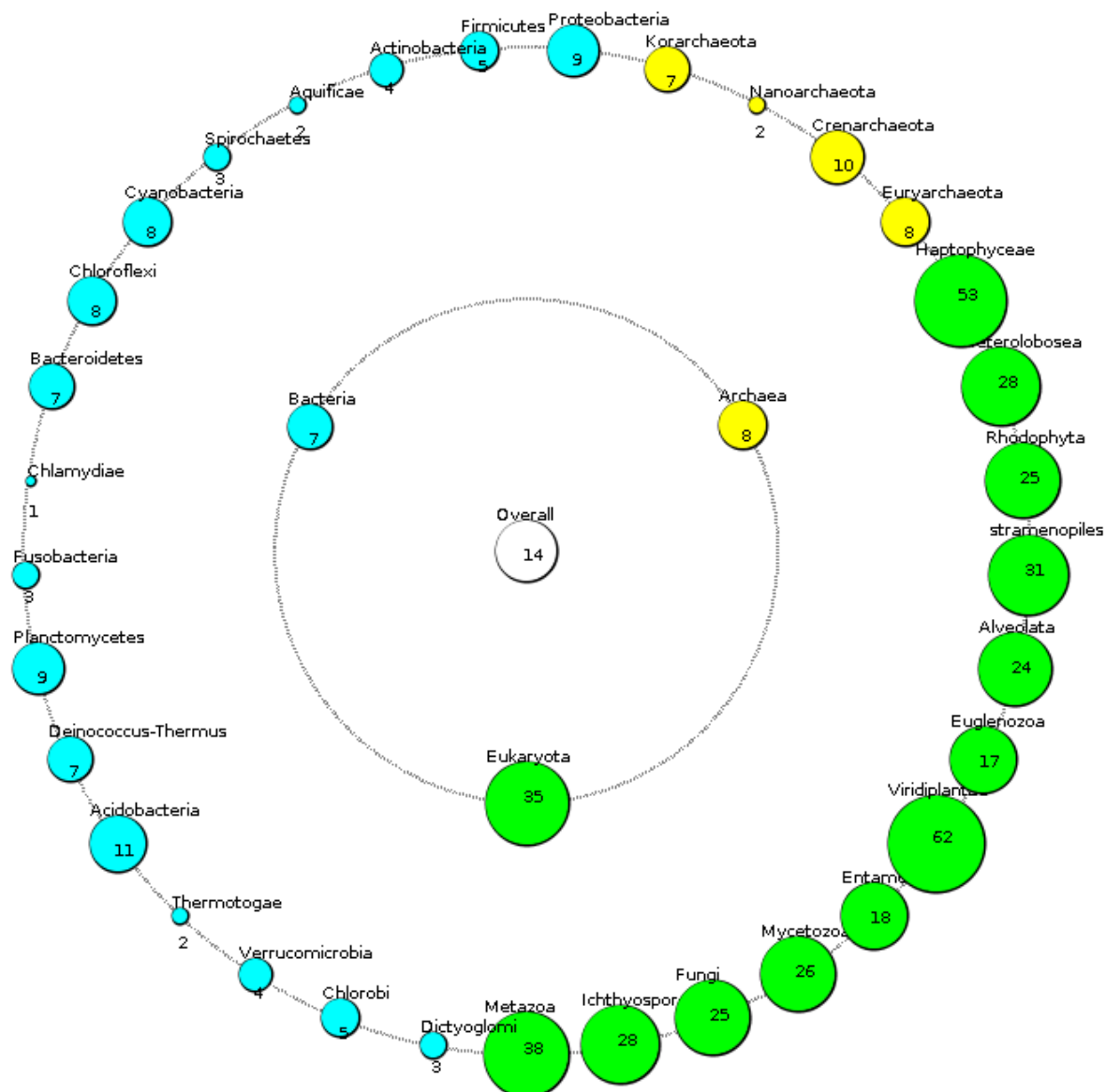


Figure 1.7: Distribution of Ntn-hydrolase enzymes across different taxonomic group. In this TaxViz representation, each node corresponds to feature of single taxonomic group. Nodes are arranged in concentric rings where parent node lies at the centre while child nodes are radiated outwards. The size of circle denotes mean number of Ntn-hydrolase domains found per organism in a given taxonomic group. The TaxViz display has been generated from Superfamily database, version 1.75.

Table 1.1: List of Ntn-hydrolase enzymes, their taxonomic distribution, domain architecture and function.

Superfamily	MEROPS Family	Enzyme	Total Domains	Clinical/Physiological importance	Structure present?	Organism*							
						B	R	P	F	L	A	V	
NtCn-hydrolase	C59	Bile Salt Hydrolase (BSH)	1	Hypercholesterolemia	Y	Y	Y	N	Y	Y	Y	Y	
		Penicillin V Acylase (PVA)	1	Antibiotics industry									
	C44	Glutamine phosphoribosylpyrophosphate amidotransferase (GPATase)	2	Purine biosynthesis	Y	Y	Y	Y	Y	Y	Y	Y	
		Glutamine-fructose-6-phosphate transaminase (GFAT)	2	Hexosamine biosynthesis, Drug target for type2-diabetes									
		Asparagine synthetase (AS)	2	Congenital microcephaly and Progressive encephalopathy									
			Glutamate synthase (GS)	3	Amino acid biosynthesis								
	C45	Acyl-CoA:isopenicillin N-acyltransferase (IPAT)	2	Penicillin biosynthetic pathway	Y	Y	Y	N	Y	Y	Y	N	
	C89	Acid Ceramidase (AC)	2	Farber lipogranulomatosis	N	Y	N	Y	Y	Y	Y	Y	Y
		N-acylethanolamine-hydrolyzing acid amidase (NAAA)	2	NAAA hydrolysis, Endocannabinoid metabolism									
	C69	Dipeptidase DA	1	Proteolytic system of Lactic acid bacteria, Useful to dairy industry	N	Y	Y	Y	Y	Y	Y	N	
C95	Lysosomal 66.3 kDa	2	Lysosomal storage disorders	N	N	N	N	N	N	Y	N		
NtSn-hydrolase	S45	Penicillin G Acylase (PGA)	2	Antibiotics industry	Y	Y	Y	Y	N	Y	N	N	
		Cephalosporin Acylase (CA)	2	Antibiotics industry									
		AHL hydrolase	2	Quorum quenching									

NtTn- hydrolase	T1	Proteasome	1	Non-lysosomal protein degradation	Y	Y	Y	Y	Y	Y	Y	N
	T2	Glycosylasparaginase	2	Aspartylglycosaminuria	Y	Y	Y	Y	Y	Y	Y	N
		Taspase-1	2	Cleaves nuclear factors to orchestrate gene expressions.								
		Isoaspartyl dipeptidase	2	Degrades proteins damaged by L-isoaspartyl residue formation								
	T3	γ -glutamyltransferase 1 (GGT)	2	Glutathione biosynthetic process	Y	Y	Y	Y	Y	Y	Y	Y
T5	Ornithine acetyltransferase	2	Amino acid biosynthesis	N	Y	Y	Y	Y	Y	Y	N	

*The organism label corresponds to B: Bacteria, R: Archaea, P: Protozoa, F: Fungi, L: Plants, A: Animals and V: Viruses. Y and N indicate enzyme present and absent, respectively in the corresponding organism.

Below mentioned is a brief description of various families belonging to Ntn-hydrolase superfamily.

1.8.1 Self-processing cysteine-dependent Ntn-hydrolase enzyme superfamily (NtCn-hydrolases)

As the name suggests, this superfamily includes Ntn-hydrolase enzymes whose N-terminal amino acid residue is a cysteine residue that acts as nucleophile and base during catalysis. The members are also characterized by their self-processing proteolytic activity for maturation. In MEROPS database clan PB(C) represents NtCn-hydrolase superfamily (Dijkstra *et al.*, 2013). It includes six structurally and/or functionally characterized families of enzymes namely family C59, C44, C45, C69, C89 and C95. These families are related to each other only through structural homology. No significant sequence homology is detected among representative enzymes of the families. The enzyme families are also different with respect to the number of functional domains. A brief description of the six families is as follows.

1.8.1.1 Family C59

This family is also known by other names such as Cholyglycine Hydrolase (CGH) or Conjugated Bile Acid Hydrolase (CBAH) or Conjugated Bile Salt Hydrolase (CBSH) family. It includes enzymes like **Bile salt hydrolase** (BSH) and **Penicillin V acylase** (PVA). Three dimensional structures have been solved for both enzymes which shows single catalytic domain having a topology similar to $\alpha\beta\beta\alpha$ Ntn-hydrolase fold (Fig. 1.8a). The former enzyme is responsible for regulating cholesterol homeostasis in mammalian gut while the later one is widely used in pharmaceutical industry for β -lactam antibiotics synthesis.

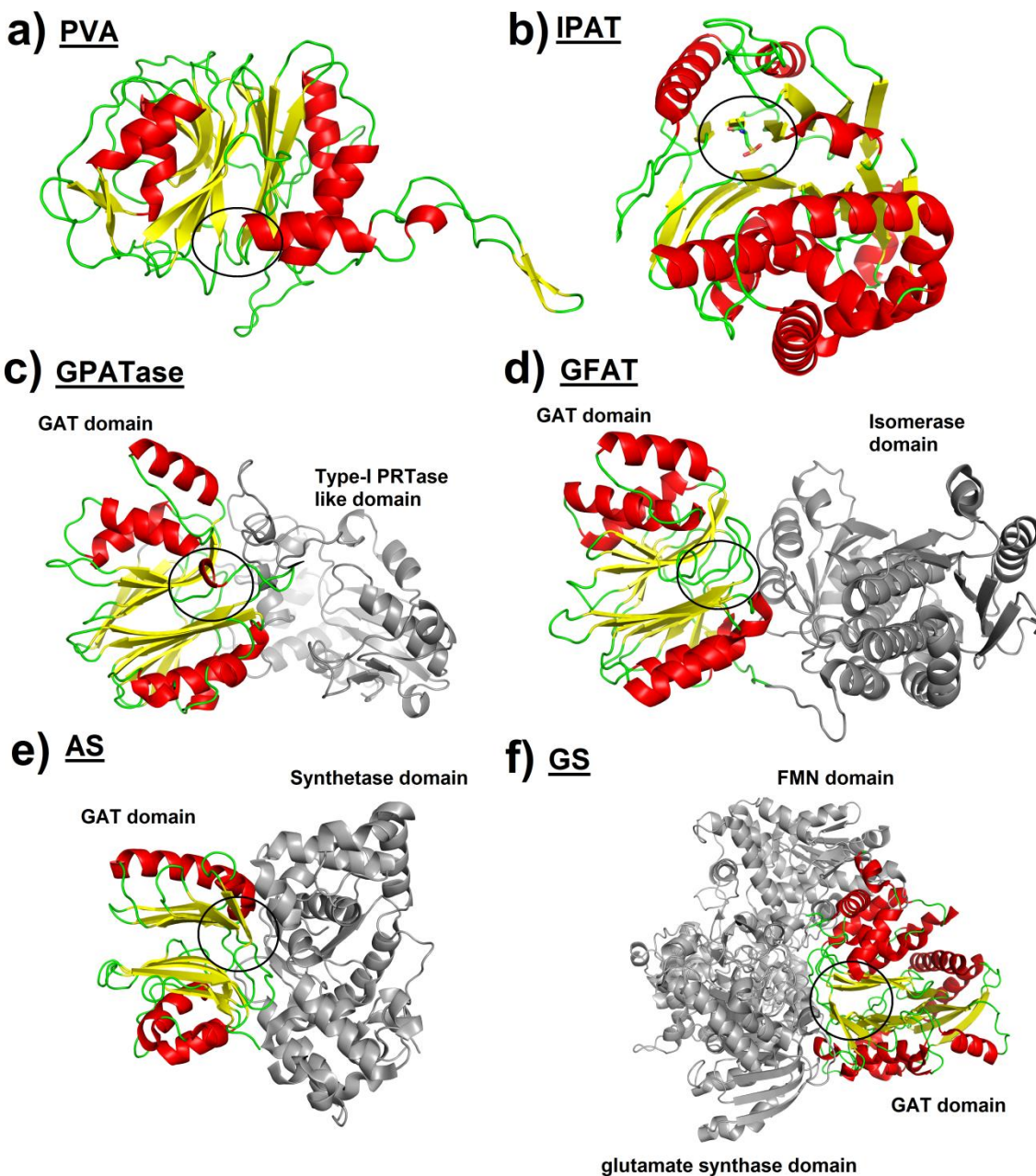


Figure 1.8: Illustrates cartoon representation of enzymes of NtCn-hydrolase superfamily namely (a) Penicillin V acylase (PVA; PDB ID 3PVA; C59 family), (b) Acyl-CoA:isopenicillin N-acyltransferase (IPAT; PDB ID 2X1D; C45 family), (c) Glutamine phosphoribosylpyrophosphate amidotransferase (GPATase; PDB ID 1GPH; C44 family), (d) Glucosamine-fructose-6-phosphate aminotransferase (GFAT; PDB ID 2J6H; C44 family), (e) Asparagine synthetase (AS; PDB ID 1CT9; C44 family) and (f) Glutamate synthase (GS, PDB ID 1EA0; C44 family). The enzymes of C59 and C45 family are composed of one domain while enzymes of C44

family are multi-domain enzymes. In case of enzymes of C59 family, each is a monomer having a topology of Ntn-hydrolase fold while in case of IPAT, Ntn-hydrolase is formed by chain α and chain β (heterodimer). In case of enzymes of C44 family, the GAT domain has topology similar to Ntn-hydrolase fold which is shown in color while other domains are shown in grey color. The circle highlights the location of catalytic site between the two β -sheets.

1.8.1.2 Family C44

This family includes key regulatory enzymes of purine, hexosamine and amino acid biosynthetic pathways such as Glutamine phosphoribosylpyrophosphate amidotransferase (GPATase), Glucosamine-fructose-6-phosphate aminotransferase (GFAT), Asparagine synthetase (AS) and Glutamate synthase (GS), respectively. These enzymes are mainly multi-domain in nature of which the N-terminal domain is always GAT (Glutamine amidotransferase) domain consisting of the Ntn-hydrolase fold. The enzymes differ with respect to their C-terminal domains. The GAT domain is responsible for transfer of nitrogen from nitrogen donor (glutamine or free ammonia of solvent) to other acceptor molecules.

The enzyme **GPATase** is a key regulatory enzyme of *de novo* purine biosynthetic pathway which transfers nitrogen from glutamine or ammonia to phosphoribosyl pyrophosphate (PRPP) acceptor (Zalkin & Smith, 1998). The C-terminal domain resembles its structural fold to Type-I PRTase (Phosphoribosyl transferase) enzymes (Fig. 1.8c). GPATase from *Bacillus subtilis* is synthesized as a pro-enzyme with 11 residues pre-peptide (Li *et al.*, 1999; Souciet *et al.*, 1988). The enzyme maturation involves both the auto-catalytic removal of pre-peptide and insertion of Fe_4S_4 cluster. It is hypothesized that the probable role of the pre-peptide is to block the catalytic cysteine residue from interfering with the Fe_4S_4 cluster insertion during the protein assembly.

The enzyme **Glutamine-fructose-6-phosphate transaminase** (GFAT), also known as Glucosamine-6-phosphate synthase, is the rate-limiting enzyme that regulates the flux of glucose into hexosamine pathway (Nakaishi *et al.*, 2009). It catalyses the first step of the *de novo* hexosamine biosynthetic pathway, producing glucosamine-6-phosphate (GlcN6P) from fructose-6-phosphate (Fru6P) and glutamine. The enzyme GFAT has been shown to play a major role in insulin resistance in cultured cells, therefore it is considered as a potential drug target for treatment of type2 diabetes (Buse, 2006). The N-terminal GAT domain (Fig. 1.8d) catalyzes the

release of NH₃ from glutamine while C-terminal isomerase domain converts Fru6P to GlcN6P utilizing the released NH₃.

The enzyme **Asparagine synthetase** synthesizes amino acid asparagine from glutamine and aspartate utilizing ATP (Van Heeke & Schuster, 1989). N-terminal domain is the GAT domain while C-terminal domain resembles asparagine synthetase domain (Fig. 1.8e). These enzymes are clinically most important because their deficiency leads to Asparagine synthetase deficiency (ASNSD) causing congenital microcephaly and progressive form of encephalopathy (Ruzzo *et al.*, 2013). **Glutamate synthase (GS)** is a complex iron-sulphur containing flavoprotein that catalyses synthesis of L-glutamate from L-glutamine (Nitrogen donor) and 2-oxoglutarate (Nitrogen acceptor). This enzyme has three domains in its structure (Vanoni & Curti, 1999), N-terminal GAT domain (419 residues), Central FMN domain (763 residues) and C-terminal glutamate synthase (270 residues) domain (Fig. 1.8f). Maturation of enzyme involves autocatalytic removal of 36-residue pre-peptide.

1.8.1.3 Family C69

The enzymes **Dipeptidase DA**, belonging to family C69 of MEROPS database show broad substrate specificity towards dipeptides other than proline-containing dipeptides (Dudley *et al.*, 1996). They are also known as dipeptidase A or pepDA or pepD. Dipeptidase was first identified in a Lactic Acid Bacteria (LAB) *Lactobacillus helveticus* CNRZ32. Although no three-dimensional structure is available, the presence of C-X(10,20)-R-X(2)-D sequence motif, as observed in C59 family members justifies their classification as Ntn-hydrolase enzymes. Dipeptidases are one of the key enzymes of complex proteolytic system utilized by lactic acid bacteria to obtain essential amino acids from milk protein casein. This property of hydrolysis of casein by lactic acid bacteria is useful for cheese ripening and flavor development (Dudley *et al.*, 1996).

1.8.1.4 Family C89

Family C89 includes enzymes of clinical importance such as **Acid Ceramidase** (EC 3.5.1.23; AC) and **N-acylethanolamine-hydrolyzing acid amidase** (NAAA). These lysosomal enzymes catalyze the hydrolysis of ceramide, a sphingolipid, to sphingosine and a free fatty acid (Park & Schuchman, 2006). Defect in these genes causes an inherited lipid storage disease called

Farber lipogranulomatosis. The disease is characterized by decreased ceramidase activity and resulting accumulation of lipid-filled nodules under the skin causing heavy pain in joints and sometimes fatal in early infancy (Park & Schuchman, 2006). The inactive precursor undergoes autocatalytic cleavage at the Ile142-Cys143 bond resulting in maturation to a heterodimeric enzyme (α and β chains) (Shtraizent *et al.*, 2008). Although no tertiary structure is available, the β chain has the characteristics of Ntn-hydrolase fold possessing catalytic Cys residue at its N-terminal and presence of C-X(10,20)-R-X(2)-D sequence motif.

1.8.1.5 Family C95

Family C95 includes a **lysosomal 66.3 kDa protein**, first discovered from mouse by lysosomal sub-proteome study. This enzyme is found to be distributed only among vertebrates and absent among prokaryotes. The enzyme is produced as a 75 kDa soluble, glycosylated precursor protein having an N-terminal signal peptide. The signal peptide guides the enzyme to lysosome where the enzyme undergoes an auto-catalytic preprocessing event in order to mature into a heterodimeric enzyme with a 28 kDa N-terminal and a 40 kDa C-terminal segment exhibiting homology with Ntn-hydrolase fold (Deuschl *et al.*, 2006).

1.8.1.6 Family C45

Family C45 of MEROPS database includes the last enzyme of penicillin biosynthetic pathway, the **Acyl-CoA:isopenicillin N-acyltransferase (IPAT)**. This enzyme is produced as a precursor-peptide which undergoes an autocatalytic cleavage of Gly102-Cys103 bond to produce a mature form of enzyme (Fig. 1.8b), a heterodimer (11 kDa α - and 29 kDa β -subunits), in which the acyl-transferase activity resides in the β -subunit resembling Ntn-hydrolase fold (Aplin *et al.*, 1993). Site-directed mutagenesis of β Cys103 residue has been shown to prevent autolysis (Tobin *et al.*, 1995). The residue β Cys103 is proposed to play different roles in auto proteolysis and substrate hydrolysis (Bokhove *et al.*, 2010b).

1.8.2 Self-processing serine-dependent Ntn-hydrolase enzyme superfamily (NtSn-hydrolases)

As the name suggests, in this superfamily of Ntn-hydrolase enzymes, the N-terminal nucleophile residue is serine. In the MEROPS database, family S45 (The prefix S denotes Ser-Ntn-hydrolases) includes two important subfamilies of enzymes, namely **penicillin G acylase** (PGA) and **cephalosporin acylase** (CA, more precisely termed as glutaryl-7-aminocephalosporanic acid acylase). Although the two enzymes are less similar in terms of their sequence but they share similar quaternary structure and the structural similarity at the active site is impressive (Oh *et al.*, 2004). Despite preferring different kinds of substrates, penicillin G and cephalosporin, respectively, they share similar mechanism of action. Although the physiological roles of these enzymes are not clear, these enzymes are very useful in industry for commercial production of antibiotics (Barends *et al.*, 2004).

Both these enzymes are periplasmic proteins, synthesized in cytoplasm as pre-pro enzyme (Signal peptide – Chain α – Spacer peptide – Chain β). The signal peptide guides the pre-pro form of enzyme to get translocated to the periplasmic space. In the periplasm, the inactive precursor enzyme (pro-enzyme) undergoes auto-catalytic processing to remove the spacer peptide or pro-peptide to attain the active form of enzyme. This auto-catalytic pre-processing is a 2-step process. The first step involves an intra-molecular cleavage towards the C-terminal side of spacer peptide leading to the generation of an inactive alpha-subunit containing the spacer-peptide and the beta-subunit. The subsequent cleavage occurs towards the N-terminal side of spacer peptide, as an inter-molecular event in which the cleavage is mediated by the newly formed serine molecule of beta-subunit of another enzyme, resulting in the removal of spacer peptide. The β -subunit in the mature enzyme (Fig. 1.9) folds into a compact structure containing the Ntn-hydrolase fold (Joon Cho *et al.*, 2013).

Three-dimensional structures have been solved for both the processed and the unprocessed forms of PGA and CA enzymes (Daumy *et al.*, 1985; Joon Cho *et al.*, 2013; Kim *et al.*, 2006; Lee *et al.*, 2000; Lee & Park, 1998). The mutation Thr263Gly in *E. coli* PGA enzyme resulted in a slow-processing precursor structure, having similar shape, unit-cell dimension and overall topology as that of mature enzyme, but the spacer peptide was observed to block the active site cleft. The removal of spacer peptide exposes the active site (Fig. 1.9d).

AHL amidohydrolases

The NtSn-hydrolase family includes another class of enzymes, the AHL amido-hydrolase (Fig. 1.9c), similar to PGA and CA in terms of their structure and mechanism of catalysis but act on different substrates. These enzymes hydrolyze the quorum sensing molecules (N-acyl homoserine lactones; AHLs) in Gram-negative pathogens; thereby playing an important role in quorum quenching. Although the overall structures of these enzymes are similar to that of PGA and CA enzymes, it has an unusually large hydrophobic pocket in order to accommodate C12 fatty acid-like chains of AHLs. In case of the enzyme from *Pseudomonas aeruginosa* (PvdQ), β Ser1 acts as the N-terminal nucleophile while β Asn269 and β Val70 acts as the oxyanion hole residues (Bokhove *et al.*, 2010a). An interesting feature of this enzyme is the presence of six highly conserved cysteine residues involved in disulfide bridge formation, a feature absent in PGA and CA enzymes.

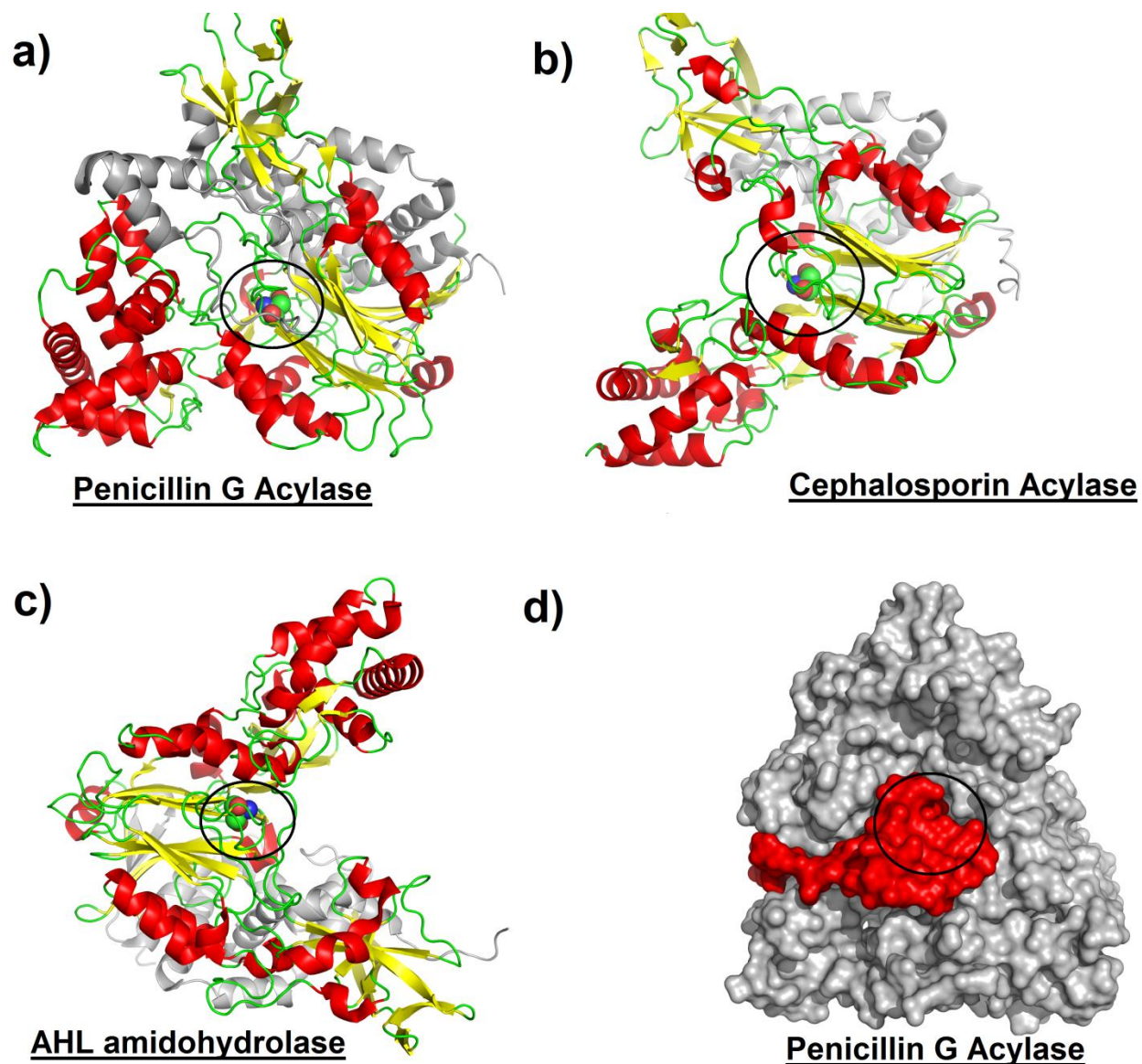


Figure 1.9: Illustrates the enzymes of NtSn-hydrolase superfamily namely (a) Penicillin G acylase (PDB ID: 1GK9), (b) Cephalosporin acylase (PDB ID: 1FM2) and (c) AHL amidohydrolase (PDB ID: 2WYB). The β -subunit which has topology similar to Ntn-hydrolase is shown in color while α -subunits are shown in grey. The N-terminal Ser residue is shown in spheres and highlighted in a circle. (d) Surface view of catalytically inactive unprocessed PGA enzyme showing the blockage of active site (shown in circle) by spacer peptide (red colored surface).

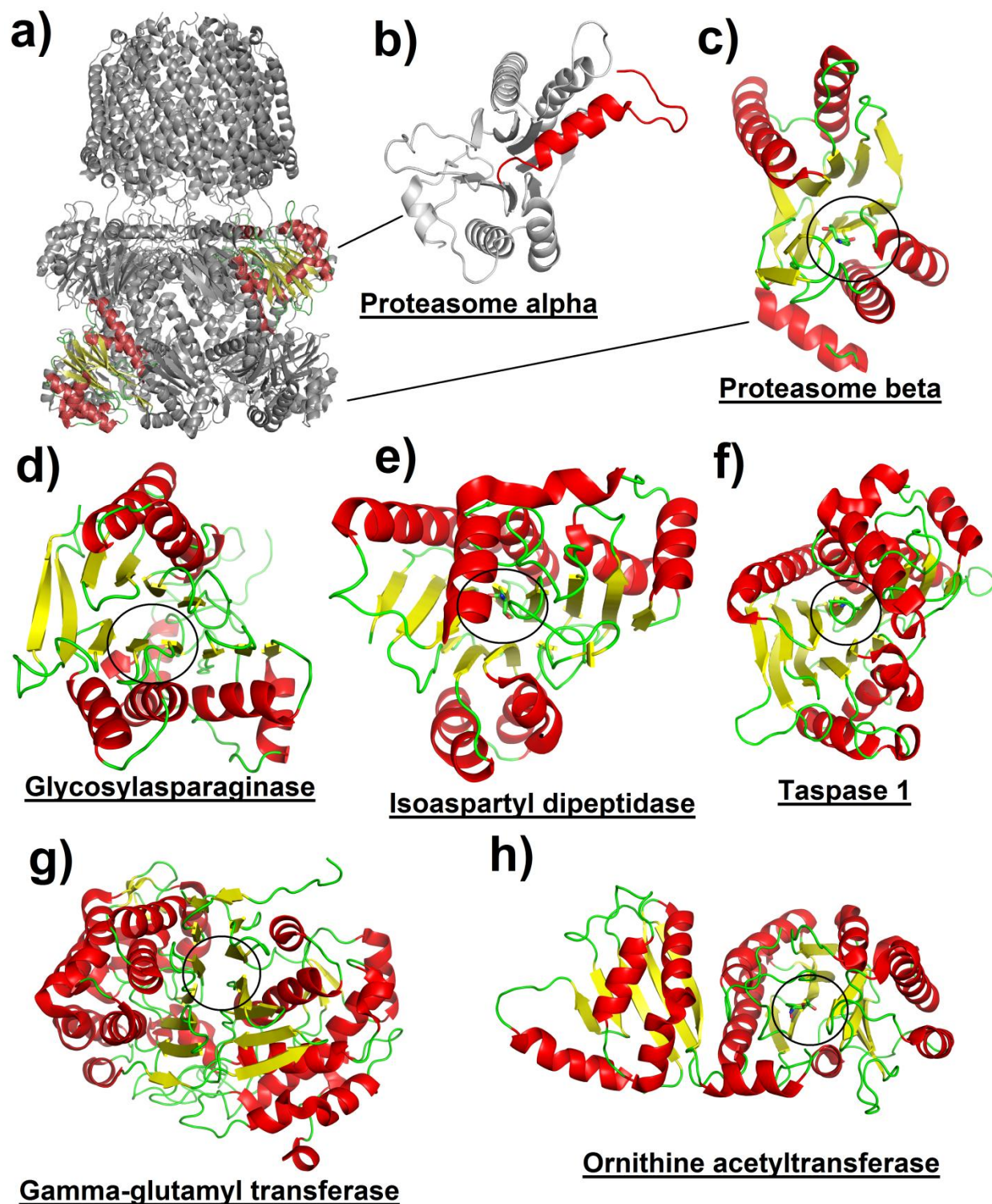


Figure 1.10: Illustrates the enzymes of NtTn-hydrolase superfamily. (a) Achaean proteasome (PDB ID: 1YAR; Family T1). (b) The inactive α -subunit of proteasome. The N-terminal propeptide which is not removed in post-translation making the subunit catalytically inactive is shown in red. (c) The catalytically

active β -subunit of proteasome. The free N-terminal Thr residue is highlighted in circle. (d) Glycosylasparaginase (PDB ID: 1APY; Family T2). (e) Isoaspartyl dipeptidase (PDB ID: 1JN9; Family T2). (f) Taspase 1 (PDB ID: 2A8I; Family T2). (g) γ -glutamyltransferase (PDB ID: 2DG5; Family T3). (h) Ornithine acetyltransferases (PDB ID 2YEP; Family T5). The location of N-terminal catalytic residue is marked by circle.

1.8.3 Self-processing threonine-dependent Ntn-hydrolase enzyme superfamily (NtTn-hydrolases)

As the name suggests these are the Ntn-hydrolase enzymes whose N-terminal residue is Thr that acts as nucleophile during substrate-hydrolysis. This superfamily includes 4 families of enzymes namely **Family T1, T2, T3** and **T5**. These enzymes are different with respect to their substrate specificities and subunit composition; however the catalytic site and overall structural folds remain similar.

1.8.3.1 Family T1

This family includes **proteasomes** from bacteria, archaea and eukaryotes. Proteasome are the central protease of ubiquitin-dependent protein degradation pathway that plays a vital role in non-lysosomal protein degradation and thus help in protein turnover (Hershko & Ciechanover, 1992). The **archaeal proteasome** is the simplest among all proteasomes in terms of their subunit composition and internal symmetry (Fig. 1.10a). The enzyme is a cylindrical shaped protein having 4 stacked rings made up of two different kinds of subunits (subunit α and β). The enzyme is having a subunit composition of $\alpha_7\beta_7\beta_7\alpha_7$ in which each ring has 7 subunits; the outer two rings are composed of α subunits while inner two rings are composed of β subunits. The subunits are arranged to form a channel traversing from one end to the other end of protein forming three inner cavities. The central cavity is lined with proteolytic active clefts of β -subunits. Although the overall structures of the α - and β -subunits were observed to be quite similar (having $\alpha\beta\beta\alpha$ Ntn-hydrolase fold), they differ at their N-terminus (Groll *et al.*, 2003; Lowe *et al.*, 1995). The β -subunits are proteolytically processed and thus are catalytically active while α -subunits do not undergo post-translational processing, thus, are catalytically inactive (Fig. 1.10b and Fig. 1.10c). The residues 1 to 35 at N-terminal of α -subunits are removed in β -subunits making β -subunit catalytically active (Lowe *et al.*, 1995). In α -subunit, the N-terminal segment forms a helical structure on top of central β -sandwich of $\alpha\beta\beta\alpha$ core and thus blocks the active site cleft making it

catalytically inactive. The **eukaryotic 20S proteasome** is a part of larger 26S proteasome complex which is composed of the 20S core particle and 19S regulatory cap (Gallastegui & Groll, 2010). The 20S core particle has similar subunit organization as that of archaeal proteasome, $\alpha_7\beta_7\beta_7\alpha_7$. Unlike archaeal proteasome, the eukaryotic proteasome α and β rings each are composed of seven distinct subunits. Like the archaeal and eukaryotic proteasome, the **bacterial proteasome** from *Rhodococcus erythropolis* (Rer) and *Mycobacterium tuberculosis* (Mtb) also have the $\alpha_7\beta_7\beta_7\alpha_7$ subunit organization (Lin *et al.*, 2006; Tamura *et al.*, 1995).

1.8.3.2 Family T2

This family includes three related enzymes with diverse functions namely Taspase 1, Isoaspartyl dipeptidase and Glycosylasparaginase. **Glycosylasparaginase** enzymes are essential for the hydrolysis of GlcNAc–Asn linkage between carbohydrate and protein in Asn-linked glycoproteins. Their genetic deficiency in humans results in prominent lysosomal disorder known as Aspartylglucosaminuria (AGU). AGU is characterized by excess glycoasparagine in the body tissues and subsequent excretion in urine. The inactive precursor molecule after removal of signal peptide undergoes intramolecular bond cleavage to produce a heterodimeric active form of enzyme (Fig. 1.10d). The enzyme **isoaspartyl dipeptidase** (Fig. 1.10e) aids in degradation of proteins damaged by L-isoaspartyl or β -Asp residues (Michalska *et al.*, 2008). Spontaneous non-enzymatic rearrangement of L-Asp and L-Asn residues into isoaspartyl residues in proteins is a significant structural modification resulting in severely impaired functionality. The isoaspartyl residues are resistant to cellular proteases, thus isoaspartyl dipeptidases play a significant role in hydrolysis of these toxic β -peptides. **Taspase 1** (Threonine aspartase 1) are endopeptidases, involved in cleavage of many nuclear factors such as MLL1 (Mixed Lineage leukemia), MLL2, TFIIA α - β and ALF proteins, that play dominant roles in gene transcription suggesting the role of taspase1 in orchestrating genetic programs. Like proteasomes, taspases are the other exception among Ntn-hydrolase enzymes which retain peptidase activity even after enzyme maturation. The structure is characterized by the presence of characteristic $\alpha\beta\beta\alpha$ fold in which central β -sandwich consisting of 13 anti-parallel β -strands is surrounded by 6 α -helices (Fig. 1.10f). The residue N-terminal Thr234 of β -chain mediates both autocatalytic processing as well as substrate cleavage (Khan *et al.*, 2005).

1.8.3.3 Family T3

Members of family T3 are **γ -glutamyltransferases (GGT)** that play a pivotal role in glutathione metabolism. These enzymes catalyze transfer of γ -glutamyl groups from various γ -glutamyl amide donors to either water (hydrolysis) or to amino acid and dipeptide acceptors (transpeptidation). The enzyme was first observed to hydrolyze glutathione (γ -L-Glu \downarrow L-Cys-Gly) in pancreas. A major physiological role of GGT is to utilize extracellular glutathione as source of Cys for biosynthesis of glutathione inside cell (Okada *et al.*, 2006; Williams *et al.*, 2009). Thus, it aids in the recovery of Cys from extracellular glutathione source. Inhibition or deficiency of GGT leads to glutathionuria and glutathionemia. Three-dimensional structures have been determined for *Escherichia coli* and *Helicobacter pylori* GGT enzymes (Okada *et al.*, 2006; Williams *et al.*, 2009). The mature form of GGT is a heterodimer with a large 40 kDa and a small 20 kDa subunit, associated noncovalently. The $\alpha\beta\alpha$ Ntn-hydrolase fold in the structure of the matured enzyme is formed with both large and small unit together (Fig. 1.10g).

1.8.3.4 Family T5

Family T5 of MEROPS database include **ornithine acetyltransferases**, enzymes that play significant role in cyclic pathway of arginine biosynthesis. They catalyze two activities in this pathway; synthesis of acetylglutamate from acetyl-CoA and glutamate, and synthesis of ornithine by transfer of an acetyl group from N-acetylornithine to glutamate. Like other Ntn-hydrolase enzymes, they are also produced as inactive proenzymes which after removal of N-terminal signal peptide undergo autocatalytic cleavage of Ala214-Thr215 bond to form heterodimeric active enzyme. The free N-terminal Thr215 residue of β -chain acts as nucleophile during catalytic reaction. Tertiary structure has been reported for *Streptomyces clavuligerus* ornithine acetyltransferases which show characteristics Ntn-hydrolase structural fold (Fig. 1.10h) in them (Iqbal *et al.*, 2011). However, due to different connectivity of secondary structure elements, they have been classified under a different clan (clan PE) in MEROPS database instead of classifying under clan PB as an Ntn-hydrolase.

1.9 Scope of the thesis

In the work presented in this thesis, an attempt has been made to understand the *sequence-structure-function* relationship among enzymes of Ntn-hydrolase superfamily. We have selected two families belonging to Ntn-hydrolase superfamily having broad range of industrial applications as well as clinical importance, namely **cholyglycine hydrolase (CGH)** family, belonging to NtCn-hydrolase superfamily, and **penicillin G acylase (PGA)** family, belonging to NtSn-hydrolase superfamily. Cholyglycine hydrolase family has been selected to study the *sequence-structure and substrate specificity* relationship while penicillin G acylase family has been selected to study *sequence-structure and stability* relationship among the family members. The work on CGH family has been described in **Chapter 2**; while **Chapter 4** and **Chapter 5** describe the work on PGA family. Besides, a web server has also been developed which automates the analysis of a large number of structure-based features that are known to influence protein stability. The development and implementation of the web server is described in **Chapter 3**.

1.9.1 Study of sequence-structure & specificity relationship in CGH family

Bile salt hydrolases (BSH) and **penicillin V acylases (PVA)**, designated by their respective substrate preference, are two closely related class of pharmaceutically important enzymes belonging to cholyglycine hydrolase (CGH) family and thus to Ntn-hydrolase superfamily. Their presence is reported in bacteria and archaea. Despite structural similarity, positional preference of groups surrounding the nucleophile atom and sequentially conserved amino acid residues participating in catalysis, BSH and PVA generally show varied preference for one of the two chemically distinct molecules as substrate, *bile salts* on one hand and *penicillin V* on the other. BSH and PVA enzymes have been shown to play a crucial role in deconjugation of bile salts and synthesis of β -lactam antibiotics, respectively. Study of the evolution of these enzymes along with better understanding of their substrate specificity variations will not only improve the current annotations of the family members but also provide better insights into their biological function.

1.9.1.1 Bile Salt Hydrolases, their physiological role and clinical importance

Bile Salt Hydrolases are enzymes secreted by gut microbes in response to bile salt toxicity (Begley *et al.*, 2006). They catalyze the deconjugation of *conjugated bile salts* (Fig. 1.11), one of the major components of bile. Bile is a predominant digestive secretion that aids in emulsification and solubilization of lipids in small intestine. It is produced in liver, stored in gall bladder and released into duodenum when food is ingested. Conjugated bile salts constitute approximately 50% of the organic components of bile. Bile salts are synthesized in liver from cholesterol by multi-enzyme process which involves *bile acid* synthesis followed by its conjugation with either glycine or taurine to form glycine- or taurine-conjugated bile salts (Johnson, 2003). Bile acids are steroid molecules composed of three six membered rings (Ring A, B and C) fused to a five-membered ring (Ring D) which is further attached to a five- or eight-carbon side chain with a terminating carboxylic acid (Fig. 1.11). All bile acids have one or more hydroxyl groups oriented in either β -orientation (up) or α -orientation (down). The bile acid is conjugated with an amide bond to one of the two amino acids, glycine and taurine, at carboxyl C24 position to form glycine- and taurine-conjugated bile salts (Fig. 1.11). This terminal step of conjugation is catalyzed by the enzyme amino acid N-acyltransferase (Johnson *et al.*, 1990).

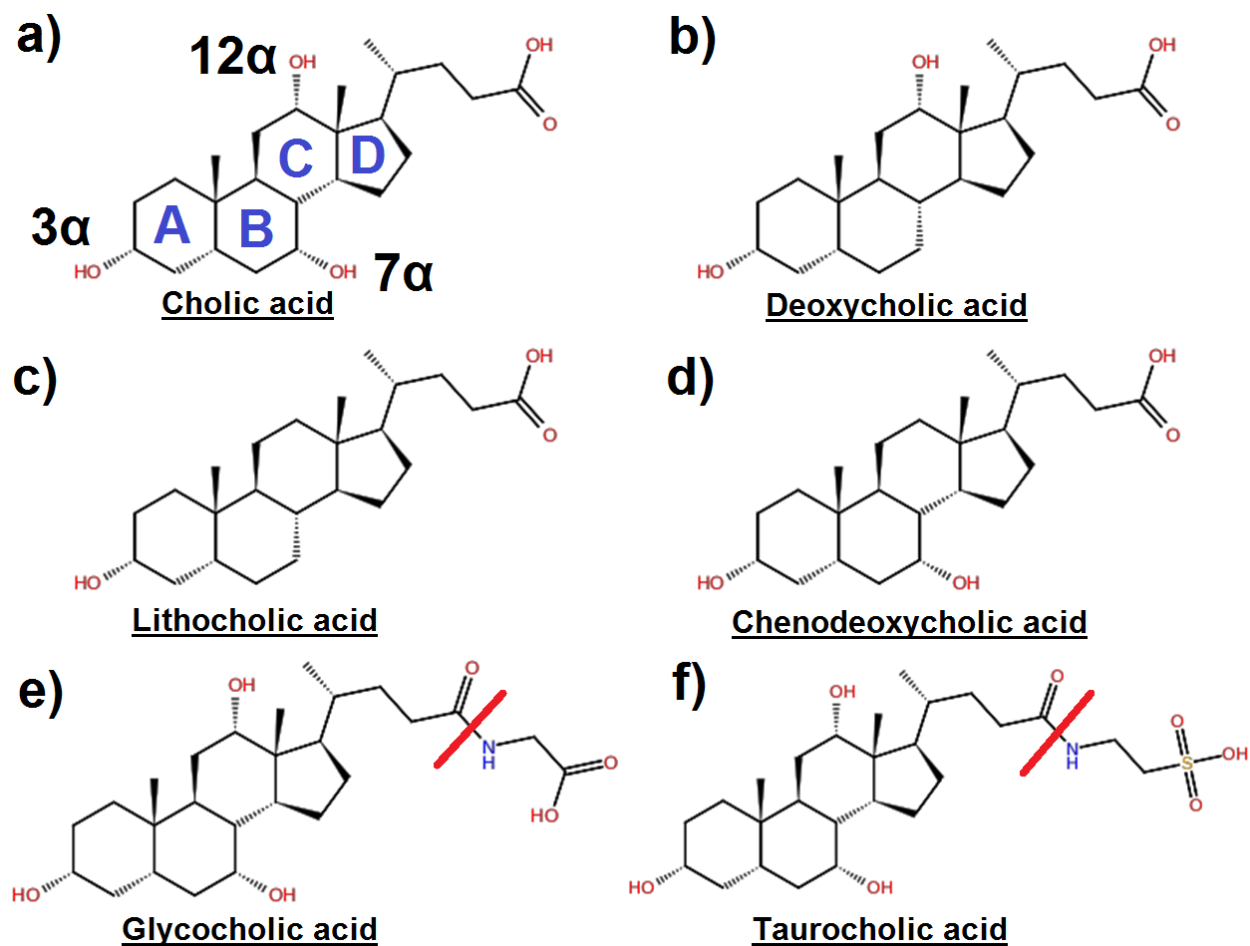


Figure 1.11: Illustrates different bile acids (a-d) and glycine or taurine conjugated bile salts (e-f). Bile salts are glycine or taurine conjugates of bile acids. Bile acids have one or more hydroxyl groups at 3 α , 7 α and 12 α positions. Experimental evidence suggests the importance of the three hydroxyl groups on their binding affinity. The four rings (A, B, C and D) of steroid moiety of bile acids are also shown in panel a. Bile Salt hydrolase (BSH) enzymes cleave the amide bond (marked in panel e and f) of bile salts to yield bile acids and glycine/taurine.

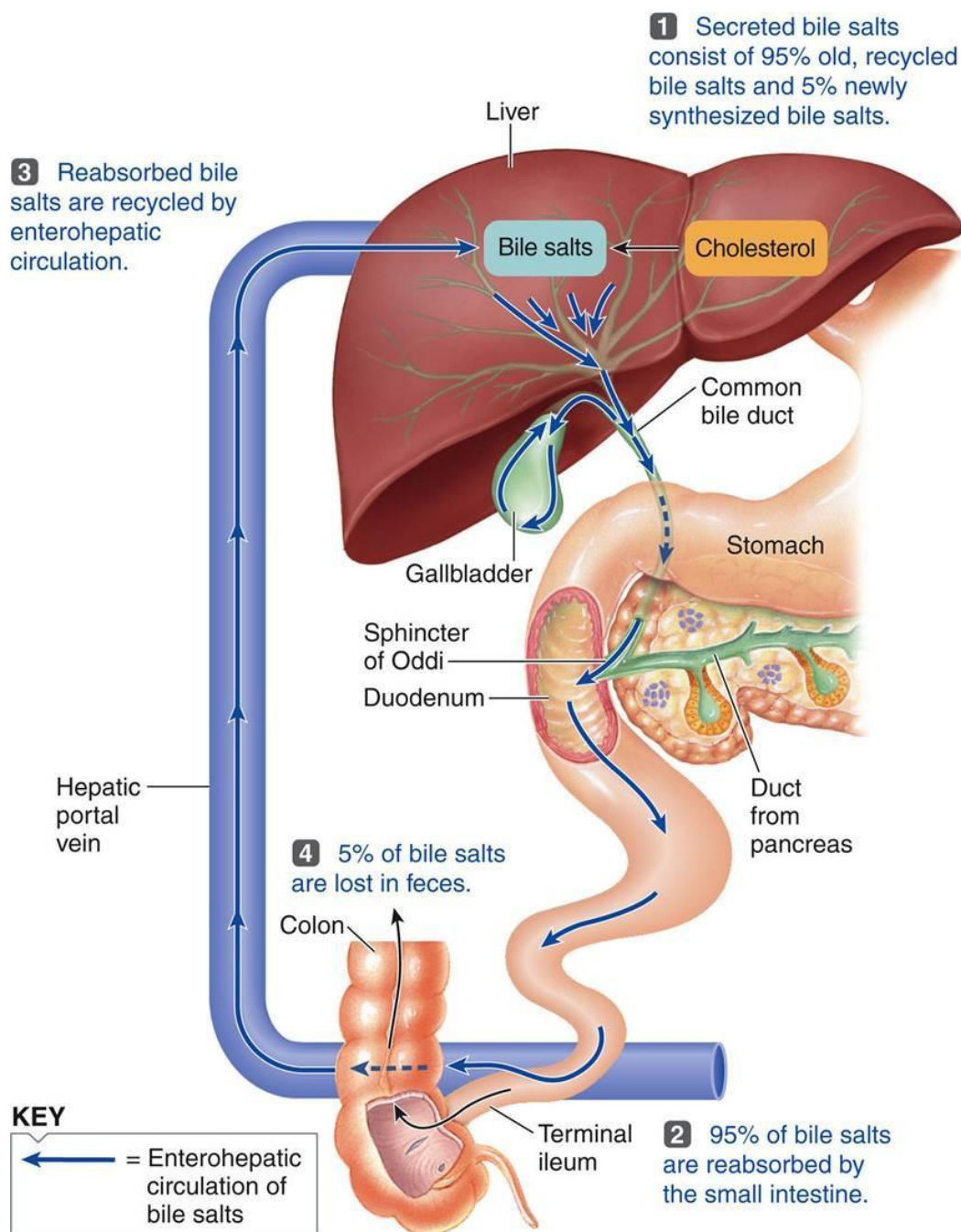


Figure 1.12: Illustrates the cholesterol homeostasis inside body. Adapted from www.gestaltreality.com/.

When food is ingested, these conjugated bile salts molecules are secreted to duodenum in the form of bile through bile duct. Conjugated bile salts being amphipathic in nature, forms spontaneous micelles that trap dietary cholesterol and fats, and break down or emulsify them into microscopic droplets. Emulsification increases the surface area of lipid molecules facilitating the lipase action. Once the emulsification is over, they are reabsorbed back to enterohepatic

circulation so that the cholesterol homeostasis can be maintained. Conjugated bile salts being comparatively more bulky, soluble and amphipathic than deconjugated bile acids, they are impermeable to cell membranes and therefore do not get absorbed in the proximal small intestine (Fig. 1.12). Instead they passed to distal ileum where they are actively reabsorbed by ileum bile acid transporters (IBAT) and ABC family of transporters. Bile salts due to their anti-microbial properties provide threat to gut microflora. The gut microbes thus secrete **BSH** enzymes which deconjugates (Drasar *et al.*, 1966) these conjugated bile salts to bile acids and free amino acids (glycine or taurine). Bile acids have less affinity for IBAT transporters and therefore passed into large intestine or cecum. In cecum, majority of bile acid transformation occurs such as 7-dehydroxylation resulting the conversion of cholic acid or chenodeoxycholic acid to deoxycholic acid and lithocholic acid, respectively (Fig. 1.11) which are reabsorbed to enterohepatic circulation. Few bile acids escape absorption and are excreted in faeces. This way BSH enzyme benefits the host for maintenance of serum cholesterol level and at the same time benefits the microbe by protecting them from bile salt toxicity. BSH enzymes are conceptualized as a competitor of IBAT for conjugated bile salt molecules.

BSH activity has been widely detected among all major divisions of gut-inhabiting bacteria (Gram-positive and Gram-negative) as well as archaea (Jones *et al.*, 2008). In human gut approximately 30% of BSH-active members belong to *Firmicutes* while the distribution is 14.4 and 8.9% amongst *Bacteroides* and *Actinobacteria*, respectively. Gut-inhabiting archaea such as *Methanobrevibacter smithii* and *Methanosphaera stadtmanae* were also observed to show BSH activity (Jones *et al.*, 2008). Interestingly BSH activity is also detected among Gram-positive gastrointestinal pathogen such as *Listeria monocytogenes* and *Enterococcus faecalis* (Begley *et al.*, 2006).

Although the exact physiological role of BSH enzymes is still not clear, they are thought to benefit the microbe with respect to bile detoxification, gastrointestinal persistence, nutritional role and membrane alterations (Begley *et al.*, 2006). Similarly in the host, they could be related to cholesterol lowering, formation of gallstones, activation of carcinogens and altered digestive functions like lipid malabsorption and weight loss (Begley *et al.*, 2006). Hypercholesterolemia is often linked to cardiovascular diseases (Levine *et al.*, 1995), thus lowering serum cholesterol level could lead to reduced chance of cardiovascular diseases. The plasma cholesterol level can

be regulated either by reducing cholesterol biosynthesis from dietary food or by triggering more excretion of bile acids in feces. Cholesterol lowering through administration of expensive drugs such as fibrates, nicotinic acid, bile acid sequestrants, and statins has many side-effects (Hay *et al.*, 1999; Kolata & Andrews, 2001). So, an alternative approach of utilizing BSH active probiotics is found to be more promising (Begley *et al.*, 2006). Oral administration of live bacterial cell therapy can lower serum cholesterol level by 22 to 33% (Jones *et al.*, 2004; Lim *et al.*, 2004). Microencapsulated *Lactobacillus plantarum* 80 (pCBH1) has been used for reduction of serum cholesterol level (Jones *et al.*, 2004). Several hypotheses have been proposed to explain the mechanism of reduction of cholesterol level by BSH-active probiotics such as co-precipitation, assimilation and enzymatic hydrolysis of conjugated bile acid (Begley *et al.*, 2006). However, the suspected role of BSH in certain intestinal disorders such as gallstones (Thomas *et al.*, 2000) and colorectal cancer (Singh *et al.*, 1997) are of concern when utilizing BSH for cholesterol control. Deconjugation of bile salts is thought to cause gall stones. They are shown to reduce the growth of chicken due to poor absorption of lipids in small intestine and are suspected to result in colorectal cancer. Reports also hint that BSH contributes to virulence factor of virulent strains in *Listeria monocytogenes* (Dussurget *et al.*, 2002). Even though BSH enzymes are widely present among many enteric bacteria, the exact physiological role of these enzymes in bacterium as well as host is still not clear. Although the precise effect of the enzymatic products of BSH on mammalian host cells is not fully deciphered at present, the enzyme has considerable pharmaceutical importance.

1.9.1.2 Penicillin V acylases, their physiological role and pharmaceutical importance

Penicillin V acylases are enzymes employed in the commercial manufacture of 6-aminopenicillanic acid (6-APA), the precursor for a large variety of semi-synthetic β -lactam antibiotics (Rathinaswamy *et al.*, 2012). β -lactam antibiotics are a class of antibiotics that can either kill bacteria (bactericidal) or arrest their growth (bacteriostatic) by inhibiting bacterial cell wall synthesis (Abraham, 1981; Demain *et al.*, 1983). Based on the structure of the core nucleus, β -lactam antibiotics can be classified (Fig. 1.13) as

- **Penicillin:** On the basis of source, penicillin can be classified further into
 - a. Natural penicillins like Penicillin G, Penicillin V, Penicillin F and Penicillin K.
 - b. Semi-synthetic penicillins such as Amoxicillin, Cloxacillin and Methicillin.

- **Cephalosporin** (Generations I, II, III, IV and Cephamycine. Together these are known as cephems).
- **Carbapenems** (Imipenem and Meropenem)
- **Monobactams** (Aztreonam)
- **β -lactam inhibitors** (Clavulanic acid and Sulbactam)

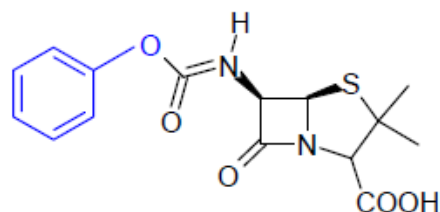
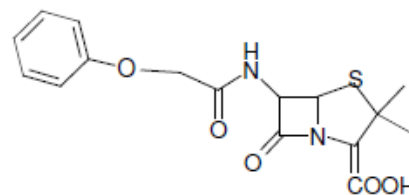
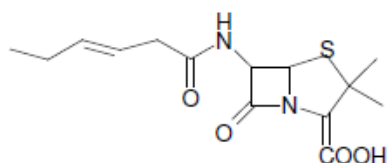
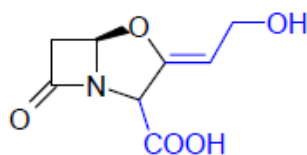
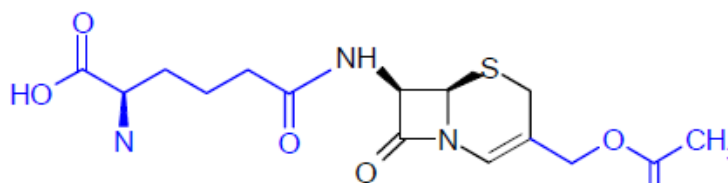
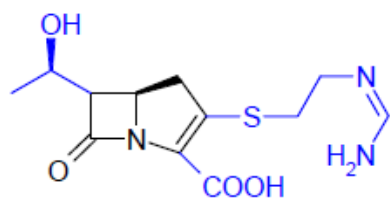
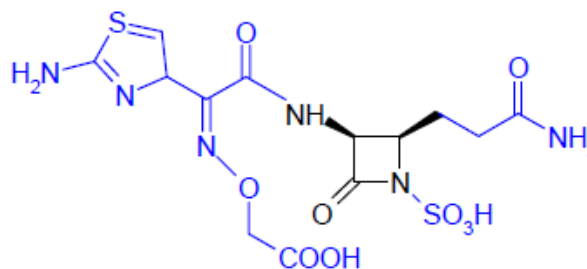
**Penicillin G (6-APA)****Penicillin V****Penicillin F****Penicillin K****Clavulanic acid (Clavam)****Cephalosporin C (Cephem)****Imipenem (Carbapenem)****Carumonam (Monobactam)**

Figure 1.13: Structures of various β -lactams showing different core nuclei (colored black and indicated in bracket) and side chains (blue).

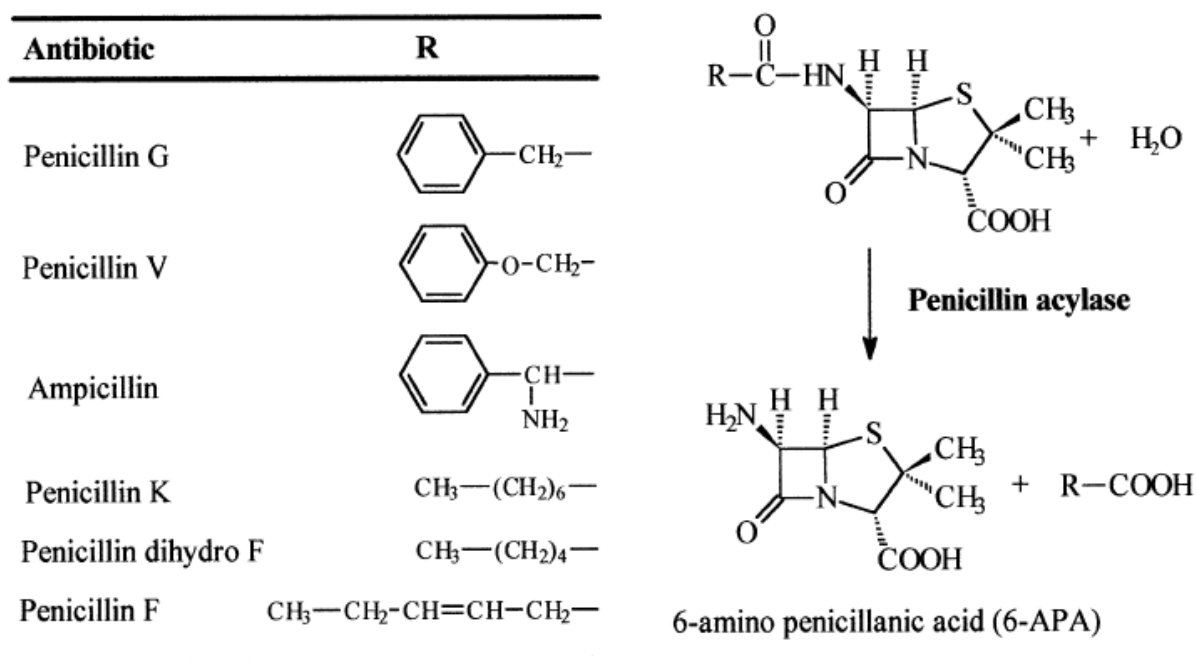


Figure 1.14: Penicillin catalyzed reaction leads to 6-APA production. Adapted from Arroyo *et al.*, 2003.

In case of penicillins, the basic component is a core β -lactam nucleus which is formed by the fusion of a 4-membered β -lactam ring with thiazolidine ring to form 6-aminopenicillanic acid (6-APA). In case of cephalosporins, the core nucleus is formed by the fusion of β -lactam ring with a six carbon ring to produce 7-amino cephalosporanic acid (7-ACA). Variation in the side chain results in different types of penicillins and cephalosporins. In case of penicillin V, the side chain is a phenoxyethyl group while in case of penicillin G, the side chain is a benzyl group (Fig. 1.13 and Fig. 1.14). Since pathogens developed resistance against natural penicillins, there was a great demand for novel synthetic antibiotics with superior bactericidal effect. It was observed that variations in side chain in natural penicillins alters the antibiotics properties, thus was considered as a strategy to produce semi-synthetic antibiotics (Abraham, 1981; Vandamme & Voets, 1974). Semi-synthetic β -lactam antibiotics were proved to be more effective against resistant pathogens compared to natural antibiotics. For preparing semi-synthetic antibiotics, a multi-step reaction has to be followed involving first the cleavage of natural penicillin molecule to yield 6-APA followed by the reverse reaction of rejoining 6-APA with the desired side chain. Chemical route for these steps is not only expensive but also generates several by-products. Alternatively by following clean and safest route, penicillin acylases were employed to cleave natural penicillin to produce 6-APA and the same enzymes were reused to catalyze reverse

reaction under different conditions (Shewale *et al.*, 1987; Vandamme & Voets, 1974). Penicillin V acylase act on the substrate penicillin V to yield 6-APA and phenoxy acetic acid as products of which 6-APA is used further for synthesis of other semi-synthetic antibiotics by varying side chains. Penicillin G acylases, distant homologs of penicillin V acylase enzymes, act on penicillin G as substrates to produce 6-APA (Fig. 1.14).

Penicillin V acylases have been reported among soil and aquatic microbes such as *Escherichia coli*, *Streptomyces lavendulae*, *Streptomyces ambofaciens*, *Pseudomonas acidovorans*, *Pseudomonas diminuta*, *Bacillus subtilis*, *Erwinia aroideae*, *Beijerinckia indica*, *Arthrobacter sp.* and *Bacillus sphaericus*. Among yeasts, *Candida*, *Rhodotorula*, *Giberella*, *Penicillium*, *Fusarium*, *Torula*, *Trichosporon* and *Saccharomyces* are examples of PVA producing ones (Ambedkar *et al.*, 1991; Batchelor *et al.*, 1959; Carlsen & Emborg, 1982; Cole, 1964; Rathinaswamy *et al.*, 2012; Vandamme & Voets, 1975). Recently PVA has been reported also in a plant pathogen *Pectobacterium atrosepticum* (Avinash *et al.*, 2013). The exact physiological role of PVA enzymes is still not clear, but evidence suggests that *pva* genes play important role in assimilation of aromatic compounds as carbon source (Valle *et al.*, 1991). It has been suggested that penicillin acylases take part in degradation of phenoxyacetylated compounds to produce phenoxy acetic acid as carbon source as well as an inducer of degradation pathway. In *E. coli*, penicillin acylase gene was observed to be located near aromatic hydroxylate encoding gene (Prieto *et al.*, 1993). Later it was further discovered that the gene cluster involved in 4-hydroxyphenylacetic acid degradative pathway is located very close to penicillin acylase encoding gene (Prieto *et al.*, 1996). Although the aromatic degradation pathway for carbon source is of little importance when *E. coli* lives as parasite, the pathway becomes very essential when *E. coli* moves into a free-living nonparasitic state in soil (Burlingame & Chapman, 1983).

The expression of penicillin V acylase in *Vibrio cholera* hinted its possible role in pathogenesis. AphA, the activator of a virulence operon has been recognized as a negative regulator of *pva* gene (Kovacikova *et al.*, 2003). It has a second binding site, virtually identical to a promoter, overlapping PVA transcriptional start site. A higher level of AphA has been observed to repress PVA expression and activation of tcpPH expression resulting in activation of virulence cascade. In El Tor strain C6706 of *V. cholera*, the PVA gene is also regulated through quorum sensing; PVA expression is reduced when cell density is low (Kovacikova *et al.*, 2003).

1.9.1.3 BSH and PVA: sequence and structural homology and difficulty in their distinction

Of the many experimentally characterized BSH and PVA enzymes, three-dimensional structures have been solved for BSH from *Bifidobacterium longum* (*B*/BSH), *Clostridium perfringens* (*Cp*BSH), *Bacillus sphaericus* (*Bsp*PVA), *Bacillus subtilis* (*Bsu*PVA) and *Bacteroides thetaiotaomicron* (*Bt*BSH). Except *Bt*BSH, all enzymes are from Gram-positive bacteria; *Bt*BSH is a Gram-negative BSH enzyme. Like all Ntn-hydrolase enzymes, BSH and PVA enzymes are also produced as inactive pro-enzymes. In case of *B*/BSH, *Cp*BSH and *Bsu*PVA, the activation of enzymes involve removal of initiator methionine residue (Table 1.2) whereas in case of *Bsp*PVA and *Bt*BSH, activation requires autocatalytic removal of 3 and 25 residues pre-peptide sequence, respectively. The N-terminal nucleophile residue is a cysteine (Cys2) which is accompanied by other catalytic residues such as Arg18, Asp21, Asn82, Asn175 and Arg228 (sequence numbering as per *Cp*BSH structure). These residues are conserved among all BSH and PVA enzymes. However, at the position corresponding to Asn82 of BSH enzymes, PVAs have Tyr residues. Using quantum mechanics/molecular mechanics free energy simulations, Lodola *et al.*, 2012, have determined the reaction mechanism of BSH and PVA catalyzed hydrolytic reactions. The study demonstrated the existence of a chair-like transition state. The Arg18 and Asp21 are important in forming hydrogen bonding interactions with the alpha-amino and sulfhydryl groups of Cys2 while Asn82 and Asn175 forms putative oxyanion hole.

Table 1.2: Sequence and structural similarity of BSH and PVA enzymes.

BSH/PVA	Mature enzyme sequence length	Pre-peptide Length	Blast2Seq with <i>B</i> /BSH		jFATCAT_flexible alignment with <i>B</i> /BSH	
			% Identity	%Similarity	Chain RMSD	Alignment Score
<i>B</i> /BSH (2HF0)	316	1	-	-	-	-
<i>Cp</i> BSH (2RLC)	328	1	35	52	1.75	811.63
<i>Bsp</i> PVA (3PVA)	335	3	28	48	1.83	800.67
<i>Bsu</i> PVA (2OQC)	327	1	29	49	1.59	767.95
<i>Bt</i> BSH (3HBC)	317	25	21	34	2.28	566.84

BSH and PVA enzymes share significant degree of sequence and structural similarity which makes their differentiation very difficult. A pair-wise sequence alignment by Blast2seq (Altschul *et al.*, 1997) program showed high degree of sequence similarity between *Bt*BSH and other BSH and PVA sequences (Table 1.2). Overall the BSH and PVA enzymes were observed to be more than 35% similar to each other. Gram-positive enzymes are more similar to each other compared to *Bt*BSH, a gram-negative BSH enzyme. The high degree of sequence similarity is also reflected in their structural similarity. When a flexible structural alignment was carried out using jFATCAT program (Ye & Godzik, 2003) the enzymes showed low RMSD values amongst each other.

Due to such high degree of sequence and structural homology, BSH and PVA enzymes have been grouped together under single family in available sequence and structure-based protein classification databases, which makes their functional differentiation very difficult. Owing to the medical importance associated with both BSH and PVA enzymes and looking at the rate at which genomes are sequenced and annotated, a high-resolution accurate sequence-based annotation method is very essential. An improved sequence-based method for substrate specificity annotation of CGH family members has been developed. The development and validation of this method has been described in **Chapter 2**. It has been suggested that BSH and PVA enzymes are evolutionary related. However, it is still not clear how the enzymes have evolved and diverged among bacteria and archaea. The evolution and physiological role of CGH family members are also discussed in Chapter 2.

1.9.2 Study of sequence-structure & stability relationship through development of iRDP web server

Engineering protein molecules with modified structure and biological function has always been a challenging problem. Rational Design of Proteins, one of the classical protein engineering strategies is a knowledge-guided process (Antikainen & Martin, 2005) widely implemented to improve the biochemical or biophysical properties of proteins such as stability (Bjørk *et al.*, 2004), altered substrate specificity (Schwarz *et al.*, 2001), catalytic activity (Mata *et al.*, 1999) as well as to design enzymes with novel or multiple functions (Béguin, 1999). The rational approach to protein designing usually involves site-directed mutagenesis (SDM) of specific residues in a protein based on available information to obtain desired changes in protein

function or properties. Since protein function is intimately linked to three-dimensional structure, the mutagenesis of one or more amino acids in protein not just alters the sequence context but also affects its structural topology (Anfinsen, 1973). The rapidly growing numbers of protein structures in the PDB and advances in homology modeling have helped to critically assess the protein structure-function relationships as well as locate key residues at the active sites, domain interfaces or hinge regions, making it increasingly possible to design proteins having the altered properties and functions. However, identification of the key residues responsible for desired changes often requires a detailed analysis of large numbers of protein structures which is time-consuming and cumbersome if carried out manually. Therefore iRDP (*in silico* Rational Design of Proteins) web server was developed, available at <http://irdp.ncl.res.in>, which aims to simplify the laborious task of exploring the vast structural space, analysis of which forms the basis of any rational protein design problem. **Chapter 3** of the thesis describes the development and implementation of the server along with different case studies for its validation.

1.9.3 Study of sequence-structure & stability relationship in PGA family.

Like penicillin V acylases (PVA) of NtCn-hydrolase superfamily, **penicillin G acylases (PGAs)** of NtSn-hydrolase superfamily are another class of hydrolytic enzymes that are widely used for commercial production of semi-synthetic antibiotics (Srirangan *et al.*, 2013). Unlike PVAs which prefer penicillin V as substrates, PGAs prefer penicillin G as substrate for the production of 6-APA, which is the starting raw material for synthesis of other semi-synthetic penicillins (Fig. 1.14). Penicillin V shows superior stability compared to penicillin G in aqueous solution at lower pH and thus could result in higher yield of 6-APA (Shewale & Sudhakaran, 1997). PVA enzymes have broader range of pH optimum compared to PGAs and therefore reduce the buffer requirement during hydrolysis. PVAs also show higher conversion rate at higher substrate concentrations compared to PGAs. Despite having these observed advantages of PVAs compared to PGAs, only 15% of world-wide 6-APA synthesis uses PVA enzymes as 6-APA source. However, if both acylases are considered together, an estimate of 9000 tons of 6-APA is enzymatically produced from penicillin V and penicillin G (Bruggink & Roy, 2001). PGA enzymes have been characterized from *Escherichia coli* (EcPGA), *Kluyvera citrophila* (KcPGA), *Providencia rettgeri* (PrPGA), *Arthrobacter viscosus* (AvPGA), *Bacillus megaterium* (BmPGA), *Alcaligenes faecalis* (AfPGA) and *Achromobacter xylosoxidans* (AxPGA), (Cai *et al.*,

2004; Duggleby *et al.*, 1995; Martin *et al.*, 1991; Martin *et al.*, 1995; McDonough *et al.*, 1999; Ohashi *et al.*, 1988; Verhaert *et al.*, 1997).

PGAs have also been employed for synthesis of peptides and their derivative which find great value as food additives (van Langen *et al.*, 2000). *Ec*PGA has been used for kinetically controlled synthesis of chiral dipeptides of phenylglycine such as D-phenylglycyl-L-phenylglycine and L-phenylglycyl-L-phenylglycine methyl esters which further undergo cyclization to corresponding diketopiperazines. These diketopiperazines are used as chitinase inhibitors, food additives and synthons for antiviral, fungicidal and anti-allergic compounds (van Langen *et al.*, 2000). The nucleophile binding site of PGAs is specific to L-isomers. This characteristics feature simplifies the chiral dipeptide synthesis via enzymatic route. The advantage of using PGAs compared to the chemical synthetic route is the reduced degree of spontaneous degradation of the synthesized dipeptide esters during enzymatic synthesis. The chemical synthesis of peptides requires both protection and activation of donor and acceptor groups whereas PGA mediated synthesis requires either simple protection or no protection at all (van Langen *et al.*, 2000). One of the classic examples of use of this strategy of using PGAs for peptide synthesis is synthesis of artificial sweetener aspartame (Fuganti *et al.*, 1986).

Another useful application of PGAs is the use of these enzymes for resolving the racemic mixtures of chiral compounds such as β -amino esters, secondary alcohols, amines, and amino acids in aqueous medium (Arroyo *et al.*, 2003). The resolved pure forms of enantiomers are often used for synthesis of other biologically active compounds. For instance, *Ec*PGA has been utilized as a biocatalyst to resolve ethyl 3-amino-4-pentynoate (R)- and (S)-enantiomers into pure form. The S-isomer is a chiral synthon that is used in the synthesis of an anti-platelet agent xemilofiban hydrochloride (Topgi *et al.*, 1999). Another important attempt of utilizing PGAs has been towards enantioselective acylation of a β -lactam intermediate in the synthesis of loracarbef, a carbacephalosporin antibiotics and a Cefaclor analogue (Zmijewski Jr *et al.*, 1991).

Although the PGA enzymes have immense industrial application they suffer from the limitation that the native soluble forms of enzymes show lesser stability under different reaction conditions such as temperature, pH and presence of organic solvents. Thus the enzymatic manufacture of 6-APA is usually carried out by employing immobilized forms of PGAs (Arroyo *et al.*, 2003). The enzyme immobilization not only improves the stability of enzyme during

conventional handling but also helps in easy separation of enzymes from products and the immobilized enzymes can be reused. Besides reducing the manufacturing cost, the immobilization of enzyme on a solid support has also been shown to modify the catalytic properties of enzymes. Several immobilization procedures have been studied for PGA enzymes for enhancement of its stability and catalytic properties (Arroyo *et al.*, 2003). For instance, Rocchietti *et al.*, 2002, have improved the enantioselectivity of PGAs during the hydrolytic resolution of racemic esters and amides of mandelic acids by using different binding technique of PGAs on matrix such as Eupergit C and agarose gel. They have observed a dependency of catalytic property of PGAs with enzyme source, immobilization technique and the substrate. They have suggested that immobilization of PGAs on different supports by varying the binding orientation and rigidity, one can modulate the catalytic property of enzymes (Rocchietti *et al.*, 2002). Katchalski-katzir *et al.*, 2000, have shown that the covalent attachment of PGAs on epoxy-activated commercial acrylic beads such as Eupergit C, leads to an increase in operational stability of enzymes (Katchalski-Katzir & Kraemer, 2000). Torres-Bacete *et al.*, 2000, have immobilized PVA from *Streptomyces lavendulae* on Eupergit C which enhanced the temperature and pH stability of enzyme (Torres-Bacete *et al.*, 2000). PGAs immobilized on Sephabeads-EP, a new epoxy-activated sephabeads, have shown improved stability compared to Eupergit C immobilized PGAs (Mateo *et al.*, 2002). Recently PGAs have also been immobilized by formation of enzyme-fatty lipid biocomposite film (Phadtare *et al.*, 2002). Whole-cell PGA immobilized derivatives were also attempted which compete well with previous immobilization techniques. Several strains of *E. coli* containing PGA enzymes have been trapped within gluten matrix, gelatin matrix and polymethacrylamide beads (Arroyo *et al.*, 2003).

Another breakthrough in antibiotics manufacturing industry was the development of CLECs (cross-linked enzyme crystals) and CLEAs (cross-linked enzyme aggregates) which are shown to improve stability of cross-linked enzymes. CLECs are prepared by crystallization followed by cross-linking of the enzymes with glutaraldehyde (Margolin, 1996) while CLEAs are produced by the aggregation of enzymes under denaturing conditions followed by glutaraldehyde cross-linking (Cao *et al.*, 2000). SynthaCLEC-PA, a cross-linked CLEC of *E. coli* PGA enzyme was observed to improve the stability of enzyme in organic solvents. CLEAs were observed to be more efficient compared to CLECs during the kinetically controlled ampicillin synthesis in both aqueous and other polar and non-polar organic solvents. Erarslan *et al.*, 1992,

had tried to improve stability of *Ec*PGA by glutaraldehyde cross-linking, however both cross-linked and wild-type *Ec*PGA showed 40-50% denaturation upon 30 min of incubation at 45 °C. Complete loss of activity was observed in both cases upon 30 min of incubation at 50 °C (Erarslan & Kocer, 1992).

Despite the above approaches towards improving the stability of enzymes under different conditions, the non-native form of enzymes shows lower turn-over rate compared to soluble enzymes. Thus alternate routes of finding novel source of PGA enzymes with improved stability property such as higher thermostability and pH stability were also attempted. In this direction one of the first successful attempts was the purification and characterization of PGA from *Alcaligenes faecalis* (*Af*PGA) which showed superior thermostability property compared to the most popular *Ec*PGA (Verhaert *et al.*, 1997). Another potential source of thermostable PGA enzyme identified was from *Achromobacter xylosoxidans* (*Ax*PGA). This enzyme is reported to be the most-thermostable PGA enzyme known so far (Cai *et al.*, 2004). The half-life of inactivation at 55 °C is four times longer compared to *Af*PGA. Identification of novel sources of PGA enzymes having a longer half-life under reaction conditions will always be beneficial. In a very recent attempt towards identification of novel sources of thermostable enzyme, a penicillin acylase from an extreme thermophile *Thermus thermophilus* *HB27* was identified which was observed to have a half-life of 9.2 hr at 75 °C. However, the enzyme's preference was for penicillin K, an octanoyl-penicillin (Fig. 1.13) and thus was named as penicillin K acylase (Torres *et al.*, 2012).

Besides stability at higher temperatures, pH stability is also an equally important parameter for a PGA enzyme to be most suitable in industry. An enzyme having broad range of pH stability will not only be useful during normal handling of the enzyme but also during other synthetic applications. Most PGA enzymes so far characterized show pH optimum and stability range between pH 5 to pH 8. In a recent attempt towards making alkaline stable PGA enzyme, Suplatov *et al.*, 2014, have carried out β Asp484→Asn mutation which showed a 9 fold increase in stability of *Ec*PGA at pH 10 (Suplatov *et al.*, 2014).

Owing to the pharmaceutical importance associated with the penicillin G acylases and the need for novel sources of potentially thermostable enzymes, we have attempted a hybrid approach. Initially a computational analysis using iRDP web server was carried out to filter few

PGA enzymes from various available putative PGA sources. Next experimental characterization was carried out for one of the potentially stable PGAs, namely PGA from *Paracoccus denitrificans* (*Pd*PGA) which exhibited features of a thermostable PGA enzyme. The computational approach followed has been described in **Chapter 4** while the experimental characterization of *Pd*PGA has been depicted in **Chapter 5**.

In summary **Chapter 2** describes the computational method developed for improved substrate specificity annotation of CGH family members, **Chapter 3** describes the development and implementation of iRDP web server, **Chapter 4** describes the computational approach that was followed towards identification of potential sources of thermostable PGA enzymes, **Chapter 5** describes the purification and characterization of *Pd*PGA enzyme and finally **Chapter 6** summarizes the results and findings of the present thesis.

1.10 Tools and techniques used in this study.

Chapter 2 includes several computational methods such as sequence alignment, phylogenetic analysis, docking, molecular dynamics simulations and binding site similarity analysis. The sequence alignment among family members has been carried out using ClustalX (Thompson *et al.*, 1997) while phylogenetic analysis has been carried out using Mega5 (Tamura *et al.*, 2011). The program Glide (Version 5.8, Schrödinger, LLC, New York, NY, 2012) has been used to study receptor-ligand binding by docking method while Gromacs 4.5 (Pronk *et al.*, 2013) was used to carry out molecular dynamics simulations of the enzyme complexes. In-house Perl script was written to obtain binding site similarity based scoring system. **Chapter 3** which deals with the development of iRDP web server has been developed using R, Perl, PHP and HTML. The web server has been hosted at the institute CSIR-National Chemical Laboratory and is freely available to public users at <http://irdp.ncl.res.in>. **Chapter 4** employs sequence and structure based approaches involving Gromacs 4.5 for molecular dynamics simulations, Prime (Version 3.1, Schrödinger, LLC, New York, NY, 2012) for homology modeling and iRDP web server for detection of various intra-molecular interactions. **Chapter 5** includes many microbiological, biochemical and biophysical methods such as cell culturing, protein expression, purification using chromatography techniques, enzyme assay, temperature, pH stability profile, fluorescence and CD spectroscopy.

Chapter 2

*Development of a sequence-based
substrate specificity annotation method
for the NtCn-hydrolase enzymes
belonging to
Cholyglycine hydrolase family
and study of their evolution*

Cholylglycine hydrolase (CGH) family, belonging to **NtCn-hydrolase** enzyme superfamily includes two pharmaceutically important classes of enzymes namely bile salt hydrolases (**BSH**) and penicillin V acylases (**PVA**). The correct annotation of such physiologically and industrially important enzymes is thus vital. Current methods relying solely on sequence homology do not always provide accurate annotations for these two members of the CGH family as BSH/PVA enzymes. Therefore we have developed an improved method [**binding site similarity (BSS)-based scoring system**] for the correct annotation of the CGH family members as BSH/PVA enzymes, which is described in this chapter.

2.1 Introduction

Although BSH and PVA enzymes prefer chemically distinct substrates (bile salts and penicillin V, respectively), both cleave a similar amide bond in their substrates (Fig. 2.1). BSH enzymes catalyze the deconjugation of glycine- and taurine-conjugated bile salts (Glycocholic acid and Taurocholic acid, respectively) into bile acids and their corresponding amino acid residues. PVA enzymes hydrolyze the amide bond in penicillin V yielding phenoxy acetic acid (PAA) and 6-aminopenicillanic acid (6-APA) as products.

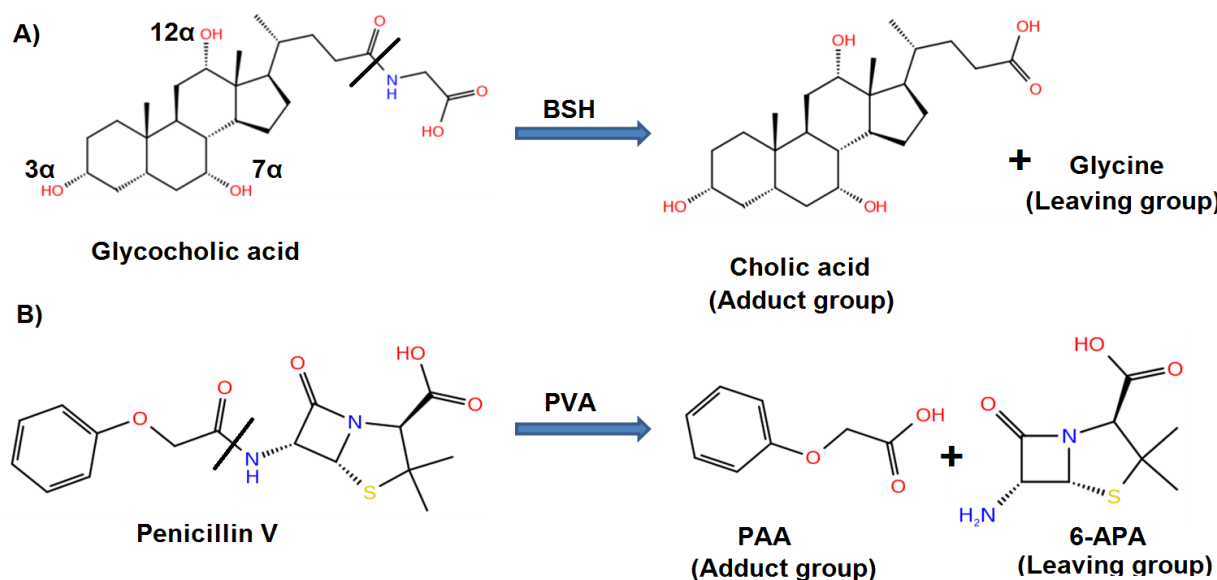


Figure 2.1: (A, B) Shown are the substrates glycocholic acid (GCA; a bile salt) and penicillin V (penV), respectively. The scissile amide bonds that are hydrolyzed by BSH and PVA enzymes are marked with a line. Upon cleavage of the amide bond, the group that is first released as product (leaving group) and the group

that remains bound to the enzyme as acyl-enzyme adduct (adduct groups) in each substrate are labeled. In case of GCA, glycine is the leaving group while cholic acid forms acyl-enzyme adduct. In case of penicillin V, 6-APA forms the leaving group while PAA remains bound to enzyme as acyl-enzyme adduct. The three polar hydroxyl groups (3 α , 7 α and 12 α -OH) of GCA are also labeled.

2.1.1 Difficulty in annotation of CGH enzymes as BSH/PVA and the need for an improved annotation method.

Both BSH and PVA enzymes contain the $\alpha\beta\beta\alpha$ Ntn-hydrolase fold. The catalytic residues Cys2, Arg18, Asp21, Asn175 and Arg228 are conserved, and therefore the mechanisms of hydrolytic reactions in both enzymes are similar (Lodola *et al.*, 2012). Kinetics and inhibition studies also show that members display a gradation of binding specificity and affinity towards bile salt and penV, rather than exclusively binding one of the molecules (Kumar *et al.*, 2006). Due to a high degree of similarity, they are annotated under a single family in public domain databases like **Pfam** (family CBAH), **CDD** (family Ntn_CGH_like) and **MEROPS** (family C59). Sometimes the family members are wrongly annotated, i.e. BSH enzymes annotated as PVA or vice versa. In the CDD database it is observed that the experimentally characterized BSH enzymes from *Bifidobacterium longum* and *Clostridium perfringens* have been annotated incorrectly as PVA enzymes (Ntn_PVA family). This issue was addressed previously by Lambert *et al.*, 2008, for Gram-positive BSH/PVA enzymes. Using phylogenetic profiling and molecular modelling, they could correctly annotate the BSH and PVA enzymes from Gram-positive bacteria. However, members from Gram-negative bacteria and archaea were not considered in their analysis (Lambert *et al.*, 2008). We extended the scope of analysis by developing an improved method for substrate specificity annotation of CGH family members including all members from Gram-positive bacteria, Gram-negative bacteria and archaea.

In the dataset used for the present analysis, we have incorporated experimentally characterized BSH members such as *B*/BSH, *Cp*BSH and *Bt*BSH, and PVA enzymes from *Bsu*PVA, *Bsp*PVA and *Pa*PVA (Table 2.1) as well as other uncharacterized BSH/PVA enzymes from Gram-positive bacteria, Gram-negative bacteria and archaea. The enzymes *B*/BSH, *Cp*BSH, *Bsp*PVA and *Bsu*PVA belongs to Gram-positive bacteria, whereas *Bt*BSH and *Pa*PVA are from Gram-negative bacteria. The entire list of sequences included in the analysis is given in Table 2.6.

The initial phylogenetic analysis failed to annotate the family members as BSH/PVA enzymes. This inaccuracy thus highlighted the need to develop a better annotation method not based solely on **phylogenetic information**, but also considering the **binding site characteristics** as well as **substrate specificity information** in order to improve the current annotations of the available sequences and to correctly annotate any new members (Fig. 2.2). Using the available structures, a comparison of the binding sites and prediction of substrate-binding modes was carried out by docking and molecular dynamics simulation studies. With the information generated by the above analysis, a binding site similarity (BSS)-based scoring system was developed which helped to annotate correctly CGH family members as BSH or PVA enzymes. The accuracy of annotation of the BSS scoring system was tested against 19 experimentally characterized CGH enzymes as well as the annotation provided previously by Lambert *et al.*, 2008. Lastly, we discuss the evolution of CGH family members and their relationship with the evolution of Gram-positive bacteria, Gram-negative bacteria and archaea.

Table 2.1: List of experimentally characterized BSH and PVA enzymes considered in the analysis. Among these six enzymes, CpBSH is a BSH enzyme with slight PVA activity while BspPVA is a PVA enzyme with slight BSH activity; other enzymes have either BSH or PVA activity.

Bacteria	Enzyme	Source	Label	PDB	Reference
Gram-positive	BSH	<i>Bifidobacterium longum</i>	B _l BSH	2HF0	(Kumar <i>et al.</i> , 2006)
		<i>Clostridium perfringens</i>	C _p BSH	2RLC	(Coleman & Hudson, 1995)
	PVA	<i>Bacillus sphaericus</i>	B _{sp} PVA	3PVA	(Olsson & Uhlen, 1986)
		<i>Bacillus subtilis</i>	B _{su} PVA	2OQC	(Rathinaswamy <i>et al.</i> , 2012)
Gram-negative	BSH	<i>Bacteroides thetaiotaomicron</i>	B _t BSH	3HBC	(Stellwag & Hylemon, 1976)
	PVA	<i>Pectobacterium atrosepticum</i>	P _a PVA	model	(Avinash <i>et al.</i> , 2013)

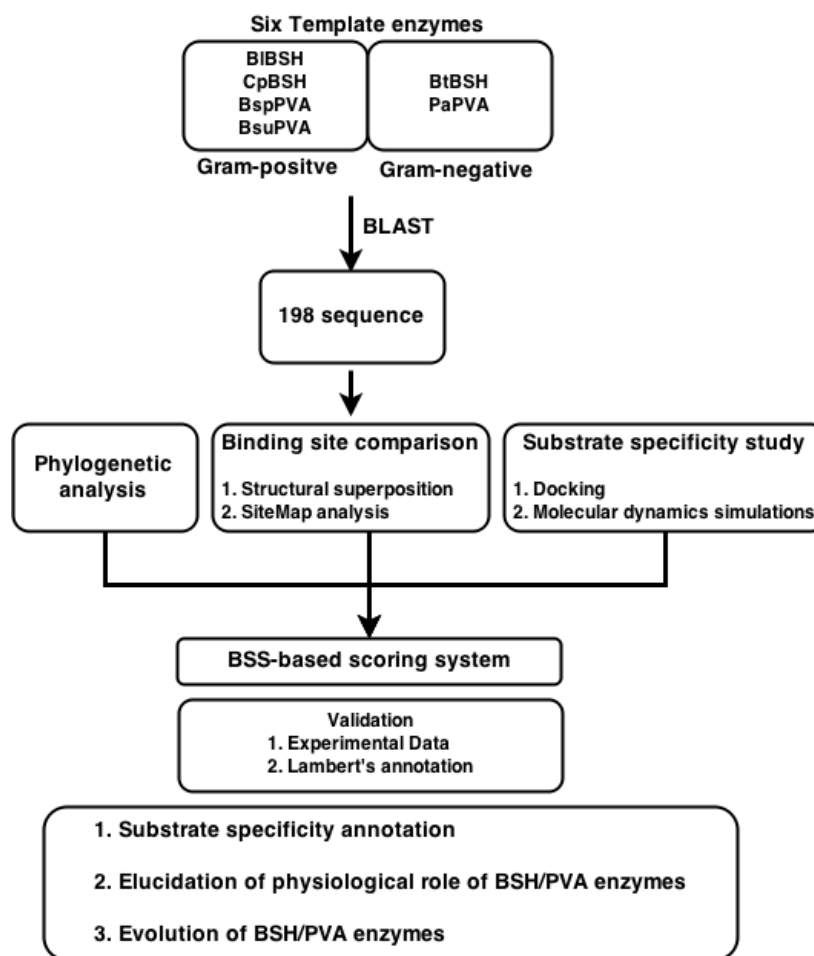


Figure 2.2: The workflow of BSS-based annotation of CGH family members.

2.2 Materials and Methods

2.2.1 Retrieval of CGH family members and phylogenetic analysis

Sequences of CGH family members were retrieved from non-redundant protein database of NCBI by performing a **Blast** search (Altschul *et al.*, 1997) using experimentally characterized *BIBSH*, *CpBSH*, *BspPVA*, *BsuPVA*, *BtBSH* and *PaPVA* protein sequences as queries. These six sequences were considered as '*template sequences*' for building the entire dataset (Table 2.1). Minimum blast score cutoff was kept at 500 and only the best hits from each organism were chosen for further analysis. Sequences lacking the conserved catalytic, nucleophilic Cys residue at their N-terminal were considered to be inactive and were excluded from the analysis. Sequences were clustered at 60% identity threshold and the non-redundant set containing **198 sequences** thus generated was used as final dataset (Table 2.6). The dataset also includes the six

template enzymes. Multiple sequence alignment was done by using **ClustalX** (Thompson *et al.*, 1997) while **Mega5.2** (Tamura *et al.*, 2011) was used to construct a phylogenetic tree of the CGH family by neighbor-joining method with a bootstrap value of 1000.

2.2.2. Structure retrieval and preprocessing

The three-dimensional structures of *BIBSH*, *CpBSH*, *BtBSH*, *BspPVA* and *BsuPVA* (PDB ID: 2HF0, 2RLC, 3HBC, 3PVA and 2OQC respectively) were downloaded from **PDB** (Berman *et al.*, 2000). Residues 48-49, 157-162 and 271-273 belonging to the substrate binding site, in the *BtBSH* structure (3HBC) were found to be missing. These regions were modeled by taking *BIBSH* structure as template using **Prime** (Version 3.0, Schrödinger, LLC, New York, NY, 2012).

2.2.3 Prediction of substrate binding modes using docking analysis

Prediction of substrate binding modes among five enzymes of CGH family with known structures (*BIBSH*, *CpBSH*, *BtBSH*, *BspPVA* and *BsuPVA*) was carried out using docking studies. The substrate binding modes in *PaPVA* homology model have already been reported (Avinash *et al.*, 2013). Glycocholic acid (GCA) and penicillin V were used as substrates. Grid based rigid receptor and flexible ligand docking program **Glide** (Friesner *et al.*, 2006) was used to predict the binding modes of ligand in the receptor binding site. Potential binding sites on each receptor were identified using **SiteMap** (Halgren, 2007).

2.2.4 Molecular dynamics simulations

Dynamics and stability of each receptor-ligand complex was evaluated by conducting explicit solvent molecular dynamics simulations of each complex on a 5 ns time scale using Amber force field in **Gromacs 4.5** (Pronk *et al.*, 2013). The topology and parameters for ligands were generated with General Amber Force Field using **acpype** (Sousa da Silva & Vranken, 2012). Each complex was solvated in a cubic box such that the complex is 10 Å away from the box boundary. System was neutralized by addition of counter ions. The system was first subjected to steepest descent energy minimization followed by conjugate gradient minimization. A maximum of 50000 minimization steps and maximum force for convergence was chosen as 10 kJ/mol/nm. The resulting system was equilibrated in NVT ensemble for 100 ps at 300 K using

V-rescale temperature coupling. The system was further equilibrated with NPT ensemble for about 100 ps at 300 K and 1 atmosphere pressure, using Parrinello-Rahman pressure coupling. The equilibrated system was finally subjected to molecular dynamics simulation using leap-frog integrator.

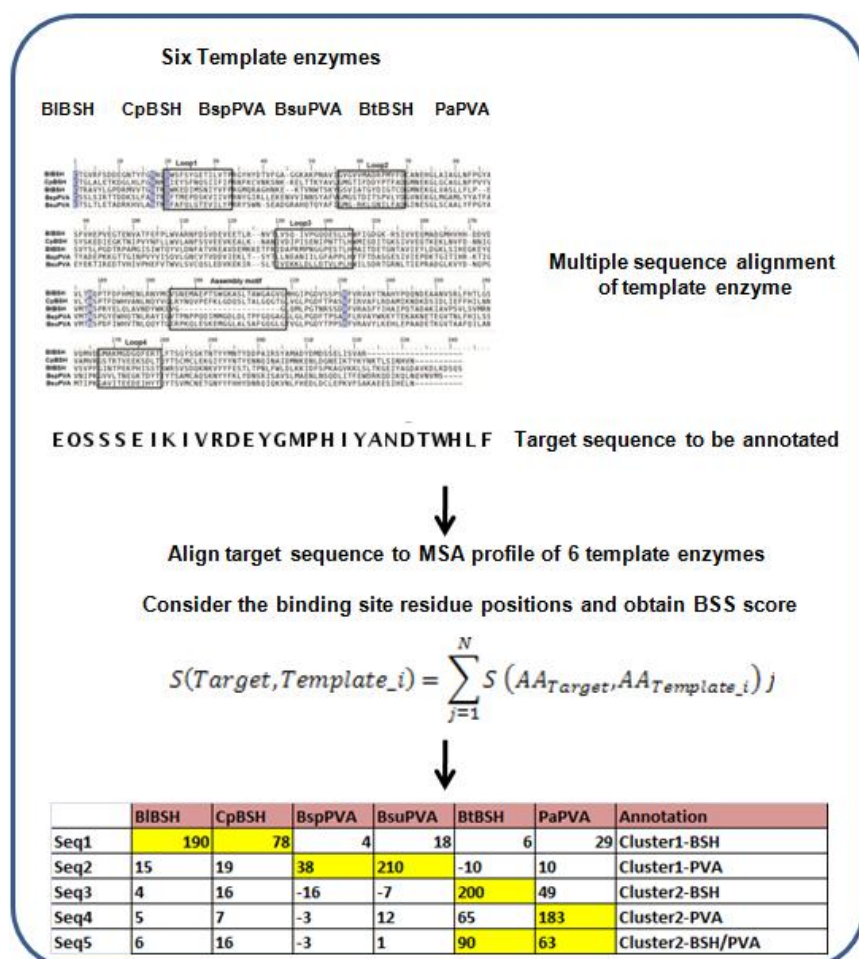


Figure 2.3: The methodology of the Binding site similarity (BSS) based scoring and annotation system. In the example shown, Seq1 has highest BSS scores with Gram-positive BSH enzymes (*BIBSH/CpBSh*) while Seq2 has highest score with Gram-positive PVA enzymes (*BsuPVA/BspPVA*). Similarly Seq3 has highest BSS scores with Gram-negative BSH (*BtBSh*) while Seq4 has highest BSS scores with Gram-negative PVA (*PaPVA*). Seq5 has highest BSS scores with both Gram-negative BSH/PVA enzymes.

2.2.5 Estimation of Binding Site Similarity (BSS) scores for all CGH sequences

A binding site profile-based scoring system (Fig. 2.3) was developed to estimate quantitatively the binding site similarity of each CGH family member within the dataset, utilizing the binding site information from each of the six template enzymes (*BIBSH*, *CpBSh*, *BtBSh*, *BspPVA*, *BsuPVA* and *PaPVA*). Each query sequence from the dataset was aligned with the multiple sequence alignment profile of above six template enzymes. Only the binding site positions (corresponding to residue 20-25, 57-67, 79-83, 102, 127-140 in *BIBSH*) of resulting

alignment were considered for scoring. Score of the query sequence with i^{th} template ($i=1$ to 6) was calculated as per the equation below.

$$S(\text{Query}, \text{Template}_i) = \sum_{j=1}^N S(\text{AA}_{\text{query}}, \text{AA}_{\text{Template}_i})_j$$

Where $S(\text{AA}_{\text{query}}, \text{AA}_{\text{Template}_i})_j$ is the similarity score between amino acid residues of query and i^{th} template sequence, at j^{th} position of binding site profile. Blosum80 scoring matrix was used for obtaining pair wise similarity score between amino acids (Henikoff & Henikoff, 1992). The scores were summed over N binding site positions ($j=1$ to N). Hence for each query sequence, a total of six scores were obtained corresponding to each template, describing the similarity of the binding site of query sequence to that of each template.

Based on the BSS scores obtained, the 198 sequences in the dataset were annotated as BSH/PVA enzymes. The BSS scoring system was tested on experimentally characterized BSH and PVA enzymes annotated earlier by Lambert *et al.*, 2008 and Jones *et al.*, 2008.

2.3 Results and Discussions

2.3.1 Dataset generation and phylogenetic analysis of the BSH/PVA sequences

A dendrogram prepared based on the phylogenetic analysis of the 198 sequences in the dataset (Table 2.6) resulted in the formation of two distinct clusters (Fig. 2.4): **Cluster1** (75 sequences) and **Cluster2** (123 sequences). The majority of the sequences in Cluster1 belonged to Gram-positive bacteria (phylum *Firmicutes*: 60; *Actinobacteria*: 12), whilst three were from archaea (*Methanobacterium formicicum*, *Methanobrevibacter smithii* and *Methanosphaera stadtmanae*). Cluster2 included a majority of the sequences from Gram-negative bacteria (phylum *Proteobacteria*: 65; *Bacteroidetes*: 24; *Cyanobacteria*: 12; *Planctomycetes*: 6; and three from other phyla), whilst 10 were from Gram-positive *Actinobacteria* and three were archaeal sequences (*Natrialba aegyptia*, *Natrinema gari* and *Methanoplanus petrolearius*). The phylum information of each sequence is given in Table 2.6. Thus, Gram-positive and archaeal members were distributed across both clusters with the majority of Gram-positive sequences grouped in Cluster1, whereas Gram-negative sequences were found only in Cluster2. Among the 10 Gram-positive *Actinobacteria* of Cluster2, eight actually belong to the order *Corynebacterineae* (Fig.

2.4), which are the intermediates between true Gram-positive and Gram-negative bacteria (Gupta, 2011). These intermediates show positive Gram staining but also have an additional highly ordered layer of mycolic acid resembling the outer membrane of true Gram-negative bacteria (Gupta, 2011). The other two Gram-positive *Actinobacteria* in Cluster2 are *Kitasatospora setae* and *Streptomyces sp. Mg1* belonging to the family *Streptomycetaceae*.

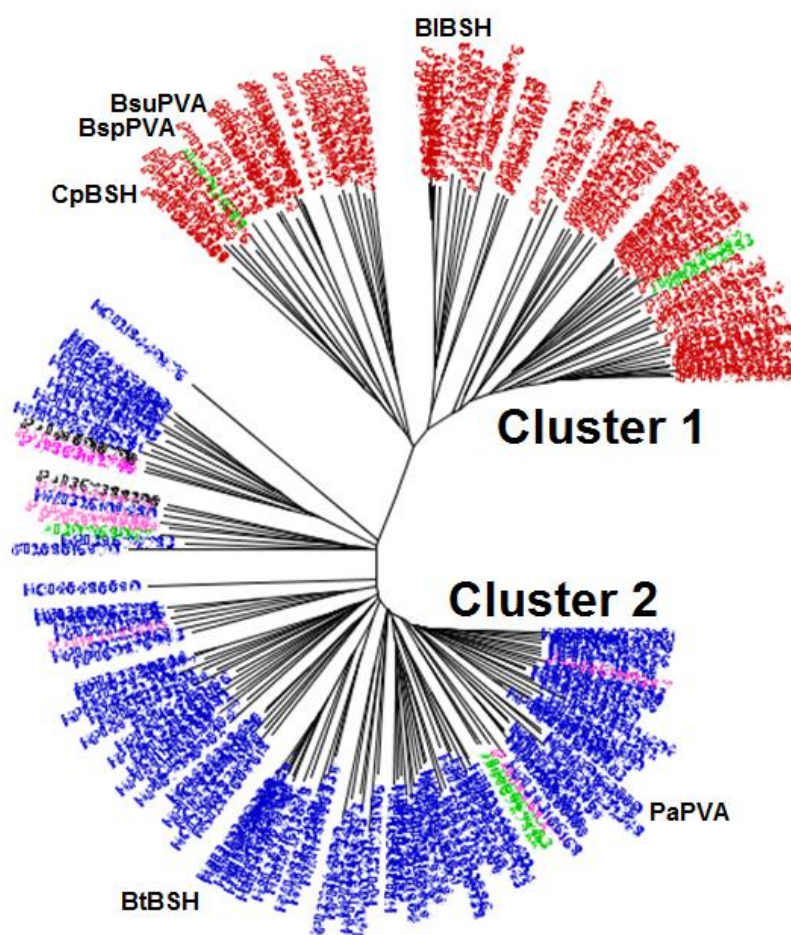


Figure 2.4: The dendrogram prepared based on the phylogenetic analysis of the sequences of CGH family. Two distinct clusters are labeled as Cluster1 and Cluster2. Members are colored according to their source (Red: Gram-positive bacteria, Blue: Gram-negative bacteria, Pink: order *Corynebacterineae* and Green: Archaea). Experimentally characterized BSH and PVA enzymes of each cluster are labeled.

The experimentally characterized BSH (*BtBSH* and *CpBSH*) and PVA (*BspPVA* and *BsuPVA*) enzymes from Gram-positive bacteria belonged to Cluster1, whereas the BSH (*BtBSH*) and PVA (*PaPVA*) enzymes from Gram-negative bacteria belonged to Cluster2 (Fig. 2.4). In Cluster1, the BSH enzymes *CpBSH* and *BtBSH* were distributed in two different branches; *CpBSH* was observed to be grouped along with PVA enzymes (*BsuPVA* and *BspPVA*). Similarly, in Cluster2, the available information was unable to annotate correctly the members of this cluster as BSH/PVA. As mere phylogenetic analysis was not enough to annotate correctly

the BSH/PVA sequences, a better annotation method was developed which not only included the **phylogenetic information**, but also took into consideration the **binding site** and **substrate specificity information** of the BSH/PVA enzymes.

2.3.2 Analysis of the substrate specificity and binding site properties of CGH enzymes

The enzymes *BIBSH*, *CpBSH*, *BtBSH*, *BspPVA*, *BsuPVA* and *PaPVA* are known to exhibit variations in their substrate specificity, i.e. they vary from being a classic BSH (no PVA activity) to a classic PVA enzyme (no BSH activity). Among the Gram-positive bacteria members, *BIBSH* is a classic enzyme with only BSH activity (Kumar *et al.*, 2006), whereas *BsuPVA* is a classic PVA enzyme with only PVA activity (Rathinaswamy *et al.*, 2012). Between these two extremes, *CpBSH* is a BSH enzyme with low PVA activity and *BspPVA* is a PVA enzyme with low BSH activity. Quantitatively, the PVA activity of *CpBSH* is 11% that of *BspPVA* and the BSH activity of *BspPVA* is 20% that of *BIBSH* (Kumar *et al.*, 2006). Among the characterized enzymes from Gram-negative bacteria, *BtBSH* has BSH activity, whilst its PVA activity has not been verified experimentally (Stellwag & Hylemon, 1976). *PaPVA* is a Gram-negative classic PVA enzyme (Avinash *et al.*, 2013). Except for *PaPVA*, the tertiary structure has been determined for the other five enzymes. Except for *CpBSH*, all determined structures are the apo-form of the enzyme without any bound substrate molecule. The *CpBSH* structure (2RLC; Fig. 2.5a) shows the enzyme bound with its product glycine and cholate (Rossocha *et al.*, 2005).

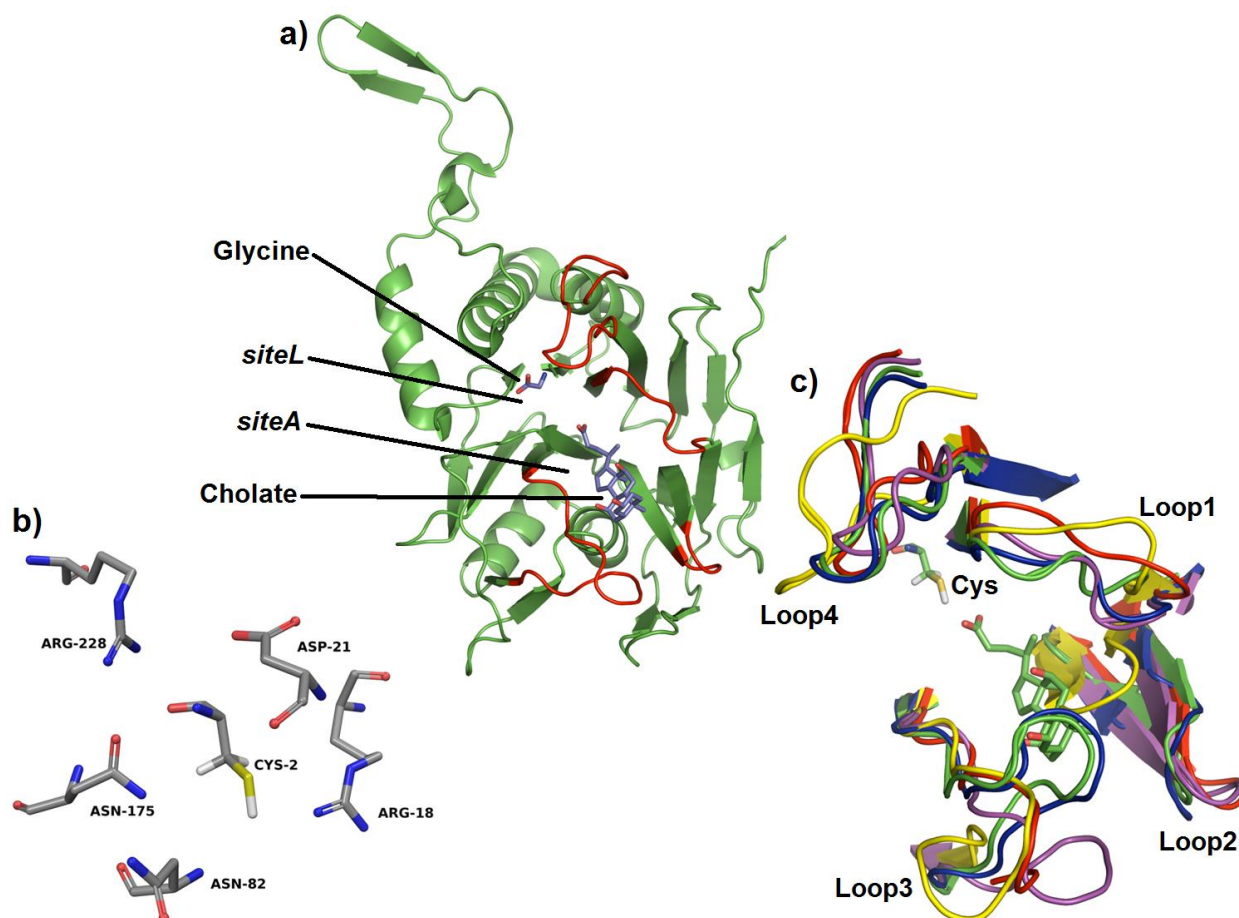


Figure 2.5: (a) Three-dimensional structure of *CpBSH* (PDB ID: 2RLC) with its bound product glycine and cholate. This structure represents the conformation of *CpBSH* after the hydrolysis of the substrate glycocholic acid. Glycine (shown in stick representation and labeled) is bound in the active site (*siteL*) where as the cholate moiety (shown in stick representation and labeled) is bound in the binding site (*siteA*; formed by four loops: Red) (b) Geometrical rearrangement of catalytic residues in *CpBSH*. During catalysis, the N-terminal Cys2 acts as both a nucleophile and base. Arg18 and Asp21 form hydrogen-bonding interactions with Cys2, whereas Asn82 and Asn175 form the putative oxyanion hole. Arg228 helps in transition-state stabilization (Lodola *et al.*, 2012). This arrangement of catalytic residues remains conserved in all CGH enzymes. (c) Illustrated here is the superposition of the four substrate binding site loops (loop1 to loop4) from *BtBSH* (Red), *CpBSH* (Magenta), *BspPVA* (Blue) and *BsuPVA* (Green). Observed is the differential folding and conformations of the above defined loops in these enzymes resulting in variance in the size of their binding site pockets. The active site Cys residue is shown and labeled.

A structural comparison of the available structures showed similar positional preference of their catalytic residues in the active site region. The catalytic framework observed in the *CpBSH* structure is shown in Fig. 2.5b. However, the enzymes show significant variation in terms of size and properties of their binding site pockets. This variation is due to differential

folding and conformations of the loops near the binding site (Fig. 2.5c). The substrate binding sites in these enzymes consist mainly of four loops, i.e. loop1–loop4, comprising residues 22-27, 58-65, 129-139 and 261-269, respectively, in *B/B*BSH. In the *Bt*BSH structure, coordinates for residues 48-49 of loop1, 157-162 of loop3 and 271-273 of loop4 were found to be missing, presumably due to the disorder of these highly dynamic loops. Therefore, these loop regions were modeled using *B/B*BSH as the template.

Table 2.2: SiteMap quantitative estimation of binding site properties of CGH enzymes

Enzyme	Volume (Å ³)*	Exposure†	Hydrophobic	Hydrophilic	Hydrophobic/ Hydrophilic#
<i>Bsp</i> PVA	153.32	0.31	3.98	0.46	8.61
<i>Bsu</i> PVA	344.37	0.46	2.55	0.59	4.34
<i>Cp</i> BSH	485.35	0.57	0.91	1.04	0.87
<i>B/B</i> BSH	535.77	0.45	0.95	1.07	0.88
<i>Bt</i> BSH	718.58	0.60	0.38	1.04	0.37

*Binding site volume (Å³) measured using shrink-wrap approach of SiteMap. †Exposure is a measure of how open the site is to solvent; higher the value more exposed it is. #Hydrophilic and hydrophobic terms are a measure of hydrophilic and hydrophobic nature of the site; the higher is the ratio of hydrophobic to hydrophilic values, more hydrophobic the site is. †, # it is calculated as ratios and do not carry any units.

Out of the four binding site loops in these enzymes, loop2-loop4 show significant differences in terms of their folding and conformation (Fig. 2.5c). Loop3 in PVA-type enzymes (*Bsp*PVA and *Bsu*PVA) is oriented more inside the cavity compared with BSH-type enzymes (*B/B*BSH, *Cp*BSH and *Bt*BSH), reducing the size of the binding site pockets. Loop2 in *Bt*BSH is oriented more into the cavity as compared with the others, thereby shifting the binding site pocket and increasing its solvent accessibility. These results are summarized quantitatively in Table 2.2, which describes the binding site volume, solvent accessibility, and hydrophobic and hydrophilic properties in these enzymes. As compared with PVA-type enzymes (*Bsp*PVA and *Bsu*PVA), the BSH-type enzymes (*B/B*BSH, *Cp*BSH and *Bt*BSH) were observed to have a larger, more exposed and hydrophilic binding site (Table 2.2), in order to accommodate the bile salt molecule, which is larger than penicillin V.

2.3.3 Mode of substrate binding among CGH enzymes

2.3.3.1 Modes of GCA binding

GCA is a bile salt molecule synthesized in the liver, formed by the conjugation of a cholic acid moiety to the amino acid glycine through an amide bond. The N-terminal Cys residue of CGH enzymes carries out a nucleophilic attack on this scissile amide bond to release glycine (leaving group) while the cholate moiety remains bound to the enzyme as the acyl-enzyme adduct (Lodola *et al.*, 2012). In the crystal structure of *Cp*BSH (PDB ID: 2RLC), the adduct group cholate occupies the binding site (*siteA*) and directs the leaving group glycine towards the active site (*siteL*) (Fig. 2.5a).

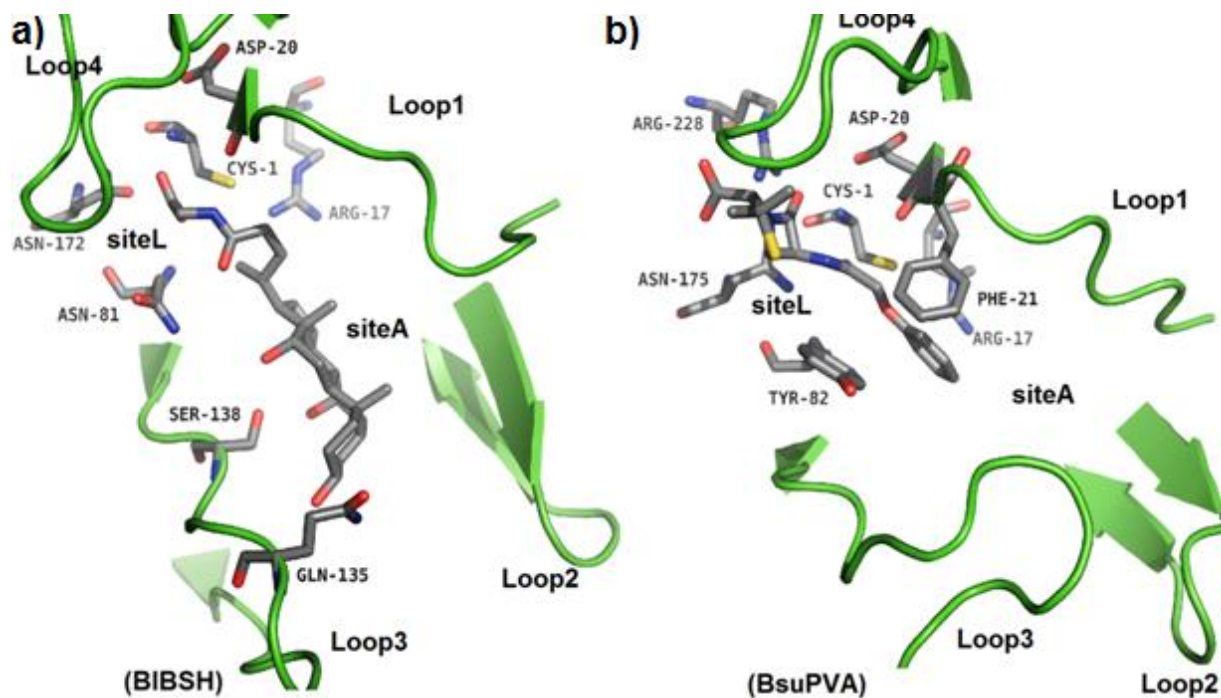


Figure 2.6: (a) Mode of binding of GCA in *BIBSH* (b) Mode of penicillin V binding in *BsuPVA*. In both complexes, the adduct groups (cholic acid in case of GCA and PAA in case of penicillin V) occupy *siteA* while directing the leaving groups (glycine in case of GCA and 6-APA in case of penicillin V) towards the active site (*siteL*), positioning the scissile amide bond just inside the cleft of the enzyme, close to the N-terminal Cys residue, in an orientation favorable for the nucleophilic attack. The four substrate binding loops are shown in cartoon representation. The modes of binding of GCA and penicillin V among other CGH enzymes are shown in Fig. 2.7 and Fig. 2.9, respectively.

In all BSH-active enzymes (*BIBSH*, *CpBSH*, *BtBSH* and *BspPVA*), the modes of GCA binding were in agreement with the binding mode seen in 2RLC wherein the adduct group cholate occupies *siteA* whilst the leaving group glycine occupies *siteL*. The *BIBSH*-GCA complex is shown in Fig. 2.6a and other complex structures are illustrated in Fig. 2.7. It was observed that GCA binding shows a directional preference for the amide bond orientation (CO–N) in the direction from *siteA* to *siteL*, with reference to the nucleophilic Cys residue. Favorable values of free energy of binding along with shorter and stable nucleophilic attack distance values during dynamics suggest the suitability of these binding modes for BSH activity in these enzymes (Table 2.3, Fig. 2.7f).

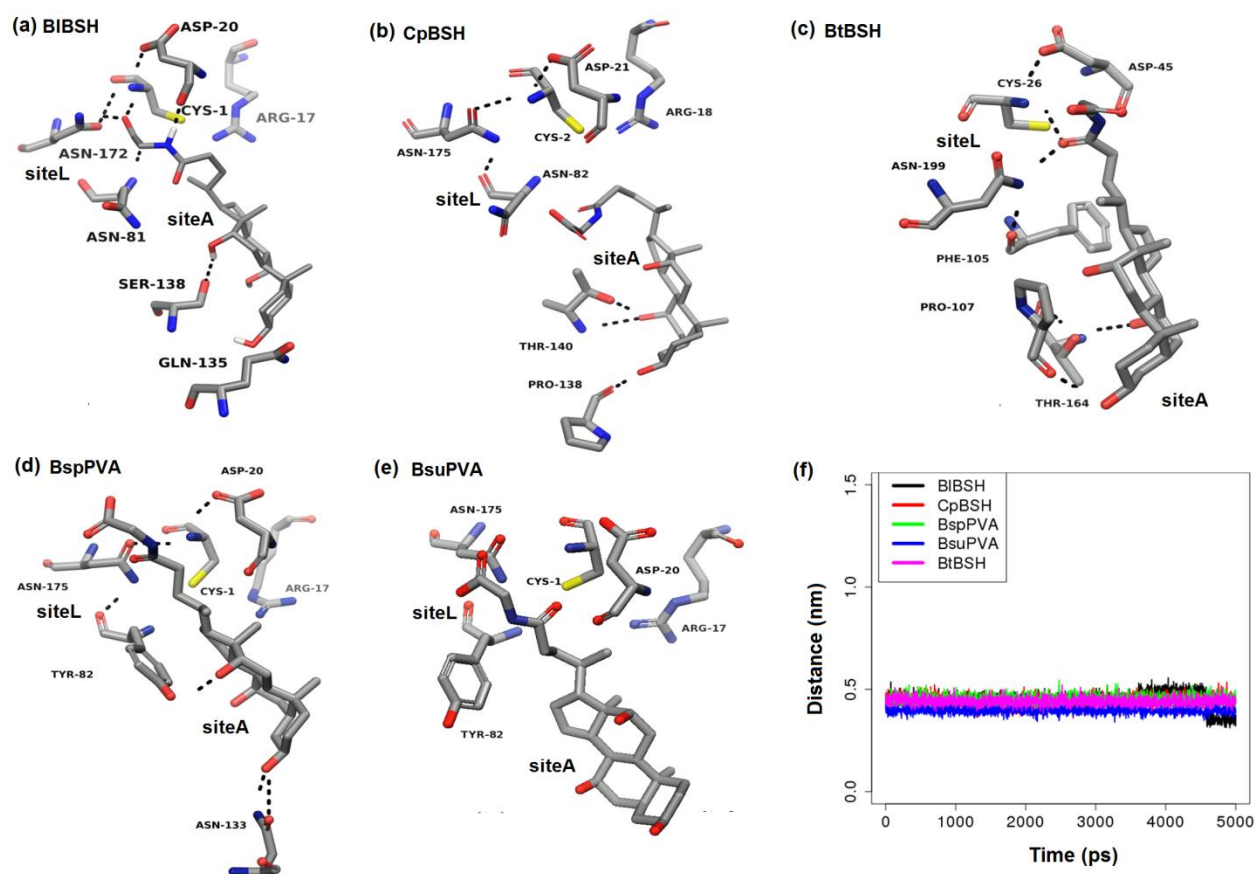


Figure 2.7: (a-e) The modes of GCA binding in *BIBSH*, *CpBSH*, *BtBSH*, *BspPVA* and *BsuPVA*, respectively. In all the enzymes the N-terminal Cys residue and other catalytic residues in its vicinity are shown in stick representation. The Loop3 residues which form hydrophilic complementarity with 3 α , 7 α and 12 α -OH groups of GCA are also shown in stick representation. Hydrogen bonding interactions are shown as black dotted line. The site of adduct group (*siteA*) and leaving group (*siteL*) binding are labeled. Except *BsuPVA* (BSH inactive), some degree of polar complementarities is seen among all BSH active enzymes. (f) Illustrates

the time evolution of Nucleophilic attack distance during the molecular dynamics simulation of each complex structure. The y-axis corresponds to the Nucleophilic attack distance values while x-axis corresponds to time scale (in ps). The Nucleophilic attack distance values remain stable and short during dynamics of all complexes suggesting suitability of these predicted poses.

Table 2.3: Summary of free energy of binding (GlideScore) of all predicted protein-ligand complex structures.

Enzyme	GCA		Penicillin V	
	GlideScore (Kcal mol ⁻¹)	Nucleophilic attack distance (Å) (mean±SD)	GlideScore (Kcal mol ⁻¹)	Nucleophilic attack distance (Å) (mean±SD)
<i>Bt</i> BSH	212.19	4.4±0.3	28.3	4.2±0.2
<i>Cp</i> BSH	29.58	4.4±0.2	24.3	4.9±0.2
<i>Bt</i> BSH	27.41	4.4±0.1	26.2	4.4±0.3
<i>Bsp</i> PVA	210.96	4.5±0.2	25.6	4.0±0.1
<i>Bsu</i> PVA	28.85	3.9±0.2	24.7	3.8±0.1

Interestingly, even in *Bsu*PVA, a BSH-inactive enzyme, the predicted mode of GCA binding was found to be similar to that of all BSH-active enzymes (Fig. 2.7e). However, in the modelled structure of the BSH-inactive *Pa*PVA enzyme, the GCA molecule was predicted to bind in a reversed amide bond orientation, from *siteL* to *siteA* (Avinash *et al.*, 2013).

2.3.3.2 Polar complementarity: probable basis for GCA specificity

GCA is a planar amphipathic molecule with its hydrophobic surface consisting of the methyl groups of the steroid ring, whilst the hydrophilic surface is formed by the three hydroxyl groups (3 α -, 7 α - and 12 α -OH; Fig. 2.1). Structural analysis of all predicted enzyme-GCA complex structures revealed an important correlation (Fig. 2.8) between GCA specificity and the degree of hydrophilic complementarity of the three hydroxyl groups. Using the radial distribution function, the maximum probability density of receptor polar atoms within 5 Å of each hydroxyl groups was estimated (Fig. 2.8, Table 2.4). Furthermore, the hydrogen bonding interaction of these hydroxyl groups with nearest receptor polar groups during dynamics was also assessed quantitatively (Table 2.4). The radial distribution function [g(r)] for the polar

complementarity was fixed at 0.5 for the analysis. Values >0.5 were considered to have polar complementarity at the particular hydroxyl group.

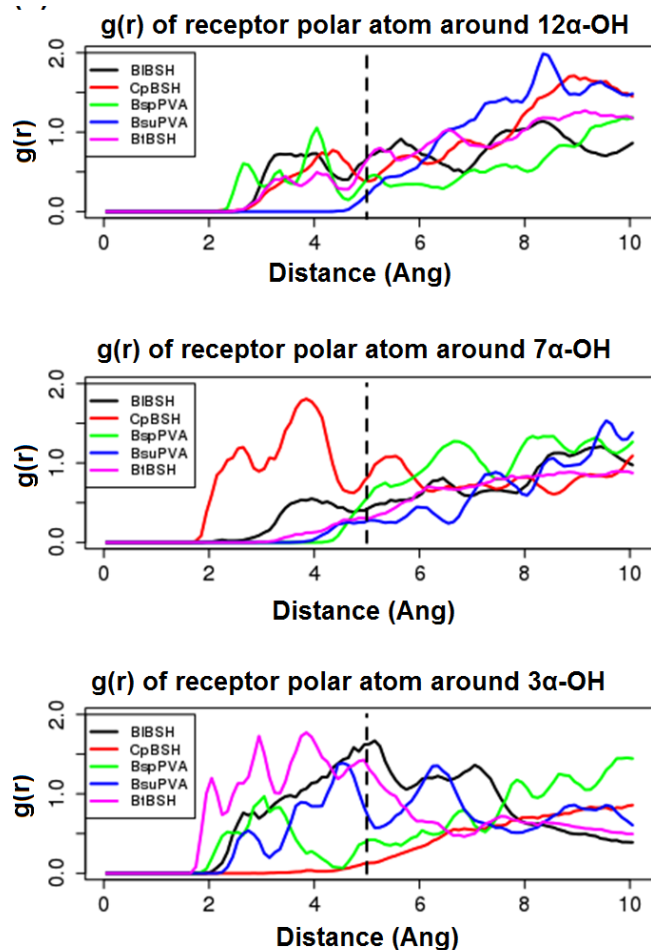


Figure 2.8: (a) Radial distribution of receptor polar atoms around three hydroxyl groups of GCA (3α -, 7α - and 12α -OH) in *BIBSH* (black), *CpBSH* (red), *BtBSH* (magenta), *BspPVA* (green) and *BsuPVA* (blue). The y-axes correspond to the probability density of receptor polar atoms [$g(r)$] and the x-axes correspond to distance (in Å). Polar complementarity at the respective hydroxyl group has been estimated at a 5 Å distance threshold as the maximum probability of finding receptor polar atoms [max. $g(r)$ within 5 Å]. Except for *BsuPVA*, a BSH-inactive enzyme, polar complementarity is observed at more than one hydroxyl group in all BSH-active enzymes. In *BsuPVA*, the polar complementarity is observed around only 3α -OH.

In *BIBSH*, polar complementarity was observed for all three hydroxyl groups along with good hydrogen-bonding interactions (Table 2.4). The values for both these factors at the hydroxyl positions 7α - and 12α -OH in *CpBSH* might support the fact that the activity of this enzyme is lower than *BIBSH* (Kumar *et al.*, 2006). In the case of *BtBSH*, although no comparative experimental evidence is available, the calculated values at 3α - and 12α -OH suggest that the activity in this case could also be lower than that of *BIBSH*. In *BspPVA*, polar complementarity values at all positions are >0.5 , but hydrogen-bonding interactions only at the 3α -OH advocate the available experimental evidence showing that it has low levels of BSH activity (Kumar *et al.*, 2006). However, in *BsuPVA*, only the 3α -OH group shows complementarity with a low percentage of hydrogen-bonding interactions at the same site, which

could be the reason for being BSH-inactive. The significance of these hydroxyl groups in contributing to binding affinity is further supported by (Batta *et al.*, 1984). Hence, polar complementarity constitutes an important factor influencing the GCA specificity among CGH enzymes. Indel mutations of polar residues in the loop3 region in PVA enzymes can be considered as an engineering strategy for producing GCA specificity in these enzymes.

Table 2.4: Quantitative estimation of polar complementarities for the three hydroxyl groups of the GCA molecule and percentage of times their involvement in hydrogen-bonding interactions (%Hbond) during molecular dynamics simulation of each enzyme-GCA complex.

Property	Hydroxyl Group	<i>B</i> /BSH	<i>Cp</i> BSH	<i>Bt</i> BSH	<i>Bsp</i> PVA	<i>Bsu</i> PVA
Maximum probability density within 5 Å (estimated by radial distribution function)	12 α	0.73	0.77	0.60	1.05	0.17
	7 α	0.55	1.80	0.30	0.51	0.25
	3 α	1.62	0.12	1.77	0.96	1.38
%Hbond	12 α	21.67	4.63	16.39	0	0
	7 α	4.83	100	0	0	0
	3 α	4.51	6.95	72.07	100	8.47

2.3.3.3 Modes of penicillin V binding among CGH enzymes

In all CGH enzymes, whether PVA-active (*Bsp*PVA, *Bsu*PVA, *Cp*BSH) or PVA-inactive (*B*/BSH), and also in *Bt*BSH, the mode of penicillin V binding was observed to be similar. The *Bsu*PVA-penicillin V complex structure is shown in Fig. 2.6b and other penicillin V complexes are depicted in Fig. 2.9. The adduct group phenoxy acetic acid occupies *siteA* and the leaving group 6-aminopenicillanic acid occupies *siteL* consequentially, establishing a directional preference for the amide bond (CO–N) with respect to the nucleophilic residue Cys1. All complexes remained stable during simulation with shorter and stable nucleophilic attack distance values (Table 2.3). During the simulation of the *Bsu*PVA-penicillin V complex, penicillin V was observed to form a hydrogen bond with Arg228, Asn175 and Cys1 residues (~98, 45 and 14%, respectively). This observation supports the proposed reaction mechanism of NtCn-hydrolases,

in which the Arg228 residue is shown to have a direct role in transition-state stabilization and Asn175 is shown to have a role for substrate recognition through hydrogen bonding (Lodola *et al.*, 2012).

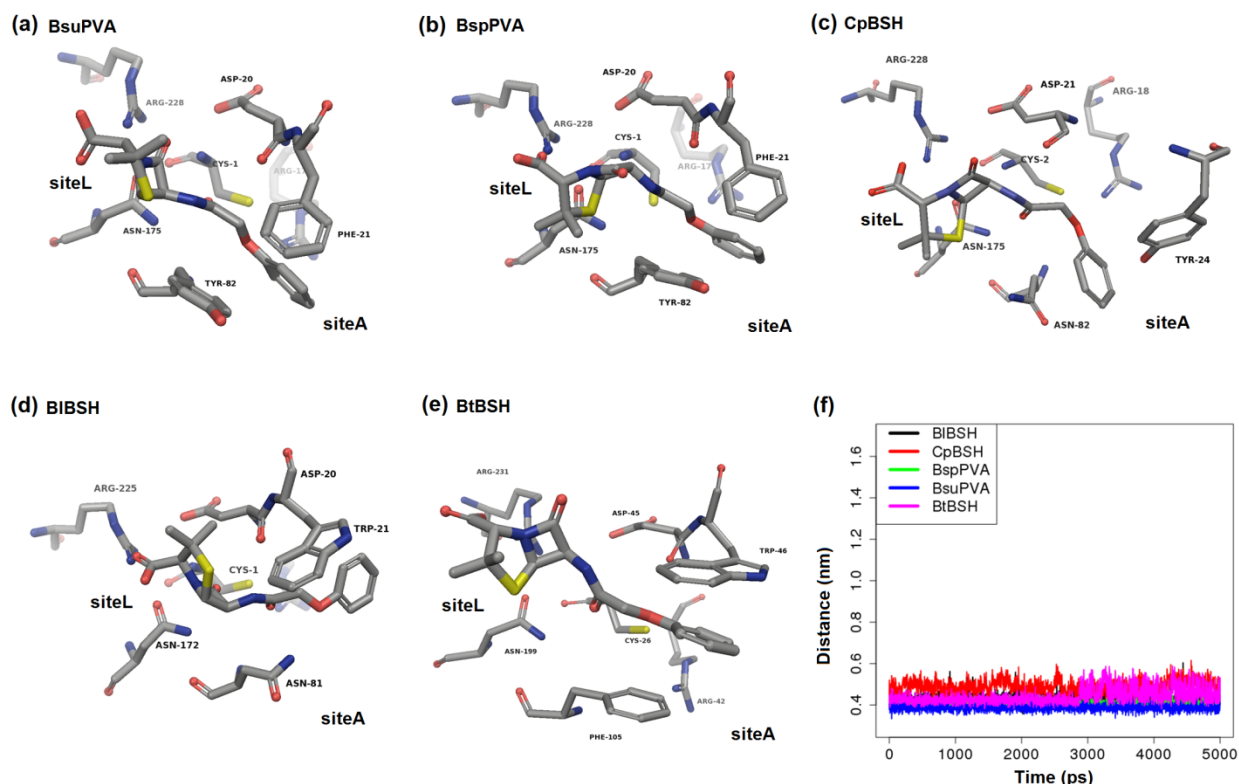


Figure 2.9: (a-e) The mode of binding of penicillin V molecule in *BsuPVA*, *BspPVA*, *CpBSH*, *BIBSH* and *BtBSH* respectively. The site of adduct group (*siteA*) and leaving group (*siteL*) binding are labeled. Like GCA binding, penicillin V binding also show a directional preference of its adduct and leaving group. Observed here are the aromatic interactions between phenyl ring of penicillin V and the residues in its vicinity. These aromatic residues might play an important role in penicillin V binding. (f) Illustrates the time evolution of Nucleophilic attack distances during the molecular dynamics simulation of each complex structure. The y-axis corresponds to the Nucleophilic attack distance values while x-axis corresponds to dynamics time scale (in ps).

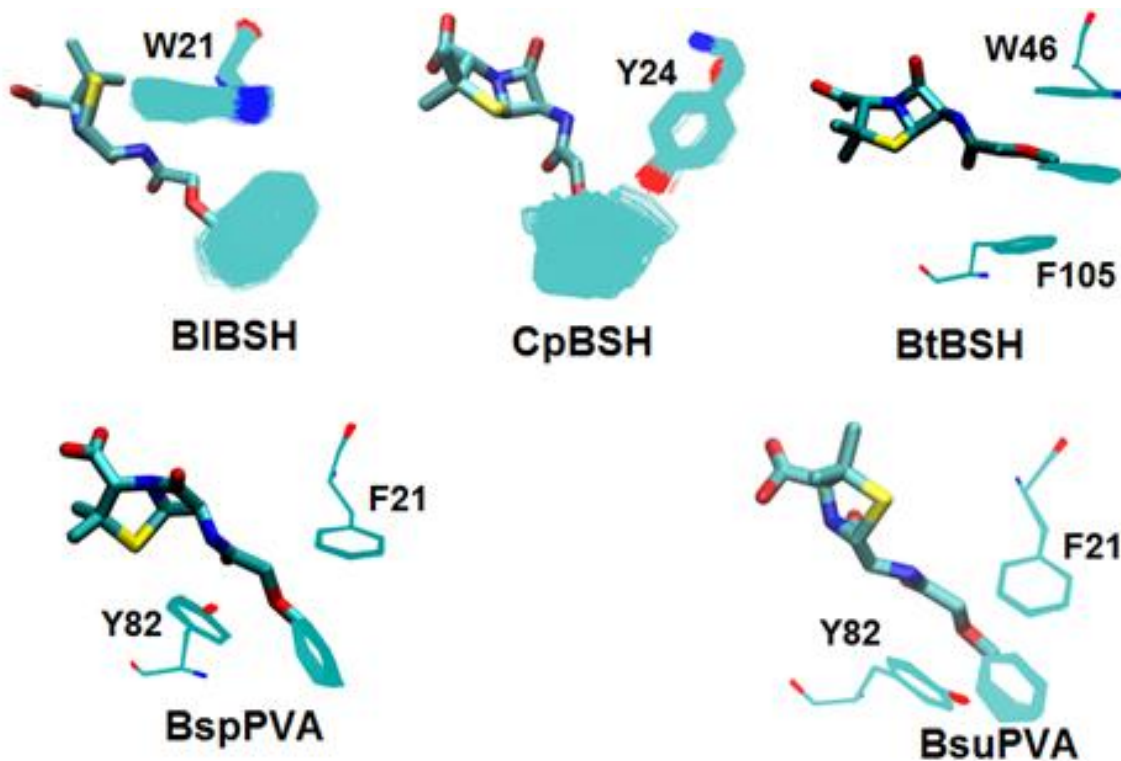


Figure 2.10: The geometrical arrangements of the aromatic planes of the residues in the vicinity of phenyl ring of substrate penicillin V, corresponding to all five enzymes during molecular dynamics simulation are shown. In case of penicillin V, the fluctuation of only the aromatic rings are shown. For the purpose of clarity, the trajectory was smoothed by window size of 10.

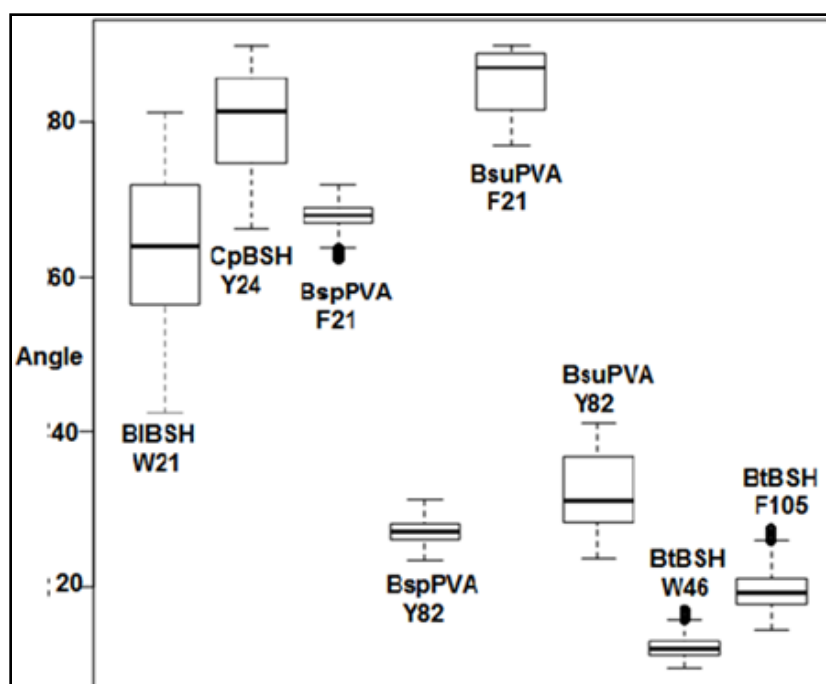


Figure 2.11: Box plot depicts the distribution of angle between the phenyl ring planes of substrate penicillin V and aromatic residues in its vicinity.

2.3.3.4 Aromatic interactions in the active site might influence penicillin V binding

The presence of aromatic-aromatic interactions between the phenyl ring of penicillin V and aromatic residues in its vicinity is deemed important in all enzyme-penicillin V complexes (Fig. 2.10). In *BspPVA* and *BsuPVA*, Phe21 and Tyr82 residues interact with the phenyl ring of penicillin V in a strict geometry. This Phe21, Tyr82 aromatic pair is substituted with (Trp21, Asn81), (Ile22, Asn82) and (Trp46, Phe105) pairs in *BtBSH*, *CpBSH* and *BtBSH*, respectively. In *BtBSH*, the possible involvement of Trp residues in penicillin V binding is evident from experimental data, which show quenching of tryptophan fluorescence as a result of penicillin V binding (Kumar *et al.*, 2006). In *CpBSH*, although Phe21 is substituted by Ile22, Tyr24 is involved in aromatic interaction with penicillin V, which compensates for the loss of Phe21. As in PVA enzymes, in *BtBSH*, both Trp46 and Phe105 interact with the phenyl ring of penicillin V in a strict geometrical arrangement, suggesting that *BtBSH* might show a low degree of PVA activity.

Participation of aromatic residues in penicillin V binding has also been shown experimentally in *PaPVA*, where the Trp23 and Trp87 aromatic pair is known to interact with the phenyl ring of penicillin V (Avinash *et al.*, 2013). Fig. 2.11 shows the distribution of the planar angle between the phenyl ring of penicillin V and the aromatic planes in its vicinity. The PVA enzymes, *BsuPVA* and *BspPVA*, show less deviation in these planar angles compared with the BSH enzymes, *BtBSH* and *CpBSH*, resulting in firm binding of the penicillin V molecule and thus showing higher PVA activity. In *BtBSH*, similar to *BspPVA*, only a slight deviation was observed, suggesting possible penicillin V binding affinity. It is probable that these aromatic residues might interact with the incoming substrate through stacking interactions and help it to initially orient favorably in the binding site, influencing the binding affinity.

In summary, it was observed that among CGH enzymes, substrate binding (bile salt or penicillin V) involves a directional and orientational preference of adduct and leaving groups, and therefore the scissile amide bond direction, with respect to the position of the nucleophile Cys residue. The polar complementarities of the three hydroxyl groups of the GCA molecule might also influence its binding affinity with these enzymes. Similarly, the presence of aromatic residues in the active site and their arrangement with respect to the phenyl ring of bound penicillin V might play a decisive role in penicillin V binding and affinity.

2.3.4 Substrate specificity annotation of family members

Enzymes possessing similar residues in their binding site pocket may have similar mechanisms of action and substrate preferences (Haupt *et al.*, 2013). Based on this assumption, each of the 198 sequences was assigned a total of six numeric scores (BSS scores) corresponding to their BSS with *BIBSH*, *CpBSH*, *BspPVA*, *BsuPVA*, *BtBSH* and *PaPVA*. Based on initial phylogenetic clustering and the calculated BSS scores, the 198 CGH enzyme sequences were annotated into five subgroups using a cut-off of score difference set at 30 between the highest scores with BSH and PVA enzymes (Fig. 2.12).

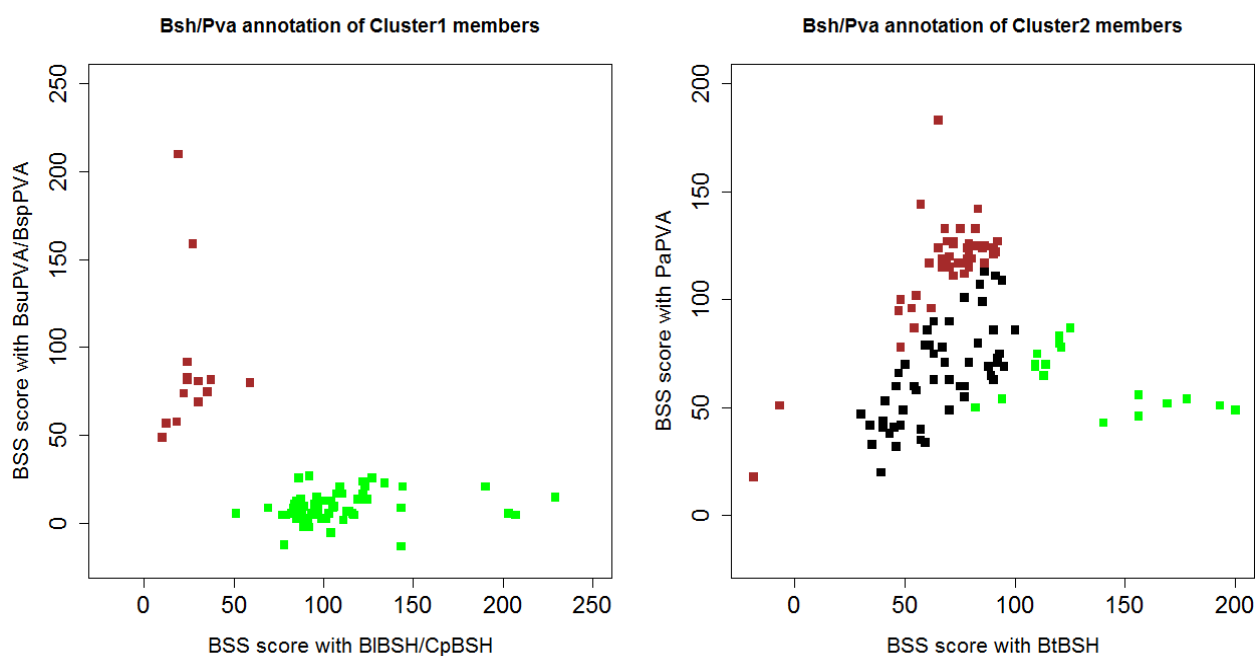


Figure 2.12: Illustrates the Binding Site Similarity (BSS) based annotation of CGH family members into BSH or PVA enzymes. Green: BSH, Brown: PVA and Black: BSH/PVA enzymes.

During the analysis, sequences belonging to Cluster1 could be annotated easily as BSH/PVA (Fig. 2.12) on the basis of the score difference of >30 . The subgroups in Cluster1 were defined as **Cluster1-BSH** (members scoring highest with Cluster1 BSH enzymes; either *BIBSH* or *CpBSH*, or both) and **Cluster1-PVA** (scoring highest with Cluster1 PVA enzymes; either *BspPVA* or *BsuPVA*, or both) (Fig. 2.12). During the annotation of Cluster2 sequences into BSH/PVA, a few sequences showed similar scores for both the Cluster2 BSH enzyme *BtBSH* as well as the Cluster2-PVA enzyme *PaPVA*. Sequences in Cluster2 were annotated into three subgroups. The first two subgroups showing a score difference of >30 were defined as

Cluster2-BSH (scoring highest with Cluster2 BSH enzyme *BtBSH*) and **Cluster2-PVA** (scoring highest with Cluster2 PVA enzyme *PaPVA*), whilst the third was defined as **Cluster2-BSH/PVA** (scoring highest with both *BtBSH* and *PaPVA*, a score difference of < 30) (Fig. 2.12). Amongst the 75 Cluster1 sequences, 59 were annotated as BSH and 16 as PVA enzymes, whilst in the 123 Cluster2 sequences, 21 were annotated as BSH, 49 as PVA and 53 as BSH/PVA enzymes. The detailed BSS scores and the resulting annotations are given in Table 2.6.

The BSS-based scoring system was validated using experimentally verified BSH/PVA enzymes from Gram-positive bacteria, Gram-negative bacteria and archaea to check the accuracy of the annotations predicted (Table 2.7). Amongst the experimentally characterized Gram-positive BSH enzymes, those from the *Firmicutes* and *Actinobacteria* were predicted correctly as Cluster1-BSH enzymes. Similarly, the PVA enzyme from *Listeria monocytogenes EGDe* was annotated correctly as a Cluster1-PVA enzyme. In the case of Gram-negative bacteria, the BSH enzymes from *Brucella abortus* and *Bacteroides vulgatus* were annotated correctly as Cluster2-BSH enzymes. Similarly, the known BSH archaeal enzymes from *Methanosphaera stadtmanae* and *Methanobrevibacter smithii* were also predicted correctly as Cluster1-BSH enzymes. The BSS scoring system was further validated against the Gram-positive CGH enzymes annotated previously as BSH/PVA by Lambert *et al.*, 2008. The BSS-based functional assignment was found to be in agreement with the earlier annotations (Table 2.8).

2.3.5 Physiological role of BSH/PVA enzymes

In Cluster1, most of the enzymes annotated as BSH enzymes were found to belong to the gut-inhabiting bacteria (*Firmicutes* and *Actinobacteria*) and archaea (*Methanobacterium formicicum*, *Methanobrevibacter smithii* and *Methanosphaera stadtmanae*) (Table 2.6). The enzyme from *Planococcus antarcticus*, an environmental bacterium isolated from cyanobacterial mat samples in the lakes of Antarctica (Reddy *et al.*, 2002), was also classified as a BSH enzyme. Interestingly, the enzymes annotated in Cluster2 as BSH enzymes were distributed widely among both gut-inhabiting bacteria and environmental microbes (e.g. *Burkholderia sp. YI23*, *Blastopirellula marina*, *Desulfovibrio fructosovorans* and *Rhodomicrobium vannielii*). In addition, the Cluster2 enzyme from *Rickettsia felis*, a pathogen causing flea-borne spotted fever in cats and which is also known to infect humans, was annotated as a BSH enzyme. The presence

of BSH enzymes among pathogenic bacteria such as *L. monocytogenes* and *Enterococcus faecalis* (an opportunistic pathogen) was reported previously (Begley *et al.*, 2006). The presence of BSH genes among the gut-inhabiting micro-organisms can be attributed to their role in bile acid resistance, thereby protecting these organisms in the host gastrointestinal tract (Jones *et al.*, 2008). However, the physiological roles of the BSH genes among pathogens such as *Rickettsia felis*, environmental bacteria such as *Burkholderia sp.* and others still need to be explored.

The enzymes annotated as PVA in the dataset were found to be distributed widely among both pathogen and environment degrading organisms (Table 2.6). Pathogenic bacteria include *Pectobacterium*, *Agrobacterium*, *Brevundimonas*, *Bordetella*, *Acinetobacter*, *Yersinia*, *Proteus*, *Providencia* and others. The involvement of *pva* genes in the virulence of the pathogen *Vibrio cholerae* was reported previously (Kovacikova *et al.*, 2003). Reports on the involvement of acyl-homoserine lactone acylase enzymes (NtSn-hydrolases, distant homologues of PVA) in quorum quenching amongst opportunistic pathogens such as *Pseudomonas* are also known (Bokhove *et al.*, 2010). However, more experimental evidence may be required to ascertain the possible role of *pva* genes in the pathogenesis in these organisms. The *pva* genes were also found to be distributed among organisms of soil and aquatic ecosystems capable of degrading compounds containing aromatic rings. Examples include *Achromobacter xylosoxidans* (haloaromatic), *Rhodococcus qingshengii* (isolated from carbendazim-contaminated soil), *Mycobacterium vanbaalenii* (polycyclic aromatic hydrocarbon-metabolizing bacteria isolated from petroleum-contaminated estuarine sediments), *Delftia acidovorans* (able to grow on chlorophenyl herbicides), etc. It has been postulated that the penicillin acylase genes are related to pathways involved in the assimilation of aromatic compounds as a carbon source by scavenging for phenylacetylated compounds in the non-parasitic environment (Valle *et al.*, 1991). However, more experimental evidence may be required to ascertain these roles of PVA genes.

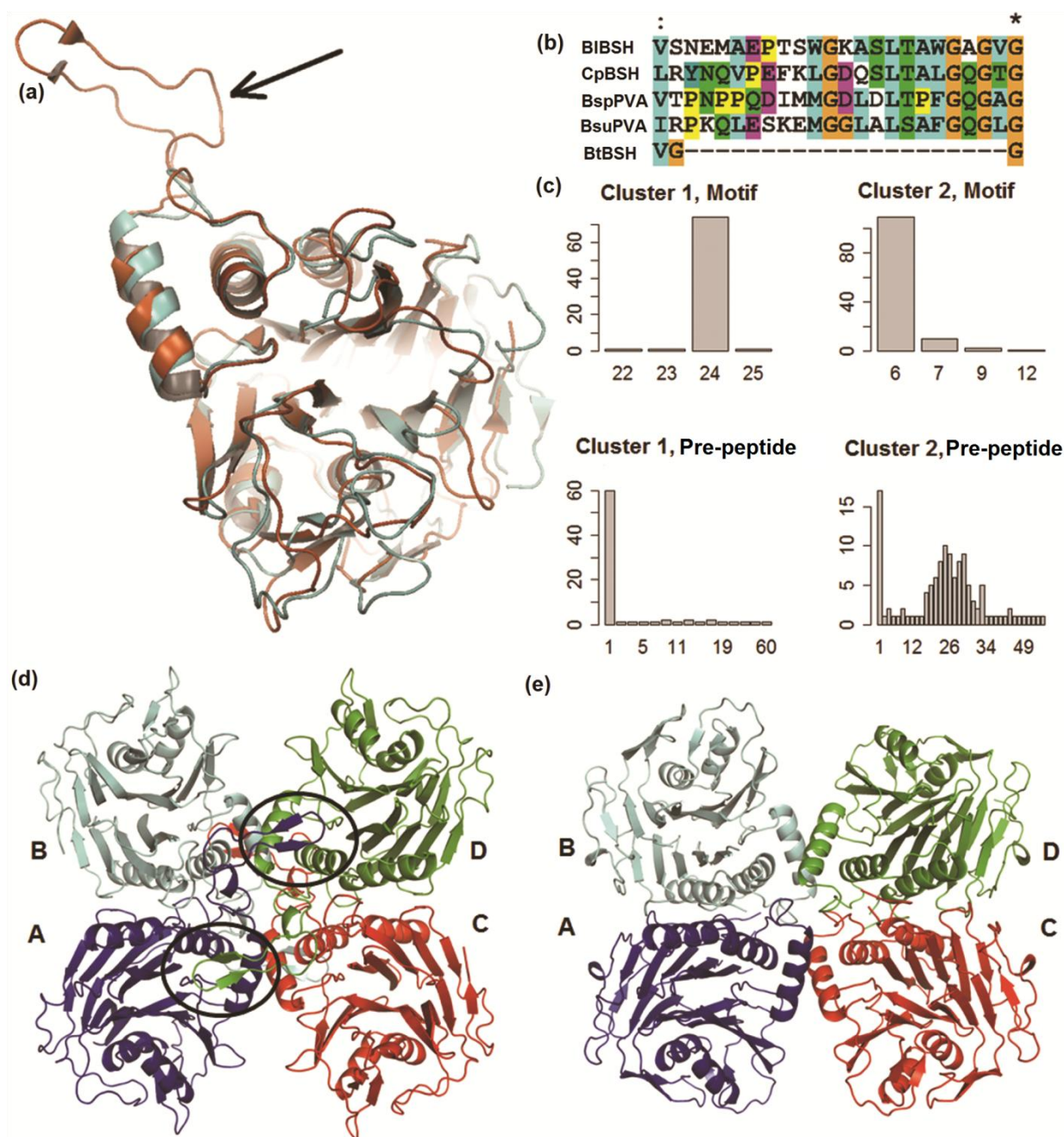


Figure 2.13: (a) Superposition of the structures of *BtBSH* (cyan) and *B/BSH* (brown) showing the missing 'assembly motif' in the *BtBSH* structure, marked by an arrow. (b) Multiple sequence alignment of five enzymes to show the absence of the 'assembly motif' in the *BtBSH* (PDB ID: 3HBC) sequence. (c) Distribution of sequence length of the assembly motif and pre-peptide sequence in Cluster1 and Cluster2 enzymes. The x-axes correspond to sequence length and the y-axes correspond to the frequency of enzymes. (d, e) The quaternary structures of the enzymes *B/BSH* (d) and *BtBSH* (e). The individual subunits of the homotetramers are labelled A-D. The 26 Å loop extensions of subunits A and D of *B/BSH*, each interacting with the neighboring subunit, are highlighted by circles.

2.3.6. Evolutionary basis for the divergence of CGH family members into two clusters

The detailed sequence analysis of the members of the dataset revealed a crucial 13-19 amino acid indel, which resulted in the separation of the members into these two distinct clusters. This indel corresponds to the presence or absence of the ‘assembly motif’ in the sequences. Irrespective of their source (Gram-positive bacteria, Gram-negative bacteria or archaea) and function (BSH or PVA), all members of Cluster1 possessed this motif, whereas those belonging to Cluster2 lacked the motif (Fig. 2.13b and Fig. 2.13c). Most CGH family members form homotetramers as quaternary structures. In *BtBSH*, this assembly motif is about 26 Å long (Fig. 2.13a) and comprises residues 188-220 (Kumar *et al.*, 2006). The long assembly motif of each monomer extends into its neighbouring monomer (diagonally opposite) helping in tetramer assembly and stabilization (Fig. 2.13d). However, in the case of *BtBSH* belonging to Cluster2, the absence of this assembly motif (Fig. 2.13a) results in the quaternary association (Fig. 2.13e) being less thermodynamically stable than that of *BtBSH*.

Theoretical estimation of thermodynamic stabilities of the tetramers by **PISA** (Krissinel & Henrick, 2007) showed that the absence of this motif in the *BtBSH* structure reduces the thermodynamic stability of its quaternary association compared with *BtBSH*. The values of ΔG^{diss} (free energy of assembly dissociation) corresponding to *BtBSH* and *BtBSH* assembly were estimated as 48.5 and 4.9 kcal mol⁻¹, respectively (positive values indicating thermodynamically stable assembly and an external driving force would be required to dissociate it). The two tetramers also differed in terms of the extent of their subunit interface area, and number of non-bonded, hydrogen bonding and salt bridge interactions between the subunits (Table 2.5). A major difference observed was in the AD and BC interface (Fig. 2.13d and Fig. 2.13e), where the absence of the assembly motif reduced significantly the interface area in the *BtBSH* structure (Table 2.5).

Table 2.5: Quantitative estimation of interface area and the number of non-bonded interactions between individual subunits of *B*/BSH and *Bt*/BSH in their quaternary structures.**

Property	Interface					
	<i>B</i> /BSH			<i>B</i> <i>t</i> /BSH		
	AB/CD	AC/BD	AD/BC	AB/CD	AC/BD	AD/BC
Interface area (Å ²)	1954.8	302.6	1506.7	1588.4	535.3	201.0
ΔG (kcal mol ⁻¹)	226.4	21.0	223.9	214.1	25.4	22.7
No. interface residues	40	6	31	29	11	5
No. hydrogen bonds	22	4	12	20	4	0
No. salt bridges	6	0	0	0	0	0
No. non-bonded contacts	280	32	178	144	53	17

Another difference observed between the members of the two clusters was the length of pre-peptide sequence preceding the N-terminal catalytic cysteine residue. The CGH enzymes undergo a post-translational modification to auto-catalytically remove this pre-peptide sequence to obtain a mature enzyme (Chandra *et al.*, 2005). Most members of Cluster2 possess a comparatively longer pre-peptide sequence than Cluster1 members (Fig. 2.13c). Among the Cluster2 members (mostly Gram-negative), in most cases the pre-peptide sequence was also found to act as a signal sequence, hinting at the possible translocation of these enzymes into the periplasmic space. It is known in Gram-negative bacteria that many enzymes participating in the inactivation of antibiotics are located in the periplasmic space (Gupta, 2011). The localization of penicillin G acylases and cephalosporin acylases in the periplasm involved in the inactivation of antibiotics is already known. As the above enzymes belong to the Ntn-hydrolase superfamily to which the CGH family belongs, it might be assumed that a similar translocation could also occur in the Gram-negative CGH family members. Of the three archaea belonging to Cluster2, only *Natrialba aegyptia* possesses a four-residue pre-peptide sequence. A majority of members of Cluster1 (mostly Gram-positive) were found to have a single methionine residue, preceding the N-terminal cysteine, which in most cases is proteolytically removed by methionyl

aminopeptidase enzyme (Ben-Bassat *et al.*, 1987). Only 16 members of Cluster1 contained a pre-peptide of more than one residue (Fig. 2.13c).

The above observations could thus help to understand the evolution of CGH family members along with the evolution of Gram-positive bacteria, Gram-negative bacteria and archaea. One hypothesis states that antibiotic selection pressure could be a major evolutionary force behind the evolution of diderms (true Gram-negative bacteria) from monoderms (true Gram-positive bacteria) (Gupta, 2011). As many CGH enzymes are known to inactivate antibiotics such as penicillin V, it could be hypothesized that during the evolution of diderms from monoderms the assembly motif could have been deleted in the case of the members of the CGH family belonging to Cluster2 (Gram-negative). The existence of eight CGH enzymes belonging to Cluster2 of the order *Corynebacterineae* (Gram-positive bacteria) shows this transition wherein the enzymes lack the assembly motif similar to that of Gram-negative bacteria (Table 2.6).

In the case of archaeal CGH enzymes, the presence of the tetramer assembly motif among the three archaeal members of Cluster1 (Table 2.6) indicates their close relation to Gram-positive members, which is also supported by a recent study proposing the emergence of archaea from Gram-positive bacteria in response to antibiotic selection pressure (Valas & Bourne, 2011). Interestingly, the three archaeal members of Cluster2 also lacked the assembly motif and were grouped phylogenetically along with the *Corynebacterineae* members, the intermediates between Gram-positive and Gram-negative (Fig. 2.4).

The possibility of CGH genes being transferred by horizontal gene transfer was also analyzed. The comparison of GC content of the available genome sequences in the dataset, their CGH genes and the region flanking the CGH genes ruled out the possibility of horizontal gene transfer as these were found to be almost similar.

Table 2.6: Describes the functional annotation of CGH family members into BSH or PVA based on binding site similarity (BSS) based scoring system. The columns from left to right correspond to NCBI GI Number, Cluster to which the enzyme belong (C1: Cluster 1, C2: Cluster2), Organism source, Phylum to which the species belong, Length of pre-peptide, Length of tetramer assembly motif, BSS scores of each enzyme with the six template enzymes, The functional annotation of each member into BSH or PVA and Organism detail, either gut-inhabiting or pathogenic or organism of soil or aquatic environment.

Gi Number	Cluster	Species	Phylum	Length of		Binding site similarity score (BSS score)						Annotated Group	Gut	Pathogen	Environment
				Pre-peptide	Assembly Motif	BBSH	CpBSH	BspPVA	BsuPVA	BtBSH	PaPVA				
345898563	C1	<i>Collinsella tanakaei</i> YIT 12063	Actinobacteria	1	23	122	77	16	24	-5	17	Cluster1-BSH	Y		
229785384	C1	<i>Bifidobacterium angulatum</i> DSM 20098 = JCM 7096	Actinobacteria	1	24	124	66	-5	14	26	8	Cluster1-BSH	Y		
404389668	C1	<i>Slackia piriformis</i> YIT 12062	Actinobacteria	1	24	143	65	-25	-13	13	1	Cluster1-BSH	Y		
83630914	C1	<i>Bifidobacterium bifidum</i>	Actinobacteria	1	24	190	75	8	21	3	26	Cluster1-BSH	Y		
489938502	C1	<i>Bifidobacterium dentium</i>	Actinobacteria	10	24	203	68	-3	6	-5	-6	Cluster1-BSH	Y		
154083794	C1	<i>Bifidobacterium adolescentis</i> L2-32	Actinobacteria	32	24	207	68	-3	5	-5	-2	Cluster1-BSH	Y		
291382017	C1	<i>Bifidobacterium breve</i> DSM 20213 = JCM 1192	Actinobacteria	18	24	229	76	6	15	-2	10	Cluster1-BSH	Y		
430438302	C1	<i>Enterococcus faecium</i> E0045	Firmicutes	1	24	69	143	-28	9	-5	-7	Cluster1-BSH	Y		
226911510	C1	<i>Clostridium</i> sp. 7_2_43FAA	Firmicutes	1	24	74	144	-4	21	-4	8	Cluster1-BSH	Y		
490131588	C1	<i>Methanobacterium formicum</i>	Euryarchaeota	1	24	50	110	1	17	-3	29	Cluster1-BSH	Y		
182378475	C1	<i>Clostridium butyricum</i> 5521	Firmicutes	1	24	65	123	9	21	-8	19	Cluster1-BSH	Y		
494496139	C1	<i>Clostridium</i>] bartlettii	Firmicutes	1	24	73	127	-2	26	-5	21	Cluster1-BSH	Y		
488447018	C1	<i>Lactobacillus acidophilus</i> La-14	Firmicutes	1	24	56	99	3	-11	-8	-13	Cluster1-BSH	Y		
326542554	C1	<i>Clostridium lentocellum</i> DSM 5427	Firmicutes	1	24	94	134	-4	23	-21	-3	Cluster1-BSH	Y		
493545525	C1	<i>Lactobacillus mucosae</i>	Firmicutes	1	24	79	113	7	6	-5	0	Cluster1-BSH	Y		
493884517	C1	<i>Planococcus antarcticus</i>	Firmicutes	1	25	83	116	4	6	-15	7	Cluster1-BSH			Y
490742377	C1	<i>Eubacterium cellulosolvens</i>	Firmicutes	1	24	88	119	-3	14	-13	-3	Cluster1-BSH	Y		
329667206	C1	<i>Lactobacillus johnsonii</i> DPC 6026	Firmicutes	1	24	72	101	3	1	-27	-14	Cluster1-BSH	Y		

295099977	C1	<i>Eubacterium] cylindroides T2-87</i>	<i>Firmicutes</i>	1	24	59	87	-4	14	-17	-3	Cluster1-BSH	Y		
257202513	C1	<i>Roseburia intestinalis L1-82</i>	<i>Firmicutes</i>	5	24	76	103	0	6	-18	-18	Cluster1-BSH	Y		
489793954	C1	<i>Lactobacillus ruminis</i>	<i>Firmicutes</i>	1	24	90	117	5	5	-4	-21	Cluster1-BSH	Y		
149830848	C1	<i>Ruminococcus obeum ATCC 29174</i>	<i>Firmicutes</i>	1	24	72	97	4	8	-17	4	Cluster1-BSH	Y		
313607689	C1	<i>Listeria monocytogenes FSL F2-208</i>	<i>Firmicutes</i>	12	24	82	106	10	-1	-20	-3	Cluster1-BSH	Y	Y	
424714946	C1	<i>Listeria monocytogenes serotype 4b str. LL195</i>	<i>Firmicutes</i>	60	24	82	106	10	-1	-20	-3	Cluster1-BSH	Y	Y	
238872917	C1	<i>Eubacterium eligens ATCC 27750</i>	<i>Firmicutes</i>	1	24	63	86	8	4	-17	-5	Cluster1-BSH	Y		
489962944	C1	<i>Eubacterium] bifforme</i>	<i>Firmicutes</i>	1	22	66	89	-13	-2	-14	-4	Cluster1-BSH	Y		
224525889	C1	<i>Catenibacterium mitsuokai DSM 15897</i>	<i>Firmicutes</i>	1	24	68	90	-2	3	-17	4	Cluster1-BSH	Y		
292646228	C1	<i>Turicibacter sanguinis PC909</i>	<i>Firmicutes</i>	18	24	56	78	-12	-16	-7	-9	Cluster1-BSH	Y		
197298802	C1	<i>Ruminococcus lactaris ATCC 29176</i>	<i>Firmicutes</i>	1	24	76	96	5	5	-15	-1	Cluster1-BSH	Y		
251848185	C1	<i>Ruminococcus sp. 5_1_39BFAA</i>	<i>Firmicutes</i>	1	24	71	91	3	-2	-21	-13	Cluster1-BSH	Y		
490135397	C1	<i>Methanobrevibacter smithii</i>	<i>Euryarchaeota</i>	1	24	73	93	6	6	13	-6	Cluster1-BSH	Y		
490988163	C1	<i>Coprococcus eutactus</i>	<i>Firmicutes</i>	1	24	63	83	-4	9	-15	1	Cluster1-BSH	Y		
84372883	C1	<i>Methanosphaera stadmanae DSM 3091</i>	<i>Euryarchaeota</i>	1	24	64	84	11	-8	-13	-4	Cluster1-BSH	Y		
227070078	C1	<i>Lactobacillus reuteri MM2-3</i>	<i>Firmicutes</i>	1	24	66	85	-1	3	-24	-17	Cluster1-BSH	Y		
495392225	C1	<i>Bacteroides] pectinophilus</i>	<i>Firmicutes</i>	10	24	69	88	7	3	-17	-4	Cluster1-BSH	Y		
495749692	C1	<i>Lactobacillus gigeriorum</i>	<i>Firmicutes</i>	1	24	86	104	-13	-5	-16	-8	Cluster1-BSH	Y		
282572108	C1	<i>Subdoligranulum variabile DSM 15176</i>	<i>Firmicutes</i>	20	24	75	92	-7	27	6	-1	Cluster1-BSH	Y		
227866194	C1	<i>Lactobacillus salivarius ATCC 11741</i>	<i>Firmicutes</i>	1	24	98	114	-6	7	-27	-11	Cluster1-BSH	Y		
489154476	C1	<i>Streptococcus equinus</i>	<i>Firmicutes</i>	1	24	92	107	4	17	-9	-17	Cluster1-BSH	Y		
312280178	C1	<i>Lactobacillus delbrueckii subsp. bulgaricus ND02</i>	<i>Firmicutes</i>	1	24	97	111	-10	2	11	-17	Cluster1-BSH	Y		
238925181	C1	<i>Eubacterium rectale ATCC 33656</i>	<i>Firmicutes</i>	1	24	77	89	4	10	-22	2	Cluster1-BSH	Y		
492023095	C1	<i>Lactobacillus crispatus</i>	<i>Firmicutes</i>	1	24	83	95	0	11	-20	-25	Cluster1-BSH	Y		
167710920	C1	<i>Clostridium sp. SS2/1</i>	<i>Firmicutes</i>	1	24	66	77	5	1	-13	10	Cluster1-BSH	Y		
354823669	C1	<i>Clostridium sp. 7_3_54FAA</i>	<i>Firmicutes</i>	1	24	69	79	-5	5	-16	-2	Cluster1-BSH	Y		
447912393	C1	<i>Enterococcus faecium NRRL B-2354</i>	<i>Firmicutes</i>	1	24	72	82	6	-9	-11	-9	Cluster1-BSH	Y		
496264670	C1	<i>Erysipelotrichaceae bacterium 5_2_54FAA</i>	<i>Firmicutes</i>	4	24	69	79	-5	5	-16	-2	Cluster1-BSH	Y		
345901428	C1	<i>Erysipelotrichaceae bacterium 2_2_44A</i>	<i>Firmicutes</i>	1	24	83	92	-5	-2	-27	-12	Cluster1-BSH	Y		

268318617	C1	<i>Lactobacillus johnsonii</i> FI9785	Firmicutes	1	24	81	86	12	26	-14	23	Cluster1-BSH	Y		
493974500	C1	<i>Lactobacillus coleohominis</i>	Firmicutes	1	24	104	105	9	4	-9	-17	Cluster1-BSH	Y		
292809624	C1	<i>Butyrivibrio crossotus</i> DSM 2876	Firmicutes	1	24	51	50	6	-1	-1	-1	Cluster1-BSH	Y		
497155328	C1	<i>Catelicoccus marimammalium</i>	Firmicutes	1	24	69	64	-1	9	-20	-10	Cluster1-BSH	Y		
493579269	C1	<i>Streptococcus infantarius</i>	Firmicutes	11	24	104	97	12	13	-15	-20	Cluster1-BSH	Y		
494199010	C1	<i>Lactobacillus antri</i>	Firmicutes	6	24	109	100	16	21	-13	2	Cluster1-BSH	Y		
270277784	C1	<i>Bifidobacterium gallicum</i> DSM 20093	Actinobacteria	1	24	96	86	-6	15	14	17	Cluster1-BSH	Y		
385701118	C1	<i>Bifidobacterium animalis</i> subsp. <i>animalis</i> ATCC 25527	Actinobacteria	1	24	99	85	-3	13	9	32	Cluster1-BSH	Y		
504295775	C1	<i>Bifidobacterium animalis</i>	Actinobacteria	19	24	99	85	-3	13	9	32	Cluster1-BSH	Y		
256791585	C1	<i>Slackia heliotrinireducens</i> DSM 20476	Actinobacteria	1	24	122	95	-2	17	-5	-10	Cluster1-BSH	Y		
291486575	C1	<i>Bacillus subtilis</i> subsp. <i>natto</i> BEST195	Firmicutes	1	24	15	19	38	210	-10	10	Cluster1-PVA			Y
504285267	C1	<i>Bacillus amyloliquefaciens</i>	Firmicutes	38	24	24	27	35	159	-6	7	Cluster1-PVA			Y
498304932	C1	<i>Lactobacillus malefermentans</i>	Firmicutes	1	24	1	24	40	92	-13	6	Cluster1-PVA			
225041277	C1	<i>Clostridium asparagiforme</i> DSM 15981	Firmicutes	1	24	14	37	43	82	7	-2	Cluster1-PVA			
495060062	C1	<i>Desulfosporosinus youngiae</i>	Firmicutes	1	24	10	35	37	75	-5	6	Cluster1-PVA			
480642171	C1	<i>Clostridium clostridioforme</i> CM201	Firmicutes	1	24	13	30	47	81	9	5	Cluster1-PVA		Y	
354814584	C1	<i>Clostridium citroniae</i> WAL-17108	Firmicutes	1	24	24	20	51	83	-1	7	Cluster1-PVA	Y		
150018493	C1	<i>Clostridium beijerinckii</i> NCIMB 8052	Firmicutes	1	24	18	12	37	58	13	9	Cluster1-PVA	Y		
228605909	C1	<i>Bacillus cereus</i> 172560W	Firmicutes	12	24	18	24	65	82	17	12	Cluster1-PVA		Y	
375285682	C1	<i>Bacillus cereus</i> NC7401	Firmicutes	1	24	12	22	58	74	20	4	Cluster1-PVA		Y	
296047217	C1	<i>Clostridium carboxidivorans</i> P7	Firmicutes	1	24	-1	10	42	49	19	26	Cluster1-PVA			Y
295319524	C1	<i>Clostridium botulinum</i> F str. 230613	Firmicutes	1	24	40	59	79	80	-9	9	Cluster1-PVA	Y		
496352295	C1	<i>Clostridium</i> sp. MSTE9	Firmicutes	1	24	5	12	57	54	8	4	Cluster1-PVA			
401268676	C1	<i>Bacillus cereus</i> VD156	Firmicutes	15	24	19	30	69	65	16	-3	Cluster1-PVA		Y	
495300065	C2	<i>Bacteroides xylanisolvens</i>	Bacteroidetes	25	6	4	16	-16	-7	200	49	Cluster2-BSH	Y		
404339926	C2	<i>Barnesiella intestinihominis</i> YIT 11860	Bacteroidetes	28	6	-3	23	-16	-7	193	51	Cluster2-BSH	Y		
317385358	C2	<i>Bacteroides</i> sp. 3_1_40A	Bacteroidetes	37	6	1	9	-13	-1	178	54	Cluster2-BSH	Y		
392644320	C2	<i>Bacteroides dorei</i> CL03T12C01	Bacteroidetes	26	6	1	9	-13	-1	178	54	Cluster2-BSH	Y		
392697615	C2	<i>Bacteroides fragilis</i> CL05T12C13	Bacteroidetes	20	6	1	9	-13	-1	178	54	Cluster2-BSH	Y		

392686197	C2	<i>Bacteroides uniformis</i> CL03T12C37	<i>Bacteroidetes</i>	25	6	0	7	-15	5	169	52	Cluster2-BSH	Y		
363642401	C2	<i>Tannerella</i> sp. 6_1_58FAA_CT1	<i>Bacteroidetes</i>	26	6	-13	-4	-14	-9	156	46	Cluster2-BSH	Y		
392662906	C2	<i>Bacteroides salyersiae</i> CL02T12C01	<i>Bacteroidetes</i>	27	6	0	8	-10	3	156	56	Cluster2-BSH	Y		
291513698	C2	<i>Alistipes shahii</i> WAL 8301	<i>Bacteroidetes</i>	23	6	-1	-3	-4	-4	140	43	Cluster2-BSH	Y		
492836855	C2	<i>Desulfovibrio fructosovorans</i>	<i>Proteobacteria</i>	27	6	5	8	3	7	125	87	Cluster2-BSH			Y
496342310	C2	<i>Pseudomonas chlororaphis</i>	<i>Proteobacteria</i>	46	6	0	5	-10	0	121	78	Cluster2-BSH			Y
502932959	C2	<i>Starkeya novella</i>	<i>Proteobacteria</i>	1	6	1	11	-6	-6	120	80	Cluster2-BSH			Y
238702377	C2	<i>Yersinia aldovae</i> ATCC 35236	<i>Proteobacteria</i>	27	6	1	7	-8	3	120	83	Cluster2-BSH			Y
145588882	C2	<i>Polynucleobacter necessarius</i> subsp. <i>asymbioticus</i> QLW-P1DMWA-1	<i>Proteobacteria</i>	1	6	-16	-11	-12	5	114	70	Cluster2-BSH			Y
377811344	C2	<i>Burkholderia</i> sp. YI23	<i>Proteobacteria</i>	29	6	-4	4	-12	13	113	65	Cluster2-BSH			Y
428771647	C2	<i>Cyanobacterium aponinum</i> PCC 10605	<i>Cyanobacteria</i>	30	6	0	3	-10	-1	110	75	Cluster2-BSH			Y
413930529	C2	<i>Burkholderia</i> sp. SJ98	<i>Proteobacteria</i>	24	6	-4	4	-16	9	109	69	Cluster2-BSH			Y
488731201	C2	<i>Blastopirellula marina</i>	<i>Planctomycetes</i>	29	6	-2	2	-2	11	109	70	Cluster2-BSH			Y
503184202	C2	<i>Rhodomicrobium vannielii</i>	<i>Proteobacteria</i>	23	6	3	-14	11	-7	94	54	Cluster2-BSH			Y
67005062	C2	<i>Rickettsia felis</i> URRWXCal2	<i>Proteobacteria</i>	42	6	-24	-13	-3	-22	82	50	Cluster2-BSH		Y	
470153660	C2	<i>Pectobacterium</i> sp. SCC3193	<i>Proteobacteria</i>	29	6	5	7	-3	12	65	183	Cluster2-PVA		Y	
497991375	C2	<i>Pectobacterium carotovorum</i>	<i>Proteobacteria</i>	28	6	9	7	4	10	57	144	Cluster2-PVA		Y	
326550145	C2	<i>Sphingobacterium</i> sp. 21	<i>Bacteroidetes</i>	22	6	-9	8	-8	-2	83	142	Cluster2-PVA			Y
85698330	C2	<i>Nitrobacter</i> sp. Nb-311A	<i>Proteobacteria</i>	43	6	-7	0	5	8	68	133	Cluster2-PVA			Y
338822159	C2	<i>Agrobacterium tumefaciens</i> F2	<i>Proteobacteria</i>	29	6	-9	1	-7	-2	75	133	Cluster2-PVA		Y	
68344683	C2	<i>Pseudomonas protegens</i> Pf-5	<i>Proteobacteria</i>	36	6	-8	6	-13	-3	82	133	Cluster2-PVA		Y	
495000397	C2	<i>Rhodococcus qingshengii</i>	<i>Actinobacteria</i>	1	6	-5	11	-15	8	69	127	Cluster2-PVA			Y
429186785	C2	<i>Brevundimonas diminuta</i> 470-4	<i>Proteobacteria</i>	21	6	-1	2	-3	-3	72	127	Cluster2-PVA		Y	
353672352	C2	<i>Commensalibacter intestini</i> A911	<i>Proteobacteria</i>	21	6	0	6	-10	2	92	127	Cluster2-PVA	Y		
81169840	C2	<i>Synechococcus elongatus</i> PCC 7942	<i>Cyanobacteria</i>	24	6	1	17	-7	15	72	126	Cluster2-PVA			Y
451921404	C2	<i>Bordetella holmesii</i> F627	<i>Proteobacteria</i>	33	6	-7	2	0	0	79	126	Cluster2-PVA		Y	
407024545	C2	<i>Alcanivorax pacificus</i> W11-5	<i>Proteobacteria</i>	28	6	-7	5	-11	-5	82	125	Cluster2-PVA			Y
480039510	C2	<i>Acinetobacter ursingii</i> ANC 3649	<i>Proteobacteria</i>	24	6	-10	11	-9	-8	86	125	Cluster2-PVA		Y	
238728065	C2	<i>Yersinia intermedia</i> ATCC 29909	<i>Proteobacteria</i>	28	6	-11	-8	-12	-1	65	124	Cluster2-PVA	Y		

491383263	C2	<i>Acinetobacter sp. CIP 64.2</i>	<i>Proteobacteria</i>	22	6	0	1	-7	-7	78	124	Cluster2-PVA		Y	
493580739	C2	<i>Proteus penneri</i>	<i>Proteobacteria</i>	27	6	-5	9	1	5	85	124	Cluster2-PVA		Y	
445654763	C2	<i>Natrialba aegyptia DSM 13077</i>	<i>Euryarchaeota</i>	4	6	4	14	-18	17	90	124	Cluster2-PVA			Y
212684636	C2	<i>Providencia alcalifaciens DSM 30120</i>	<i>Proteobacteria</i>	52	6	-7	7	-1	3	91	122	Cluster2-PVA		Y	
115421851	C2	<i>Bordetella avium 197N</i>	<i>Proteobacteria</i>	31	6	-11	-1	-5	1	79	121	Cluster2-PVA		Y	
310759557	C2	<i>Achromobacter xylosoxidans A8</i>	<i>Proteobacteria</i>	29	6	-10	5	-5	1	79	121	Cluster2-PVA			Y
491051621	C2	<i>Providencia rettgeri</i>	<i>Proteobacteria</i>	26	6	-8	7	-2	8	90	121	Cluster2-PVA		Y	
480106373	C2	<i>Acinetobacter baumannii NIPH 335</i>	<i>Proteobacteria</i>	23	6	-7	5	-6	-4	70	120	Cluster2-PVA		Y	
404607402	C2	<i>Myroides odoratimimus CCUG 3837</i>	<i>Bacteroidetes</i>	26	6	-18	3	-10	-4	67	119	Cluster2-PVA		Y	
226835088	C2	<i>Acinetobacter sp. ATCC 27244</i>	<i>Proteobacteria</i>	9	6	0	6	-8	-7	78	119	Cluster2-PVA		Y	
119959277	C2	<i>Mycobacterium vanbaalenii PYR-1</i>	<i>Actinobacteria</i>	11	6	3	10	-11	4	79	119	Cluster2-PVA			Y
187719716	C2	<i>Burkholderia phytofirmans PsJN</i>	<i>Proteobacteria</i>	27	6	-2	4	-7	-5	80	119	Cluster2-PVA			Y
492303970	C2	<i>Acinetobacter pittii</i>	<i>Proteobacteria</i>	32	6	-10	5	-7	-7	69	118	Cluster2-PVA		Y	
197285965	C2	<i>Proteus mirabilis HI4320</i>	<i>Proteobacteria</i>	28	6	-5	9	7	11	79	118	Cluster2-PVA		Y	
505079989	C2	<i>Echinicola vietnamensis</i>	<i>Bacteroidetes</i>	21	6	-3	-1	-15	-2	61	117	Cluster2-PVA			Y
496272799	C2	<i>Alishewanella agri</i>	<i>Proteobacteria</i>	23	6	0	19	-11	8	74	117	Cluster2-PVA			Y
494983306	C2	<i>Sphingobium sp. AP49</i>	<i>Proteobacteria</i>	26	6	-6	3	-10	-7	76	117	Cluster2-PVA			Y
282566166	C2	<i>Providencia rustigianii DSM 4541</i>	<i>Proteobacteria</i>	41	6	-7	2	0	4	86	117	Cluster2-PVA		Y	
430759621	C2	<i>Thioalkalivibrio nitratreducens DSM 14787</i>	<i>Proteobacteria</i>	10	6	-2	15	-14	12	68	116	Cluster2-PVA			Y
496382143	C2	<i>Elizabethkingia anophelis</i>	<i>Bacteroidetes</i>	28	6	3	10	-15	4	67	115	Cluster2-PVA	Y		
300761964	C2	<i>Sphingobacterium spiritivorum ATCC 33861</i>	<i>Bacteroidetes</i>	33	6	-3	7	-19	8	70	115	Cluster2-PVA		Y	
160898906	C2	<i>Delftia acidovorans SPH-1</i>	<i>Proteobacteria</i>	49	6	-10	4	-8	-5	79	115	Cluster2-PVA			Y
495727822	C2	<i>Natrinema gari</i>	<i>Euryarchaeota</i>	1	6	0	11	-3	-1	77	112	Cluster2-PVA			Y
264677961	C2	<i>Comamonas testosteroni CNB-2</i>	<i>Proteobacteria</i>	30	6	-12	-5	-8	-12	72	111	Cluster2-PVA			Y
489153403	C2	<i>Comamonas testosteroni</i>	<i>Proteobacteria</i>	6	6	-12	-5	-8	-12	72	111	Cluster2-PVA			Y
506432348	C2	<i>Methylobacterium extorquens</i>	<i>Proteobacteria</i>	29	6	5	14	-12	-3	55	102	Cluster2-PVA			Y
325109480	C2	<i>Planctomyces brasiliensis DSM 5305</i>	<i>Planctomycetes</i>	28	6	-14	-5	-10	1	48	100	Cluster2-PVA			Y
261837273	C2	<i>Halothiobacillus neapolitanus c2</i>	<i>Proteobacteria</i>	30	6	-2	4	-21	4	53	96	Cluster2-PVA			Y
300502137	C2	<i>Chryseobacterium gleum ATCC 35910</i>	<i>Bacteroidetes</i>	23	6	-6	7	-19	7	62	96	Cluster2-PVA		Y	
119373429	C2	<i>Paracoccus denitrificans PDI222</i>	<i>Proteobacteria</i>	22	6	-9	-2	-5	-6	47	95	Cluster2-PVA			Y

493506038	C2	<i>Thioalkalimicrobium aerophilum</i>	<i>Proteobacteria</i>	34	6	-16	8	-4	19	54	87	Cluster2-PVA			Y
336282685	C2	<i>Idiomarina sp. A28L</i>	<i>Proteobacteria</i>	24	6	-12	6	2	10	48	78	Cluster2-PVA			Y
268615970	C2	<i>Sebaldella termitidis ATCC 33386</i>	<i>Fusobacteria</i>	23	6	0	-14	-15	-4	-7	51	Cluster2-PVA	Y		
218440975	C2	<i>Cyanothece sp. PCC 7424</i>	<i>Cyanobacteria</i>	1	6	-26	-29	-26	-10	-19	18	Cluster2-PVA			Y
158329534	C2	<i>Azorhizobium caulinodans ORS 571</i>	<i>Proteobacteria</i>	43	6	6	16	-3	1	90	63	Cluster2-BSH/PVA			Y
434391542	C2	<i>Gloeocapsa sp. PCC 7428</i>	<i>Cyanobacteria</i>	31	6	1	20	4	5	95	69	Cluster2-BSH/PVA			Y
428772046	C2	<i>Cyanobacterium stanieri PCC 7202</i>	<i>Cyanobacteria</i>	12	6	-4	2	-14	-9	59	34	Cluster2-BSH/PVA			
498360106	C2	<i>Aeromonas caviae</i>	<i>Proteobacteria</i>	28	6	0	19	-7	2	89	65	Cluster2-BSH/PVA		Y	
373942010	C2	<i>Ectothiorhodospira sp. PHS-1</i>	<i>Proteobacteria</i>	26	6	-4	6	-11	-12	77	55	Cluster2-BSH/PVA			Y
427376666	C2	<i>Synechococcus sp. PCC 6312</i>	<i>Cyanobacteria</i>	32	7	-4	9	1	-11	57	35	Cluster2-BSH/PVA			Y
148845339	C2	<i>Planctomyces maris DSM 8797</i>	<i>Planctomycetes</i>	1	6	-8	10	-1	8	92	71	Cluster2-BSH/PVA			Y
492547731	C2	<i>Oxalobacter formigenes</i>	<i>Proteobacteria</i>	29	6	19	-7	11	-15	70	49	Cluster2-BSH/PVA	Y		
297620761	C2	<i>Waddlia chondrophila WSU 86-1044</i>	<i>Chlamydiae</i>	1	7	-6	-11	4	-18	39	20	Cluster2-BSH/PVA		Y	
434398426	C2	<i>Staniera cyanosphaera PCC 7437</i>	<i>Cyanobacteria</i>	25	6	-2	23	5	8	92	73	Cluster2-BSH/PVA			Y
498270648	C2	<i>Schlesneria paludicola</i>	<i>Planctomycetes</i>	30	6	-3	10	2	4	88	69	Cluster2-BSH/PVA			Y
113880097	C2	<i>Synechococcus sp. CC9311</i>	<i>Cyanobacteria</i>	24	6	0	14	2	12	93	75	Cluster2-BSH/PVA			Y
296122135	C2	<i>Planctomyces limnophilus DSM 3776</i>	<i>Planctomycetes</i>	10	6	-10	-6	-12	-6	77	60	Cluster2-BSH/PVA			Y
492729669	C2	<i>Gordonia hirsuta</i>	<i>Actinobacteria</i>	1	6	-9	12	11	7	57	40	Cluster2-BSH/PVA			
333918774	C2	<i>Amycolicococcus subflavus DQS3-9A1</i>	<i>Actinobacteria</i>	1	6	6	14	-11	-19	75	60	Cluster2-BSH/PVA			Y
154160155	C2	<i>Xanthobacter autotrophicus Py2</i>	<i>Proteobacteria</i>	91	6	-6	7	2	10	100	86	Cluster2-BSH/PVA			Y
307545218	C2	<i>Halomonas elongata DSM 2581</i>	<i>Proteobacteria</i>	27	7	-11	-23	-7	-16	46	32	Cluster2-BSH/PVA			Y
356454792	C2	<i>Vibrio cholerae HC-61A1</i>	<i>Proteobacteria</i>	31	6	-5	-2	-5	15	79	71	Cluster2-BSH/PVA		Y	
360037775	C2	<i>Vibrio cholerae O1 str. 2010EL-1786</i>	<i>Proteobacteria</i>	33	6	-5	-2	-5	15	79	71	Cluster2-BSH/PVA		Y	
402773859	C2	<i>Methylocystis sp. SC2</i>	<i>Proteobacteria</i>	25	6	-7	-2	2	1	70	63	Cluster2-BSH/PVA			Y
207089003	C2	<i>Nitrosococcus oceani AFC27</i>	<i>Proteobacteria</i>	48	9	-17	-16	-6	0	48	42	Cluster2-BSH/PVA			Y
76884707	C2	<i>Nitrosococcus oceani ATCC 19707</i>	<i>Proteobacteria</i>	29	9	-17	-16	-6	0	48	42	Cluster2-BSH/PVA			Y
494325418	C2	<i>Burkholderia sp. Ch1-1</i>	<i>Proteobacteria</i>	30	7	-7	5	11	-1	43	38	Cluster2-BSH/PVA			Y
242121098	C2	<i>Desulfovibrio salexigens DSM 2638</i>	<i>Proteobacteria</i>	33	6	8	7	-16	3	90	86	Cluster2-BSH/PVA			Y
499173823	C2	<i>Synechocystis sp. PCC 6803</i>	<i>Cyanobacteria</i>	53	7	-4	4	-11	-4	45	41	Cluster2-BSH/PVA			Y
491364988	C2	<i>Marichromatium purpuratum</i>	<i>Proteobacteria</i>	25	6	10	23	-3	8	83	80	Cluster2-BSH/PVA			Y

238715731	C2	<i>Yersinia bercovieri</i> ATCC 43970	Proteobacteria	60	7	-12	10	0	-15	35	33	Cluster2-BSH/PVA		Y	
183982580	C2	<i>Mycobacterium marinum</i> M	Actinobacteria	1	6	1	7	-6	-7	63	63	Cluster2-BSH/PVA			Y
46486690	C2	<i>Lyngbya majuscula</i>	Cyanobacteria	33	7	-21	-9	0	-12	49	49	Cluster2-BSH/PVA			Y
494546223	C2	<i>Rhodopirellula baltica</i>	Planctomycetes	29	7	8	3	19	-5	40	41	Cluster2-BSH/PVA			Y
373545965	C2	<i>Mycobacterium tusciae</i> JS617	Actinobacteria	1	6	4	12	-28	-11	55	58	Cluster2-BSH/PVA		Y	
39574783	C2	<i>Bdellovibrio bacteriovorus</i> HD100	Proteobacteria	18	6	0	6	-9	-12	68	71	Cluster2-BSH/PVA			Y
389826809	C2	<i>Microcystis aeruginosa</i> PCC 9808	Cyanobacteria	25	7	-24	-10	-12	0	40	44	Cluster2-BSH/PVA			Y
494446686	C2	<i>Gordonia otitidis</i>	Actinobacteria	1	6	-10	7	17	5	54	60	Cluster2-BSH/PVA		Y	
353192745	C2	<i>Mycobacterium rhodesiae</i> JS60	Actinobacteria	1	6	-4	-2	-11	-15	34	42	Cluster2-BSH/PVA			Y
288569518	C2	<i>Dethiosulfovibrio peptidovorans</i> DSM 11002	Synergistetes	25	6	-3	1	-5	-13	67	78	Cluster2-BSH/PVA			Y
357388206	C2	<i>Kitasatospora setae</i> KM-6054	Actinobacteria	4	6	11	16	-9	-15	63	75	Cluster2-BSH/PVA			Y
410881228	C2	<i>Afipia felis</i> ATCC 53690	Proteobacteria	24	7	-5	1	-4	-14	41	53	Cluster2-BSH/PVA		Y	
306983931	C2	<i>Cyanothece</i> sp. PCC 7822	Cyanobacteria	1	6	-21	7	-11	-16	46	60	Cluster2-BSH/PVA			Y
493797143	C2	<i>Bacteroides coprosuis</i>	Bacteroidetes	21	6	4	-4	-3	1	85	99	Cluster2-BSH/PVA	Y		
393165158	C2	<i>Alcaligenes faecalis</i> subsp. <i>faecalis</i> NCIB 8687	Proteobacteria	26	6	-7	9	-14	15	94	109	Cluster2-BSH/PVA			Y
496016706	C2	<i>Streptomyces</i> sp. <i>Mg1</i>	Actinobacteria	1	12	-17	-3	12	-9	30	47	Cluster2-BSH/PVA			Y
110281458	C2	<i>Cytophaga hutchinsonii</i> ATCC 33406	Bacteroidetes	26	6	-17	-6	-17	-6	61	79	Cluster2-BSH/PVA			Y
307158047	C2	<i>Methanoplanus petrolearius</i> DSM 11571	Euryarchaeota	1	6	-11	1	-18	-12	47	66	Cluster2-BSH/PVA			Y
146154999	C2	<i>Flavobacterium johnsoniae</i> UW101	Bacteroidetes	24	6	-17	-8	-19	2	59	79	Cluster2-BSH/PVA			Y
190012029	C2	<i>Stenotrophomonas maltophilia</i> K279a	Proteobacteria	25	6	-3	9	-7	0	91	111	Cluster2-BSH/PVA		Y	
229449422	C2	<i>Bacteroides</i> sp. <i>2_2_4</i>	Bacteroidetes	1	6	18	13	-3	4	70	90	Cluster2-BSH/PVA	Y		
50875093	C2	<i>Desulfotalea psychrophila</i> Lsv54	Proteobacteria	24	6	15	-3	-3	4	50	70	Cluster2-BSH/PVA			Y
493853916	C2	<i>Dysgonomonas gadei</i>	Bacteroidetes	25	6	5	2	8	-9	84	107	Cluster2-BSH/PVA	Y		
325106261	C2	<i>Pedobacter saltans</i> DSM 12145	Bacteroidetes	22	6	8	16	-19	5	77	101	Cluster2-BSH/PVA			Y
198284583	C2	<i>Acidithiobacillus ferrooxidans</i> ATCC 53993	Proteobacteria	25	6	3	10	2	2	60	86	Cluster2-BSH/PVA			Y
332885804	C2	<i>Dysgonomonas mossii</i> DSM 22836	Bacteroidetes	22	6	6	2	0	-3	63	90	Cluster2-BSH/PVA	Y		
494414169	C2	<i>Bacteroides cellulosilyticus</i>	Bacteroidetes	26	6	3	6	-3	0	86	113	Cluster2-BSH/PVA	Y		

Table 2.7: Binding site similarity (BSS) based functional annotations for those members for which experimental evidence of their BSH or PVA activity is known.

Accession	Cluster	Organism	Phylum	BSS scores						Experimental annotation	BSS based annotation	Reference
				BIB SH	CpB SH	Bsp PVA	BsuP VA	BtB SH	PaPVA			
UP:Q6R974	C1	<i>Bifidobacterium bifidum</i> ATCC 11863	Actinobacteria	190	78	4	18	6	29	BSH	C1-BSH	(Kim <i>et al.</i> , 2004)
UP:Q83YZ2	C1	<i>Enterococcus faecium</i> FAIRE-E 345	Firmicutes	72	82	6	-9	-11	-9	BSH	C1-BSH	(Wijaya <i>et al.</i> , 2004)
UP:Q9F660	C1	<i>Lactobacillus johnsonii</i> 100-100	Firmicutes	72	106	3	0	-22	-9	BSH	C1-BSH	(Elkins & Savage, 1998)
UP:P97038	C1	<i>Lactobacillus johnsonii</i> 100-100	Firmicutes	81	86	12	26	-14	23	BSH	C1-BSH	(Elkins & Savage, 1998)
UP:Q8Y5J3	C1	<i>Listeria monocytogenes</i> EGDe	Firmicutes	82	106	10	-1	-20	-3	BSH	C1-BSH	(Dussurget <i>et al.</i> , 2002)
UP:Q06115	C1	<i>Lactobacillus plantarum</i> WCFS1	Firmicutes	63	88	12	-3	-24	0	BSH	C1-BSH	(Lambert <i>et al.</i> , 2007)
GP:262676	C1	<i>Lactobacillus plantarum</i> LP80	Firmicutes	63	88	12	-3	-24	0	BSH	C1-BSH	(Christiaens <i>et al.</i> , 1992)
GP:AAX86039	C1	<i>Bifidobacterium adolescentis</i> ATCC15705	Actinobacteria	207	68	-3	5	-5	-2	BSH	C1-BSH	(Kim <i>et al.</i> , 2005)
GP:AAV42751	C1	<i>Lactobacillus acidophilus</i> NCFM	Firmicutes	83	105	-14	-4	-13	-20	BSH	C1-BSH	(McAuliffe <i>et al.</i> , 2005)
GP:AAV42923	C1	<i>Lactobacillus acidophilus</i> NCFM	Firmicutes	56	99	3	-11	-8	-13	BSH	C1-BSH	(McAuliffe <i>et al.</i> , 2005)
GP:AAD03709	C1	<i>Lactobacillus acidophilus</i> KS-13	Firmicutes	82	87	10	25	-14	24	BSH	C1-BSH	(Moser & Savage, 2001)
GP:ZP_01771587.1	C1	<i>Collinsella aerofaciens</i> ATCC25986	Actinobacteria	105	72	-18	16	14	3	BSH	C1-BSH	(Jones <i>et al.</i> , 2008)
GP:EDN82839.1	C1	<i>Bifidobacterium adolescentis</i> L2-32	Actinobacteria	207	68	-3	5	-5	-2	BSH	C1-BSH	(Jones <i>et al.</i> , 2008)
GP:518092358	C1	<i>Methanobrevibacter smithii</i>	Euryarchaeota	73	93	6	6	13	-6	BSH	C1-BSH	(Jones <i>et al.</i> , 2008)
GP:84489564	C1	<i>Methanosphaera stadmanae</i> DSM 3091	Euryarchaeota	64	84	11	-8	-13	-4	BSH	C1-BSH	(Jones <i>et al.</i> , 2008)
GP:WP_005851448.1	C2	<i>Brucella abortus</i> 2308	Proteobacteria	1	9	-13	-1	178	54	BSH	C2-BSH	(Delpino <i>et al.</i> , 2007)
GP:WP_005839662.1	C2	<i>B. vulgatus</i>	Bacteroidetes	1	9	-13	-1	178	54	BSH	C2-BSH	(Kawamoto <i>et al.</i> , 1989)
GP:Gi:20089640	C2	<i>Methanosarcina acetivorans</i> C2A	Euryarchaeota	-6	5	-4	7	64	54	BSH-inactive	C2-BSH/PVA	(Jones <i>et al.</i> , 2008)
UP:Q8Y9S7	C1	<i>Listeria monocytogenes</i> EGDe	Firmicutes	19	45	38	87	12	12	PVA	C1-PVA	(Begley <i>et al.</i> , 2005)

*The prefix UP and GP in Accession column indicates the sequences from Uniprot and Genpept database respectively. From left to right the columns corresponds to Accession number, Cluster to which the enzymes belong (C1: Cluster1 and C2: Cluster2), Organism, Phylum, BSS scores of each enzyme with six template enzymes, Experimental annotation of enzymes as BSH/PVA, BSS based annotation as either BSH/PVA and References.

Table 2.8: Binding site similarity (BSS) based annotation of the Gram-positive members that were previously annotated by Lambert *et al.*, 2008.

Accession	Cluster	Organism	Phylum	BSS scores						Annotation by Lambert <i>et al.</i> , 2008	BSS based annotation
				<i>B</i> BSSH	<i>Cp</i> BSSH	<i>Bsp</i> PVA	<i>Bsu</i> PVA	<i>Bt</i> BSSH	<i>Pa</i> PVA		
16804106	C1	<i>Listeria monocytogenes</i>	<i>Firmicutes</i>	82	106	10	-1	-20	-3	BSH	Cluster1-BSH
29375146	C1	<i>Enterococcus faecalis</i> V583	<i>Firmicutes</i>	77	87	8	-3	-13	-4	BSH	Cluster1-BSH
46908302	C1	<i>Listeria monocytogenes</i> 4b F2365	<i>Firmicutes</i>	82	106	10	-1	-20	-3	BSH	Cluster1-BSH
58337197	C1	<i>Lactobacillus acidophilus</i> NCFM	<i>Firmicutes</i>	83	105	-14	-4	-13	-20	BSH	Cluster1-BSH
42519282	C1	<i>Lactobacillus johnsonii</i> NCC 533	<i>Firmicutes</i>	72	101	3	1	-27	-14	BSH	Cluster1-BSH
58337369	C1	<i>Lactobacillus acidophilus</i> NCFM	<i>Firmicutes</i>	56	99	3	-11	-8	-13	BSH	Cluster1-BSH
28379847	C1	<i>Lactobacillus plantarum</i>	<i>Firmicutes</i>	63	88	12	-3	-24	0	BSH	Cluster1-BSH
116629611	C1	<i>Lactobacillus gasserii</i> ATCC 33323	<i>Firmicutes</i>	58	84	-1	-1	7	-9	BSH	Cluster1-BSH
42519073	C1	<i>Lactobacillus johnsonii</i> NCC 533	<i>Firmicutes</i>	69	102	-14	-2	-1	-13	BSH	Cluster1-BSH
90962773	C1	<i>Lactobacillus salivarius</i> UCC118	<i>Firmicutes</i>	98	114	-6	7	-27	-11	BSH	Cluster1-BSH
18309691	C1	<i>Clostridium perfringens</i>	<i>Firmicutes</i>	76	230	-19	19	9	5	BSH	Cluster1-BSH
110800687	C1	<i>Clostridium perfringens</i> ATCC 13124	<i>Firmicutes</i>	76	230	-19	19	9	5	BSH	Cluster1-BSH
23465372	C1	<i>Bifidobacterium longum</i>	<i>Actinobacteria</i>	229	76	6	15	-2	10	BSH	Cluster1-BSH
119025874	C1	<i>Bifidobacterium adolescentis</i> ATCC 15703	<i>Actinobacteria</i>	207	68	-3	5	-5	-2	BSH	Cluster1-BSH
42518142	C1	<i>Lactobacillus johnsonii</i> NCC 533	<i>Firmicutes</i>	81	86	12	26	-14	23	BSH	Cluster1-BSH
116628735	C1	<i>Lactobacillus gasserii</i> ATCC 33323	<i>Firmicutes</i>	82	87	10	25	-14	24	BSH	Cluster1-BSH
52082498	C1	<i>Bacillus licheniformis</i> ATCC_14580	<i>Firmicutes</i>	31	53	38	138	-6	14	PVA	Cluster1-PVA
16802490	C1	<i>Listeria monocytogenes</i>	<i>Firmicutes</i>	19	45	38	87	12	12	PVA	Cluster1-PVA
49478365	C1	<i>Bacillus thuringiensis</i> konkukian	<i>Firmicutes</i>	11	21	64	74	19	3	PVA	Cluster1-PVA
30019170	C1	<i>Bacillus cereus</i> ATCC14579	<i>Firmicutes</i>	21	25	59	79	16	13	PVA	Cluster1-PVA
118478988	C1	<i>Bacillus thuringiensis</i> Al_Hakam	<i>Firmicutes</i>	12	22	57	76	20	4	PVA	Cluster1-PVA
42782851	C1	<i>Bacillus cereus</i> ATCC_10987	<i>Firmicutes</i>	12	22	58	74	20	4	PVA	Cluster1-PVA

30263768	C1	<i>Bacillus_anthraxis_Ames</i>	<i>Firmicutes</i>	12	22	58	74	20	4	PVA	Cluster1-PVA
49186612	C1	<i>Bacillus_anthraxis_str_Sterne</i>	<i>Firmicutes</i>	12	22	58	74	20	4	PVA	Cluster1-PVA
47529187	C1	<i>Bacillus_anthraxis_Ames_0581</i>	<i>Firmicutes</i>	12	22	58	74	20	4	PVA	Cluster1-PVA
81427820	C1	<i>Lactobacillus_sakei_23K</i>	<i>Firmicutes</i>	14	20	50	76	-5	-1	PVA	Cluster1-PVA
116334525	C1	<i>Lactobacillus_brevis_ATCC_367</i>	<i>Firmicutes</i>	10	50	50	92	-8	-4	PVA	Cluster1-PVA
15673817	C1	<i>Lactococcus_lactis</i>	<i>Firmicutes</i>	19	22	99	62	11	-1	PVA	Cluster1-PVA
57866161	C1	<i>Staphylococcus_epidermidis_RP62A</i>	<i>Firmicutes</i>	-5	3	49	55	-31	-15	PVA	Cluster1-PVA
73661455	C1	<i>Staphylococcus_saprophyticus</i>	<i>Firmicutes</i>	-4	3	24	72	-16	8	PVA	Cluster1-PVA
116491067	C1	<i>Oenococcus_oeni_PSU-1</i>	<i>Firmicutes</i>	33	27	79	63	13	-4	PVA	Cluster1-PVA
27467268	C1	<i>Staphylococcus_epidermidis_ATCC_12228</i>	<i>Firmicutes</i>	-5	3	49	55	-31	-15	PVA	Cluster1-PVA
88194054	C1	<i>Staphylococcus_aureus_NCTC_8325</i>	<i>Firmicutes</i>	2	15	15	58	-6	-2	PVA	Cluster1-PVA
15925977	C1	<i>Staphylococcus_aureus_N315</i>	<i>Firmicutes</i>	2	15	15	58	-6	-2	PVA	Cluster1-PVA
15923265	C1	<i>Staphylococcus_aureus_Mu50</i>	<i>Firmicutes</i>	2	15	15	58	-6	-2	PVA	Cluster1-PVA
148266699	C1	<i>Staphylococcus_aureus_JH9</i>	<i>Firmicutes</i>	2	15	15	58	-6	-2	PVA	Cluster1-PVA
82749980	C1	<i>Staphylococcus_aureus_RF122</i>	<i>Firmicutes</i>	3	15	15	59	-7	-1	PVA	Cluster1-PVA
87161761	C1	<i>Staphylococcus_aureus_USA300</i>	<i>Firmicutes</i>	2	15	15	58	-6	-2	PVA	Cluster1-PVA
49485155	C1	<i>Staphylococcus_aureus_aureus_MSSA476</i>	<i>Firmicutes</i>	2	15	15	58	-6	-2	PVA	Cluster1-PVA
21281980	C1	<i>Staphylococcus_aureus_MW2</i>	<i>Firmicutes</i>	2	15	15	58	-6	-2	PVA	Cluster1-PVA
70725298	C1	<i>Staphylococcus_haemolyticus</i>	<i>Firmicutes</i>	-3	2	17	48	-21	-4	PVA	Cluster1-PVA
49482512	C1	<i>Staphylococcus_aureus_aureus_MRSA252</i>	<i>Firmicutes</i>	2	15	15	58	-6	-2	PVA	Cluster1-PVA
57652536	C1	<i>Staphylococcus_aureus_COL</i>	<i>Firmicutes</i>	2	15	15	58	-6	-2	PVA	Cluster1-PVA
116334689	C1	<i>Lactobacillus_brevis_ATCC_367</i>	<i>Firmicutes</i>	11	18	17	45	-21	2	PVA	Cluster1-PVA

2.4. Summary

Extending from the annotation system developed by Lambert *et al.*, 2008, we present an improved method for the annotation of members of the CGH family based on the phylogenetic, substrate specificity and binding site information of the enzymes. The method presented here could annotate correctly the BSH/PVA sequences in the dataset into five distinct groups, based on the BSS scores. Based solely on sequence information, this method could thus be used to annotate correctly any putative CGH family members.

The presence of *bsh* genes among gut-inhabiting microbes supports the role of BSH in the protection of microbes in the host gastrointestinal tract. The occurrence of *pva* genes in pathogens and organisms degrading molecules with aromatic rings suggests the need for further exploration of the physiological roles of PVA enzymes.

The emergence of diderms from monoderms represents a crucial juncture in the evolutionary history of microbes. Using the method described here, we have identified two distinct subfamilies within the CGH family showing divergent evolution. Most members of Cluster1 are Gram-positive bacteria, whereas Cluster2 is rich in Gram-negative members and archaeal members are distributed across both subfamilies. The detailed sequence analysis of the CGH family members reveals that the members of two subfamilies differ not only by a 19-23 aa indel signature, which alters the thermodynamic stabilities of their quaternary structure, but also in terms of the length of their pre-peptide sequence.

The above analysis thus provides a supporting explanation for the antibiotic selection pressure theory, whilst also opening new dimensions for exploration of the true significance of tetramer assembly loops.

Chapter 3

iRDP: An integrated web-platform for rational design, analysis and engineering of proteins, featuring investigations into factors that impart thermostability to proteins

Engineering protein molecules with desired structures and biological functions has been an elusive goal. Development of industrially viable proteins with improved properties such as stability, catalytic activity and altered specificity by modifying the structure of an existing protein has widely been targeted through rational protein engineering. This is often a knowledge-driven process that requires cycling between structural analyses, generation of a large number of potential mutants and their evaluation before proceeding to the experimental stage (Fig. 3.1).

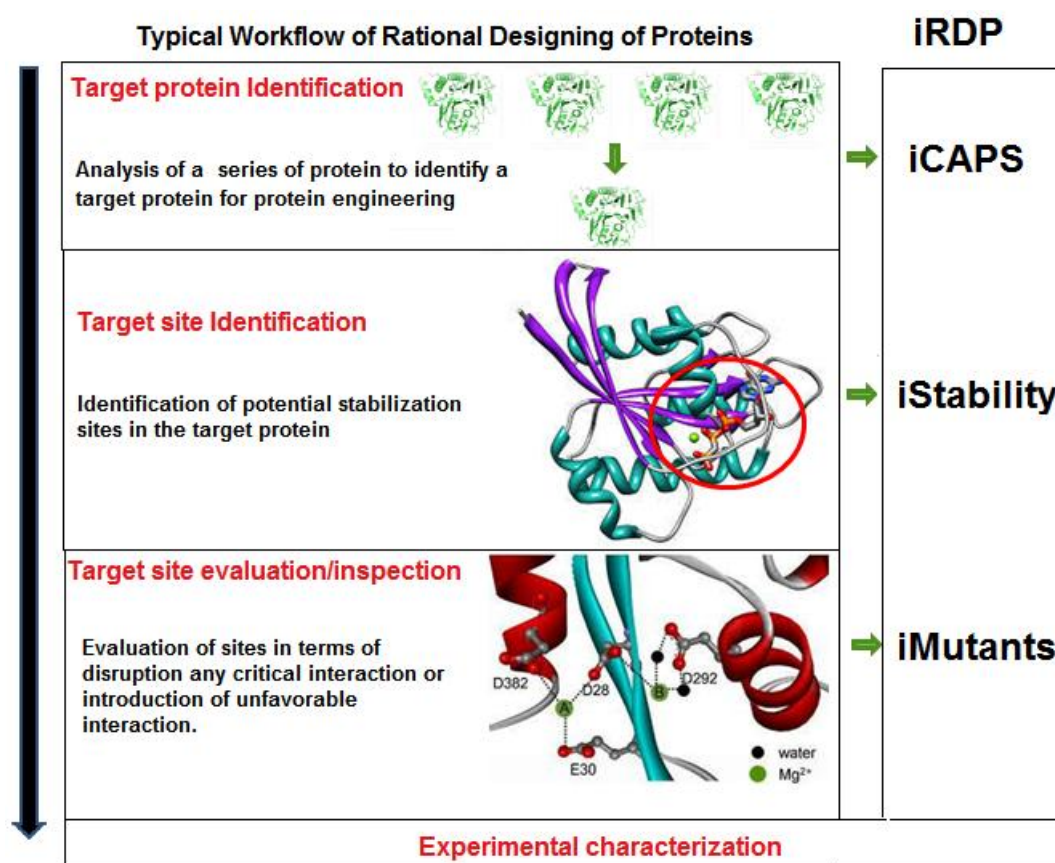


Figure 3.1: The workflow of rational protein engineering experiments which usually begins with *target protein identification* through comparative structural analysis of a series of proteins followed by *identification of potential mutant sites* in the targeted protein, ending with detailed *evaluation/inspection of the identified mutation sites* in terms of loss or gain of local interactions, before proceeding to the experimental stage. Each of these steps normally involves large scale analysis, which is time-consuming, cumbersome and sometimes error prone, if carried out manually. iRDP web server, comprising of iCAPS, iStability and iMutants module, was developed to automate these steps and assist the protein engineering studies by providing a single web platform.

Although a range of factors contributing to thermal stability have been identified and widely researched, the *in silico* implementation of these features as strategies directed towards enhancement of protein stability has not yet been explored extensively. A wide range of structural analysis tools is currently available for *in silico* protein engineering. However, these tools concentrate only on a limited number of factors or individual protein structures, resulting in cumbersome and time-consuming analysis. The current chapter describes the development and implementation of iRDP, a web server that was developed to act as a single platform that simplifies these extensive tasks leading to effective rational engineering of proteins.

3.1 Introduction

Thermophiles and hyperthermophiles are organisms that grow at extreme temperatures (50 to 110 °C). Enzymes from these organisms are inherently stable and active at high temperatures, offering a major industrial advantage over their mesophilic homologues with respect to their storage, resistance against chemical denaturants and risk of microbial contaminations. Thermal stability is an important parameter that determines economic feasibility of applying an enzyme in any industrial process. Understanding the molecular determinants of thermostability can not only provide useful insights into the evolution of such enzymes but the application of these through rational protein engineering to existing mesophilic proteins can also lead to development of more efficient and thermally stable biocatalysts for varied industrial applications (Zamost *et al.*, 1991).

3.1.1 Molecular determinants of protein thermostability

Studies have revealed several trends of residue preference towards thermostabilization of thermophilic proteins, such as lower content of uncharged polar residues, preference of arginine over lysine residues and higher charged residue contents (Deckert *et al.*, 1998). Shortening of loop regions is a known mechanism of protein thermostabilization in hyperthermophilic proteins (Russell *et al.*, 1997; Thompson & Eisenberg, 1999; Usher *et al.*, 1998). Various non-bonded interactions such as hydrogen bonds, ionic, aromatic-aromatic, aromatic-sulphur and cation-pi interactions are known to play a vital role in thermostabilization of proteins (Vieille & Zeikus, 2001; Vogt & Argos, 1997; Vogt *et al.*, 1997). Disulfide bridges are covalent interactions known

to provide stability to protein by entropic effect (Betz, 1993; Matsumura *et al.*, 1989; Zhang *et al.*, 1994).

The presence of thermolabile residues and bonds involving asparagine and glutamine are known to introduce instability to the protein backbone by undergoing deamidation at elevated temperatures (Robinson, 2002). Residues in left-handed helical conformation on mutation to glycine are known to contribute favorably to protein thermal stability (Kawamura *et al.*, 1996; Kimura *et al.*, 1992). Marshall *et al.*, 2002, have studied the interactions of the α -helix dipole with side chains of sequentially charged residues and found it to contribute favorably to stability (Marshall *et al.*, 2002; Nicholson *et al.*, 1988). Proline residues being conformationally most rigid are thought to provide stability to proteins by entropic effects (Bogin *et al.*, 1998; Watanabe *et al.*, 1994). The hydrophobic effect is understood to be one of the primary driving forces of protein folding (Dill, 1990). Decrease in hydrophobic surface area, as a stabilization mechanism has been studied in superoxide dismutase from *S. acidocaldarius* (Knapp *et al.*, 1999). Bound metal is also vital for stability and functioning of many proteins (Kasumi *et al.*, 1982; Marg & Clark, 1990; Smith *et al.*, 1999).

Table 3.1: List of currently available protein structural analysis tools, their usefulness and need for further improvement.

Tools	Usefulness	Current limitations	Reference
PIC	Using these tools user can analyze various non-bonded interactions in protein.	Does not allow simultaneous analysis of multiple structures. Though efficient in detection of isolated non-bonded interactions, these tools do not identify interaction-networks in proteins.	(Tina et al., 2007)
ESBRI			(Costantini et al., 2008)
Capture			(Gallivan & Dougherty, 1999)
WHAT IF		Except non-bonded interaction analysis, these tools do not analyze other mechanisms of protein thermostabilization.	(Vriend, 1990)

iCAPS (in silico Comparative Analysis of Protein Structures)

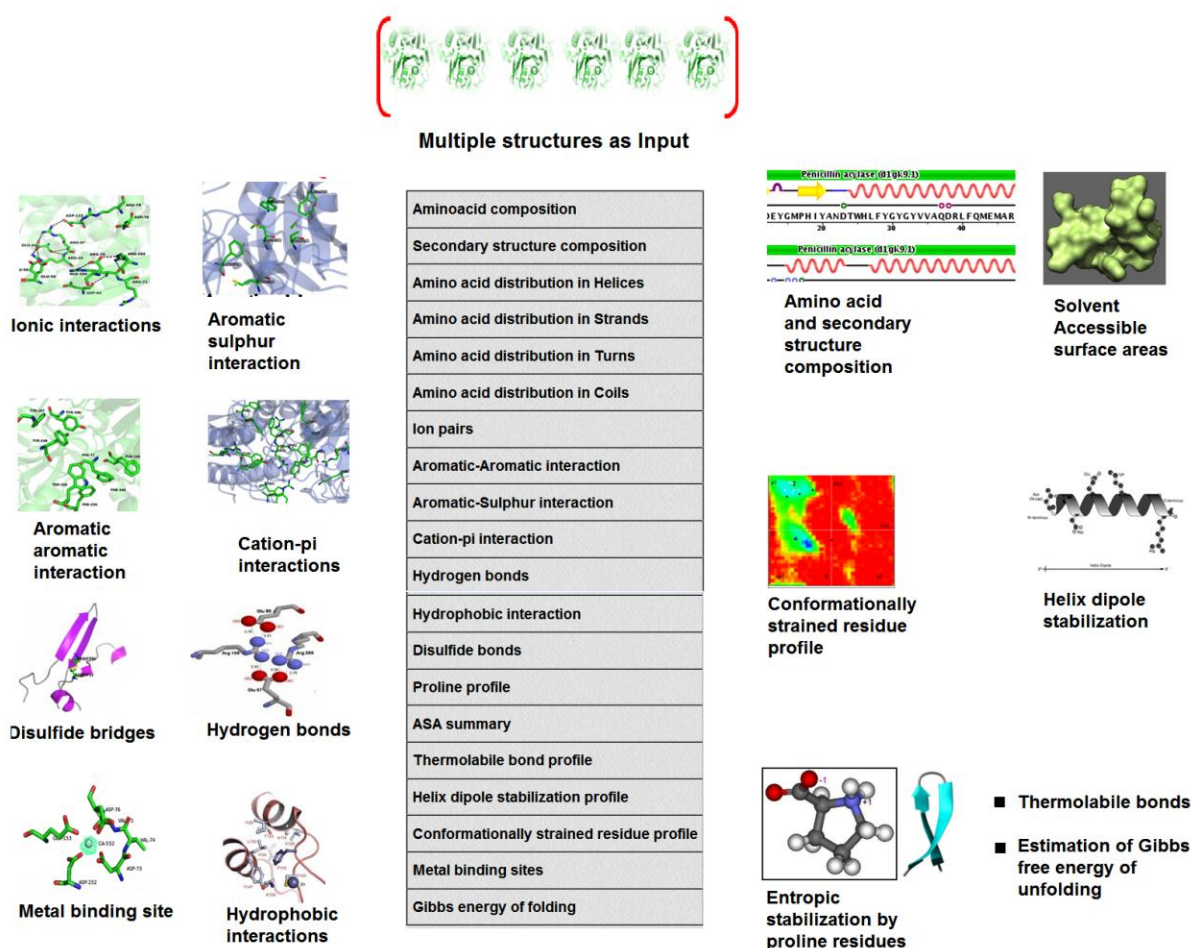


Figure 3.2: List of sequence- and structure-based features that can be analyzed using iCAPS.

3.1.2 Development of iCAPS (*in silico* Comparative Analysis of Protein Structures) module

The rapid addition of protein structures in PDB (Berman *et al.*, 2000), has made manual analysis of combination of such a large number of factors (section 3.1.1) extremely time-consuming and sometimes error-prone. Although a variety of computational tools such as WHAT IF (Vriend, 1990), PIC (Tina *et al.*, 2007), Capture (Gallivan & Dougherty, 1999) are available for structural analysis, most of these are limited by their ability to analyze only a single

structure at a time (Table 3.1). These tools primarily focus on analysis of non-covalent interactions as stabilizing mechanisms ignoring most molecular determinants listed above (section 3.1.1). However, most protein engineering studies necessitate simultaneous analysis of several structural mechanisms amongst a vast set of protein structures for improved selection of target protein and potential mutation sites. In view of this, the **iCAPS** (*in silico* Comparative Analysis of Protein Structures) module was developed to simplify the comparison process for a large number of protein structures in terms of features listed above known to affect protein stability. iCAPS aims to help the user to compare a series of proteins in order to select a target protein most suitable for initiation of protein engineering studies (Fig. 3.2).

3.1.3 Identification of potential sites for structural stabilization.

Once a target protein is selected for engineering, the next task is to identify potential target mutation sites. The Suzuki group has conclusively proved entropic stabilization by proline insertion as a protein thermostabilization mechanism through their work on oligo 1,6 glucosidase from *Bacillus cereus*. Through this work, they were able to identify that **proline insertion at second position of β -turns and N-cap of helix** enhanced thermostability of the protein. This mechanism has been defined as the “The Proline Rule” (Suzuki *et al.*, 1987; Suzuki, 1989). Loop stabilization by proline has been successfully implemented for thermostabilization of proteins such as cold shock protein, ubiquitin, ribonuclease Sa2 and guanyl specific ribonuclease Sa3, bacteriophage T4 lysozyme and human lysozyme (Fu *et al.*, 2009; Herning *et al.*, 1992; Nicholson *et al.*, 1992). Proline insertion at helix N-cap has also been used to enhance the stability of proteins like alcohol dehydrogenase, α -parvalbumin and triosephosphate isomerase (Agah *et al.*, 2003; Bogin *et al.*, 1998; Mainfroid *et al.*, 1996). Studies on ribonuclease HI show that **release of conformational strain** caused due to **left-handed helical residue Lys95 on mutation to Gly**, results in considerable increase in thermostability of the protein (Kimura *et al.*, 1992). This strategy has been used to improve the stability of proteins such as Drosophila adapter protein Drk, barnase, lysozyme and Pin1 WW (Bezsonova *et al.*, 2005; Jäger *et al.*, 2009; Serrano *et al.*, 1992; Takano *et al.*, 1999; Takano *et al.*, 2001). In their classic work on bacteriophage T4 lysozyme, Matsumura *et al.*, 1989, have not only elucidated the **role of disulfide bridges in protein stability** but have also shown that the effect of introduction of a combination of disulfide bridges on protein stability is additive in nature (Matsumura *et al.*,

1989). This mechanism has been used successfully for improving the stability of proteins like T4 lysozyme (Perry & Wetzel, 1984), subtilisin BPN (Pantoliano *et al.*, 1987), xylanase (Davoodi *et al.*, 2007), lipase (Han *et al.*, 2009), lipase B (Le *et al.*, 2012) and glucose 1-dehydrogenase (Ding *et al.*, 2013).

3.1.4 Development of iStability (*in silico* Analysis of Stability Change in Protein Structures) module

Currently computational tools such as CUPSAT (Parthiban *et al.*, 2006), SDM (Worth *et al.*, 2011), PopMusic (Dehouck *et al.*, 2011) and Rosetta Design (Liu & Kuhlman, 2006) are available that predict the effect of mutation on protein stability. However, these tools require the user to input the mutations and do not suggest potential stabilizing mutations using specific strategies. Hence the **iStability** module was developed which not only aids in identification of stabilizing mutation sites through the application of protein design strategies described above (section 3.1.3) for improvement of thermal stability but also assesses the stability of any mutants (Fig. 3.3).

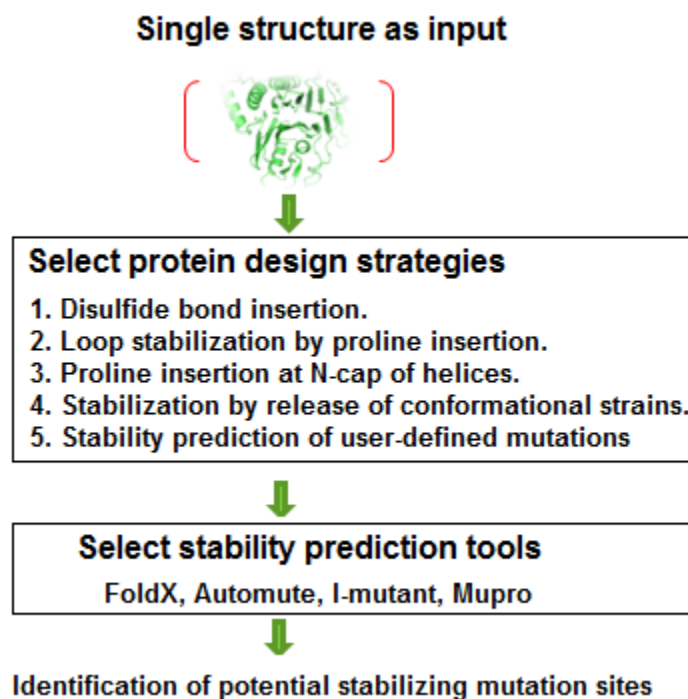


Figure 3.3: iStability, a module that predicts potential sites of thermostabilization in the input protein structure based on user's choice of protein-design strategies. Currently it supports implementation of four

well-established strategies. It also supports stability prediction of any user-specific mutations. Stabilities are predicted based on the selected stability prediction tools.

3.1.5 Evaluating potential thermostabilization sites using molecular interactions

Once a stabilizing mutation site is selected in the target protein, it is important to evaluate the mutation in terms of its effects on neighboring residues, which is vital to any protein engineering experiment. Serrano *et al.*, 1992, in their work with barnase enzyme have revealed that loss of buried salt bridges and hydrogen bonds due to mutations affects protein stability significantly. Studies carried out on the *arc repressor* protein of *bacteriophage P22* have shown deleterious effects of mutations on protein stability due to disruptions in hydrogen bonds and salt bridges (Milla *et al.*, 1994).

3.1.6 Development of iMutants (*in silico* Comparative Analysis of Interactions in Protein Mutants) module.

Computational tools such as CUPSAT, SDM, POPMUSIC, Rosetta Design and others (Table 3.2) are mainly of predictive nature. Although users are usually informed of the effects of mutations on protein stability by these tools in the form of stability scores, underlying details of interaction rearrangements at the mutation site are currently not provided. Along with the stability scores, the information regarding the change in interactions could provide a better evaluation of the mutations being considered. Understanding this, we have developed a mutation evaluation/inspection tool called **iMutants**, which assesses the change in local interactions at mutation sites through comparison of wild-type and mutant proteins (Fig. 3.4). Evolutionary conservation analysis is crucial to successful protein design since highly conserved positions typically play important structural or functional roles, i.e. mutation could adversely affect protein function (Suemori, 2013). Therefore, iMutants also evaluates the conserved nature of the mutation sites.

The above three modules, iCAPS, iStability and iMutants, addressing different facets of the protein engineering problem, are integrated on a single platform in the form of a web server, iRDP (*in silico* Rational Design of Proteins), available at <http://irdp.ncl.res.in>.

Table 3.2: List of available stability prediction tools, their usefulness, limitations and the need for further improvements.

Tools	Usefulness	Limitations	Reference
SDM	Protein stability prediction	Supports the analysis of single mutants i.e. does not predict stabilities of double/multiple mutants.	(Worth et al., 2011)
NeEMO			(Giollo et al., 2014)
POPMUSIC		(Dehouck <i>et al.</i> , 2011)	
CUPSAT		Does not support the analysis of large number of mutations at a time.	(Parthiban et al., 2006)
FoldX		Do not estimate evolutionary conservation of the mutation site.	(Guerois et al., 2002)
AUTO-MUTE		Undertake user-specified mutations for analysis while failing to suggest potential mutation sites by implementing known protein-design strategies.	(Masso & Vaisman, 2011)
I-Mutant 2.0			(Capriotti et al., 2005)
Mupro			(Cheng et al., 2006)

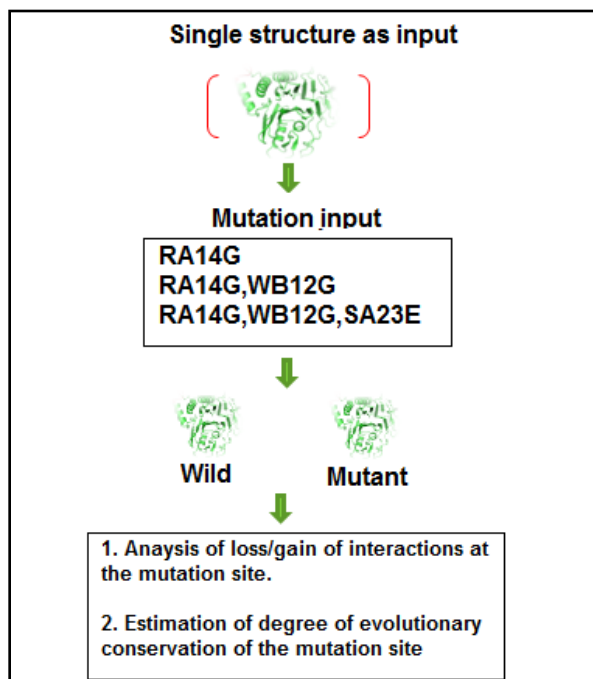


Figure 3.4: iMutants, a module which aids in the analysis of change in local interactions due to mutation near the mutation site by comparing the wild-type and mutant proteins. It supports simultaneous analysis of large number of mutations which can be either single or double or multiple mutations. In addition, it also supports the estimation of degree of evolutionary conservation of the mutation site.

3.2 Materials and Methods

The iRDP server is built on a Linux platform using R, Perl, HTML and PHP. The Bio3d (Grant *et al.*, 2006) and iGraph (Csardi & Nepusz, 2006) packages form the core of all iRDP modules. The vast analysis carried out by modules of iRDP server use both in-house developed scripts and established tools (Table 3.3). Described below is the detailed workflow of each module in iRDP web server (Fig. 3.5).

Table 3.3: List of tools used by iRDP web server for estimation of few structural parameters.

Tools	Purpose	Reference
DSSP	For assignment of secondary structures	(Kabsch & Sander, 1983)
NACCESS	For estimation of residue solvent accessibility	(Hubbard & Thornton, 1993)
Promotif	For detection of β -turns	(Hutchinson & Thornton, 1996)
Procheck	For calculation of conformational parameters	(Laskowski <i>et al.</i> , 1993)
HBPLUS	For identification of hydrogen bonds	(McDonald & Thornton, 1994)
SSBOND	For identification of residue pairs for disulfide bond insertion	(Hazes & Dijkstra, 1988)
MODELLER	For generating <i>in-silico</i> mutants	(Eswar <i>et al.</i> , 2006)
FindGeo	For analysis of metal binding sites	(Andreini <i>et al.</i> , 2012)
FoldX AUTO- MUTE I-Mutant 2.0 Mupro	For prediction of mutant stability	(Capriotti <i>et al.</i> , 2005; Cheng <i>et al.</i> , 2006; Guerois <i>et al.</i> , 2002; Masso & Vaisman, 2011)

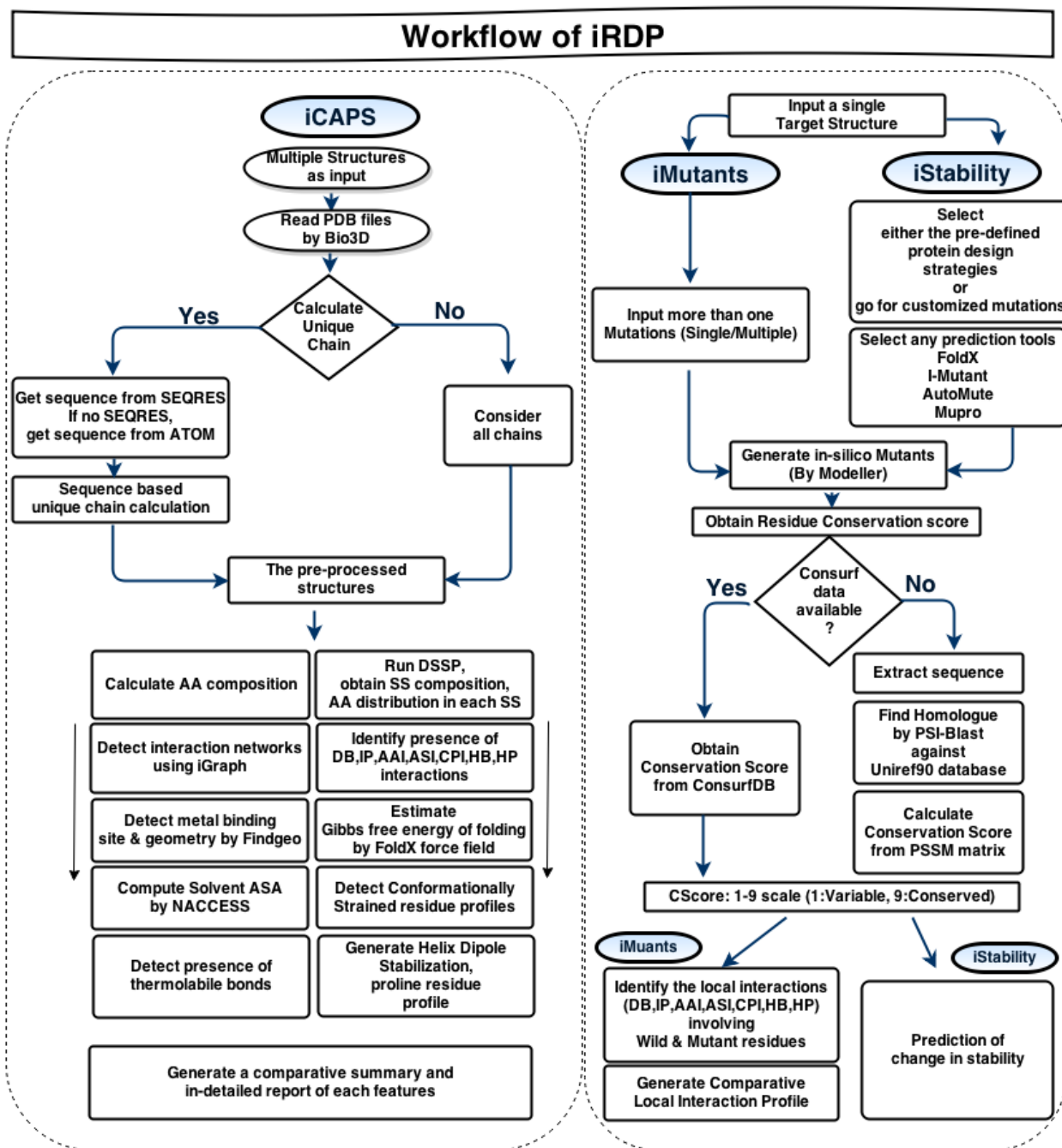
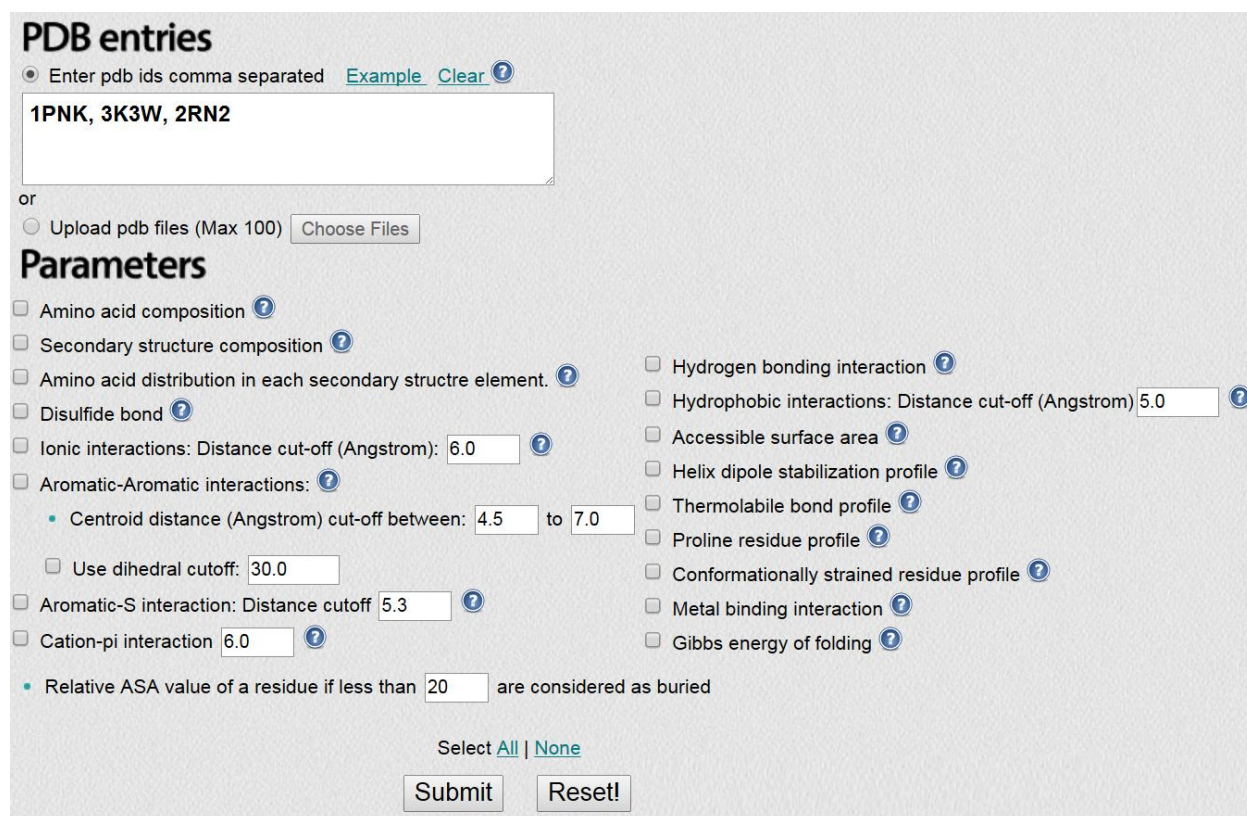


Figure 3.5: The workflow of the working modules implemented in the iRDP web server. The abbreviations AA, SS, DB, IP, AAI, ASI, CPI, HB, HP, ASA, PSSM and CScore corresponds to amino acid, secondary structure, disulfide bridges, ion-pairs, aromatic-aromatic interactions, aromatic-sulphur interactions, cation-pi interactions, hydrogen bonds, hydrophobic interactions, accessible surface areas, position-specific scoring matrix and conservation scores, respectively.

3.2.1 *In silico* Comparative Analysis of Protein Structures (iCAPS)

Multiple structures serve as input to iCAPS. Input can be a list of PDB entries separated by commas, or files in valid PDB format can be uploaded. The user can select the structural features to be analyzed, modify various interaction cutoffs and relative accessible surface area (ASA) value before submitting the job (Fig. 3.6). The results page contains a unique job identification number for each job being submitted. Users can bookmark this page and return to view and retrieve the results later. An extensive help file has been prepared and provided in the website which explains the importance of every parameter generated and their calculation, along with relevant references.



The screenshot displays the iCAPS web interface. At the top, under "PDB entries", there are two options: "Enter pdb ids comma separated" (selected) with a text input field containing "1PNK, 3K3W, 2RN2", and "Upload pdb files (Max 100)" with a "Choose Files" button. Below this is the "Parameters" section, which includes a list of checkboxes for various structural features and interaction types. Some parameters have associated input fields for numerical values. At the bottom, there are "Submit" and "Reset!" buttons, along with "Select All" and "None" options.

PDB entries

Enter pdb ids comma separated [Example](#) [Clear](#)

1PNK, 3K3W, 2RN2

or

Upload pdb files (Max 100) [Choose Files](#)

Parameters

Amino acid composition

Secondary structure composition

Amino acid distribution in each secondary structure element.

Disulfide bond

Ionic interactions: Distance cut-off (Angstrom): 6.0

Aromatic-Aromatic interactions:

- Centroid distance (Angstrom) cut-off between: 4.5 to 7.0
- Use dihedral cutoff: 30.0

Aromatic-S interaction: Distance cutoff 5.3

Cation-pi interaction 6.0

Relative ASA value of a residue if less than 20 are considered as buried

Hydrogen bonding interaction

Hydrophobic interactions: Distance cut-off (Angstrom) 5.0

Accessible surface area

Helix dipole stabilization profile

Thermolabile bond profile

Proline residue profile

Conformationally strained residue profile

Metal binding interaction

Gibbs energy of folding

Select [All](#) | [None](#)

[Submit](#) [Reset!](#)

Figure 3.6: The user interface of iCAPS module.

Analysis begins with primary structural features like amino acid composition, secondary structure content, information such as helix/strand/turn/coil composition and then proceeds to calculation of non-covalent interactions. The program DSSP (Kabsch & Sander, 1983) is used to detect secondary structures in the input proteins while NACCESS is used for estimation of

residue solvent accessibility and accessible surface areas (ASA) (Hubbard & Thornton, 1993). The non-covalent interactions and disulfide bonds are identified using the standard criteria reported in literature. Users are provided options to change the criteria of interaction calculations. In-house scripts are written for estimation of parameters such as proline residue distribution profile, thermolabile bond profile and helix dipole stabilization profile. Conformationally strained residues are identified by using Procheck (Laskowski *et al.*, 1993) while β -turn and N-cap proline residues are identified using Promotif (Hutchinson & Thornton, 1996) and DSSP, respectively. The program FindGeo has been employed for analysis of metal binding sites and geometry (Andreini *et al.*, 2012). Gibbs free energy of unfolding is calculated using FoldX (Guerois *et al.*, 2002).

For the validation of iCAPS module, 16 thermophilic-mesophilic protein pairs were used. Each pair was submitted to iCAPS module and raw values of various thermostability parameters were calculated. Raw values were first normalized by methods similar to that of Kumar *et al.*, 2000, in order to estimate the percentage change of these parameters between mesophilic and thermophilic proteins (Kumar *et al.*, 2000). Percentage change was calculated by using difference of normalized values between thermophilic and mesophilic protein, divided by the corresponding normalized value of mesophilic protein. For normalization of parameters such as total percentage of aromatic (Aro), uncharged polar (UP), proline (Pro), hydrophobic or aliphatic (ALI), charged (CHG) residues, total percentage of ion-pairs (IP), aromatic-aromatic (AAI), aromatic-sulphur (ASI), cation-pi (CPI), hydrogen bonding (HB), hydrophobic (HP) interactions, total percentage of conformationally strained residues (CS) and percentage of residues in loop regions (Loop), the raw values obtained were normalized using sequence length. In case of parameters such as total percentage of proline residues occurring at 2nd position of beta turns (Bt2P) and Ncap helix positions (NCap), normalization was carried out using total number of proline residues. Similarly, the total percentage of dipole-stabilized helices parameter was normalized with total number of helices. In case of normalization for total percentage of thermolabile bonds (TL), raw values were normalized using total number of Asn and Gln residues. The parameters such as ratio of Nonpolar to Polar accessible surface areas (NP/P) and Arg to Lys ratio (R/K); raw values were directly considered for the calculation of percentage change. Since the protein families considered for analysis in the dataset were highly diverse in

terms of sequence and structure, the normalization process was focused on the pairs rather than the entire dataset.

PDB Entries

Enter pdb id [?](#)

or

Upload a pdb file No file chosen

Identify all possible residue pairs that are likely to form disulfide bonds

Loop stabilization by proline insertion at 2nd position of Beta-turns

Stabilization by proline insertion at N-cap position of helices

Stabilization by release of Conformational strain

Input your own mutations [?](#)

Parameters

• Mutation Analysis Tool:

• Automute Classification model:

• I-mutant pH: Temperature (C):

• FoldX pH: Temp (K): Ion strength (M):

Use conservation analysis (Slow if you upload pdb file !!!) [?](#)

Figure 3.7: The user interface of iStability module. The “input your own mutations” feature of iStability allows user to predict stability of any user-defined mutations. Any number of mutations can be analyzed simultaneously. Each mutation should be on separate line and must be input in a standard format (*Wild-type residue* followed by *Chain* followed by *Residue number* followed by *Mutant residue*). The YA20G mutation in this figure corresponds to mutation of Tyr20 residue of chain A to Gly.

3.2.2 *In silico* Analysis of Stability Change in Protein Structures (iStability)

This module accepts either a PDB ID or PDB formatted file as input. The user can select any of the four pre-defined protein design strategies or provide their own mutations in the specified format (Fig. 3.7). Once the design strategy is selected, the user has the choice of trying out different stability prediction tools, which are based on empirical potential energy functions or machine learning methods with provision to modify input. Currently iStability implements four freely available tools for stability prediction such as FoldX, Auto-mute, I-mutant and Mupro.

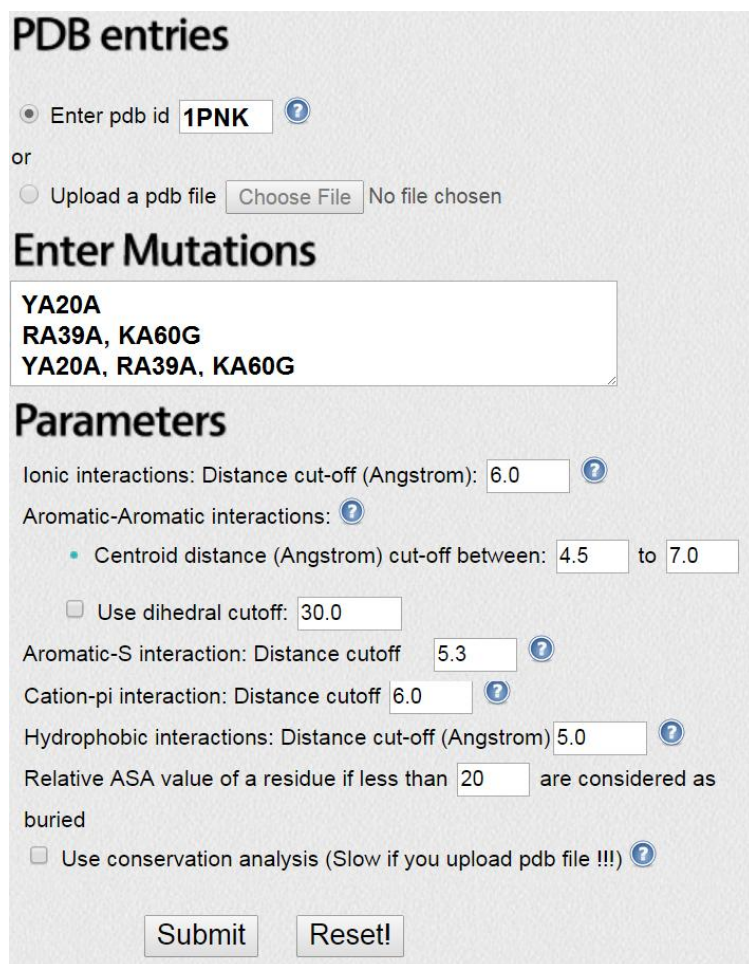
The user can choose residue conservation analysis, if required. If input is a PDB entry, then the evolutionary residue conservation score is derived from ConSurf-DB (Goldenberg *et al.*, 2009), on a scale of 1 – 9 (1 is an indication of least conserved/highly variable and 9 highly conserved/least variable). If input is a structure uploaded by the user, then the extracted sequence is used to search for homologues against the UniRef90 database using PSI-BLAST (2 iterations and e-value cut-off 1) (Altschul *et al.*, 1997). Weighted observed percentages from generated Position Specific Scoring Matrix (PSSM) are scaled from 1 to 9 as already defined and are presented as conservation scores.

For validation of iStability module, a total of 17, 6, 10 and 15 proteins were selected for the analysis of beta-turn proline insertion, N-cap proline insertion, conformational strain release and disulfide bond insertion strategies, respectively. Protein structures were analyzed using iStability module by selecting each strategy and predictions obtained were compared with experimental observations.

3.2.3 *In silico* Comparative Analysis of Interactions in Protein Mutants (iMutants)

iMutants takes a single structure as input similar to iStability. The user must also provide mutations in the specified format. A large number of mutations can be analyzed simultaneously in iMutants. Each mutation must be provided in a separate line. For double or multiple mutants, mutations should be provided in a comma-separated format on a single line (Fig. 3.8). Users can modify interaction cutoffs and relative ASA value before submitting a job (Fig. 3.8). Similar to iStability, an option is provided for residue conservation analysis at the mutation site.

For validation of iMutants module, a total of 51 mutations were analyzed in arc repressor protein of bacteriophage P22 (PDB ID: 1ARR). Mutants were generated using MODELLER, energy minimized using steepest descent and finally the structure thus generated was compared with wild-type structure to analyze the change in interactions at the mutation site.



PDB entries

Enter pdb id ?

or

Upload a pdb file No file chosen

Enter Mutations

YA20A
RA39A, KA60G
YA20A, RA39A, KA60G

Parameters

Ionic interactions: Distance cut-off (Angstrom): ?

Aromatic-Aromatic interactions: ?

- Centroid distance (Angstrom) cut-off between: to
- Use dihedral cutoff:

Aromatic-S interaction: Distance cutoff ?

Cation-pi interaction: Distance cutoff ?

Hydrophobic interactions: Distance cut-off (Angstrom) ?

Relative ASA value of a residue if less than are considered as buried

Use conservation analysis (Slow if you upload pdb file !!!) ?

Figure 3.8: The user interface of iMutants module. The mutation input format is same as that of iStability module, *Wild type residue* followed by *Chain* followed by *Residue number* followed by *Mutant residue*. In this example, three mutations are analyzed in protein 1PNK by iMutants module, a single mutant (YA20A), a double mutant (RA39A and KA60G) and a triple mutant (YA20A, RA39A and KA60G).

3.3 Results and Discussion

Below is the description of each module of iRDP web server along with case studies conducted to validate each module. At the end, description is given for iATMs database that have been developed as an additional information resource to users, providing detailed structural perspective for large number of experimentally characterized known mutations.

3.3.1 Analysis of structure stabilization mechanisms using iCAPS

iCAPS has been designed to carry out a comparative analysis of protein structures in terms of structural features and interactions that are known to contribute to thermodynamic stability. iCAPS supports investigation of 20 different stabilization mechanisms, as described below, estimating more than 250 parameters (Table 3.4) analyzed simultaneously for a maximum of 100 structures.

Table 3.4: List of various quantitative parameters (total 288) generated by the iCAPS module.

Property	Quantitative Features	Number of Features	Calculation
Amino acid (aa) Composition	Sequence length	1	(Vieille & Zeikus, 2001)
	Frequency of 20 Natural amino acids	20	
	Frequency of Unnatural amino acids	1	
	Frequency of Aromatic residues [FWY]	1	
	Frequency of Uncharged polar residues [NQST]	1	
	Frequency of Positively charged residues [RKH]	1	
	Frequency of Negatively charged residues [DE]	1	
	Ratio of Arg to Lys residue content (Arg/Lys ratio)	1	
Secondary structure Composition	Percentage of residues in isolated beta-bridge [B]	1	DSSP is used to calculate secondary structures in proteins.
	Percentage of residues in extended strands [E]	1	
	Percentage of residues in 3-helix (3/10 helix) [G]	1	
	Percentage of residues in pi helix [I]	1	
	Percentage of residues in alpha helix [H]	1	
	Percentage of residues in hydrogen bonded turn [T]	1	
	Percentage of residues in bend [S]	1	
	Percentage of residues in random coil [C]	1	
Helix Composition	Total residues in helix [DSSP notation: H/G/I]	1	(Kabsch & Sander, 1983)
	Frequency of each of the 20 residues in helix.	20	
Strand Composition	Total residues in strand [DSSP notation: B/E]	1	
	Frequency of each of the 20 residues in strand.	20	
Turn Composition	Total residues in turn [DSSP notation: T]	1	
	Frequency of each of the 20 residues in turn	20	
Coil Composition	Total residues in coil [DSSP notation: S/C]	1	
	Frequency of each of the 20 residues in coil	20	
Ion pairs (IPs)	Total number of IPs	1	A salt bridge is considered to be formed if the distance between any of the oxygen atoms of acidic residues and the nitrogen atoms of basic residues are within the cut-off distance (default 6 Å).
	Total number of intra-subunit IPs	1	
	Total number of inter-subunit IPs	1	
	Total IPs involving D/E/R/K/H residues	5	
	Total IPs of type DR/DK/DH/ER/EK/EH	6	
	Total number of buried IPs	1	
	Total number of exposed IPs	1	
	Percentage of isolated IPs, not involved in any network	1	
	Total number of IP-networks	1	
IP-network details	1		
Aromatic pairs (APs)	Total number of APs	1	Aromatic residues interact with each other if the distance between their phenyl ring centroids lies between 4.5 Å- 7.0 Å. A cut-off of dihedral angle between the planes of such interacting aromatic residues
	Total number of intra-subunit APs	1	
	Total number of inter-subunit APs	1	
	Total APs involving F/W/Y residues	3	
	Total APs of type FF/FY/FW/YY/YW/WW	6	
	Total number of buried APs	1	
	Total number of exposed APs	1	
	Percentage of isolated APs, not involved in network	1	

	Total number of AP-networks	1	can be set between 30° and 90°.
	AP-network details	1	(Burley & Petsko, 1985)
Aromatic-sulphur interactions (ASI)	Total number of ASI	1	Distance between the sulphur atoms of Cys/Met and the aromatic rings of Phe/Tyr/Trp if lie within 5.3 Å (default), they account for aromatic-sulphur interactions. (Reid <i>et al.</i> , 1985)
	Total number of intra-subunit ASI	1	
	Total number of inter-subunit ASI	1	
	Total ASI involving F/W/Y/C/M	5	
	Total ASI of type FC/YC/WC/FM/YM/WM	6	
	Total number of buried ASI	1	
	Total number of exposed ASI	1	
	Percentage of isolated ASI, not involved in any network	1	
	Total number of ASI-networks	1	
ASI-network details	1		
Cation-pi interactions (CPI)	Total number of CPI	1	A cationic side chain (Lys/Arg) if nearer to an aromatic side chain (Phe/Tyr/Trp) within 6 Å (default) separation, they account for cation-pi interactions. (Sathyapriya & Vishveshwara, 2004)
	Total number of intra-subunit CPI	1	
	Total number of inter-subunit CPI	1	
	Total CPI involving KF/KY/KW/RF/RX/RW	5	
	Total number of buried CPI	1	
	Total number of exposed CPI	1	
	Percentage of isolated CPI, not involved in any network	1	
	Total number of CPI-networks	1	
CPI-network details	1		
Disulfide bridges (DB)	Total number of DB	1	Pairs of cysteines (sulphur atoms) if fall within 2.2 Å (default) are accounted as disulphide bridges. (Matsumura <i>et al.</i> , 1989)
	Total number of intra-subunit DB	1	
	Total number of inter-subunit DB	1	
	Total number of buried DB	1	
	Total number of exposed DB	1	
	Size of loops connecting cys residues [Loop size between 0-10/10-20/20-30/30-40/40-50/>50]	6	
	Number of DB connecting two periodic (PP) secondary structures (DSSP notation: H/G/I/E)	1	
	Number of DB connecting two non-periodic (NN) secondary structures (DSSP notation: B/T/C/S)	1	
Number of DB connecting periodic and non-periodic secondary structures (NP)	1		
Hydrogen bonds (HB)	Total number of HB	1	HBPLUS is used to detect the hydrogen bonds. (Baker & Hubbard, 1984)
	Total number of intra-subunit HB	1	
	Total number of inter-subunit HB	1	
	Total number of Mainchain-Mainchain HB (MM)	1	
	Total number of Mainchain-Sidechain HB (MS or SM)	2	
	Total number of Sidechain-Sidechain HB (SS)	1	
	Total number of Charged-Neutral HB (CNHB)	1	
	Total number of Neutral-Neutral HB (NNHB)	1	

Hydrophobic interactions (HP)	Total number of HP	1	The residues ALA, VAL, LEU, ILE, MET, PHE, TRP, PRO and TYR are considered to interact if they fall within 5Å (default) range. (Pace, 1992)
	Total number of intra-subunit HP	1	
	Total number of inter-subunit HP	1	
	Total number of buried HP	1	
	Total number of exposed HP	1	
Proline residue profile	Total number of proline residue	1	(Li <i>et al.</i> , 1999) (Suzuki <i>et al.</i> , 1987)
	Frequency of proline residues in helices/strands/turns/coils	4	
	Total number of buried proline residues	1	
	Total number of exposed proline residues	1	
	Total number of prolines present in beta-turns	1	
	Total number of prolines, present at 2 nd -position of beta-turns	1	
	Total number of prolines, at N-cap position of helices.	1	
Analysis of Solvent Accessible Surface Area (ASA)	ASA of all-atoms.	1	NACCESS is used to calculate the ASA values. (Hubbard & Thornton, 1993)
	ASA of all side-chains atoms.	1	
	ASA of all main-chain atoms.	1	
	ASA of non-polar side-chain atoms (NP)	1	
	ASA of polar side-chain atoms (P)	1	
	Ratio of Non-polar to polar ASA of side-chain atoms (NP/P ratio)	1	
Thermolabile bond profile	ASA of all C/N/O/S atoms	4	
	Total number of thermolabile (TL) bonds.	1	(Robinson, 2002)
	Number of TL bonds of type NG/NA/NS/QG/QA/QS	6	
	Number of TL bonds where the nucleophilic attack distance is < 4Å	1	
Helix dipole stabilization profile	Total number of Helices	1	Helices are identified using DSSP. Identified helices are checked for presence of charged residues at their N- and C- terminals. (Vieille & Zeikus, 2001)
	Number of dipole-stabilized helices.	1	
	Number of helices stabilized at their N/C/NC terminal.	3	
	Number of helices whose dipoles are stabilized at N-2/N-1/N/N+1/N+2 positions.	5	
	Number of helices whose dipoles are stabilized at C-2/C-1/C/C+1/C+2 positions	5	
Conformationally strained residue profile	Total number of conformationally strained (CS) residues.	1	Procheck is used to identify conformationally strained residues. (Kimura <i>et al.</i> , 1992)
	Number of CS residues in L/I/~I region of Ramachandran plot	3	
	Details of the CS residues	1	
Metal binding summary	Total number of metals.	1	Findgeo (Andreini <i>et al.</i> , 2012) program is used to identify metal binding sites.
	Details of metal binding sites.	1	
Gibbs energy of folding	Gibbs energy of folding decomposed into individual energies	23	(Guerois <i>et al.</i> , 2002)

Below described are the structural parameters that can be analyzed by iCAPS.

1. Amino acid composition: iCAPS generates a comparative summary for a set of proteins in terms of their amino acid composition and its property-wise classification into different categories like positively charged, negatively charged, uncharged polar and aromatic residues.

2. Secondary structure information: Comparative summary generated for overall secondary structure (SS) content as well as the residue composition of each type of SS (Helix/Strand/Turn/Coil) of proteins, provides better understanding of the contribution of SS to protein thermostabilization.

3. Non-bonded interactions: iCAPS is distinct in its ability to calculate non-bonded interactions such as ion-pairs (IP), aromatic-aromatic interactions (AAI), aromatic-sulphur interactions (ASI), cation- π interactions (CPI), hydrogen bonds (HB) and hydrophobic interactions (HP). It also offers identification of interaction networks that are energetically more favorable compared to isolated interactions.

4. Disulfide bridges: iCAPS identifies disulfide bridges in input protein structures and provides useful insights by classifying them in terms of their expected entropic effect (i.e. based on the number of residues between bridged Cys residues) while providing other details such as solvent accessibility and SS preference of Cys residues, revealing the contribution of these bonds to structural stabilization.

5. Thermolabile bond profile: iCAPS studies spatial distributions of thermolabile bonds as potential target sites for stability enhancement in input structures involving asparagine and glutamine along with additional details such as SS preference and solvent accessible nature of the residues forming these bonds.

6. Conformationally strained residue profile: Conformationally strained residue detection feature in iCAPS not only identifies conformationally strained residues in input structures which could be considered for mutation to glycine for improving thermal stability of proteins but also provides additional information such as conformational geometry, SS preference and strain distance (distance between the C_{β} and main chain O atom) of the strained residues. While mutation of such residues to Gly remain the most established strategy, the void generated due to

the lack of side chain in the Gly residue remains a viable concern (Borgo & Havranek, 2012). In such cases it is advisable to explore non-glycine substitutions using the customized mutation option in iStability.

7. Helix dipole stabilization profile: The helix dipole stabilization feature of iCAPS identifies dipole-stabilized helices in input protein structures along-with position-wise classification of dipole stabilizing charged residues.

8. Proline residue profile: iCAPS reports the distribution of proline residues in various secondary structures along with specific identification of prolines occurring at the second position of beta-turns and at N-terminus of helices. Since the solvent exposed loops and intrinsically disordered regions of proteins are often found to be proline-rich (Theillet *et al.*, 2013), the secondary structure wise proline distribution provided warrants careful analysis.

9. Accessible Surface Area analysis: iCAPS measures the total, main-chain, side-chain, polar and non-polar accessible surface areas of proteins. ASA analysis also classifies all 20 amino acids by their solvent accessible nature as buried or exposed.

10. Metal binding analysis: The module identifies residues involved in metal binding sites along-with determination of metal co-ordination geometries.

11. Estimation of Gibbs free energy of unfolding: Protein stabilization energies for input structures are computed by iCAPS using the FoldX energy function (Guerois *et al.*, 2002). The total energy is considered as an approximation of overall stability of the protein. This comparative report gives a comprehensive overview of energies involving various structure stabilization mechanisms amongst proteins under study.

The above results are presented in a formatted web page. Besides this, for further downstream analysis the user can download a zipped file containing all results in the form of tab-delimited text files.

3.3.2 Demonstration of applicability of iCAPS module.

A diverse non-redundant dataset of thermophilic-mesophilic (TS-MS) protein pairs, from organisms that are moderately thermophilic to hyperthermophilic as well as their mesophilic counterparts, were investigated (Table 3.5). The pairs comprising of structures having resolution ≤ 2.5 Å were selected from a diverse set of families (Table 3.6). The selected TS-MS pairs were observed to be highly similar to each other with RMSD of the structures in the range of 0.69 – 1.68 Å while sequence identity in the range of 24 – 73%. The thermophilic and mesophilic protein sets among themselves were found to be highly dissimilar with the sequence identity ranging from 1-12% and 2-13% respectively, suggesting the diverse families considered for the analysis. In terms of structural diversity, 2 families were found to belong to all-alpha class, 3 to all beta class, 1 belonging to small proteins while the rest belonged to alpha-beta class according to the SCOP classification (Murzin *et al.*, 1995). In most cases the oligomeric state of the pairs selected was found to be the same. Table 3.5 shows the values of percentage change between TS-MS pairs with respect to various structural parameters estimated by iCAPS. The values of percentage change can be correlated to the extent of contribution of each of the factor towards thermostability of the proteins in the dataset. A positive value indicates higher occurrence of a particular parameter in thermophilic proteins while a negative value corresponds to higher occurrence in their mesophilic counterparts.

Table 3.5: Comparative analysis of various thermostability factors among 16 thermophilic-mesophilic pairs of protein.

TS*	Thermophilic (TS) and Mesophilic (MS) protein pairs**																Total positive values ***
	1AJ8	1BD M	1CA A	1CIU	1GTM	1LDN	1LN F	1PH P	1TM Y	1XGS	1YN A	1ZIN	1IQZ	2PR D	3MD S	3PFK	
MS*	1CS H	4MD H	8RX N	1CD G	1HRD	1LDG	1NP C	1QP G	3CH Y	1MAT	1XN B	1AK Y	1FC A	1INO	1QN M	2PFK	
Aro	0.17	-0.03	-0.02	0.15	-0.05	0.81	0.11	-0.06	-0.41	0.25	-0.05	0.32	1.38	-0.18	0.16	-0.22	8
UP	-0.37	-0.23	-0.22	0.12	-0.09	-0.19	-0.09	-0.39	-0.21	-0.28	-0.28	-0.26	-0.22	-0.28	-0.26	0.17	2
Pro	-0.17	0.26	-0.18	-0.10	0.02	-0.25	0.34	-0.07	0.78	0.14	-0.05	-0.30	0.02	-0.15	0.27	-0.12	7
ALI	0.09	-0.03	0.12	0.05	0.04	-0.14	0.14	0.06	0.12	0.05	0.03	0.11	0.05	0.10	0.07	-0.07	13
CHG	0.32	-0.05	0.26	-0.05	0.15	0.10	0.04	0.12	0.10	-0.01	0.45	0.01	0.89	0.01	0.07	0.01	13
R/K	-0.38	2.50	#	0.26	-0.07	4.26	0.60	1.14	-0.19	-0.25	0.91	1.83	#	1.28	0.59	-0.32	9
IP	0.54	-0.01	5.87	-0.01	0.57	0.20	0.00	0.70	0.51	0.45	1.45	-0.12	3.07	0.08	-0.27	0.50	11
AAI	0.24	0.92	-0.02	0.00	-0.20	6.50	0.06	0.05	-3.13	0.79	-0.05	0.92	0.00	0.01	-0.37	-0.16	8
ASI	-0.61	5.13	-0.02	0.23	-0.54	0.00	-0.32	2.16	0.07	-0.85	-0.52	0.92	0.02	-0.50	-0.51	-0.33	6
CPI	0.94	-0.42	0.23	0.07	-0.45	6.00	0.08	0.05	-0.47	-0.62	-0.15	-0.32	3.07	-0.06	-0.21	0.94	8
HB	-0.05	0.10	0.16	0.03	0.08	-0.14	0.00	0.06	-0.04	0.01	0.04	0.09	0.58	0.14	0.01	0.05	12
HP	0.17	0.16	0.01	0.01	0.11	0.10	0.16	0.08	-0.06	0.47	-0.09	-0.29	-0.16	-0.06	0.23	-0.11	10
Bt2P	-0.65	-0.19	-0.20	0.12	-0.40	0.33	0.50	-1.00	-0.40	-0.61	0.00	0.44	0.00	0.48	-0.49	0.50	6
NCap	-0.19	~	0.20	-0.72	1.11	0.33	-0.25	0.51	~	-1.00	#	-0.04	#	#	~	-0.50	7
Hdip	-0.11	0.03	0.00	0.26	0.23	0.01	-0.20	0.06	-0.17	0.07	0.00	-0.11	0.33	0.25	-0.27	-0.07	8
TL	-0.04	0.83	0.00	0.05	0.01	0.89	-0.30	-0.39	1.86	0.39	-0.15	-0.07	-0.50	-1.00	0.26	1.17	8
CS	1.35	-0.23	#	-0.11	-0.25	-0.33	-0.14	-0.12	~	-0.40	-0.05	-1.00	-0.15	-0.50	#	-0.14	2
NP/P	0.00	-0.08	0.00	-0.07	-0.01	-0.15	-0.10	-0.09	0.16	-0.22	-0.04	-0.25	-0.16	-0.11	0.01	0.01	3
Loop	0.09	-0.12	-0.05	0.07	-0.01	0.06	0.08	-0.01	-0.22	-0.12	-0.03	-0.05	-0.12	-0.17	0.15	0.10	6

* The PDB IDs of Thermophilic (TS), Mesophilic (MS) pair, starting from Column 2, belong to family Citrate Synthase, Malate dehydrogenase, Rubredoxin, Cyclodextrin, Glutamate dehydrogenase, L-Lactate dehydrogenase, Thermolysin, 3-Phosphoglycerate kinase, Chey protein, Methionine aminopeptidase, Endo-1,4-Beta-Xylanase, Adenylate kinase, Ferredoxin, Pyrophosphate phosphohydrolase, Manganese superoxide dismutase and Phosphofructokinase. The parameters listed in Column 1 correspond to aromatic (Aro: residues FWY), uncharged polar (UP: residues NQST), proline (Pro), hydrophobic or aliphatic (ALI: residues VILM), charged (CHG: residues DERKH) residue contents, Arg to Lys ratio (R/K), total percentage of ion-pairs (IP), aromatic-aromatic (AAI), aromatic-sulphur (ASI), cation-pi (CPI), hydrogen bonding (HB), hydrophobic (HP) interactions, proline residue percentages occurring at 2nd position of beta turns (Bt2P) and Ncap helix positions (NCap), percentage of dipole stabilized helices (Hdip), thermolabile bonds (TL) and conformationally strained residues (CS), ratio of nonpolar to polar accessible surface areas (NP/P) and percentage of loop region (Loop). **The value shown in # represents the case in which both MS and TS proteins show absence of the corresponding parameters while the values shown in ~ represents the case in which only the MS protein shows absence of the corresponding features. Therefore numbers of ~ values are also considered while counting total number of positive values. Detailed results can be found at http://irdp.ncl.res.in/cgi-bin/result_fetch.php?ID=iCAPScase.

Table 3.6: Details of proteins considered in iCAPS validation.

Protein family	TS								MS								R.M. S.D (in Å)	Seq. ID (in %)
	Source	TL (in °C)	PDB entry	Seq. Length	Resolution (Å)	SCOP Class	Mol. Wt. (in Da)	Oligomeric state	Source	TL (in °C)*	PDB entry	Seq. Length	Resolution (Å)	SCOP Class	Mol. Wt. (in Da)	Oligomeric state		
Citrate synthase	<i>Pyrococcus furiosus</i>	100	1AJ8	371	1.9	All alpha	42340.4	Dimer	<i>Gallus gallus</i>	37	1CSH	435	1.6	All alpha	48175.4	Dimer	1.68	26.2
Malate dehydrogenase	<i>Thermus flavus</i>	70–75	1BDM	327	2.5	Alpha and beta	35465	Dimer	<i>Sus scrofa</i>	37	4MDH	334	2.5	Alpha and beta	36394.3	Dimer	0.94	54.1
Rubredoxin	<i>Pyrococcus furiosus</i>	100	1CAA	53	1.8	Small proteins	5900.58	Monomer	<i>Desulfovibrio vulgaris</i>	34–37	8RXN	52	1	Small proteins	5578.21	Monomer	0.69	66.7
Cyclodextrin	<i>Thermoanaerobacterium thermosulfurigenes</i>	60	1CIU	683	2.3	All beta	75498.8	Monomer	<i>Bacillus circulans</i>	30–40	1CDG	686	2	All beta	74576.2	Monomer	0.7	70.5
Glutamate dehydrogenase	<i>Pyrococcus furiosus</i>	75–100	1GTM	419	2.2	Alpha and beta	46983.1	Hexamer	<i>Clostridium symbiosum</i>	30–37	1HRD	449	1.96	Alpha and beta	49216.1	Hexamer	1.38	34.3
Lactate dehydrogenase	<i>Bacillus stearothermophilus</i>	40–65	1LDN	316	2.5	Alpha and beta	34745.8	Tetramer	<i>Plasmodium falciparum</i>	37	1LDG	316	1.74	Alpha and beta	34163	Tetramer	1.25	28.4
Thermolysin and neutral	<i>Bacillus thermoproteolyticus</i>	52.5	1LNF	316	1.7	Alpha and beta	34362.6	Monomer	<i>Bacillus cereus</i>	30	1NPC	317	2	Alpha and beta	33816.8	Monomer	0.86	73.3
3-Phosphoglycerate kinase	<i>Bacillus stearothermophilus</i>	40–65	1PHP	394	1.65	Alpha and beta	42790.5	Monomer	<i>Saccharomyces cerevisiae</i>	25–30	1QPG	415	2.4	Alpha and beta	44641.6	Monomer	1.28	51.4
CheY	<i>Thermotoga maritima</i>	90	1TMY	120	1.9	Alpha and beta	13234.8	Monomer	<i>Escherichia coli</i>	37	3CHY	128	1.66	Alpha and beta	13981.2	Monomer	1.39	28.6

Methionine aminopeptidase	<i>Pyrococcus furiosus</i>	100	1XGS	295	1.75	Alpha and beta	32888.7	Dimer	<i>Escherichia coli</i>	37	1MAT	264	2.4	Alpha and beta	29371	Monomer	1.39	30.6
Endo-1,4-b Xylanase	<i>Thermomyces lanuginosus</i>	50	1YNA	194	1.55	All beta	21312	Monomer	<i>Bacillus circulans</i>	30–40	1XNB	185	1.49	All beta	20409.2	Monomer	1.14	50.9
Adenylate kinase	<i>Bacillus stearothermophilus</i>	40–65	1ZIN	217	1.65	Alpha and beta	24175	Monomer	<i>Saccharomyces cerevisiae</i>	25–30	1AKY	220	1.63	Alpha and beta	24068.7	Monomer	1.22	42
Ferredoxin	<i>Bacillus thermoproteolyticus</i>	52.5	1IQZ	81	2.3	Alpha and beta	8773.65	Monomer	<i>Clostridium acidurici</i>	19–37	1FCA	55	1.8	Alpha and beta	5496.18	Monomer	1.27	24
Inorganic pyrophosphatase	<i>Thermus thermophilus</i>	70–75	2PRD	174	2	All beta	19110	Hexamer	<i>Escherichia coli</i>	37	1INO	175	2.2	All beta	19597.5	Hexamer	1.1	48.5
Manganese superoxide dismutase	<i>Thermus thermophilus</i>	70–75	3MDS	203	1.8	All alpha	23129.5	Tetramer	<i>Homo sapiens</i>	37	1QNM	198	2.3	All alpha	22219.3	Tetramer	1.17	53.2
Phosphofructokinase	<i>Bacillus stearothermophilus</i>	40–65	3PFK	319	2.4	Alpha and beta	34167.1	Tetramer	<i>Escherichia coli</i>	37	2PFK	320	2.4	Alpha and beta	34885.3	Tetramer	0.87	57.1

* TL corresponds to living temperature.

Comparative amino acid composition analysis revealed 13 families showing higher preference of charged (CHG) amino acids while 14 families displayed a lesser content of uncharged polar amino acids (UP) in the thermophilic proteins (Table 3.5). Of the 16 families, Ferredoxin from *Bacillus thermoproteolyticus* (1IQZ) was observed to show highest preference for charged residue content compared to its mesophilic partner from *Clostridium acidurici*. Similarly, 3-Phosphoglycerate kinase from *Geobacillus stearothermophilus* (1PHP) showed lowest preference for UP content compared to its mesophilic homolog from *Saccharomyces cerevisiae*. This preference for charged residues compared to uncharged polar residues, a thermostabilization trend (Chakravarty & Varadarajan, 2000), also affected other parameters such as ion-pairs (IP), and the R/K ratio. It was found that 11 families had higher numbers of ion-pairs. The percentage change of ion-pairs was observed to be highest in case of Rubredoxin, a 53-residue protein. Rubredoxin from thermophilic *Pyrococcus furiosus* (1CAA), has 7 ion-pairs in its structure (5 ion-pairs form a network) while its mesophilic counterpart from *Desulfovibrio vulgaris* showed only one ion-pair. Similarly, 9 thermophilic proteins showed higher preference for Arginine than Lysine with highest preference observed in case of L-Lactate dehydrogenase enzyme family (1LDN). In terms of hydrogen bonding interactions (HB), 13 families contained a higher number of hydrogen bonds in the thermophilic set as compared to their mesophilic counterparts, thereby revealing hydrogen bonds as a contributing factor towards better protein stability. For this dataset aromatic residue content (Aro) and interactions involving aromatic amino acids (AAI, ASI and CPI) showed lower contribution towards stability. Hydrophobic residue content (ALI) and hydrophobic interactions (HP) were observed to be higher in 13 and 10 families of thermophilic proteins, respectively. Among all pairs in the dataset, Methionine aminopeptidase from *Pyrococcus furiosus* (1XGS) showed highest hydrophobic interactions compared to its mesophilic counterpart from *Escherichia coli*. Reduction in hydrophobic surface area of a protein is a known thermostabilization mechanism. It was seen that in 11 cases the change in NP/P ratio was found to be negative (highest in case of Adenylate kinase family). While 7 families in the thermophilic set showed higher Pro content, in the current dataset only 6 and 7 thermophilic proteins respectively show beta-turn (Bt2P), NCap proline insertion parameters to be a contributing factor. Contribution by shortening of loop regions (Loop) in proteins towards thermostability was observed in 10 thermophilic proteins. In case of CheY protein family, loop percentage was observed to be lowest in case of thermophilic

protein (1TMY) than its mesophilic partner. It was observed that 12 thermophilic proteins contained fewer conformationally strained residues (CS), a factor contributing positively towards thermostability.

Although it was difficult to observe a generalized rule for protein thermostabilization, the analysis highlighted few parameters such as charged residue preference, increased ion-pairs and hydrogen bonding interactions, decreased non-polar accessible surface area and conformationally strained residues, and shortening of loops to contribute positively to thermostability of proteins in this dataset.

3.3.3 Identification of potential stabilizing mutations in a protein using iStability module

i. Increasing protein thermostability through release of conformational strain by mutation to Glycine.

For the *release of conformational strain strategy*, iStability identifies conformationally strained residues, mutates them to glycine to release the strain and predicts their effects on stability. The results constitute the stability score, stability prediction (I: increasing stability and D: decreasing stability) and conservation score of the residue being mutated. Figure 3.9 shows sample output for conformational strain release strategy applied on *Ribonuclease HI* from *E. coli*.

MutantPDB	Chain	Res.No	Wild_Residue	Mut_Residue	Score	Stability	CScore
mutantpdb	A	90	W	G	3.85	D	6
mutantpdb	A	95	K	G	-1.51	I	4
mutantpdb	A	100	N	G	0.49	D	8

Figure 3.9: Using the *Stabilization by release of conformational strain* strategy with the FoldX as stability prediction tool, three residues Trp90, Lys95 and Asn100 were predicted to impose a conformational strain on the protein (Ribonuclease HI from *E. coli*; PDB ID 2RN2) due to their left-handed helical conformation. Of the three, only the K95G mutation was predicted to increase stability. This result correlated with the increase of 6.8 °C in the Δt_m along-with an increase in stability of 1.9 kcal/mol in the free energy of unfolding for the K95G mutation reported by (Kimura *et al.*, 1992). The CScore represents the evolutionary conservation score on a scale of 1-9 (9: Highly conserved). Using the hyperlink of first column, user can download the mutant PDB files for downstream analysis.

To check for the validity of this strategy a total 14 conformational strained residues in 10 proteins from 5 organisms (Table 3.7) were studied using iStability and the predictions were compared with experimentally validated results. The stability of 11 mutants was predicted accurately by iStability. Only 3 cases (R21G in 1LZ1, N30G in 1PIN and K136G in 1STN) were predicted incorrectly by the iStability module. Experimentally these three mutants were found to be thermostable (Jäger *et al.*, 2009; Stites *et al.*, 1994; Takano *et al.*, 2001) while iStability predicted decreased stability.

ii. Improvement of protein thermostability by entropic reduction due to Proline introduction.

In case of entropic stabilization strategy by *insertion of proline residues*, iStability identifies the beta-turns in the protein containing non-proline residues at second position and helices containing non-proline residues at the N-cap position. Residues in the identified positions are then mutated to proline followed by prediction of the mutant stability.

Using iStability we have studied 28, second position β -turn proline insertions in 17 proteins from 13 organisms for which experimental stability results were available (Table 3.7). Of the 28, iStability could accurately predict the stabilities for 22 β -turn insertions. In 20 cases, upon proline insertion an increase in stability was observed both experimentally and in iStability results. In the case of *Protein G* from *Streptococcus sp. GX7805*, the mutation K10P was predicted to decrease the stability by iStability which correlated with experimental results showing a decrease of 8.4 °C in the T_m value of the mutant. For the G68P mutation in 2IMM, experimentally a significant decrease in stability was observed which iStability also predicted accurately. For three cases (A93P in 2RN2, G13P and A206P in 3MBP) showing near wild-type stability experimentally, iStability predicted an increase in stability. For 3 other cases (A48P in 1PGA, L15P in 1LVE and A21P in 1RTP), the module was unable to predict the stability of the mutants correctly as experimentally they were observed to have decreased stability whereas iStability predicted them to have increased stability.

11 proline insertions at the N-cap position of helices were analysed in 6 proteins from 5 organisms (Table 3.7) for which experimentally determined stability results were available. Of the 11 mutations only 1 mutation, namely L316P carried out for *alcohol dehydrogenase* was predicted inaccurately by iStability. The experimental results (Bogin *et al.*, 1998) for this mutant

indicate the mutant to have higher stability (ΔT_m : +10.8 °C) than wild-type while iStability predicts a decreased stability for the mutant.

iii. Reducing entropy for enhancement of thermostability by introduction of disulfide bridges.

For the *insertion of disulfide bonds* strategy, iStability invokes SSBOND software (Hazes & Dijkstra, 1988), which identifies and ranks residue pairs that on mutation to cysteines could form stable disulfide bridges. Based on identified residue pairs, disulfide bonds are inserted and their effect on stability predicted.

A set of 28 double Cysteine mutations (Table 3.7) was studied for insertion of disulfide bonds for enhancement of protein stability in 15 proteins from 9 organisms. For this strategy, though iStability detected all the 28 residues pairs as potential insertion sites, stability was predicted correctly in 11 cases on comparison with experimentally determined stabilities. In one case (T72C, A471C in 3GLY), experimental evidence showed near wild-type stability while iStability predicted increased stability.

Apart from implementation of these strategies, iStability reads user-defined mutations through the customized mutation feature and predicts the mutant stability. For further downstream analysis, generated mutant structures can be downloaded. The identification of mutation sites along with stability predictions and residue conservation makes iStability a unique *in silico* protein-engineering tool.

Of the total 81 predictions studied, 47 were true-positives (Both experiment and predictions showed increased stability), 8 were true-negatives (Both experiment and predictions showed decreased stability), 9 were false-positives (Experiment showed increase while prediction showed decrease of stability) and 17 were false-negatives (Experiment showed decrease while prediction showed increase of stability). Thus, the true-positive rate (Sensitivity) calculated was 0.73 while true-negative rate (Specificity) was 0.47. While the sensitivity and specificity calculated actually test the accuracy of the underlying stability prediction programs, the values shown above also reflect the importance of the use of protein design strategies for better prediction of mutation sites. In those cases where iStability prediction differed from that observed experimentally, further analysis was carried out using iMutants. In case of A48P, the beta-turn proline insertion strategy in 1PGA, iMutants analysis revealed the loss of A48 (N) –

(OD1) 46D hydrogen bond in mutant protein. This loss of hydrogen bond could result in decrease in stability observed experimentally. Similar changes in interactions near the mutation site were also noted in other cases, suggesting the need for further evaluation of the identified mutants using iMutants.

iStability currently relies on use of stability prediction that is based either on empirical potential energy functions or machine learning methods. Since these tools are based on defined training dataset, the predictions on mutations that are distant to the training dataset are a cause for concern.

Table 3.7: Validation of iStability using the four protein engineering strategies.

PDB ID	Protein	Organisms	Mutation	Experiment *	iStability*	Other stable sites**	PUBMED Id
Stabilization by insertion of Proline residues at 2nd position of Beta-turns							
1CSP	Cold shock protein	<i>Bacillus subtilis</i>	N55P	I (1.0 kcal/mol)	I	3	19626709
1ZW7	Ubiquitin	<i>Saccharomyces cerevisiae</i>	S19P	I (0.9 kcal/mol)	I	2	
1PYL	Ribonuclease Sa2	<i>Streptomyces aureofaciens</i>	N33P	I (0.5 kcal/mol)	I	1	
			N51P	I (0.7 kcal/mol)	I		
1MGR	Guanyl-specific ribonuclease Sa3		S34P	I (0.9 kcal/mol)	I	3	
			T52P	I (0.5 kcal/mol)	I		
9RNT	Ribonuclease T1	<i>Aspergillus oryzae</i>	S63P	I (0.8 kcal/mol)	I	1	
2RN2	RibonucleaseH	<i>Escherichia coli</i>	A93P	N (-0.1 kcal/mol)	I	4	
			G123P	I (0.3 kcal/mol)	I		
3MBP	Maltose Binding Protein		G13P	N (0 kcal/mol)	I	3	
		A206P	N (-0.1 kcal/mol)	I			
1RGG	Ribonuclease (RNase) Sa	<i>Streptomyces aureofaciens</i>	S31P	I (0.7 kcal/mol)	I	3	17765922
			T76P	I (1 kcal/mol)	I		
1PGA	Protein G	<i>Streptococcus sp. GX7805</i>	K10P	D (-8.4 °C)	D	1	16549401
			A48P	D (-6.8 °C)	I		
2LZM	Bacteriophage T4 Lysozyme	<i>Enterobacteria phage T4</i>	A82P	I (0.8 °C)	I	1	1457724
1UOK	Oligo-1, 6-glucosidase	<i>Bacillus cereus</i>	K121P	I (4.6 kJ/mol)	I	11	8001545
			E208P	I (11.7 kJ/mol)	I		
			E290P	I	I		
2IMM	IgA-Kappa MCPC603 FV (Light chain)	<i>Mus musculus</i>	A15P	I	I	2	9007995
			S56P	I	I		
			D60P	I	I		
			G68P	D	D		
1LZ1	Lysozyme	<i>Homo sapiens</i>	A47P	I (0.3 °C)	I	2	1643041
1KEV	Alcohol dehydrogenase	<i>Clostridium</i>	S24P	I (3.9 °C)	I	10	9836874

		<i>beijerinckii</i>					
1LVE	Immunoglobulin K-4 light chain Len	<i>Homo sapiens</i>	L15P	D (-1.15 kcal/mol)	I	4	10091653
1RTP	Alpha-Parvalbumin	<i>Rattus rattus</i>	A21P	D (-8.5 °C)	I	1	12974622
3GLY	Glucoamylase	<i>Aspergillus awamori</i>	S30P	I (1.6 kJ/mol)	I	13	9796827
Stabilization by insertion of Proline residues at N-cap of Helices							
1HTI	Triosephosphate isomerase	<i>Homo sapiens</i>	A215P	I	I	0	8672446
1LZ1	Lysozyme	<i>Homo sapiens</i>	V110P	I	I	2	1911779
1RTP	Alpha-parvalbumin	<i>Rattus rattus</i>	H26P	I(5.6 °C)	I	5	12974622
1UOK	oligo-1, 6-glucosidase	<i>Bacillus cereus</i>	N109P	I	I	6	8001545
			E175P	I	I		
			T261P	I	I		
			E270P	I	I		
			I403P	I	I		
1KEV	Alcohol dehydrogenase	<i>Clostridium beijerinckii</i>	A177P	I(0.5 °C)	I	4	9836874
			L316P	I(10.8 °C)	D		
2LZM	Bacteriophage T4 Lysozyme	<i>Enterobacteria phage T4</i>	K60P	I (0.3 °C)	I	2	1457724
Stabilization by Conformational Strain release strategy							
1A5E	Cyclin-dependent kinase inhibitor	<i>Homo sapiens</i>	L78G	I (0.9 °C)	I	2	12614625
1LZ1	Lysozyme		R50G	I (0.9 °C)	I		
			Q58G	I (5.7 °C)	I		
			R21G	I (3.7 °C)	D		
			N118G	I (0.2 °C)	I		
			N30G	I (6.4 °C)	D		
1PIN	Pin1 WW domain		S18G	I (0.02 kcal/mol)	I	0	19565466
1STN	Staphylococcal nuclease	<i>Staphylococcus aureus</i>	K136G	I (0.1 kcal/mol)	D	1	8289248
2AFG	Acidic fibroblast growth facto	<i>Homo sapiens</i>	N106G	I (0.38 kcal/mo	I	1	12729767

				l)			
2RN2	Ribonuclease HI	<i>Escherichia coli</i>	K95G	I (5.7 °C)	I	0	1331044
1BNI	Barnase	<i>Bacillus amyloliquefaciens</i>	H18G	I (0.51 kcal/mo l)	I	2	1569555
2AFG	Acidic fibroblast growth factor	<i>Homo sapiens</i>	H93G	I (1.32 kcal/mo l)	I	1	7692436
1ROP	Rop	<i>Escherichia coli</i>	D30G	I (11.6 °C)	I	0	8548455
2A36	Drk	<i>Drosophila</i>	T22G	I (3.6 kcal/mo l)	I	1	16300404
Stabilization by insertion of Disulfide bridges							
1BNI	Barnase	<i>Bacillus amyloliquefaciens</i>	A43C,S80C	I	D	-	8476861
			T70C,S92C	D	D	-	
5AZU	Azurin	<i>Pseudomonas aeruginosa</i>	D62C,K74C	I	D	-	15449946
1BCX	Xylanase	<i>Bacillus circulans</i>	V98C,A152C	I	I	-	17141401
			S100C,N148C	I	D	-	
1CAH	Carbonic anhydrase II	<i>Homo sapiens</i>	L60C,S173C	I	I	-	10794421
			A38C,A258C	D	D	-	
			S99C,V242C	I	D	-	
1PLC	Plastocyanin	<i>Populus nigra</i>	I21C,E25C	I	D	-	11679761
1WE4	β -lactamase	<i>Escherichia coli</i>	C69C,G238C	I	D	-	15595829
1LTA	Cholera toxin	<i>Vibrio cholerae</i>	N40C,G166C	I	D	-	9416616
3CI2	Chymotrypsin inhibitor-2	<i>Homo sapiens</i>	T22C,V82C	I	I	-	11045611
4DFR	Dihydrofolatereductase	<i>Escherichia coli</i>	P39C,C85C	I	D	-	3304420
3GLY	Glucoamylase	<i>Aspergillus awamori</i>	T72C,A471C	N	I	-	9749918
			T246C,C320C	I	D	-	8679632
1PII	Indoleglycerol-phosphate synthase	<i>Escherichia coli</i>	T3C,R189C	I	D	-	11856350
1FYH	Interferon-gamma	<i>Homo sapiens</i>	E7C,S69C	I	D	-	8931130

1EYA	Nuclease V8	<i>Staphylococcus aureus</i>	Q80C,K116C	I	D	-	8756688
			N118C,D77C	D	D	-	
1EY0	Nuclease V8		N118C,G79C	I	I	-	
1Z7X	Ribonuclease I	<i>Homo sapiens</i>	A4C,V118C	I	D	-	10920260
1SBT	Subtilisin BPN	<i>Bacillus amyloliquefaciens</i>	A26C,A232C	D	D	-	2504281
			D41C,G80C	D	I	-	
T22C,S87C			I	I	-	3476160	
1SUE			A29C,M119C	D	D	-	2504281
			D36C,P210C	D	I	-	
			V148C,N243C	D	D	-	
			A22C,A87C	I	I	-	3476160

* The labels I, D and N correspond to an increase, decrease and no change in stability, respectively for the mutations as inferred from experiment. iStability predicts two states denoted I and D for the mutations. The values with unit kcal/mol represent $\Delta\Delta G$ value (Change in free energy of unfolding, Mutant-Wild) while those with unit $^{\circ}\text{C}$ represent ΔT_m value (Change in midpoint temperature of the thermal unfolding, Mutant-Wild) as inferred from the experiment. A positive value represents an increase in stability. ** The value represents the total number of other mutation sites that have been identified by iStability module as potential stabilizing sites, which can be considered for thermostabilization of respective proteins.

3.3.4 Evaluating mutations through interaction framework and evolutionary residue conservation at mutation sites using iMutants

A comparative local interaction profile generated from the detailed atomic interactions in a model is unique to iMutants. It offers a quantitative measure of structural changes in mutants, through loss or gain of interactions at the mutation site. The module provides a comparative interaction analysis of wild-type and mutant residues summarized in the form of a local interaction profile comprising a number of interactions and their networks (Fig. 3.10). Hyperlinks provide details of calculated interactions. iMutants also supplements the interaction profile with estimated evolutionary conservation scores of the wild-type residues being mutated. In addition, mutant structures generated can also be downloaded for further downstream analysis.

Local Interactions	Download	Type	Residue details				Interaction profile											
			Chain	Res.No	Res.ID	CScore	IP	IP.Net	AP	AP.Net	AS	AS.Net	HB	Disul	Cat-pi	Cat-pi.Net	Hphob	
1	Mutant	wild	A	20	Y	3	0	0	0	0	0	0	0	2	0	1	0	2
		mut	A	20	A	-	0	0	0	0	0	0	0	2	0	0	0	0
2	Mutant	wild	A	39	R	9	1	1	0	0	0	0	7	0	1	1	0	
		mut	A	39	A	-	0	0	0	0	0	0	3	0	0	0	3	
		wild	A	60	K	2	0	0	0	0	0	0	1	0	0	0	0	
		mut	A	60	G	-	0	0	0	0	0	0	1	0	0	0	0	

Figure 3.10: Illustrates sample output of iMutants module where a single mutation (Y20A) and a double mutation (R39A/K60G) were carried out on *penicillin amidase* (PDB ID: 1PNK). In case of Y20A mutation, the *interaction profile column* shows a possible loss of 1 cation-pi interaction (Cat-pi) and 2 hydrophobic interactions (Hphob) while it showed no change in hydrogen bonding interactions (HB). In case of R39A/K60G double mutant, a loss of 1 ionic interaction (IP), 1 ion-pair network (IP.Net), 4 hydrogen bonds (HB), 1 cation-pi interaction (Cat-pi) and 1 cation-pi interaction network (Cat-pi.Net) was observed due to R39A mutation while K60G mutation has not much effect at the mutation site. Mutant structures can be downloaded by clicking the hyperlink on *Download column*. Clicking the hyperlink on *Local interactions column* opens another web page showing the detailed local interaction changes. The *CScore column* indicates conservative nature of the mutation site; R39 position is highly conserved with CScore 9 while Y20 and K60 positions are variable in nature.

3.3.5 Demonstration of the utility of iMutants module

The equilibrium stabilities of 51 mutants for the *Arc Repressor* protein of *bacteriophage P22* (PDB ID: 1ARR) have been studied experimentally by Milla *et al.*, 1994, using thermal and urea denaturation (Milla *et al.*, 1994). These 51 mutations were analysed using the iMutants module and the change in various non-bonded interactions was recorded. The mutations were divided into four groups as established by Milla *et al.*, 1994, for analysis purposes. The first group consisted of 5 mutants (Table 3.8; V22A, I37A, V41A, F45A, E36A), which were experimentally determined to be **highly unstable** as they were unable to form dimers, and remained in an unfolded state. Of the five, the first four mutations showed a dramatic decrease of 5, 6, 3 and 4 hydrophobic interactions, respectively. These interactions affect the hydrophobic core of the protein and the loss of interactions coincides with the experimental instability observed. Since the E36A mutation involves alteration of a buried polar residue, iMutants

recorded drastic loss of two ionic interactions, one ionic network along with one hydrogen bond, which was established by Milla *et al.*, 1994, as a possible cause of instability of the mutant.

The next set analysed comprised 20 mutants (Table 3.8), which experimentally exhibited **reduced stability** with t_m values ranging from 30-50 °C as compared to the wild-type protein (T_m : 57.9 °C). Since three mutations R31A, R40A and R50A involved polar residues, iMutants recorded a loss of one ionic interaction and one ionic interaction network for R31A, two ionic interactions, two hydrogen bonds and one ionic interaction network for R40A and finally one ionic interaction for R50A mutants that could explain the instability of these mutants (T_m : 37.1 °C, 31.2 °C and 47.9 °C, respectively). For mutants W14A, L21A, N29A, V33A and Y38A, changes in hydrogen bonding interactions were observed (Table 3.8). The mutation W14A also showed a loss of two aromatic pair interactions, one aromatic pair network as well as five hydrophobic interactions. This change in interactions could thus explain the decrease of T_m to 31.5 °C observed for this mutant. Mutants F10A, L12A, P15A, L19A, L21A, Y38A and M42A showed a loss of 6, 4, 2, 4, 1, 3, and 2 hydrophobic interactions, respectively, which could contribute to the instability observed.

The set of 25 mutants (Table 3.8) analysed next displayed **near wild-type stability** with their T_m ranging between 55-63 °C. Most mutations in this set showed marginal or no change in their hydrogen bonding, ionic and hydrophobic interactions. P8A, the only mutant with **increased stability** (T_m : 74.1 °C), showed a change in just one hydrophobic interaction in the iMutants analysis (Table 3.8). The stabilization of this particular mutant could be due to the extension of β -sheets or relief from unfavourable packing interactions as postulated by Milla *et al.*, 1994.

Table 3.8: The iMutants analysis on 51 mutations in *Arc Repressor* protein of *bacteriophage P22*.

No	Type	Chain	Res No *	Res ID *	Local Interaction Profile**											T _m (obs) °C
					IP	IP.N et	AP	AP.N et	AS	AS.N et	HB	Disul	Ca t-pi	Cat-pi.N et	Hph ob	
Highly Unstable Mutations																
1	wild	A	22	V	-	-	-	-	-	-	2	-	-	-	7	<20
	mut	A	22	A	-	-	-	-	-	-	2	-	-	-	2	
2	wild	A	36	E	2	1	-	-	-	-	2	-	-	-	-	<20
	mut	A	36	A	-	-	-	-	-	-	1	-	-	-	2	
3	wild	A	37	I	-	-	-	-	-	-	3	-	-	-	9	<20
	mut	A	37	A	-	-	-	-	-	-	2	-	-	-	3	
4	wild	A	41	V	-	-	-	-	-	-	2	-	-	-	6	<20
	mut	A	41	A	-	-	-	-	-	-	1	-	-	-	3	
5	wild	A	45	F	-	-	-	-	-	-	-	-	-	-	6	<20
	mut	A	45	A	-	-	-	-	-	-	-	-	-	-	2	
Unstable Mutations																
6	wild	A	10	F	-	-	1	1	-	-	2	-	-	-	6	40.6
	mut	A	10	A	-	-	-	-	-	-	2	-	-	-	-	
7	wild	A	12	L	-	-	-	-	-	-	2	-	-	-	6	42.3
	mut	A	12	A	-	-	-	-	-	-	2	-	-	-	2	
8	wild	A	14	W	-	-	2	1	-	-	2	-	-	-	1	31.5
	mut	A	14	A	-	-	-	-	-	-	1	-	-	-	5	
9	wild	A	15	P	-	-	-	-	-	-	2	-	-	-	4	46.6
	mut	A	15	A	-	-	-	-	-	-	2	-	-	-	2	
10	wild	A	19	L	-	-	-	-	-	-	2	-	-	-	4	48.3
	mut	A	19	A	-	-	-	-	-	-	2	-	-	-	-	
11	wild	A	21	L	-	-	-	-	-	-	3	-	-	-	2	39.6
	mut	A	21	A	-	-	-	-	-	-	2	-	-	-	1	
12	wild	A	29	N	-	-	-	-	-	-	2	-	-	-	-	45.3
	mut	A	29	A	-	-	-	-	-	-	1	-	-	-	-	
13	wild	A	30	G	-	-	-	-	-	-	2	-	-	-	-	47.9
	mut	A	30	A	-	-	-	-	-	-	2	-	-	-	-	
14	wild	A	31	R	1	1	-	-	-	-	1	-	-	-	-	37.1
	mut	A	31	A	-	-	-	-	-	-	1	-	-	-	-	
15	wild	A	32	S	-	-	-	-	-	-	1	-	-	-	-	33.5
	mut	A	32	A	-	-	-	-	-	-	1	-	-	-	-	
16	wild	A	33	V	-	-	-	-	-	-	2	-	-	-	3	44.1
	mut	A	33	A	-	-	-	-	-	-	1	-	-	-	3	
17	wild	A	38	Y	-	-	1	1	1	-	3	-	-	-	4	33
	mut	A	38	A	-	-	-	-	-	-	2	-	-	-	1	
18	wild	A	40	R	2	1	-	-	-	-	3	-	-	-	-	31.2
	mut	A	40	A	-	-	-	-	-	-	1	-	-	-	2	
19	wild	A	42	M	-	-	-	-	1	-	1	-	-	-	3	35.6
	mut	A	42	A	-	-	-	-	-	-	1	-	-	-	1	
20	wild	A	44	S	-	-	-	-	-	-	1	-	-	-	-	46.3
	mut	A	44	A	-	-	-	-	-	-	2	-	-	-	2	
21	wild	A	47	K	-	-	-	-	-	-	3	-	-	-	-	47.2

	mut	A	47	A	-	-	-	-	-	-	1	-	-	-	-	
22	wild	A	48	E	-	-	-	-	-	-	1	-	-	-	-	43.2
	mut	A	48	A	-	-	-	-	-	-	1	-	-	-	1	
23	wild	A	49	G	-	-	-	-	-	-	-	-	-	-	-	48.7
	mut	A	49	A	-	-	-	-	-	-	-	-	-	-	-	
24	wild	A	50	R	1	-	-	-	-	-	3	-	-	-	-	47.9
	mut	A	50	A	-	-	-	-	-	-	3	-	-	-	1	
25	wild	A	51	I	-	-	-	-	-	-	1	-	-	-	2	50.9
	mut	A	51	A	-	-	-	-	-	-	1	-	-	-	2	
Mutations having Near Wild-type Stability																
26	wild	A	1	M	-	-	-	-	-	-	2	-	-	-	-	58
	mut	A	1	A	-	-	-	-	-	-	-	-	-	-	1	
27	wild	A	2	K	-	-	-	-	-	-	-	-	-	-	-	58.7
	mut	A	2	A	-	-	-	-	-	-	-	-	-	-	1	
28	wild	A	3	G	-	-	-	-	-	-	-	-	-	-	-	58.1
	mut	A	3	A	-	-	-	-	-	-	-	-	-	-	1	
29	wild	A	4	M	-	-	-	-	-	-	2	-	-	-	-	59.2
	mut	A	4	A	-	-	-	-	-	-	1	-	-	-	-	
30	wild	A	5	S	-	-	-	-	-	-	2	-	-	-	-	57.5
	mut	A	5	A	-	-	-	-	-	-	-	-	-	-	-	
31	wild	A	6	K	1	1	-	-	-	-	2	-	-	-	-	59.6
	mut	A	6	A	-	-	-	-	-	-	-	-	-	-	-	
32	wild	A	7	M	-	-	-	-	-	-	1	-	-	-	2	55.5
	mut	A	7	A	-	-	-	-	-	-	1	-	-	-	1	
33	wild	A	9	Q	-	-	-	-	-	-	1	-	-	-	-	58.4
	mut	A	9	A	-	-	-	-	-	-	-	-	-	-	-	
34	wild	A	11	N	-	-	-	-	-	-	1	-	-	-	-	62.1
	mut	A	11	A	-	-	-	-	-	-	-	-	-	-	-	
35	wild	A	13	R	-	-	-	-	-	-	3	-	-	-	-	57.3
	mut	A	13	A	-	-	-	-	-	-	2	-	-	-	1	
36	wild	A	16	R	1	1	-	-	-	-	2	-	-	-	-	59.5
	mut	A	16	A	-	-	-	-	-	-	2	-	-	-	1	
37	wild	A	17	E	1	-	-	-	-	-	3	-	-	-	-	57
	mut	A	17	A	-	-	-	-	-	-	1	-	-	-	-	
38	wild	A	18	V	-	-	-	-	-	-	3	-	-	-	7	56.9
	mut	A	18	A	-	-	-	-	-	-	2	-	-	-	4	
39	wild	A	20	D	2	1	-	-	-	-	2	-	-	-	-	55.3
	mut	A	20	A	-	-	-	-	-	-	2	-	-	-	-	
40	wild	A	23	R	1	1	-	-	-	-	2	-	1	-	-	56.7
	mut	A	23	A	-	-	-	-	-	-	2	-	-	-	1	
41	wild	A	24	K	1	1	-	-	-	-	2	-	-	-	-	56.3
	mut	A	24	A	-	-	-	-	-	-	2	-	-	-	1	
42	wild	A	25	V	-	-	-	-	-	-	2	-	-	-	3	59.3
	mut	A	25	A	-	-	-	-	-	-	2	-	-	-	2	
43	wild	A	27	E	-	-	-	-	-	-	2	-	-	-	-	58.8
	mut	A	27	A	-	-	-	-	-	-	1	-	-	-	-	
44	wild	A	28	E	2	1	-	-	-	-	1	-	-	-	-	55.7
	mut	A	28	A	-	-	-	-	-	-	1	-	-	-	-	
45	wild	A	34	N	-	-	-	-	-	-	3	-	-	-	-	63
	mut	A	34	A	-	-	-	-	-	-	1	-	-	-	2	

46	wild	A	35	S	-	-	-	-	-	-	2	-	-	-	-	63.4
	mut	A	35	A	-	-	-	-	-	-	2	-	-	-	-	
47	wild	A	39	Q	-	-	-	-	-	-	2	-	-	-	-	61.4
	mut	A	39	A	-	-	-	-	-	-	2	-	-	-	1	
48	wild	A	43	E	1	1	-	-	-	-	4	-	-	-	-	56.1
	mut	A	43	A	-	-	-	-	-	-	2	-	-	-	-	
49	wild	A	46	K	-	-	-	-	-	-	1	-	-	-	-	57.1
	mut	A	46	A	-	-	-	-	-	-	1	-	-	-	-	
50	wild	A	52	G	-	-	-	-	-	-	2	-	-	-	-	60.9
	mut	A	52	A	-	-	-	-	-	-	1	-	-	-	-	
Stable Mutations																
51	wild	A	8	P	-	-	-	-	-	-	1	-	-	-	2	74.1
	mut	A	8	A	-	-	-	-	-	-	1	-	-	-	1	

* ResNo and ResID correspond to residue number and residue name respectively. **Local interaction profile represents number of various interactions and interaction networks of wild-type (wild) and mutant (mut) residues. The label corresponds to number of IP: ion-pair, IP.Net: ion-pair networks, AP: aromatic-aromatic interaction, AP.Net: aromatic-aromatic interaction network, AS: aromatic sulphur interactions, AS.Net: aromatic-sulphur interaction network, HB: hydrogen bonds, Disul: disulfide bonds, Cat-pi: cation-pi interactions, Cat-pi.Net: cation-pi interaction networks, Hphob: hydrophobic interactions. The - (hyphen) corresponds to no interaction or interaction networks detected. The Tm(obs) value corresponds to the Tm value of Mutant. The results can be accessed following the link http://irdp.ncl.res.in/cgi-bin/result_fetch_MutAna.php?ID=iMutcase.

3.3.6 iATMs (*in silico* Analysis of Thermally stable Mutants): An information resource.

The local interaction analysis approach of iMutants was extended to analyse experimentally validated mutations listed in the ProTherm database and is provided in the form of iATMs (*in silico* Analysis of Thermally stable Mutants), as a supplementary information resource to ProTherm (Kumar *et al.*, 2006). Although ProTherm contains a vast resource of experimental information, no information is available describing the changes in the structure and atomic interactions due to the mutations carried out. iATMs is organized in three sections based on the type of mutation as single, double or multiple. Within these, the sections are further classified into those containing crystal structures for both wild-type and mutant proteins and those where only the wild-type crystal structures are available. Wherever wild-type and mutants structures were known, interaction profiles were generated using those structures. In cases where crystal structures for mutants were absent, generation of local interaction profiles was carried out using known wild-type and modelled mutant structure. Information provided in iATMs could provide a better understanding of correlation between experimental observations and interaction

rearrangements due to mutations, leading to better application of derived knowledge towards efficient protein engineering.

3.4 Summary

Despite the availability of a large number of structural analysis tools, to the best of our knowledge there is currently no unified platform addressing the rational protein design problem. The web platform iRDP uniquely offers investigators a multi-faceted approach for carrying out rational protein engineering by integrating protein structure and mutation analysis tools. The modules of iRDP server can either be used separately for various independent analyses or as a systematically directed strategy encompassing the steps involved in rational protein engineering. Applications of modules do not limit themselves to the protein stability problem since the information generated comprises of diverse structural features, which can correlate with a wide range of properties in proteins. Investigations carried out using iRDP act as a guide for analyzing varied structural features that relate to problems such as pH stability, protein active site analysis, crystallizability, analysis of frames from molecular simulations and protein structure-function relationships.

The future direction of the iRDP web server aspires towards implementation of sequence-based inputs complementing the existing structure-based input followed by visualization of interaction networks and mutation sites, thereby providing a better structural perspective.

Chapter 4

*A comparative study of
Penicillin G Acylases,
to assess the thermostability using
computational approaches*

Penicillin G Acylase (PGA) are enzyme of NtSn-hydrolase superfamily. They are widely used and commercially exploited in pharmaceutical industry for synthesis of many semi-synthetic antibiotics. Identification of novel sources of PGA enzymes that are stable under wide-range of reaction conditions such as temperature and pH would be beneficial to industry. The current chapter describes a multiple strategy based computational approach towards identification of potentially thermostable enzymes from PGA family.

4.1 Introduction

Penicillin G acylase (PGA, E.C. 3.5.1.11) is an important biocatalyst, commonly employed in pharmaceutical industry for the enzymatic deacylation of Penicillin G to yield phenylacetic acid and 6-aminopenicillanic acid (6-APA). The product 6-APA is a key intermediate in the large scale production of many semi-synthetic penicillins, which are more effective against resistant pathogens compared to natural penicillins (Arroyo *et al.*, 2003). These enzymes are also employed in resolving the racemic mixtures of chiral compounds such as secondary alcohols (Fuganti *et al.*, 1986b) and protection of amino and hydroxyl groups in peptide synthesis (Fuganti *et al.*, 1986a). Currently the most widely used source of PGA for the manufacture of β -lactam antibiotics is *Escherichia coli* (*EcPGA*). However, since this enzyme is found to be unstable beyond 30 °C, its industrial application requires this enzyme to be present in an immobilized form (Sio & Quax, 2004). As the rate of an enzymatic reaction increases with temperature, the role of PGAs as commercial biocatalyst would become far more attractive if their stability at higher temperatures can be improved. Several investigations in the past have focused on enhancing *EcPGA*'s thermostability through cross-linking the enzyme with glutaraldehyde (Erarslan & Kocer, 1992) or site-directed mutagenesis of few carefully selected amino acid residues (Polizzi *et al.*, 2006). However, these efforts seem to result only in minor enhancement of thermostability. Hence, efforts have also been undertaken to identify novel sources of thermostable PGAs namely, PGA from *Alcaligenes faecalis* (*AfPGA*) (Verhaert *et al.*, 1997) and *Achromobacter xylosoxidans* (*AxPGA*) (Cai *et al.*, 2004), which are comparatively more stable than *EcPGA*. In case of *AfPGA* and *AxPGA*, the half life of the enzymes at 55 °C has been observed to be 15 min and 55 min, respectively. The higher thermostability of *AfPGA* has been experimentally attributed to the presence of a disulfide bond in its structure (Varshney *et al.*, 2012; Verhaert *et al.*, 1997). Interestingly, *AxPGA*, which is even more thermostable than

*Af*PGA, lacks a disulfide bond. Factors such as preference of Arg residues over Lys, decrease in the number of thermolabile amino acids and bonds, increase in proline residue content and the presence of more stable ion-pairs have been suggested to play role in *Ax*PGA thermostability (Cai *et al.*, 2004). Understanding the mechanism of increased thermostability among PGAs can not only help in identification of novel sources of thermostable PGA but can also aid in the tailoring thermostability amongst mesostable PGAs via protein engineering.

With the advances in high throughput genome sequencing techniques and the availability of powerful functional annotation tools, PGAs have been predicted to occur in a wide range of microbial sources. However, the annotation of their biochemical and biophysical characteristics are currently unavailable. A preliminary computational screening of these putative PGAs for the identification of a thermostable candidate preceding their experimental characterization could prove to be beneficial with respect to both time and cost. In this work the presence of a disulfide bond was considered as a specific criterion for the selection of putative PGA enzymes from MEROPS (Rawlings *et al.*, 2012) database. Enzymes from three different sources, namely *Sphingomonas wittichii* (*Sw*PGA), *Paracoccus denitrificans* (*Pd*PGA) and *Acinetobacter oleivorans* (*Ao*PGA) were identified as probable thermostable candidates based on the above mentioned selection criterion. These three were then compared with the already well characterized PGAs, that is, *Ec*PGA (Least thermostable/Mesostable), *Af*PGA (Moderately thermostable) and *Ax*PGA (Most thermostable). Together the six PGAs were first compared using a *sequence-based consensus* approach followed by *structure-based* comparative analysis, in an attempt to explore the various known protein thermo-stabilization mechanisms. The former analysis revealed the three candidate enzymes could, at best, be mesostable in nature. However, the later structural analysis revealed that *Pd*PGA might be a potential thermostable enzyme. The iRDP web server developed (Chapter 3) was used for the comparative analysis of the PGA enzymes in terms of various structural parameters.

4. 2 Materials and Methods

4.2.1 Computational screening strategy for obtain of ptPGAs (putative thermostable PGAs)

Three-dimensional structures of *Ec*PGA (PDB ID: 1GK9) and *Af*PGA (PDB ID: 3K3W) were extracted from PDB (Berman *et al.*, 2000). The subfamily S45.001 (Penicillin G Acylase precursor subfamily) of MEROPS database was selected for computational screening of putative thermostable PGAs (ptPGAs). Each of the PGA sequence from the database was aligned with *Af*PGA and only those sequences having Cys residues at equivalent positions facilitating disulfide bond formation were selected.

4.2.2 Removal of signal and spacer peptide

The selected putative sequences (ptPGAs: *Sw*PGA, *Pd*PGA and *Ao*PGA) from MEROPS database were in their precursor form and the locations of their signal and spacer peptides were unknown. Since the active form of PGAs does not contain the signal and spacer peptides, these regions were identified by comparing the sequences with active and processed form of *Ec*PGA and *Af*PGA and were removed from ptPGAs.

4.2.3 Homology modeling

In the absence of three-dimensional structures for *Ax*PGA and ptPGAs, high-resolution crystal structure of *Ec*PGA was used for building three-dimensional homology models using Prime 3.0 (Version 3.1, Schrödinger, LLC, New York, NY, 2012) and validated by model validation programs. Comparative Modeling approach of Prime 3.0 was applied, which modeled the target sequences using template structure by a two-step process of target-template alignment followed by model building. The model building step involved first copying the coordinates of backbone atoms for the aligned regions along with side chains of conserved residues. This is followed by optimization of side chains, minimization of non-template residues and finally building insertions and removing deletions. OPLS_2005 all-atom force field was used for energy scoring of protein and Surface Generalized Born (SGB) continuum solvation model for treating solvation energies. PGA being a heterodimer, each chain was modeled separately by two different runs, and the heterodimer was assembled from individual chain models. The stereochemical quality and geometry of each of these final models was evaluated using Ramachandran

plot computed using PROCHECK (Laskowski *et al.*, 1993). Similarly the models were also validated for their quality using the model validation programs such as Errat (Colovos & Yeates, 1993), Verify3D (Eisenberg *et al.*, 1997) and Prosa (Wiederstein & Sippl, 2007).

4.2.4 Molecular dynamics simulations

Effects of temperature on PGA stabilities were studied by explicit solvent molecular dynamics simulations at three temperature scales (330K, 400K and 500K respectively) for a time scale of 15 ns using OPLS-AA force field in Gromacs 4.5 (Pronk *et al.*, 2013). Tip4p solvent model was used for modeling the solvent and appropriate Na⁺ and Cl⁻ ions were added to neutralize the system. The system was first subjected to steepest descent followed by conjugate gradient energy minimization. Resulting system was equilibrated in NVT ensemble for 100 ps at respective temperatures using V-rescale temperature coupling. The system was further equilibrated with NPT ensemble for 100 ps at 1 atm pressure and respective temperatures, using Parrinello-Rahman pressure coupling. The equilibrated system was finally subjected to molecular dynamics simulation using leap-frog integrator. For computing non-bonded interaction, grid based search algorithm was employed to generate the pair list and the list was updated in every 5 step. Short range neighbor list, electrostatic and VdW cutoff each was chosen as 10 Å. Particle Mesh Ewald (PME) method with 0.16 nm Fourier spacing and cubic interpolation was employed to treat the long-range electrostatic interactions in a periodic boundary condition.

4.2.5 Estimation of non-bonded interactions

Multiple sequence alignment of PGA sequences was carried out using ClustalX (Thompson *et al.*, 1997). Intra molecular interactions in PGA enzyme structures were identified using iCAPS module of iRDP web server (<http://irdp.ncl.res.in>) using default parameters. The interactions analyzed were disulfide bond, ion-pairs, aromatic-aromatic interactions, aromatic-sulphur interactions, cation-pi interactions and hydrogen bonding interactions. Cys SG atoms if lie within 2.2 Å, they were considered to form disulfide bond. If the distance between any of the oxygen atoms of acidic residues and the nitrogen atoms of basic residues were within the cut-off distance, they were considered to interact by ionic interactions. Three ranges, *short-range*, *medium-range* and *long-range* ionic interactions were estimated based on 4, 6 and 8 Å distance cut-offs. If distance between phenyl ring centroids of aromatic residues lie between 4.5-7.0 Å,

they were considered to interact via aromatic-aromatic interactions. If the distance between sulphur atoms of Cys/Met and aromatic ring centroids of Phe/Tyr/Trp lie within 5.3 Å, the residues were considered to interact via aromatic-sulphur interactions. If the cationic side chains of Lys/Arg residues lie within 6 Å of aromatic ring centroids, they were considered to interact via cation- π interactions. Hydrogen bonds are estimated using HBPLUS (McDonald & Thornton, 1994).

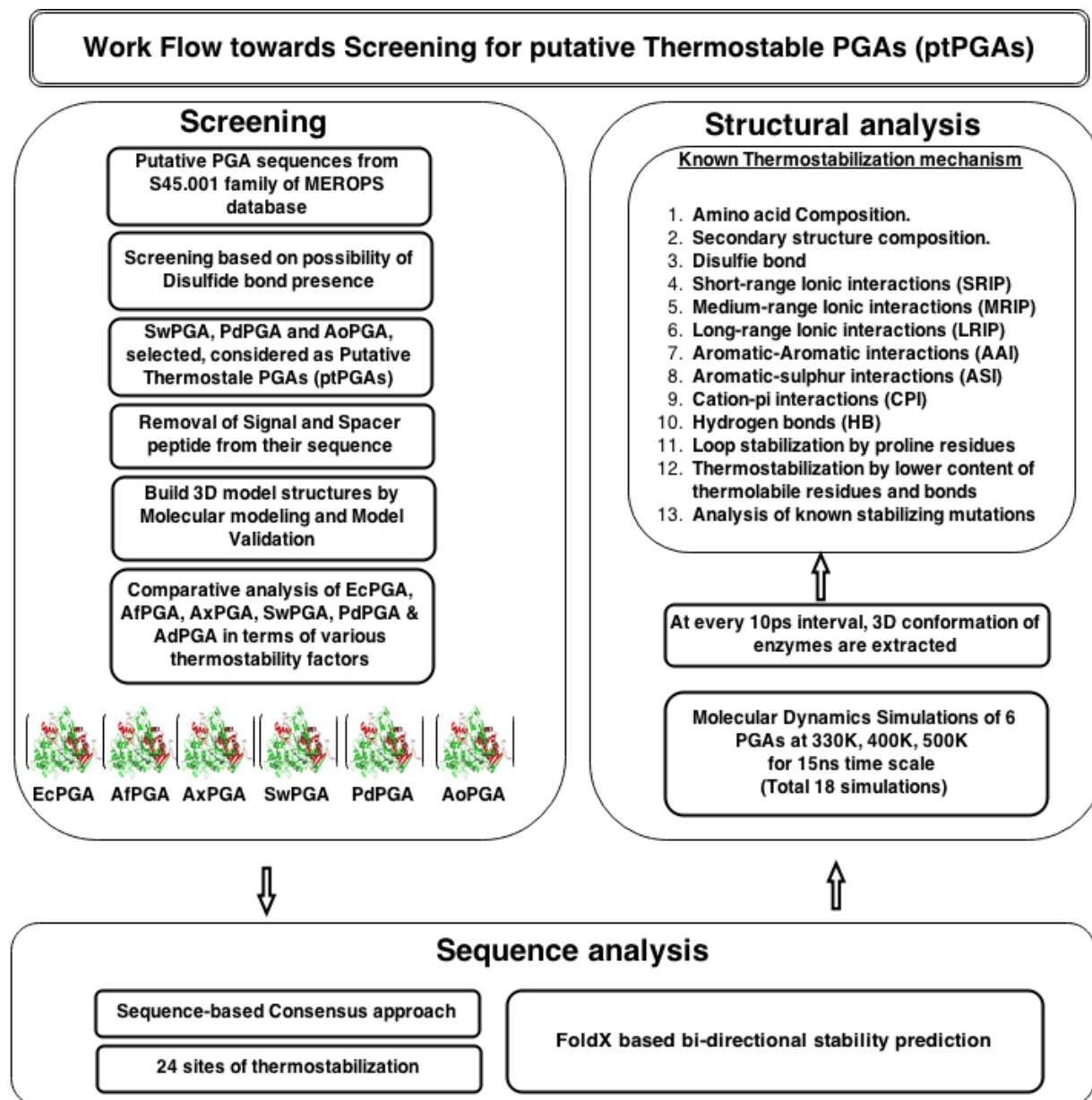


Figure 4.1: The workflow that was followed towards identification of potential thermostable PGA enzymes.

4.3 Results and Discussions

The workflow towards obtaining potential thermostable PGA enzymes involved computational screening of ptPGAs followed by their comparison with known PGAs in terms of both their sequence and structure (Fig. 4.1). In the following section first the *sequence-based consensus* approach is discussed followed by *structure-based comparative analysis* of PGA enzymes.

Table 4.1: PGAs from S45.001 family of MEROPS database containing Cys residues are listed.

MEROPS ID	Organism**	Cys residue position*
MER059680	<i>Sphingomonas wittichii</i> (SwPGA)	<u>6</u> , 27 , 766 , 799
MER074190	<i>Paracoccus denitrificans</i> (PdPGA)	<u>21</u> , 675 , 733 , 766
MER219436	<i>Acinetobacter oleivorans</i> (AoPGA)	<u>27</u> , 765 , 798
MER312023	<i>Acinetobacter baumannii</i>	<u>27</u> , 765 , 798
MER081546	<i>Acinetobacter baumannii</i>	<u>52</u> , 790 , 823
MER285059	<i>Acinetobacter calcoaceticus</i>	<u>27</u> , 765 , 798
MER238699	<i>Achromobacter piechaudii</i>	<u>14</u>
MER107787	<i>uncultured γ proteobacterium</i>	<u>25</u>
MER312015	<i>Marinobacterium stanieri</i>	<u>242</u>
MER087144	<i>Serratia proteamaculans</i>	<u>19</u>
MER107793	<i>Achromobacter sp. CCM 4824</i>	<u>16</u> , 782
MER311995	<i>Cupriavidus basilensis</i>	<u>12</u> , 358
MER238517	<i>Enterobacter cloacae</i>	<u>18</u> , 724
MER003307	<i>Kluyvera citrophila</i>	<u>18</u> , <u>19</u>
MER169879	<i>Luminiphilus syltensis NOR5-1B</i>	<u>7</u> , <u>19</u> , 211
MER311986	<i>Shigella sp. D9</i>	<u>11</u> , <u>18</u> , <u>19</u>

*The Cys positions that are in bold form disulfide bond while the underlined ones are removed with signal peptide during post translational processing. Residue numbers are according to their positions in precursor sequence (with signal and spacer peptide). ** PGA sequences from *Acinetobacter* genus were more than 94% identical to each other, therefore only one of them (*Acinetobacter oleivorans*; AoPGA) was used in the analysis.

4.3.1. Putative Thermostable PGAs (ptPGAs): Screening

Presence of disulfide bond has been experimentally shown to provide thermostability in case of *Af*PGA. The subfamily S45.001 of MEROPS database consists of several PGA sequences of which *Sw*PGA, *Pd*PGA and *Ao*PGA (NCBI GI numbers 148553843, 119378111 and 299769954 respectively) were selected due to presence of Cys residues at equivalent positions as that of *Af*PGA, which could lead to eventual formation of disulfide bond and possibly contribute towards thermostability of the enzymes (Fig. 4.2). These three PGAs would be referred together as **ptPGAs (putative thermostable PGAs)**. Although other PGA sequences in the MEROPS database were found to contain Cys residues, these were mainly found to occur in the signal peptide which would be cleaved during the post-translational processing stage (Table 4.1).



Figure 4.2: Sequence alignment showing 34 residue regions involved in disulfide bond formation in *Af*PGA. The Cys residues are observed to be conserved among ptPGAs while *Ec*PGA and *Ax*PGA have variable residues.

4.3.2 Removal of signal and spacer peptide

The sequence renumbering of the mature form of PGA enzymes after removal of signal and spacer peptide are given in Table 4.2. In case of *Ec*PGA and *Af*PGA, for which three-dimensional structures of matured forms are available, the residue numbering is considered as assigned in their respective PDB files, 1GK9 and 3K3W, respectively. However, in case of *Ax*PGA, *Sw*PGA, *Pd*PGA and *Ao*PGA, since no experimental data are available for the mature

enzyme forms, the signal and spacer peptides have been predicted by comparing with mature *Ec*PGA and *Af*PGA structures. For example in precursor *Ax*PGA, the regions 1-20, 21-231, 232-286 and 287-843 correspond to signal peptide, chain α , spacer peptide and chain β , respectively. Thus, in mature *Ax*PGA, the chain α (1-211) and chain β (1-557), corresponds to regions 21-231 and 287-843, respectively in precursor *Ax*PGA.

Table 4.2: The residue renumbering after the removal of signal and spacer peptide among the PGA enzymes under study.

PGA	Chain α		Chain β	
	precursor	Mature form	precursor	Mature form
<i>Ec</i> PGA	27-235	1-209	290-846	1-557
<i>Af</i> PGA	27-222	1-196	266-816	1-551
<i>Ax</i> PGA	21-231	1-211	287-843	1-557
<i>Sw</i> PGA	44-234	1-191	278-825	1-548
<i>Pd</i> PGA	25-213	1-189	251-790	1-540
<i>Ao</i> PGA	38-224	1-187	274-820	1-547

Table 4.3: Evaluation statistics for molecular models of *Ax*PGA, *Sw*PGA, *Pd*PGA and *Ao*PGA sequences by various model validation tools.

PGA	PROCHECK					Errat score	Verify3 D ^a	Prosa Z-score, chain α	Prosa Z-score chain β
	Ramachandran plot statistics in percentage				G-factor				
	Core	Allowed	Generously allowed	Disallowed					
<i>Ax</i> PGA	90.3	9.3	0.3	0.2	-0.26	89.72	100	-5.25	-9.29
<i>Sw</i> PGA	89.1	10.4	0.2	0.3	-0.28	90.58	100	-4.61	-8.86
<i>Pd</i> PGA	89.6	9.8	0.5	0.2	-0.28	92.57	99.04	-6.13	-8.15
<i>Ao</i> PGA	88.9	9.7	0.9	0.5	-0.29	88.30	100	-6.03	-9.12

^a Percentage of residues with positive Verify3D score. Positive score represents a good quality residue.

4.3.3 Homology modeling and model validation

Due to absence of any three-dimensional structures for the *Ax*PGA and *pt*PGAs, 3D homology models were built using *Ec*PGA structure as template. Sequence comparison revealed that α -chains of target sequences were more than 41% identical to *Ec*PGA and while β -chains were more than 34% identical. The key residues involved in catalysis such as β Ser1, β Gln23, β Ala69 and β Asn241 of *Ec*PGA are well conserved among all PGAs. Similarly the residues involved in binding of substrate penicillin G namely, α Phe146, β Ile177, β Pro49 and β Trp154 were also found to be conserved in all enzymes. Other residues in the penicillin side chain-binding pocket such as β Val56, β Thr32 and β Phe24 were observed to be partially conserved. Thus the tertiary structure models built for *Ax*PGA and *pt*PGAs were observed to be of high quality as validated by model validation programs (Table 4.3).

Ramachandran plot calculated using PROCHECK estimated that more than 99% of the residues of each model lie in the core and allowed regions. Errat program evaluates the quality of a model by analyzing the statistics of non-bonded atom-atom interactions. All models were having overall quality factor >88%, close to the value for a high quality structure (Table 4.3). The Verify3D analysis of the models found no conformational error and more than 99% of residues of every model were having score above zero. The overall Z-score using Prosa were within the range of scores usually observed for native proteins of similar size, further validating the quality of modeled structures. Thus, the validation analysis confirmed the acceptable quality of all models. Also, each of the model structures superposed very well with the template structure *Ec*PGA. The overall RMSD of all atom positions between the models and the template calculated using iterative magic fit tool of SPDBV (Guex & Peitsch, 1997) were 0.19, 0.19, 0.25, 0.21 Å, respectively, for *Ax*PGA, *Sw*PGA, *Pd*PGA and *Ao*PGA. Figure 4.3 illustrates sample model validation result for *Pd*PGA model.

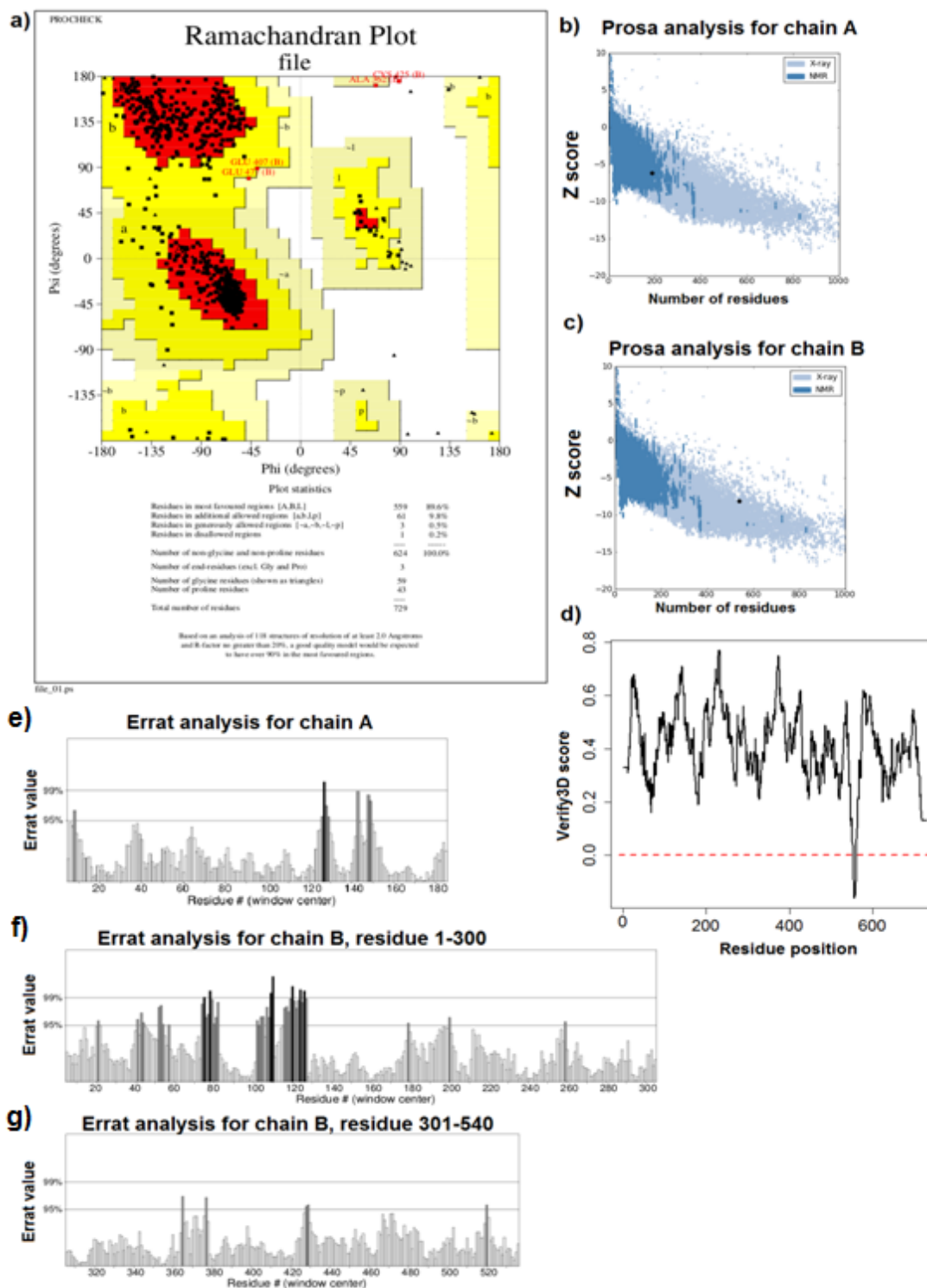


Figure 4.3: Illustrates the results of model validation programs such as Procheck (a), Prosa (b-c), Verify3d (d) and Errat (e-g), for *PdPGA* homology model.

Table 4.4: List of 11 experimentally characterized mutations known to enhance thermostability in *Ec*PGA. The amino acid residues present at these 11 sites amongst other PGAs are listed.

Stabilizing mutations carried out in <i>Ec</i> PGA (Polizzi <i>et al.</i> , 2006)	Corresponding residues present in other PGAs				
	<i>Ax</i> PGA	<i>Aj</i> PGA	<i>Sw</i> PGA	<i>Pd</i> PGA	<i>Ao</i> PGA
$\beta 84A \rightarrow P$	P	P	P	P	P
$\beta 400V \rightarrow L$	L	L	L	L	L
$\alpha 25W \rightarrow Y$	Y	Y	Y	H	Y
$\beta 359V \rightarrow L$	L	L	V	L	L
$\alpha 150T \rightarrow N$	N	N	N	S	S
$\beta 311T \rightarrow P / \beta 312Q \rightarrow A$	P/A	P/A	P/A	D/P	N/A
$\beta 100L \rightarrow E$	E	E	D	T	E
$\alpha 80A \rightarrow R$	R	R	A	S	Q
$\beta 305A \rightarrow S$	R	K	S	D	Q
$\beta 348N \rightarrow D$	D	K	A	Q	R
Total number of conserved stabilizing mutations	9	8	5	3	5

4.3.4 Sequence-based consensus approach for thermostability analysis

Rational protein engineering or directed evolution methods have shown that careful selection of residues for site-directed mutagenesis could result in an enhancement of thermostability among mesostable enzymes (Blundell *et al.*, 1989). In case of larger proteins such as PGA, identification of potential thermostabilization sites by these methods remains a challenging problem due to wide range of possible substitutions. The consensus method on the other hand has been proven to be successful in such cases where the potential mutation sites are identified solely based on sequence comparison (Lehmann *et al.*, 2002).

Polizzi *et al.*, 2006, have combined the sequence-based consensus approach with the structural information by examining 21 amino acid positions in *Ec*PGA in an effort to increase its thermostability by site directed mutagenesis experiments (Polizzi *et al.*, 2006). Approximately 50% (11 mutations) mutations were found to show positive effect towards *Ec*PGA

thermostability (Table 4.4). These 11 mutation positions upon analysis in *Ax*PGA, *Af*PGA and *pt*PGAs revealed the natural presence of some of these stabilizing mutations (Table 4.4). Of the 11, a total of 9 stabilizing mutations were observed to be conserved in *Ax*PGA (most thermostable) while *Af*PGA, *Sw*PGA, *Pd*PGA, and *Ao*PGA showed 8, 5, 3 and 5 conserved mutations respectively. Natural occurrence of these stabilizing amino acid residues in *pt*PGAs hinted, these *pt*PGAs might have higher thermostability as compared to *Ec*PGA but lower compared to that of *Ax*PGA and *Af*PGA.

In an effort to identify more such stabilizing sites, a modified consensus approach was designed which involved position-wise comparison of mesostable PGAs with their thermostable homologues. Mesostable *Ec*PGA and *Kc*PGA (PGA from *Kluyvera cryocrescens*; Genbank accession AID61747) sequences were aligned with thermostable *Af*PGA and *Ax*PGA using multiple sequence alignment. It was assumed that a residue position conserved among all thermostable enzymes (*Af*PGA and *Ax*PGA) while variable among mesostable homologs (*Ec*PGA and *Kc*PGA) could be considered as potential site selected by nature for protein thermostabilization. For example, Ala was found at α 80 position and Leu at β 100 in both mesostable *Ec*PGA and *Kc*PGA. However, in case of the thermostable *Af*PGA and *Ax*PGA, at this position charged residues Arg and Glu were found to be conserved respectively since these residues are better suited for higher temperatures (Table 4.5). Similarly at β 311 and β 312 positions both thermostable *Af*PGA and *Ax*PGA contains the conserved Pro and Ala residues, respectively, whereas the mesostable *Kc*PGA possesses Ala and Glu residues and *Ec*PGA was observed to have Thr and Gln at the corresponding positions (Table 4.5). Mutation experiments carried out by Polizzi *et al.*, 2006, in *Ec*PGA showed that the α 80A \rightarrow R substitution resulted in a 2.7 fold increase of half-life of *Ec*PGA at 50 °C, and the β 100L \rightarrow Glu position, where uncharged non-polar residue Leu was substituted with negatively charged residue Glu, showed 1.2 fold increase in half-life at 50 °C. Similarly, the *Ec*PGA double mutant (β 311T \rightarrow P/ β 312Q \rightarrow A) was observed to enhance the enzyme half-life to nearly 2 fold.

Table 4.5: List of 24 sites identified as thermostabilization sites by *sequence-based consensus* approach. The residues at these 24 sites among the putative thermostable PGAs are listed.

<i>Ec</i> PGA Resno*	Mesostable		Thermostable		Putative Thermostable**		
	<i>Ec</i> PGA	<i>Kc</i> PGA	<i>Af</i> PGA	<i>Ax</i> PGA	<i>Pd</i> PGA	<i>Sw</i> PGA	<i>Ao</i> PGA
α 79	R	R	Q	Q	R	A	K
α80	A	A	R	R	<i>S</i>	A	<i>Q</i>
α 108	N	N	R	R	L	L	L
α 121	T	T	D	D	D	E	D
β 98	K	K	T	T	E	Q	P
β100	L	L	E	E	T	<i>D</i>	E
β 112	Q	Q	A	A	P	<i>S</i>	A
β 129	T	T	F	F	I	F	W
β 133	T	T	Q	Q	T	Q	N
β 218	K	K	L	L	<i>Q</i>	L	<i>Q</i>
β 234	S	S	Q	Q	Q	<i>G</i>	K
β 280	D	D	Q	Q	<i>D</i>	D	Q
β 308	S	A	Q	Q	<i>E</i>	<i>G</i>	Q
β311	T	A	P	P	D	P	<i>N</i>
β312	Q	E	A	A	P	A	A
β 313	S	N	D	D	Q	<i>H</i>	S
β 337	K	K	G	G	R	-	K
β 404	K	K	A	A	R	Q	Q
β 432	E	Q	A	A	A	A	<i>T</i>
β 436	K	K	Q	Q	<i>D</i>	A	K
β 443	S	T	A	A	A	D	N
β 457	N	N	K	K	<i>S</i>	L	K
β 519	K	K	P	P	P	P	<i>V</i>
β 544	E	D	R	R	E	E	<i>D</i>
Total number of sites where different residues are present than observed in thermostable <i>Af</i>PGA/<i>Ax</i>PGA					19	17	17
Total potential stabilizing sites predicted to be thermostabilizing by FoldX stability prediction analysis using iStability module of iRDP web server.					6	8	6

*Resno corresponds to residue number. The residue positions in bold correspond to the sites for which experimental evidence is available as stabilization sites. **The bold and italic residues are the sites having higher potential of thermostabilization as identified by FoldX stability prediction analysis. The hyphen in *Sw*PGA column represents gap in the sequence alignment.

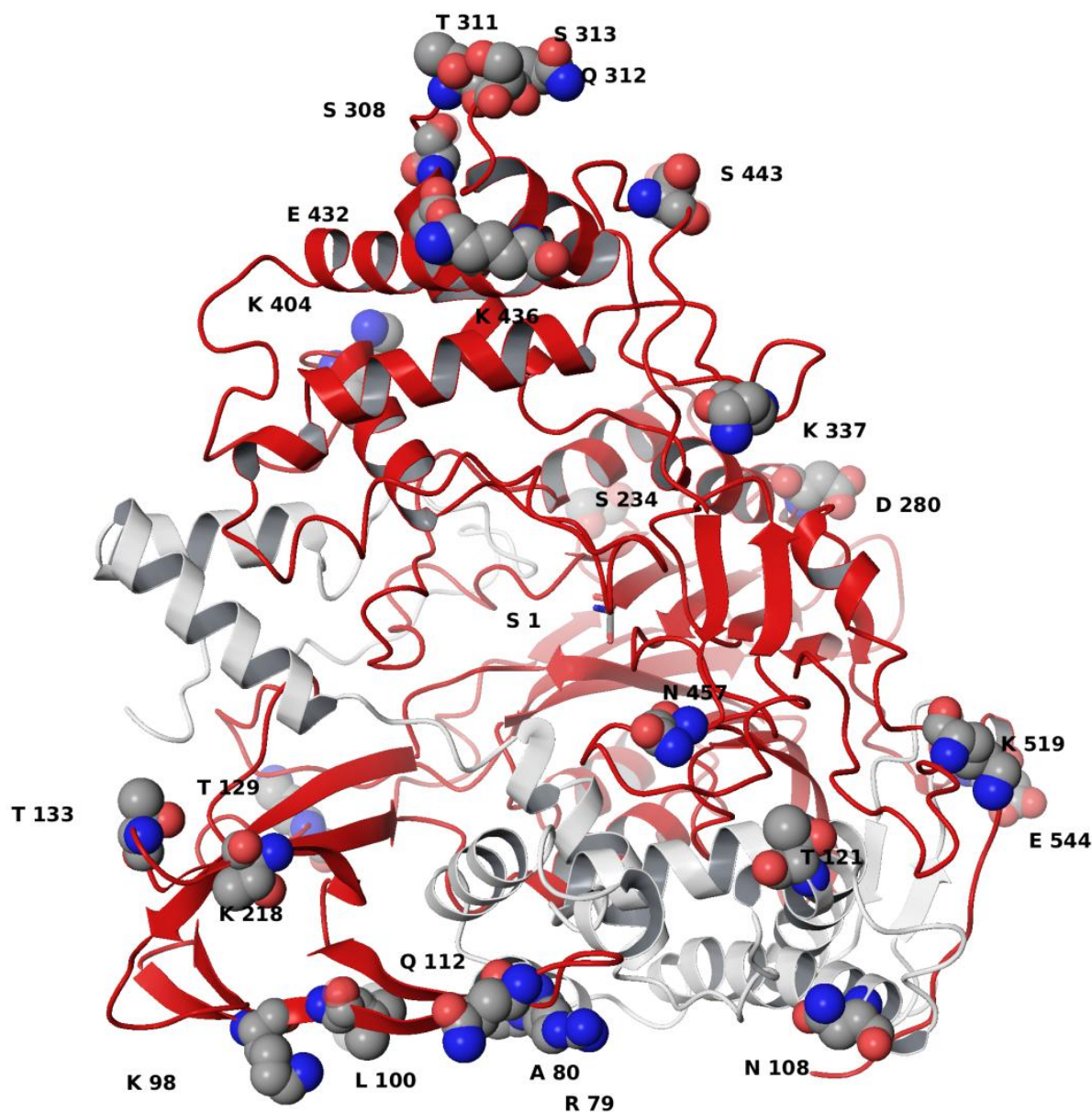


Figure 4.4: The location of 24 sites on three-dimensional structure of *EcPGA*. The residues are shown in CPK format and labeled. The active site residue β S1 is represented in stick. The chain α and β are colored in grey and red, respectively.

A total of 24 stabilization sites (Table 4.5) were identified at which mesostable *EcPGA* and *KcPGA* have different residues compared to the conserved residue in thermostable *AfPGA* and *AxPGA*. Sites were selected such that the residues are surface-exposed and lie outside a 10 Å radius cut-off from the active site so that substitution of these residues would have minimum effect on the catalytic efficiency of enzymes (Fig. 4.4). Based on residue preference at these 24

sites, ptPGAs were assigned into either mesostable or thermostable groups. Analysis revealed that 19 sites in *Pd*PGA and 17 in each case of *Sw*PGA and *Ao*PGA were found to possess different residues than observed in case of thermostable *Ax*PGA/*Af*PGA (Table 4.5). At α 80 position, amongst ptPGAs, *Sw*PGA was found to contain a non-polar Ala while *Pd*PGA and *Ao*PGA contain uncharged-polar Ser and Gln respectively in lieu of the positively charged Arg residues found in *Af*PGA/*Ax*PGA. Similarly at β 100 position ptPGAs possess residues different from that observed in *Ax*PGA (Table 4.5). At positions β 311/ β 312, only *Sw*PGA was found to maintain residues similar to *Af*PGA/*Ax*PGA (P/A) while others have different residues (D/P in *Pd*PGA and N/A in *Ao*PGA). Thus more than 70% of the sites in ptPGAs behave similar to mesostable enzymes.

In order to further filter highly potential sites amongst these 24 sites, each site was analyzed for its effect on enzyme's stability from a structural point of view. For the *Af*PGA-*Sw*PGA pair of enzymes, the stabilizing residues present in *Af*PGA at the 24 sites, were introduced in *Sw*PGA and similarly the reverse substitutions were carried out in *Af*PGA by introducing the residues of *Sw*PGA. The sites where a change may lead to an enhancement of thermostability in *Sw*PGA and the reverse mutations at same positions in *Af*PGA would result in its destabilization, were considered as sites on which substitution has higher chance of enhancing thermostability. The mutation effect on protein stability was estimated using iStability module with FoldX as stability prediction tool which not only scores for favorable interactions but also penalizes for unfavorable clashes. Similar analyses were also carried out for *Af*PGA-*Pd*PGA and *Af*PGA-*Ao*PGA pairs of enzymes. Of the 24 sites, a total of 8, 6 and 6 sites, respectively, in case of *Sw*PGA, *Pd*PGA and *Ao*PGA were filtered, which showed higher potential for thermostabilization (Table 4.5). For example, at site corresponding to α 80 position of *Ec*PGA, Ser was substituted with Arg in *Sw*PGA while the reverse i.e. Arg was substituted with Ser in *Af*PGA. Stability prediction showed that Ser \rightarrow Arg mutation in *Sw*PGA enhances its stability while Arg \rightarrow Ser mutation in *Af*PGA destabilizes *Af*PGA. Thus it hints at higher preference of Arg residue compared to Ser at α 80 site towards thermostabilization. Interestingly, of the 24 sites, the α 80 site was found to be the most potent site which upon mutation to Arg could increase thermostability in all ptPGAs; and this has also been shown as the prime site for thermostabilization by Polizzi *et al.*, 2006. Other sites corresponding to β 308 and β 436 in *Ec*PGA were also found to be important since substitutions at these positions were also shown to

enhance stability in at least two ptPGAs (Table 4.5). Overall the stability prediction analysis helped to narrow down the choices for potential substitution sites, aimed at enhancement of thermostability, to 3 possible positions for the selected ptPGAs.

4.3.5 Structure-based approach of thermostability analysis

In this structure-based approach the modeled structures of *Sw*PGA, *Pd*PGA and *Ao*PGA were compared with crystal structure of *Ec*PGA (Less thermostable), *Af*PGA (Moderately thermostable) and the modeled structure of *Ax*PGA (Most thermostable) in terms of several factors known to influence thermodynamic stability of proteins such as disulfide bridges, ion-pairs, hydrogen bonds, aromatic-aromatic, aromatic-sulphur and cation-pi interactions, entropic stabilization due to proline residues and reduction of deamidation damages. The effect of temperature on structural stability was monitored for all the six PGAs by subjecting them to molecular dynamics simulations at 330K, 400K and 500K temperatures. The average RMSD values of C α atoms of the enzymes during all the three temperature-dependent simulations were monitored (Table 4.6). Lower average RMSD with lower standard deviation values at 330K and 400K suggest that the systems were stable at these temperatures. However, the deviations were observed to be higher at 500K compared to 400K and 330K suggesting structural instability at higher temperatures. At every 10 ps intervals, enzyme conformations were extracted and above mentioned structural parameters were monitored. Next described is the comparative analysis of six PGA enzymes.

Table 4.6: The average and standard deviations of the RMSD (Root Mean Square Deviations) of C α atoms of PGA enzyme conformations during the production phase of molecular dynamics simulations (330, 400 and 500K).

PGA	C α RMSD Values (nm) (Mean \pm Standard deviation)		
	330K	400K	500K
<i>Ec</i> PGA	0.14 \pm 0.01	0.23 \pm 0.07	0.69 \pm 0.17
<i>Af</i> PGA	0.22 \pm 0.03	0.28 \pm 0.03	0.88 \pm 0.21
<i>Ax</i> PGA	0.2 \pm 0.01	0.26 \pm 0.03	0.66 \pm 0.14
<i>Sw</i> PGA	0.2 \pm 0.01	0.34 \pm 0.03	0.78 \pm 0.17
<i>Pd</i> PGA	0.21 \pm 0.02	0.3 \pm 0.04	0.83 \pm 0.19
<i>Ao</i> PGA	0.21 \pm 0.01	0.32 \pm 0.05	0.61 \pm 0.1

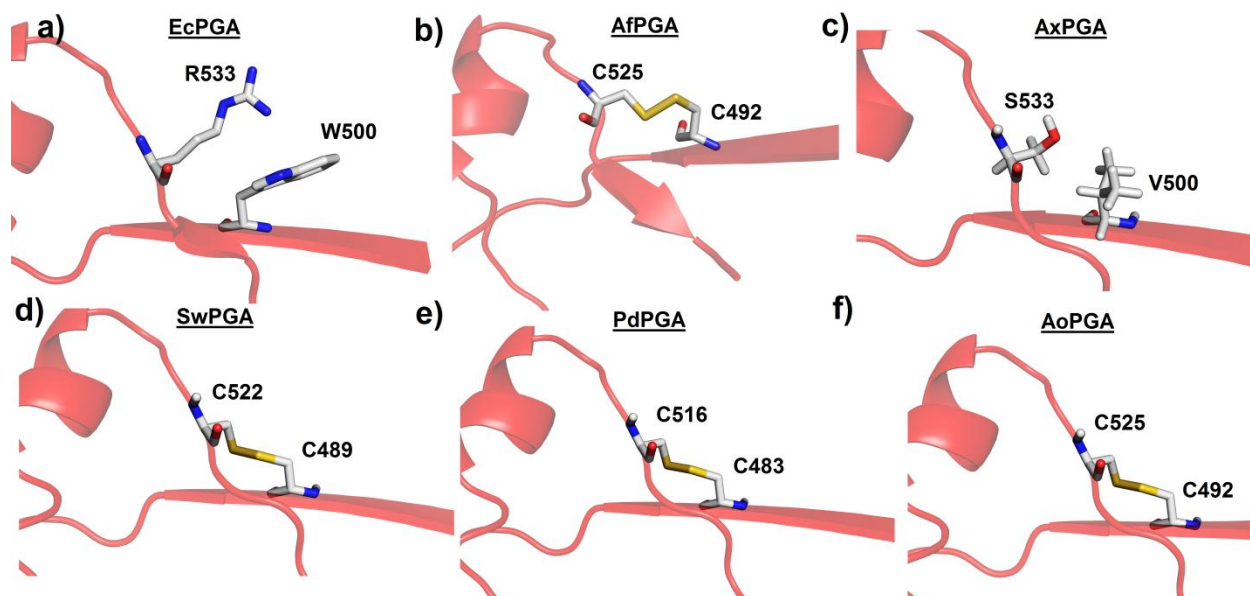


Figure 4.5: (a-f) The residues corresponding to the disulfide bridge forming Cys residue positions of AfPGA is shown among all PGA enzymes under study.

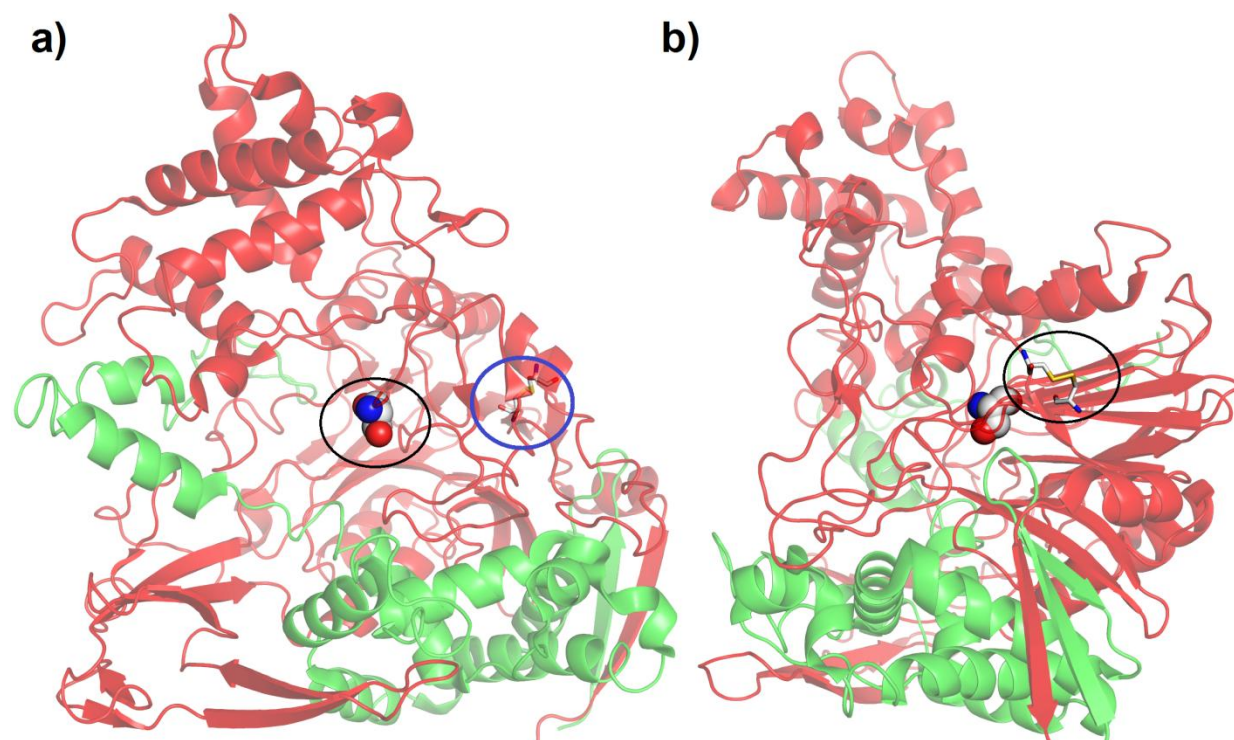


Figure 4.6: (a) AfPGA (PDB ID: 3K3W) structure showing the N-terminal nucleophilic β Ser1 residue in spheres (encircled in black) and the disulfide bond (encircled in blue). (b) AfPGA structure rotated

counterclockwise by 90° about y-axes to illustrate the disulfide bond linking the layers of $\alpha\beta\beta\alpha$ Ntn-hydrolase fold.

4.3.5.1 Stabilization by Disulfide Bridges

Disulfide bridges are known to stabilize protein structure by reducing the conformational entropy of its unfolded state (Matsumura *et al.*, 1989). *Af*PGA contains one disulfide bond between residues β Cys492 and β Cys525 in its structure (Varshney *et al.*, 2012). Verhaert *et al.*, 1997, have shown that the reduction of this disulfide bridge with 10 mM DTT resulted in decrease in stability of *Af*PGA. Both *Ec*PGA and *Ax*PGA lack this disulfide bridge; in the place of residues β Cys492 and β Cys525 of *Af*PGA, the residues β Trp500 and β Arg533 were present in *Ec*PGA while *Ax*PGA had β Val500 and β Ser533 in those positions. The modeled structures of ptPGAs showed their potential to form disulfide bond between a pair of cysteine residues of β -chain (*Sw*PGA: β Cys489- β Cys522, *Pd*PGA: β Cys483- β Cys516 and *Ao*PGA: β Cys492- β Cys525; Fig. 4.5). Previous studies have shown that the contribution of entropy to the free energy of stabilization increases proportionately with the number of residues separating the two cysteine residues of the disulfide bond. The number of residues separating the two Cys residues in all ptPGAs is same as that in *Af*PGA (32 residues), suggesting a similar entropic effect. Since the two Cys residues connect the layers of $\alpha\beta\beta\alpha$ Ntn-hydrolase fold (Fig. 4.6), the possible role of disulfide bond could be to maintain the structural integrity of the Ntn-hydrolase fold. Thus presence of disulfide bridges in ptPGAs is expected to contribute positively to their thermostability, as seen in case of *Af*PGA.

4.3.5.2 Preference of Arg over Lys: A mechanism for thermostabilization

*Ax*PGA has been observed to be more thermostable than *Af*PGA though it lacks a disulphide bond suggesting the higher influence of other factors towards its thermostability. Preference of Arg over Lys residue and higher number of ion-pairs are proposed to be contributing factors of thermostability in *Ax*PGA (Cai *et al.*, 2004). Arg residues are better favored for higher temperatures than Lys due to resonance stabilization of their side chain along with higher surface area for charged interactions. Thermostable proteins are known to maintain their Arg/Lys ratio greater than one (Vieille & Zeikus, 2001). Although the total content of Arg and Lys residues together was found to be almost equal across all six PGAs (Table 4.7), Arg/Lys ratios of individual enzymes were found to differ considerably, correlating with the thermostable

nature. While in less thermostable *Ec*PGA, the ratio was observed to be 0.78 (Lys preferred over Arg), the ratio is changed to 1.6 (Arg preferred over Lys) in moderately-thermostable *Af*PGA and in case of most-thermostable *Ax*PGA, the ratio was found to be even higher (2.29). Among the selected ptPGA enzymes, *Sw*PGA and *Pd*PGA have Arg/Lys ratio 1.26 and 3.23, respectively suggesting an increase in thermal stability while in *Ao*PGA, this ratio was found to be the least (0.38). Thus the overall analysis suggested that Arg/Lys ratio contributed maximally towards thermostability in case of the enzyme *Pd*PGA.

Table 4.7: Amino acid composition of six PGA enzymes under study estimated using iCAPS module of iRDP web server.

Amino acid	<i>Ec</i> PGA	<i>Af</i> PGA	<i>Ax</i> PGA	<i>Sw</i> PGA	<i>Pd</i> PGA	<i>Ao</i> PGA
Length	765	747	768	739	729	734
Val	6.54	6.56	5.99	7.17	6.17	5.18
Ile	3.92	3.08	3.13	4.33	5.08	5.18
Leu	7.32	6.43	7.16	6.63	6.45	8.45
Met	2.35	3.48	2.87	2.30	1.78	1.23
Phe	3.92	4.42	4.69	4.06	4.80	3.41
Trp	3.66	2.68	2.87	2.44	2.74	2.73
Tyr	4.05	4.95	4.17	4.47	3.02	4.91
Ser	5.49	5.49	3.78	5.55	3.98	7.36
Thr	7.19	5.62	5.47	6.23	5.21	5.59
Asn	5.49	5.09	4.30	4.06	4.66	5.86
Gln	6.41	7.36	4.69	2.98	4.39	6.54
His	1.70	1.74	1.95	1.89	1.65	1.77
Lys	5.36	3.75	3.13	4.60	2.33	7.08
Arg	4.18	6.02	7.16	5.82	7.55	2.73
Asp	6.28	6.29	7.16	7.44	7.13	5.59
Glu	4.44	5.09	3.26	4.20	7.41	4.77
Ala	9.02	8.70	13.15	10.69	11.25	9.26
Gly	7.45	6.96	8.59	8.93	8.09	6.40
Pro	5.23	6.02	6.51	5.95	5.90	5.72
Cys	0.00	0.27	0.00	0.27	0.41	0.27

Unnatural	0.00	0.00	0.00	0.00	0.00	0.00
Aromatic	11.63	12.05	11.72	10.96	10.56	11.04
Uncharged Polar	24.58	23.56	18.23	18.81	18.24	25.34
Positive	9.54	9.77	10.29	10.42	9.88	9.81
Negative	10.72	11.38	10.42	11.64	14.54	10.35
Arg/Lys ratio	0.78	1.61	2.29	1.27	3.24	0.39

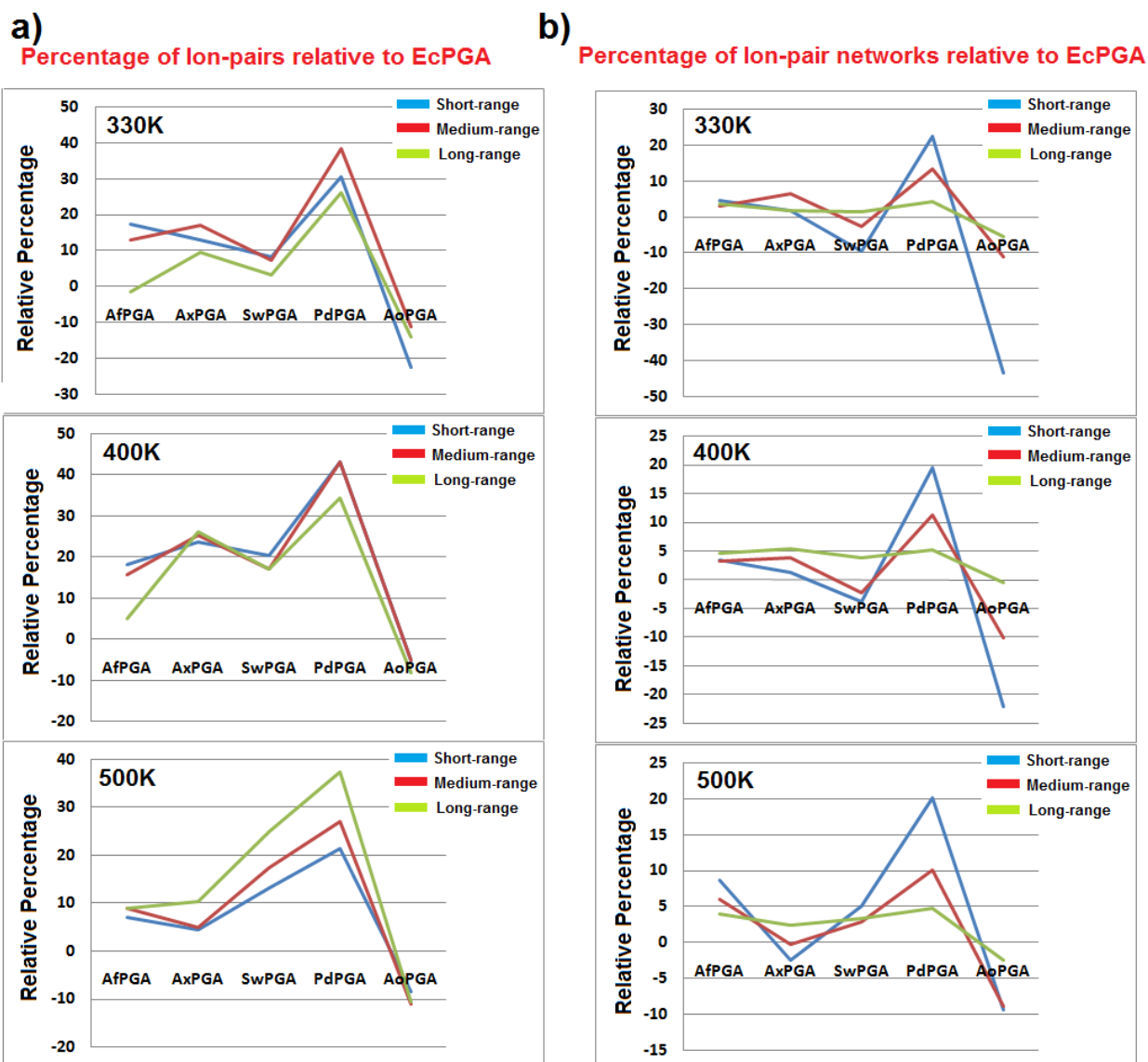


Figure 4.7: Line plots depicting the average percentages of ion-pairs (panel a) and ion-pair networks (panel b) of AfPGA, AxPGA, SwPGA, PdPGA and AoPGA relative to EcPGA (Relative percentage values in y-axes), during 330K (top panel), 400K (middle panel) and 500K (bottom panel) of molecular dynamics simulations.

The blue, red and green lines in the plots correspond to *short-range*, *medium-range* and *long-range* electrostatics interactions.

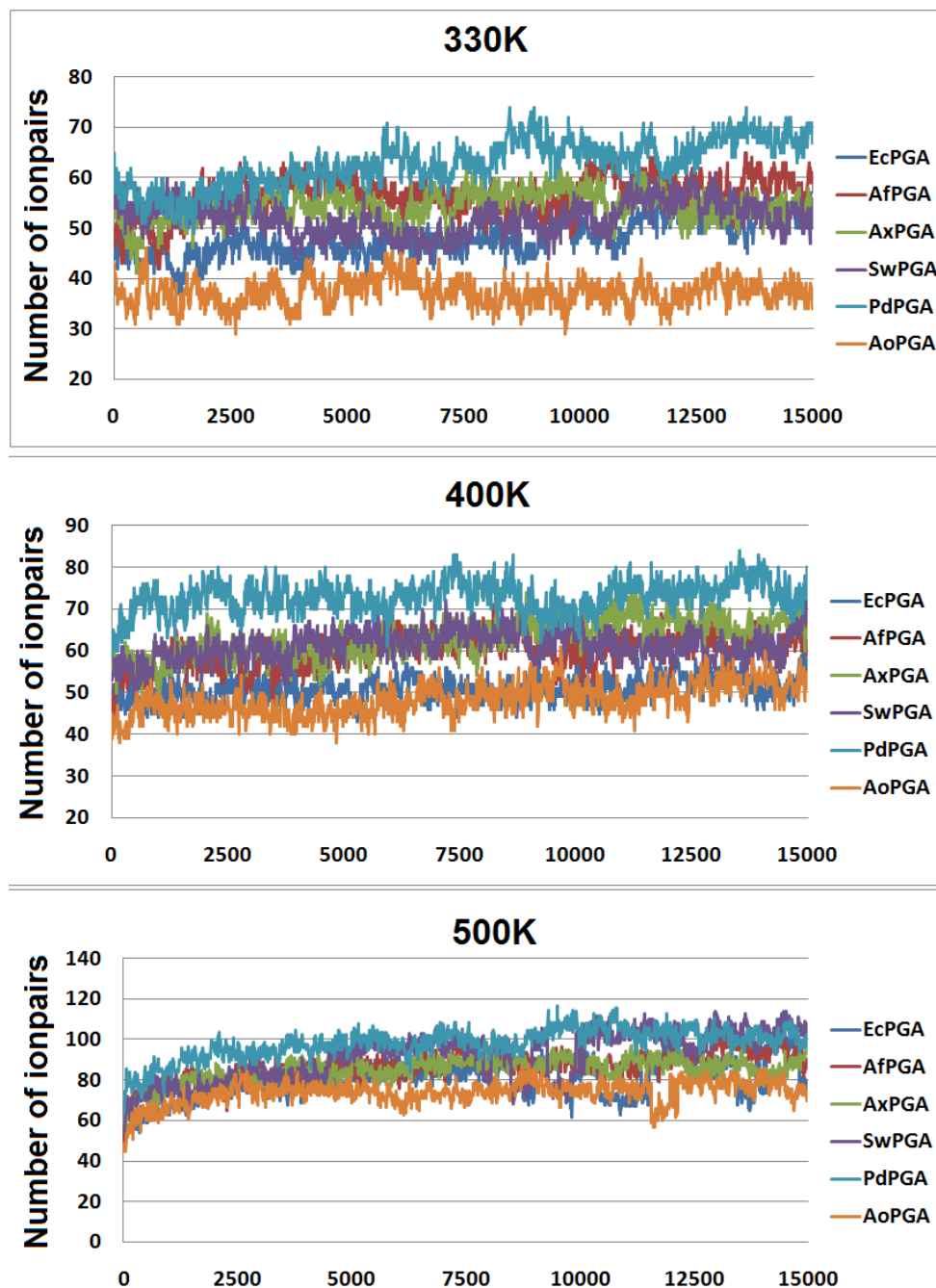


Figure 4.8: Illustrates the time evolution of number of *short-range* ion-pairs amongst six PGAs at 330K, 400K and 500K molecular dynamics simulations.

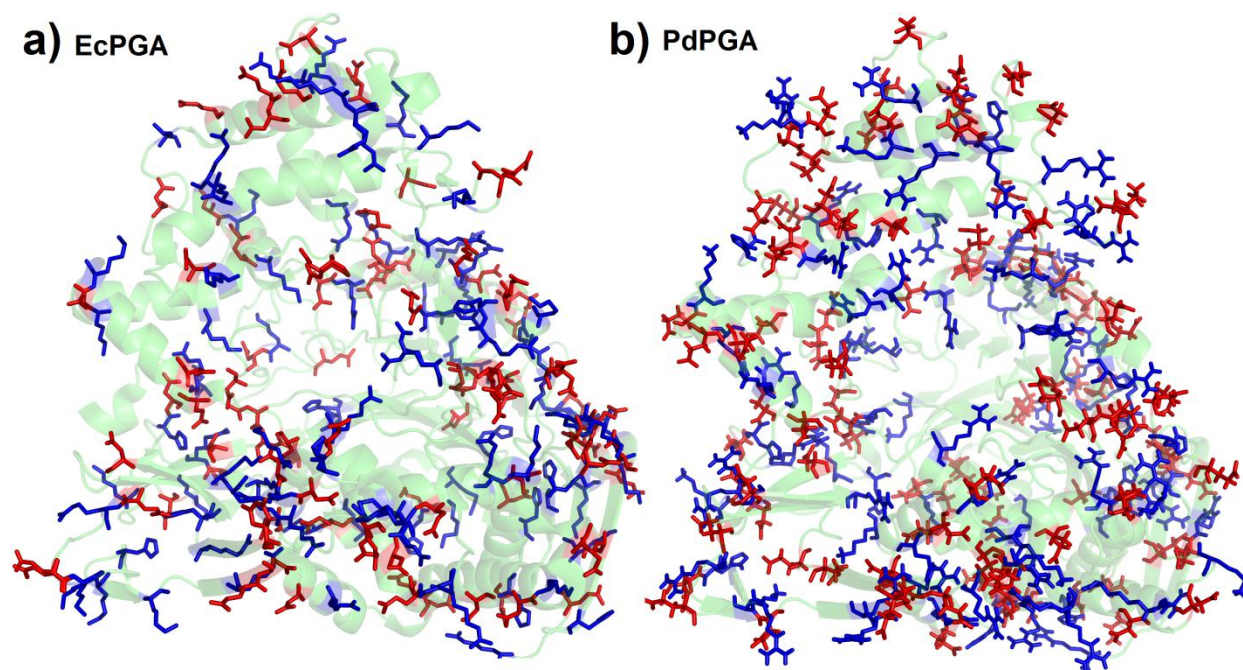


Figure 4.9: The three-dimensional structures of *EcPGA* and *PdPGA* showing the ionic interactions between the acidic (red) and basic (blue) residues. Higher number of ionic interactions in *PdPGA* is observed compared to *EcPGA*.

4.3.5.3 Presence of higher ion-pair networks: A stabilizing mechanism

Thermostable enzymes *AfPGA* and *AxPGA* were found to maintain comparatively higher percentage of ion-pairs and ion-pair networks in their tertiary structures than mesostable *EcPGA*. During the course of simulations at 330K, 400K and 500K, at every 10ps interval, enzyme conformations were extracted and total number of ion-pairs (**IPs**) and percentage of ion-pairs that are involved in network formation (**IPnets**) were estimated. The average values of number of IPs and percentage of IP-networks of *EcPGA* was considered as reference and relative percentages were calculated for other PGAs with respect to *EcPGA*. A relative percentage value > 0 shows higher percentage of IPs and IPnets as compared to *EcPGA*. Three ranges of electrostatics interactions were estimated, that is: *short-range*, *medium-range* and *long-range*, based on the distance-cutoffs between the acidic and basic residue side chains as 4, 6 and 8 Å, respectively.

Figure 4.7 shows the relative percentage values of IPs and IPnets of *AfPGA*, *AxPGA*, *SwPGA*, *PdPGA* and *AoPGA* compared to *EcPGA* monitored at 330K, 400K and 500K of

simulations. At 330K, both thermostable *Af*PGA and *Ax*PGA were found to contain higher percentage (17.48 and 13.03% higher respectively) of *short-range* IPs compared to *Ec*PGA. Similar observations were recorded for *Sw*PGA and *Pd*PGA which maintained 8.27 and 30.61% more *short-range* IPs as compared to *Ec*PGA. *Ao*PGA was found to contain the least percentage of *short-range* IPs (22.53% less as compared to *Ec*PGA) amongst all PGAs. At 400 and 500K, *Pd*PGA was observed to dominate all PGAs in terms of *short-range* IPs (Fig. 4.8). Similarly, *Pd*PGA was also observed to dominantly maintain *medium-* and *long-range* IPs at 330K, 400K and 500K.

Since interaction networks are comparatively stable and energetically more favorable than isolated interactions, PGA enzymes were also compared in terms of percentage of occurrence of ion-pair networks (IPnets). At 330K, compared to *Ec*PGA, thermostable *Af*PGA and *Ax*PGA were found to maintain slightly higher percentage (4.7 and 1.6%, respectively) of *short-range* IPnets. While among the ptPGAs, *Sw*PGA and *Ao*PGA maintained lower percentage (9.42 and 43.35% lower respectively) of IPnets. Notably, *Pd*PGA was found to contain the highest ion-pair network percentage (22.42% higher than *Ec*PGA) amongst all PGA enzymes. Interestingly at 400K and 500K, *Pd*PGA was again found to dominate all PGAs with respect to IPnet values. When *medium-* and *long-range* ionic interactions were monitored, *Pd*PGA consistently dominated all PGAs in terms of IPnet values at 300K, 400K and 500K (Fig. 4.7). The observation of larger number of IPs and IPnets (Fig. 4.9) along-with the possible existence of a disulfide bridge suggests a possibility of higher thermostability in case of *Pd*PGA enzyme.

4.3.5.4 Loop stabilization by proline residues

Proline residues being conformationally rigid are known to provide stability by entropic effect (Watanabe *et al.*, 1994). Many thermophilic and hyperthermophilic proteins have shown to adopt this stabilization mechanism by maintaining higher proline content as compared to their mesophilic homologues. A positive correlation between proline content and thermostability was observed among the PGA enzymes under study. The known thermostable *Af*PGA (6.02%) and *Ax*PGA (6.51%) enzymes were observed to have higher proline content compared to mesostable *Ec*PGA (5.22%). All the selected ptPGA enzymes were found to have higher proline content than *Ec*PGA (*Sw*PGA: 5.95%, *Pd*PGA: 5.89% and *Ao*PGA: 5.72%).

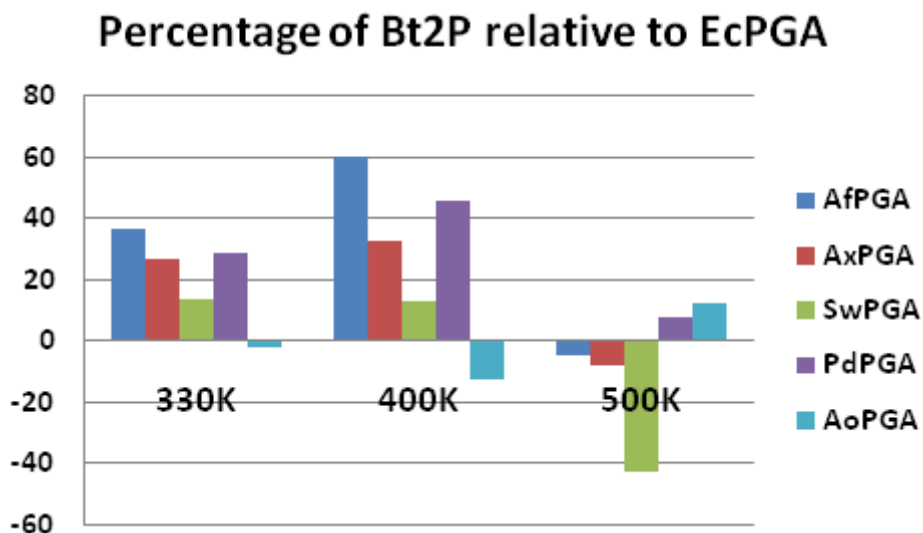


Figure 4.10: Bar plot depicting the average percentages of Bt2P residues among all PGAs relative to *EcPGA* during the 300K, 400K and 500K temperature scales of molecular dynamics simulations.

Introduction of proline residues at 2nd position of β -turns (Bt2P) has been considered as one of the protein engineering strategies for enhancement of protein thermostability (Watanabe *et al.*, 1997). The average number of Bt2P residues among the six PGA enzymes was monitored during all temperature scales of simulation (Fig. 4.10) and relative percentage values were compared. At 330K, thermostable *AfPGA* and *AxPGA* were observed to maintain higher average percentage of Bt2P residues (36.99 and 27.05% higher, respectively) than *EcPGA*. Among the ptPGAs, the behavior of *PdPGA* was found to be similar to that of thermostable PGAs having higher average Bt2P percentages (28.87% higher) than *EcPGA*. *PdPGA* was also found to maintain stable β -turns even at higher temperatures of 400K and 500K suggesting the loop stabilization by proline residues, as another possible contributing factor towards *PdPGA* thermostability.

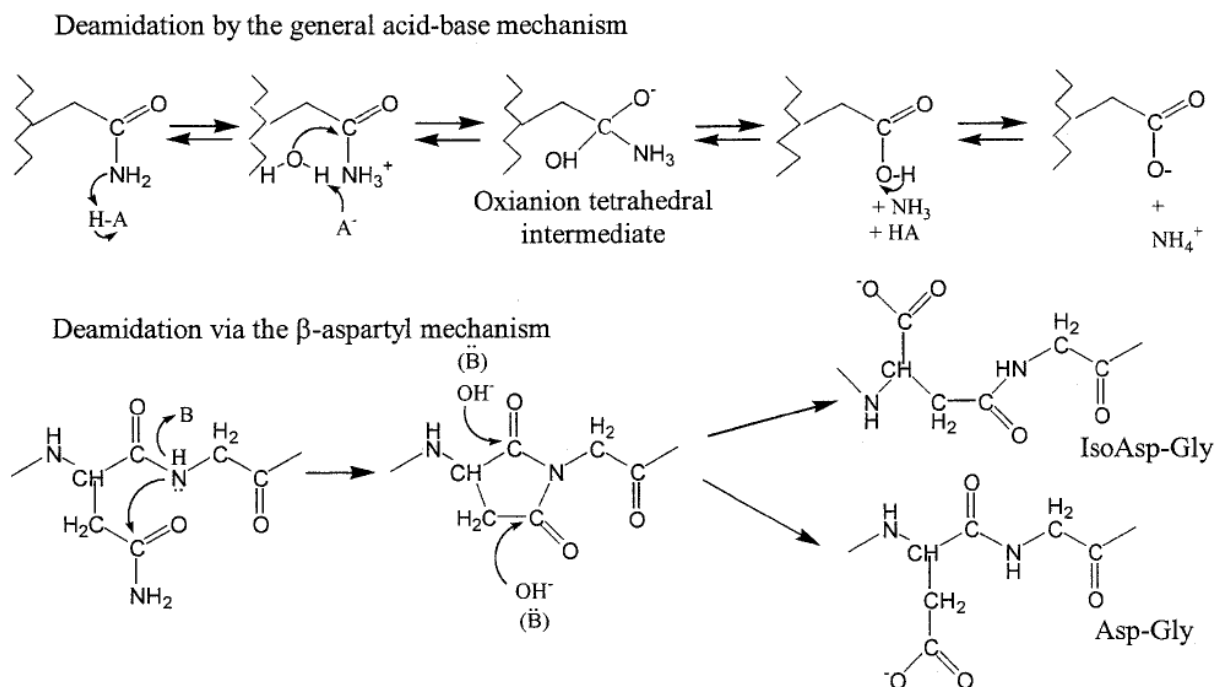


Figure 4.11: The acid-base and β -aspartyl shift mechanism of protein deamidation. Image adapted from (Vieille & Zeikus, 2001)

Table 4.8: List of thermolabile Asn-Gly bond positions in *Ec*PGA and the corresponding amino acids in other PGAs.

<i>Ec</i> PGA position	CD*	Substitutions in Asn-Gly bond in other PGA**				
		<i>Ax</i> PGA	<i>Af</i> PGA	<i>Sw</i> PGA	<i>Pd</i> PGA	<i>Ao</i> PGA
β 20- β 21	1.224	NG	NG	NG	NG	NG
β 60- β 61	1.134	NA	NS	NG	NG	NG
β 93- β 94	4.074	DG	NN	NG	HD	KG
β 110- β 110	0.414	DA	GQ	DG	GQ	NQ
β 185- β 186	3.699	RG	HG	SG	DG	KD

* CD: coefficient of deamidation. ** The favorable substitutions are shown in bold.

4.3.5.5 Decreased content of thermolabile residues

Asparagine (Asn) and Glutamine (Gln) are considered as thermolabile amino acids since they undergo deamidation at higher temperatures. Spontaneous deamidation of Asn leads to formation of aspartic or iso-aspartic residues resulting in important functional and biological damage in peptides and protein structures. In acid-base mechanism of deamidation, residues Ser and Thr are thought to act as acid groups, which protonate the leaving side chain amide group of Asn (Fig. 4.11). Therefore, thermophilic proteins tend to have less uncharged polar amino acid content (NQST content: Asn+Gln+Ser+Thr) to reduce the heat induced damage (Robinson, 2002). Both *Aj*PGA and *Ax*PGA are observed to have lower percentage of NQST residues than *Ec*PGA (23.6, 18.22 and 24.5, respectively; Table 4.7). Among the ptPGAs, *Sw*PGA and *Pd*PGA were found to behave similar to thermostable PGAs by lowering their NQST content as compared to *Ec*PGA (18.8% and 18.24%, respectively). In contrast, *Ao*PGA has highest NQST content among all PGA enzymes (25.3%).

Asn-Gly bonds are considered as thermolabile in nature since they undergo deamidation at higher temperature through β -aspartyl shift mechanism (Robinson, 2002) (Fig. 4.11). Based on the predicted coefficient of deamidation values, *Ec*PGA was found to contain five thermolabile Asn-Gly bonds (Table 4.8). Of these, 4 bonds were observed to be substituted either at Asn or Gly position amongst the thermostable *Aj*PGA and *Ax*PGA. Among ptPGAs, a total of 2, 3 and 3 such bonds are substituted among *Sw*PGA, *Pd*PGA and *Ao*PGA, respectively. The reduction in thermolabile amino acid content and thermolabile bond substitutions could be another stabilization strategy in PGA enzyme family.

4.3.5.6 Contribution from Hydrogen bonds

The contribution of hydrogen bonds towards structural stability had been studied in *RNase T1*, by mutagenesis and unfolding experiments (Shirley *et al.*, 1992). Tanner *et al.*, 1996 have identified a strong correlation between the number of charged-neutral hydrogen bond (CNHB) and thermostability in the case of *GADPH*. Similar correlation is also seen in case of *T. maritime ferredoxin* stability (Macedo-Ribeiro *et al.*, 1996). A CNHB is considered as the interaction between an atom of a charged side chain with atom of either a main chain or side chain of a neutral residue (Tanner *et al.*, 1996). A possible reason for this correlation might be because of the less desolvation penalty to be paid in order to bury a CNHB compared to a

charged-charged hydrogen bond (CCHB). Same way, to bury a CNHB, protein gains higher enthalpy than to bury a neutral-neutral hydrogen bond (NNHB) (Macedo-Ribeiro *et al.*, 1996).

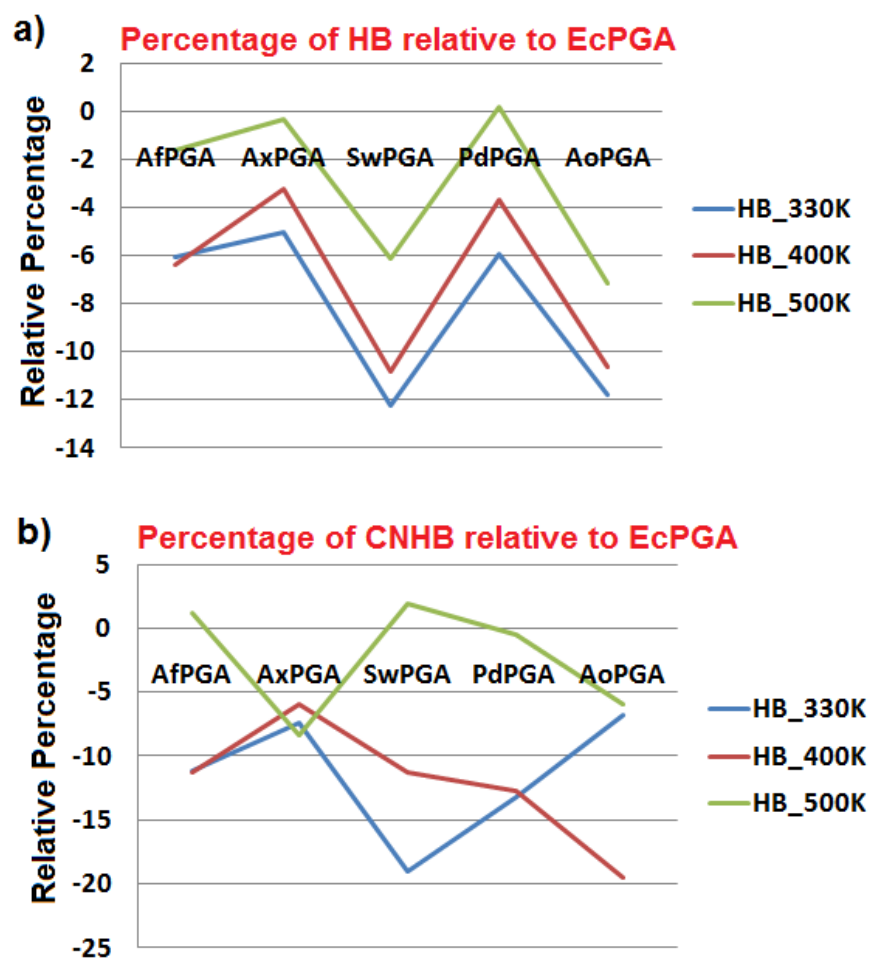


Figure 4.12: The average percentage of total hydrogen bonds (a) and charged-neutral hydrogen bonds (b) among PGA enzymes relative to *EcPGA*. The blue, red and green lines correspond to hydrogen bonding interactions measured at 330K, 400K and 500K of molecular dynamics simulations respectively.

Among the PGA enzymes under study, *EcPGA* was observed to be dominating in terms of both number of hydrogen bonds and percentage of charged-neutral hydrogen bonds (Fig. 4.12). The thermostable *AfPGA* and *AxPGA* were observed to have lesser relative percentages of HB and CNHB compared to *EcPGA*. The observation was consistent at all temperatures of molecular dynamics simulations. The behavior of ptPGAs was observed similar to *AfPGA/AxPGA* having comparatively lower HB and CNHB percentages than *EcPGA*. However, no direct correlation was observed with respect to HB and CNHB percentage values and PGA thermostabilities.

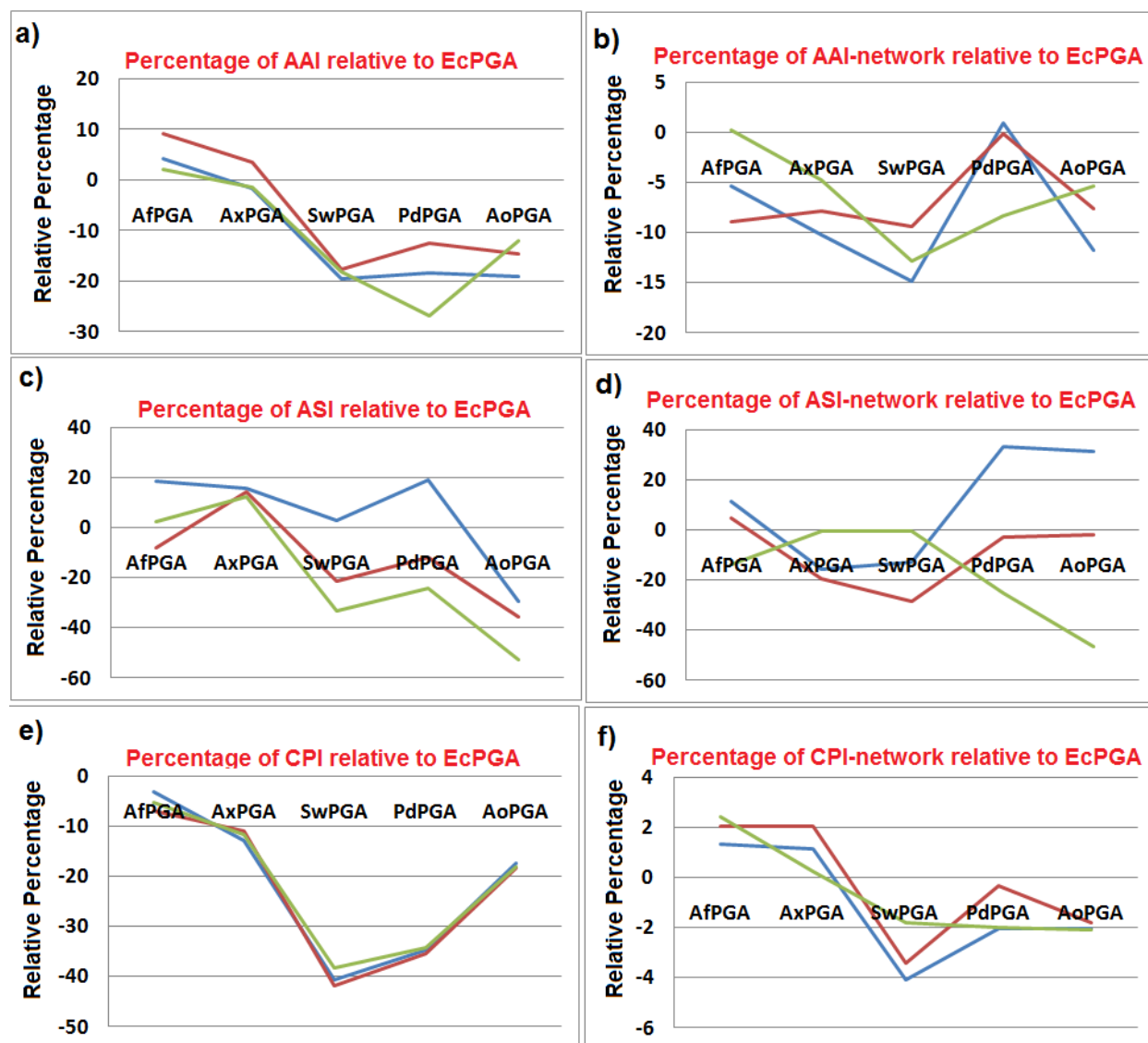


Figure 4.13: Line plots depicting various statistics of interactions involving aromatic residues among the six PGA enzymes during molecular dynamics simulations. Panel a, c and e corresponds to average percentages of Aromatic-aromatic (AAI), Aromatic-sulphur (ASI) and Cation-pi (CPI) interactions of all PGAs relative to *EcPGA* while panel b, d and f corresponds to the average percentages of the respective interaction networks relative to *EcPGA*. The blue, red and green lines correspond to the values during 330, 400 and 500K molecular dynamics simulations.

4.3.5.7 Contribution from interactions involving aromatic residues

It is experimentally shown in an enzyme like *B. amyloliquifaciens* RNase that a typical aromatic-aromatic interaction contributes between -0.6 and -1.3 kcal/mol energy towards protein stability (Serrano *et al.*, 1991). Burley *et al.*, 1985 analyzed 272 aromatic pairs in 34 high

resolution structures of mesophilic proteins and showed that the aromatic-aromatic interactions are possible among aromatic residues such as Phe, Tyr, Trp if their phenyl ring centroid separation are between 4.5 to 7.0 Å and the dihedral angles are within 30 to 90°. Although aromatic interactions, like aromatic-sulphur and cation-pi interactions have not been explored in relation to thermostability, these interactions are known to exist in proteins. In cation-pi interactions, positively charged residues such as Arg or Lys or metal cations interact with negatively charged centre of aromatic rings while in aromatic-sulphur interactions, the sulphur atom interacts with the aromatic ring centre.

The PGAs considered here have little variation in their content of aromatic residues (Table 4.7). No direct correlations were observed for interactions involving aromatic residues between mesostable and thermostable PGA enzymes (Fig. 4.13), thus it was difficult to assess the degree of contribution of these factors towards PGA thermostability.

4.4 Summary

In summary, the screened ptPGA enzymes were compared with the experimentally characterized *Ec*PGA, *Af*PGA, and *Ax*PGA (in increasing order of thermostability) first in terms of their sequence followed by their structural comparison. In the *sequence-based consensus* approach PGAs were compared in terms of their residue preference at 24 thermostabilization sites while in the *structure-based* approach enzymes were compared in terms of various structural features known to contribute to protein thermostability. The *sequence-based* analysis revealed that ptPGA enzymes could have higher thermostability as compared to *Ec*PGA while also highlighting few additional potential sites which on mutation could improve their thermostability. The structural analysis of the enzymes emphasized that, of all the ptPGAs, *Pd*PGA could behave like thermostable *Af*PGA and *Ax*PGA owing to the presence of a disulfide bond, higher number of stable ion-pair networks, greater proline content, higher percentages of proline residues in beta-turns and lower content of thermolabile residues and bonds. Finally based on the above results, it was decided to explore the structure-function relationship of the *Pd*PGA enzyme through experimental characterization of its stability in comparison with other PGAs.

Chapter 5

*Cloning, expression, purification and
assessing the thermostability of
PGA enzyme from
Paracoccus denitrificans*

The computational analysis described in Chapter 4 highlighted *Pd*PGA (PGA from *Paracoccus denitrificans* PD1222) as a potential thermostable enzyme candidate among the three identified putative thermostable PGAs. This chapter describes cloning of *Paracoccus denitrificans* penicillin G acylase gene, its expression, purification followed by characterization and evaluation of its thermostability. A combinatorial approach using both biochemical and biophysical tools was undertaken to better understand the structure-function relationship of the enzyme.

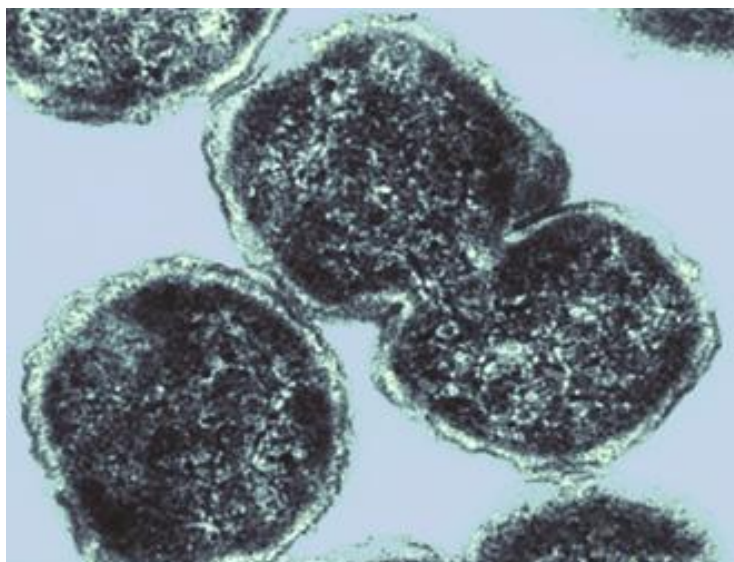


Figure 5.1: *Paracoccus denitrificans*. The image is adapted from Joint genome institute genome portal (<http://genome.jgi-psf.org/>).

5.1 Introduction

5.1.1 *Paracoccus denitrificans*

Paracoccus denitrificans, a non-motile coccoid soil microbe belonging to class alpha proteobacteria, is a model organism for the study of denitrification (Fig. 5.1). Its taxonomic lineage is *Bacteria* (superkingdom), *Proteobacteria* (phylum), *Alphaproteobacteria* (class), *Rhodobacterales* (order), *Rhodobacteraceae* (family), *Paracoccus* (genus) and *Paracoccus denitrificans* (species). It was first isolated by Martinus Beijerinck in 1908 based on its capability to reduce nitrate to di-nitrogen under anaerobic growth conditions (denitrification). The four oxido reductases and their corresponding regulatory genes of the denitrification pathway have been well characterized (Stouthamer, 1991). The organism can also grow well under aerobic conditions due to presence of a respiratory chain, similar to eukaryotic mitochondria. Evidence suggests *Paracoccus* as a close relative of evolutionary pre-cursor of eukaryotic mitochondrion,

and therefore it has been exploited as a model organism for the study of mitochondrial respiratory chain (John & Whatley, 1975).

P. denitrificans relies on compounds such as methanol, n-pentanol and n-propanol as carbon source for its growth. It oxidizes these compounds to carbon dioxide, assimilates it through Calvin cycle and produces aromatic compounds such as poly(3-hydroxybutyrate), poly(3-hydroxyvalerate) and a mixed polymer of the two, respectively, as the storage compounds (Baker *et al.*, 1998). The organism can also grow as a chemolithotroph by utilizing inorganic compounds such as sulphur and hydrogen as energy sources and carbon dioxide as carbon source (Friedrich & Mitrenga, 1981). Therefore, *P. denitrificans* has been considered a popular model organism for investigators with interests in sulfur compound transformations.

The genus *Paracoccus* is a biochemically versatile genus capable of degrading wide range of chemicals and therefore very useful in bioremediation of polluted environments. Members of this genus have been known to be involved in the bioremediation of waste water from coke-oven factories contaminated with thiocyanate by utilizing thiocyanate as an energy source (Katayama *et al.*, 1995). Under anaerobic growth condition, strains have been isolated that are known to degrade halobenzoates, sulfonates such as taurine, cysteate, sulfoacetate, 2-hydroxyethanesulfonate and 3-aminopropanesulfonate. Some of the other strains depend on carbon disulfide or carbonyl sulfide as energy sources (Jordan *et al.*, 1997). Isolated strains from activated sludge have been observed to degrade quaternary carbon compounds such as dimethylmalonate under denitrifying conditions while other strains are capable of degrading variety of methylated amines under both aerobic and anaerobic conditions. Interesting thing about *P. denitrificans* is its unusual heterotrophic nitrification activity which involves oxidation of ammonia to nitrite during growth on organic energy sources. Thus the coupled activity of denitrification and heterotrophic nitrification results in complete transformation of ammonia to di-nitrogen by a single organism (Baumann *et al.*, 1996).

The genome of *P. denitrificans* PDI222 strain consists of two circular chromosomes (Genbank Accession: CP000489 and CP000490 respectively) and a 653815 bp plasmid (Genbank Accession: CP000491). The plasmid encodes the *PdPGA* gene starting from position 308495 to 310867 in the 5'-3' orientation, whose reverse complement yields a 790 residue *PdPGA* protein (Genbank Accession: ABL72874.1). The genome sequence shows presence of

many genes related to polycyclic aromatic hydrocarbon degradation pathway. Although the physiological role of *pga* gene is not clear, hypothesis suggests its relation to aromatic compound degradation pathway (Valle *et al.*, 1991). Despite the fact that the preferred substrate of PGAs for industrial application is penicillin G, the natural substrates of choice for these enzymes are as yet unclear.

5.2 Materials and Methods

5.2.1 Cloning of *PdPGA* gene

The *PdPGA* gene clone was gifted by Dr. Sureshkumar Ramasamy. The *PdPGA* gene, encoded in the plasmid of *Paracoccus denitrificans* *PDI222*, was cloned in pETKat vector (Gift from Dr. Katrin Tiemann of Dr. Clemons Lab, Caltech). The insert has been amplified using gene specific primers (forward primer: GAA AAC CTG TAC TTC CAG AGC ATG GGC ACC CAG GTC GAG and reverse primer: CCC TGA AAC AAG ACT TCC AAC CGC GGA ACG GCA AGG GTT T) and inserted into pETKat vector by PIPE cloning protocol (Klock & Lesley, 2009). In short the pETKat vectors were PCR amplified with vector PIPE-for (5'-TTGGAAGTCTTGTTTCAGGGACCA-3') and vector PIPE-rev (5'-GCTCTGGAAGTACAGGTTTTTCACC-3'). 1.2 µl of insert and vector were mixed on ice and 50 µl NovaBlue (Invitrogen) competent cells were added. This pETKat vector contained a suicide cassette, derived from pDest53 (Invitrogen) for better cloning efficacy. The gene cloning was confirmed by DNA sequencing.

5.2.2 Preparation of competent cell

E. coli NovaBlue strain was used as the host for maintenance of plasmid containing *PdPGA* gene while *E. coli* BL21-Gold (*DE3*) strain was considered for *PdPGA* expression. Cells were made competent by growing in 5 ml Luria-Bertani (LB) media overnight at 37 °C. Next day, 50 ml of fresh LB media was inoculated with 1 ml of overnight culture and incubated at 37 °C for 4 hr until OD₆₀₀ reached to 0.5. The culture grown was then cooled down in ice for 15 min and centrifuged at 4000 rpm for 15 min at 4 °C. The supernatant was discarded and the pellet was resuspended in approximately 16 ml of RF-I buffer (100 mM RbCl₂, 50 mM MnCl₂, 30 mM KAc and 10 mM CaCl₂) and incubated in ice for 1 hr. Cells were then centrifuged at 4000 rpm for 15 min at 4 °C. Pellet obtained was resuspended in approximately 3 ml of RF-II buffer (10

mM RbCl, 10 mM MOPS, 30 mM CaCl₂, 15% Glycerol, pH 7.8) and incubated in ice for 15 min. Cells were then aliquoted in 1.5 ml microfuge tubes with 20% sterile glycerol and stored at -80 °C.

5.2.3 Transformation of plasmid

Plasmid containing *PdPGA* gene was transferred into the maintenance and expression hosts by first thawing the competent cells in ice. 5 µl of plasmid was added gently to 50 µl of competent cells, mixed and kept on ice for 30 min. The cells were then heat shocked at 42 °C for exactly 60 sec and immediately kept back on ice for 5 min. 1 ml of LB media was added and cell cultures were incubated at 37 °C for 1 hr at 180 rpm. Culture was then centrifuged at 3000 rpm. The supernatant of around 800 µl was discarded and the remaining 200 µl of cells were spread on LB agar plate containing kanamycin (34 µg/ml) as selection marker. The agar plate was incubated overnight at 37 °C. Next day several colonies were obtained. Single isolated colony was picked and inoculated in 5 ml LB media and grown overnight at 37 °C. Subsequently the culture was streaked on LB agar plate containing kanamycin. These plates were re-streaked every month and used throughout the studies.

5.2.4 Cryopreservation of Bacterial culture

Transformed cell cultures containing plasmid with *PdPGA* were preserved in glycerol at ultra low temperature for further use. In a microfuge tube (1.5 mL), 20% sterile glycerol was mixed thoroughly with the overnight grown cultures by pipeting. Aliquots were frozen in liquid nitrogen and stored at -80 °C. Fresh stocks were prepared every 3 months.

5.2.5 Expression of *PdPGA*

From the LB agar plate of transformed expression host, cells were picked and inoculated in 5 mL liquid LB medium containing kanamycin (34 µg/ml). Culture was grown overnight with shaking at 180 rpm and 37 °C. 1% of the overnight grown culture was used to inoculate 300 ml of liquid LB containing 300 µl of kanamycin. Once the cultures reached OD₆₀₀ 0.6-0.8, cells were induced by the addition of 0.75 mM isopropyl β-D-thiogalactopyranoside (IPTG) as inducer. After induction the cultures were grown for 16-18 hr at 16 °C and 180 rpm.

5.2.6 Cell lysis

Cells were harvested by centrifugation at 5000 rpm for 30 min at 4 °C. Cell pellet obtained was resuspended in minimum volume of Lysis buffer (25 mM Tris-HCL pH 7.5, 100 mM NaCl and 20 mM Imidazole). The cell suspension was disrupted by sonication on ice using Ultrasonic homogenizer of Esquine Biotech. A total of three 5 min cycles of 6 sec pulse-on, 8 sec pulse-off at 60% power was performed. Cells were then centrifuged at 12000 rpm for 30 min at 4 °C. The crude lysate was used further for purification.

5.2.7 Purification of *Pd*PGA by Immobilized Metal Ion Affinity chromatography (IMAC)

Immobilized metal-affinity chromatography (IMAC) is a powerful method for the purification of recombinant protein containing poly-histidine affinity tag (Bornhorst & Falke, 2000). The principle behind the purification is based on the interaction between the affinity-tag and immobilized metal ion matrices. Histidine residues acts as electron donor group and coordinates with the metal ion (Zn^{2+} , Cu^{2+} , Co^{2+} , Ni^{2+}) immobilized on a matrix. Widely used and commercially available immobilized metal matrices include Co^{2+} -carboxymethylaspartate (Co^{2+} -CMA) and Ni^{2+} -nickel-nitrilotriacetic acid (Ni^{2+} -NTA). Of the six coordinate sites of the metal ions, four are used for coordinating with the matrix while two remaining sites are exposed for their interactions with the affinity-tag. The advantages of using these resins lie with their tolerance to a wide range of conditions and their ability to get regenerated and enabling reuse several times. However, the major problem of IMAC method is the non-specific binding of proteins containing two or more adjacent histidine residues (Bornhorst & Falke, 2000).

The C-terminal His-tagged *Pd*PGA protein was purified by Ni-NTA affinity chromatography by loading the supernatant in a column containing Ni^{+2} -sepharose beads pre-equilibrated with equilibration buffer (25 mM Tris-HCl pH 7.5, 100 mM NaCl and 20 mM Imidazole). The matrix was then washed with equilibration buffer followed by removal of non-specific and weakly bound proteins by washing with wash-buffer (25 mM Tris-HCl pH 7.5, 100 mM NaCl and 50 mM Imidazole). The matrix-bound *Pd*PGA enzyme was eluted by passing elution buffer (25 mM Tris-HCl pH 7.5, 100 mM NaCl and 500 mM Imidazole). Eluted fractions were checked for PGA enzyme activity and the fractions showing PGA activity were pooled together.

5.2.8 Removal of imidazole by desalting

The high concentration of imidazole in the eluted fractions from Ni-NTA affinity chromatography step was removed by passing through PD10 desalting column and by exchanging with Glycine-NaOH (25 mM) buffer, pH 10 containing 150 mM NaCl. PD-10 desalting column consists of Sephadex G-25 medium which is used for removal of excess salts from any sample. The desalting method is based on the principle of gel filtration where molecules are separated on the basis of their difference in size. Molecules which are larger than the pore size, do not enter matrix pores, thereby get eluted first. Smaller molecules enter the pores, and thus elute after the elution of larger molecules present in the void volume.

5.2.9 Purification by size exclusion chromatography

Size exclusion chromatography is a protein purification method which separates molecules based on their size and shape. The sample is passed through a column containing the matrix/beads. Molecules diffuse into the beads to varying extent. Smaller molecules diffuse inside the pores, thus move through the matrix slowly, while larger molecule cannot penetrate the pore, and move quickly. The sample containing *PdPGA* protein after purification and desalting step was concentrated with Amicon centrifugal concentrator (Millipore, USA) with cutoff range of 30 kDa and passed through Gel filtration column (Sephacryl S-200) connected to AKTA Explorer and fractions were eluted with 25 mM Glycine NaOH buffer pH 10 containing 150 mM NaCl. The aliquots of the fractions were checked for the presence of enzyme activity using standard assay. Purity and homogeneity of fractions were checked using 12% SDS-PAGE and *PdPGA* protein band was detected using Western blot.

5.2.10 SDS - Polyacrylamide Gel Electrophoresis (SDS-PAGE)

Polyacrylamide gel electrophoresis is a widely used analytical method for the separation of proteins on the basis of their size and shape. It is based on the principle that charged molecules migrate towards oppositely charged electrode in an electric field. Protein molecules with differential size and shape are first denatured by heating and treating with SDS (Sodium dodecyl sulfate). SDS, an anionic detergent, provides negative charge to all component of a given protein mixture. Therefore proteins migrate solely on the basis of their size towards the anode. Proteins can be visualized by staining with Coomassie brilliant blue, a protein-specific dye. The size of a

protein is derived by comparing its migration distance with the migration distance of known molecular weight markers. The purity and homogeneity of *Pd*PGA was checked on 12% sodium dodecyl sulphate polyacrylamide gel (SDS-PAGE). The denatured protein samples were first stacked in a stacking gel before entering into the separating gel. Gel was prepared using the BioRad SDS-PAGE apparatus with 1.5 mm spacers and ran at a voltage of 150 V. Each sample was loaded onto separate wells and electrophoresed alongside low-range molecular weight markers. Protein bands in the gel were stained with 75-100 ml of Coomassie brilliant blue staining solution (0.2% Coomassie brilliant blue R-250 in 25% propane-2-ol and 10% Acetic acid). The staining solution was discarded after 2 hour, and the gel was washed twice with de-ionized water and further washed three times with approximately 100-150 ml of fresh destaining solution (5% Propan-2-ol, 7% Glacial Acetic acid).

5.2.11 Western blot

Western blot is a method for detecting specific protein from a complex mixture of proteins. The steps involved are separation of proteins first by gel electrophoresis, followed by their transfer from gel onto a membrane and eventual visualization of target protein using primary and secondary antibodies. Since the antibodies binds only to the protein of interest, only the band corresponding to target protein will be visible; thickness of band corresponds to the amount of protein present. Blotting or transfer of proteins from gel can be done by using either nitrocellulose or PVDF membranes. Nitrocellulose membrane has advantage of high affinity and retention abilities for protein but it is brittle, thus does not allow for re-probing. On the contrary, PVDF membrane has advantage of higher mechanical rigidity and thus allows the membrane to be re-probed and stored. However the background is higher in case of PVDF membrane, thus thorough washing is necessary (Mahmood & Yang, 2012).

In case of *Pd*PGA, SDS-PAGE gels were first electroblotted onto the PVDF membrane. Blotting was done by placing the PVDF membrane between gel and positive electrode so that proteins move out of gel onto the membrane due to the electric field perpendicular to the gel surface. This electrophoretic transfer or blotting was then followed by blocking of membrane with 5% skimmed milk in PBS for about 1 hr followed by washing in PBS containing 0.05% Tween 20, 3 times 5 min each. Blocking is an important step in the western blotting procedure as it prevents non-specific binding of antibodies on the membrane. Due to presence of poly-his tag

at C-terminal of *Pd*PGA, the membrane was subsequently incubated with primary monoclonal anti-polyhistidine antibody (Sigma-Aldrich, Cat# H1029) at 1:1000 dilutions in PBS containing 1% BSA. Next day, the membrane was washed with PBS containing 0.05% Tween 20, three times for 5 min each and incubated with anti-mouse Ig-G (Fc specific) peroxidase conjugate secondary antibody (Sigma-Aldrich, Cat#A0168) at 1:6000 dilutions in PBS and Tween 20. To visualize the band, membrane was finally treated with the substrate Novex HRP (Invitrogen Cat# 20004).

5.2.12 Protein estimation

Protein concentrations in samples were estimated by using Bradford (BioRad) method using bovine serum albumin (BSA) as calibration standard. In this calibration standard, 20 μ l of different concentration of BSA ranging from 0.05-1 mg/ml was incubated with 1 ml of Bradford solution for 5 min. Absorbance was recorded at 595 nm. For *Pd*PGA samples, same method was followed and concentration of protein was determined by comparing with the standard curve of BSA. Buffer blanks were taken into consideration for analysis.

5.2.13 Penicillin G Acylase activity

The enzyme activity of *Pd*PGA was determined according to Bomstein and Evans method (Bomstein & Evans, 1965), modified by (Shewale *et al.*, 1987), by allowing the purified enzyme preparation (0.04 mg/ml) to interact with its substrate Penicillin G (20 mg/ml) in 50 mM phosphate buffer, pH 7.5, at 45 °C for 10 min. The reaction was quenched by the addition of 1 ml of Citrate Phosphate Buffer (300 mM citric acid in 50 mM Phosphate Buffer, pH 2.5). The product 6-APA released was estimated spectrophotometrically at 415 nm by reacting it with 2 ml of 0.6% (w/v) of p-dimethylaminobenzaldehyde in methanol. The 6-amino group of the product forms colored Schiff base which is estimated. The optimal temperature for PGA activity of *Pd*PGA was determined in the range of 25 to 60 °C while the optimum pH determined was between pH 3 to 12. For studying the thermal stability of *Pd*PGA the enzyme samples were heated in the range of 35 to 50 °C for up to 30 min followed by estimation of enzyme activity. The pH stability profile of *Pd*PGA was measured by incubating the enzyme in suitable buffers in the pH range of 3-12 for up to 3 hrs, followed by estimation of its enzyme activity. Buffers used were 25 mM Glycine-HCl buffer (pH 1-3), Acetate buffer (pH 4-5), Phosphate buffer (pH 6-7),

Tris-HCl buffer (pH 8-9) and Glycine-NaOH buffer (pH 10-12). To understand the effect of reducing disulfide bond on *Pd*PGA thermostability, *Pd*PGA was treated with 10, 50 and 100 mM of reducing agent DTT (Dithiothreitol) and incubated at 40 °C for 10 min. Enzyme samples without the addition of DTT was used as control and residual activity was measured in each case. In order to understand the influence of various modulators such as ions (10 mM Mn²⁺, Co²⁺, Ni²⁺, Ag⁺, Zn²⁺, Ca²⁺, Cs⁺, Fe²⁺, Mg²⁺, Cu²⁺, and EDTA), detergents (1% Tween-20, Triton-X, IGEPAL CA-630 and Tween-80) and organic solvents (5-10% Methanol, ethanol, propanol, butanol and isoamyl alcohol) on *Pd*PGA thermostability, enzyme was incubated with the modulators at 40-50 °C for 10 min followed by measurement of residual activity. Purified *Pd*PGA was checked for its kinetic behavior under previously optimized conditions of its enzyme activity, pH 10 and 45 °C. Enzyme was incubated with a range of substrate concentrations, from 0.1 mM to 50 mM. The effect of increasing concentration of substrate (PenG) on *Pd*PGA activity was analyzed using Michaelis-Menten plot.

5.2.14 Thermal unfolding measured using steady state fluorescence

Fluorescence spectroscopy is a popular method for studying proteins in terms of their biochemical and biophysical characteristics such as conformational changes, metal-binding information, protein-protein interactions, membrane localization, long-range distance measurements and kinetic/dynamic parameters (Vogel & Weljie, 2002). This method has many advantages when applying to biological systems since it detects the occurrence of natural fluorophores in proteins such as Tryptophan and Tyrosine, it requires relatively low sample concentrations. Tryptophan is a stronger intrinsic fluorophore than Tyrosine due to the indole ring's high sensitivity to electronic excitation. The indole ring has high electron density of its extended pi system that allows electronic transition to high-energy excited state upon absorption of photon. The electronic excited state returns to the ground state by the emission of photon having comparatively longer wavelength or lesser energy than the excitation. The fluorophore is excited over a range of wavelengths and the corresponding emission spectra are recorded. The power of fluorescence of a fluorophore is influenced by its microenvironment such as hydrophobicity, viscosity and mobility.

The intrinsic fluorescence of *Pd*PGA was recorded using a Perkin Elmer LS50 fluorescence spectrophotometer attached to a Julabo F20 water bath. The enzyme (0.04 mg/ml)

was excited at 295 nm followed by measurement of the emission spectra between 310 and 400 nm, while keeping the speed at 100 nm min⁻¹ and slit width at 7 nm. Background emission due to buffer was subtracted from the results. Thermal unfolding of *Pd*PGA was monitored by incubating the enzyme between 25 to 70 °C for 10 min followed by measurement of emission spectra. Thermal aggregation of *Pd*PGA was monitored between 25 to 70 °C also by Rayleigh scattering measurements on the same instrument.

5.2.15 Estimation of thermal unfolding using CD spectroscopy

Circular dichroism is a powerful spectroscopic technique, widely used to study chiral molecules of varied size and types. It has its most important application in analyzing secondary structures or conformations of large biological molecules. Circular dichroism (CD) is due to the differential absorption of left-handed and right-handed circularly polarized lights (CPL) by a molecule containing one or more chiral chromophores.

$$CD=A(\lambda)_{LCPL} - A(\lambda)_{RCPL} \text{ where } \lambda \text{ is wave length and A is absorption.}$$

If the molecule of study is a chiral molecule, then it will absorb one CPL state higher than the other CPL state. If left-CPL is absorbed more, the CD-signal will be positive while if right-CPL is absorbed more, CD-signal will be negative. The variation of CD as a function of wavelength is measured in the form of CD-spectra in a circular dichroism spectrometer. Vast majority of biomolecules are mainly chiral in nature. For example 19 of the 20 amino acids in protein are chiral and therefore protein molecule can be studied using CD. Protein secondary structure is sensitive to its environment, thus CD can be used to monitor the change in its secondary structure with respect to the change in environment such as temperature or pH change. From the CD spectra, structural, kinetic and thermodynamic parameters of a protein can be derived.

CD measurements of the purified *Pd*PGA were recorded using a Jasco J-815-150S (Jasco, Tokyo, Japan) spectropolarimeter connected to a Peltier CDFL cell circulating water bath. Far-UV spectra were recorded in a rectangular quartz cell of 1-mm path length in the range of 200–250 nm at a scan speed of 100 nm/min with a response time of 1 s and a slit width of 1 nm. Purified *Pd*PGA at a concentration of 0.05 mg/ml was used for all the samples. Each spectrum was recorded as an average of five scanned spectra. Thermal denaturation studies of *Pd*PGA

were carried out by incubating enzyme at temperatures ranging between 25-70 °C for 10 minutes. Results were expressed as mean residue ellipticity (MRE) in deg cm²/dmol defined as

$$\text{MRE} = M\theta_{\lambda}/10dcr$$

Where M is the molecular weight of the protein, θ_{λ} is CD in millidegree, d is the path length in cm, c is the protein concentration in mg/ml, and r is the average number of amino acid residues in the protein. The relative content of various secondary structure elements was calculated by using CDPro software (<http://lamar.colostate.edu/~sreeram/CDPro/main.html>). Low NRMSD values were observed for analysis with CONTINLL.

5.2.16 Hydrophobic dye binding

8-Anilino-1-naphthalene sulfonic acid (ANS), a hydrophobe-selective dye and a fluorescent molecular probe is widely used to study conformational transition in proteins. ANS's fluorescent property changes when it binds to the hydrophobic regions of proteins and thus it has been used to study ligand induced conformational changes in proteins. The naphthalene backbone and aniline ring provides hydrophobicity while its sulfonate group provides negative charge to ANS. The amide group on aniline ring also provides electron for hydrogen bonding interactions with protein. Thus ANS can interact with both the hydrophobic regions as well as positively charged residues of protein. This property is very useful to study the protein folding. A correctly folded protein buries most of its hydrophobic regions in the core. When protein unfolds under denaturing conditions, hydrophobic residues are exposed, thus allows the binding of ANS. The λ_{max} of emission spectra of ANS in water is 500 nm. When it binds to hydrophobic patches on the surface of protein, a blue shift along with huge increase in emission intensity is usually observed. The degree of blue-shift depends on the surrounding microenvironment of dye in protein.

Binding of ANS was studied by exciting the sample at 375 nm followed by recording the emission spectra between 400-550 nm using a steady state spectrofluorimeter. Protein solutions were incubated at different temperatures (35 to 60 °C) to study the thermal unfolding of *Pd*PGA. 5 μ l of 15 mM ANS was mixed with 2 ml of protein solution (0.05 mg/ml). Spectrum of buffer with ANS was subtracted in each case for further analysis.

5.3 Results and Discussion

5.3.1 Confirmation of *PdPGA* gene clone

Cloning of *PdPGA* gene in pETKat vector (Fig. 5.2) was confirmed by sequencing approximately 900 bases from both 5' and 3' end using standard T7 forward and T7 reverse primers. Figure 5.3 and 5.4 shows 5' and 3' sequence of *PdPGA* gene respectively, their translation product and sequence alignment with the *PdPGA* sequence available at Genbank database (ABL72874.1). The 5' and 3' encoded *PdPGA* protein regions were found to be 100% identical with the N-terminal and C-terminal regions of reported *PdPGA* protein sequence of Genbank database.

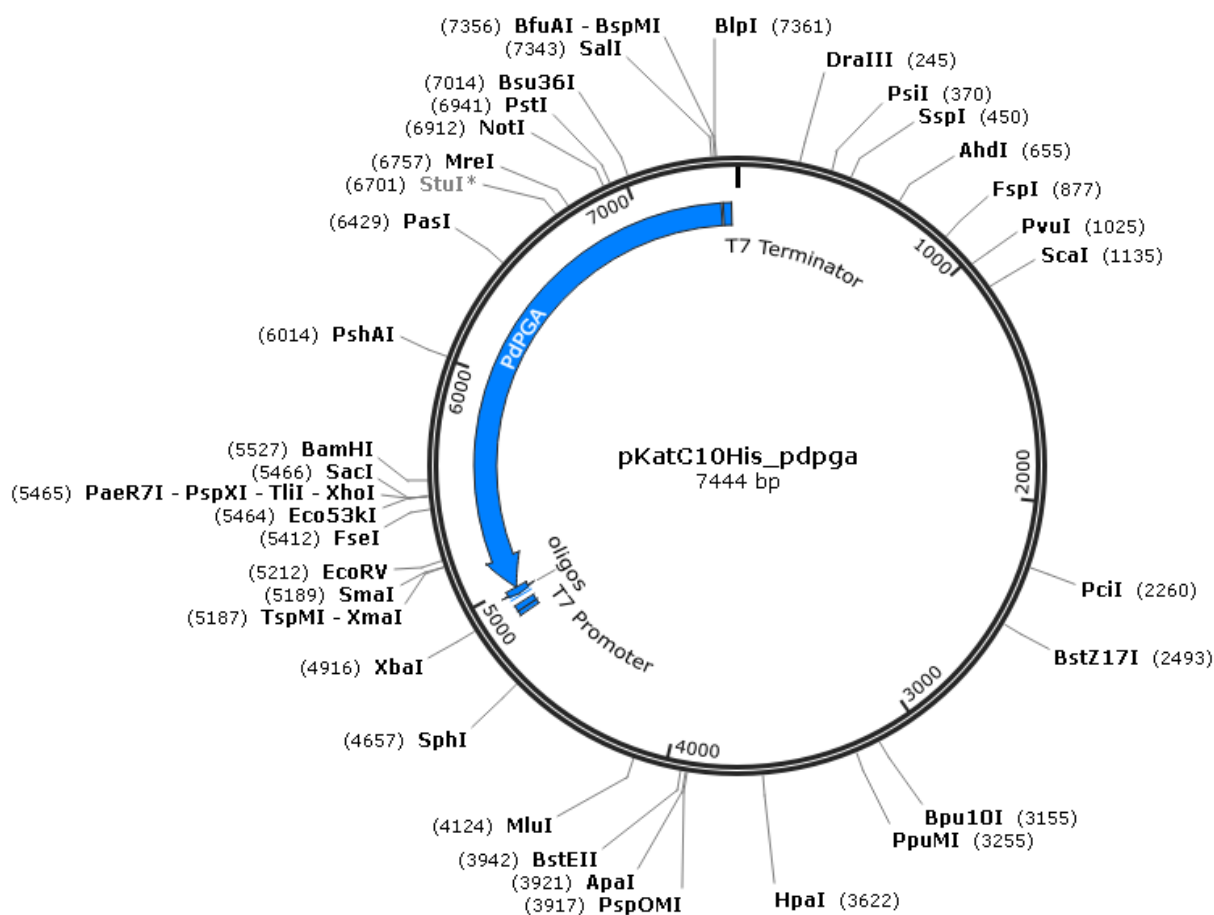


Figure 5.2: The vector map of pETKat vector containing the *PdPGA* gene.



Figure 5.3: (a) Approximately 900 bases of *PdPGA* gene sequenced by T7 forward primer. The bases highlighted in red encode N-terminal region of *PdPGA* protein. (b) The translated *PdPGA* gene sequence in panel a, using standard genetic code. The residues shown in red correspond to *PdPGA*. (c) Pairwise sequence alignment of *PdPGA* amino acid sequence in panel b (Query) with the available *PdPGA* sequence of Genbank database (Sbjct: ABL72874.1).

Like the other PGAs, *Pd*PGA is also produced as an inactive precursor (pre-pro-*Pd*PGA) having the typical polypeptide organization which consists of signal peptide, chain α , spacer peptide and chain β . The signal peptide directs the enzyme to periplasmic space. In periplasmic space, enzyme (pro-*Pd*PGA) gets activated by autocatalytic removal of spacer peptide. The mature enzyme is a heterodimer with Chain α and chain β . The SignalP 2.0 HMM model (Nielsen & Krogh, 1998) identified 26 residues from N-terminal of *Pd*PGA as signal peptide with signal peptide probability of 1 and cleavage site probability of 0.9 at residue 26. Therefore in the *Pd*PGA gene cloning, initial 78 bases corresponding to signal peptide were not included. Thus the recombinant *Pd*PGA is produced as a pro-*Pd*PGA protein (Chain α , Spacer peptide and Chain β). The length of spacer peptide usually varies from species to species. Since the N-terminal residue of β -chain is always a serine residue, the C-terminal end of spacer peptide corresponds to the residue preceding β Ser1. However, owing to the difference in lengths of Chain α from species to species (*Ec*PGA: 209 and *Af*PGA: 196), N-terminal of spacer peptide is difficult to identify solely based on sequence comparison. In case of 790 residue native *Pd*PGA protein, the signal peptide is of 26 residue while Chain β is of 540 residue (theoretical molecular weight 60.1 kDa) thus leaving the Chain α and spacer peptide together as 224 residue (theoretical molecular weight 25.1 kDa).

5.3.2 Purification of *Pd*PGA

The over expressed *Pd*PGA protein was purified from *E. coli* BL21-Gold (DE3) cells by cell lysis followed by a 3-step purification protocol of Ni-NTA affinity chromatography, desalting and size exclusion chromatography. The elution profile of gel-filtration chromatography in Figure 5.5 shows 3 peaks in OD₂₈₀ reading. While most of the aggregated high-molecular weight proteins were removed already (Peak1), low-molecular weight proteins were present in Peak2 and Peak3. Of the three peaks obtained, the fractions belonging to second peak (Fractions 17-20) alone showed PGA activity.

The SDS-PAGE and western blot experiments confirmed the expression of *Pd*PGA where the purified enzyme showed 2 bands on SDS-PAGE, corresponding to α and β chains, along with the detection of β -chain by western blot with monoclonal anti-His antibodies (Fig. 5.6).

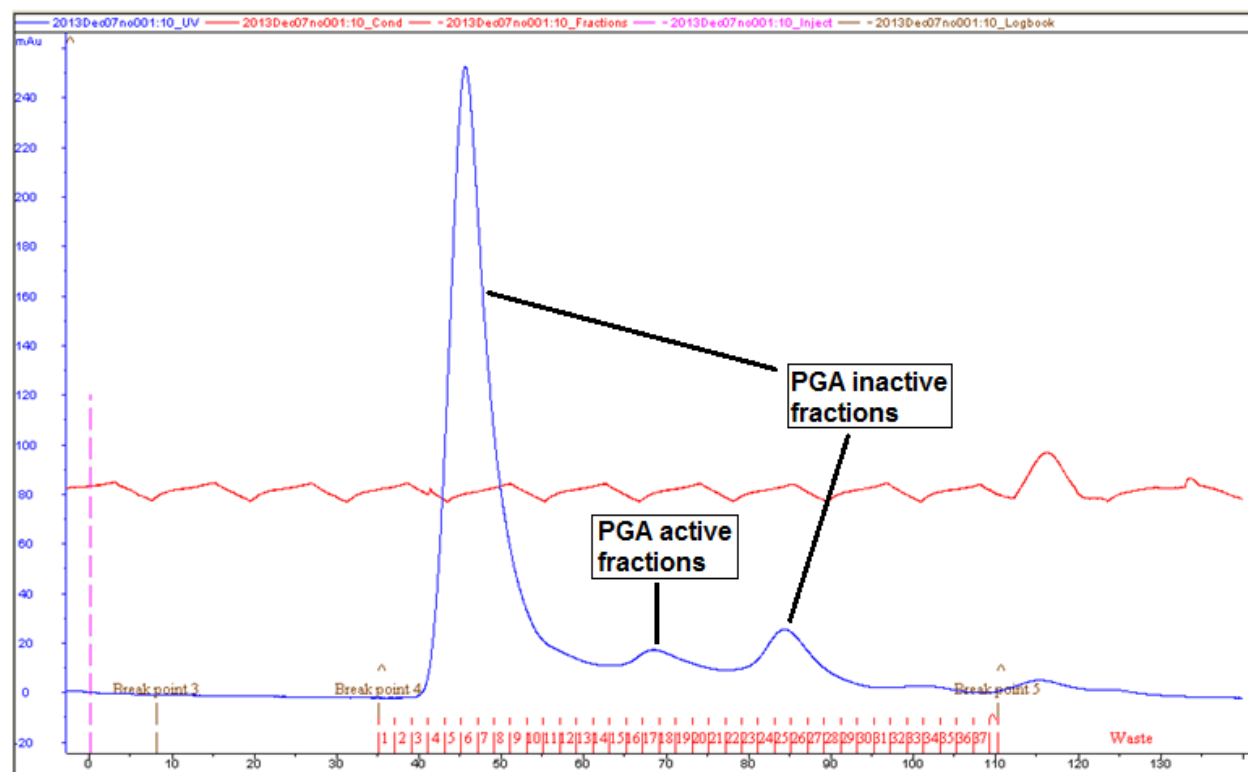


Figure 5.5: Elution profile of Gel filtration chromatography. OD_{280} is shown in blue color.

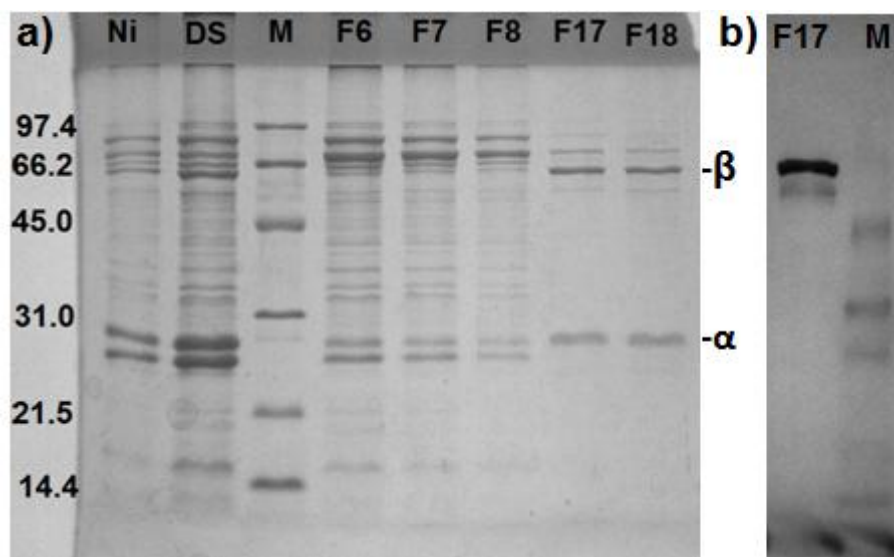


Figure 5.6: a) 12% SDS-PAGE of various fractions obtained during 3-step *Pd*PGA purification protocol. The lanes in the gel labeled as Ni, DS, M, F6, F7, F8, F17 and F18 corresponds to elution fractions from Ni-NTA affinity chromatography, desalting column, molecular-weight marker, and Peak1 fractions (F6, F7, F8), Peak2 fractions (F17 and F18) of gel filtration chromatography, respectively. Of the gel filtration fractions, F17 and F18 fractions containing *Pd*PGA shows PGA activity. The molecular weights of markers are also

shown. b) Western blot of F17 fraction containing *Pd*PGA protein. The β -chain of *Pd*PGA with C-terminal His-tag was detected using anti-His monoclonal antibody.

5.3.3 Enzyme kinetics

Kinetic parameters of the enzyme were estimated using Michaelis Menten plot (Fig. 5.7). The V_{\max} and K_m values were obtained by linear regression curve fitting to the available data-points using Lineweaver-Burk plot. V_{\max} and K_m of *Pd*PGA obtained were 0.378 mM of PenG degraded /mg of protein/minute and 3.97 mM respectively.

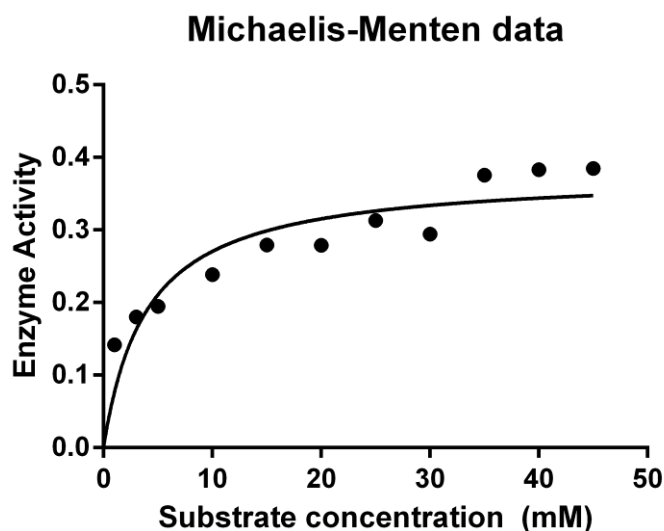


Figure 5.7: The Michaelis-Menten plot depicting the rate of enzyme activity at different substrate concentrations.

5.3.4 Alkaline stability of *Pd*PGA

The optimum pH of *Pd*PGA enzyme activity was observed to be pH 10 (Fig. 5.8b). The pH stability profile of *Pd*PGA showed that the enzyme is stable over a wide range of pH 5-11 (Fig. 5.8c). The enzyme was found to be most stable in alkaline pH 10. The enzyme retained almost 60% of its activity at pH 10 even after 3 hrs of incubation. The enzyme was also found to be stable at pH 11, retaining 50% of its activity after 3 hrs of incubation. However, it was found to be unstable at pH 12. Towards the acidic pH scale, *Pd*PGA was found to be stable till pH 5 with 50% of enzyme activity being retained after 3 hrs of incubation but at pH 4, a 70% reduction in enzyme activity was observed after 3 hrs of incubation.

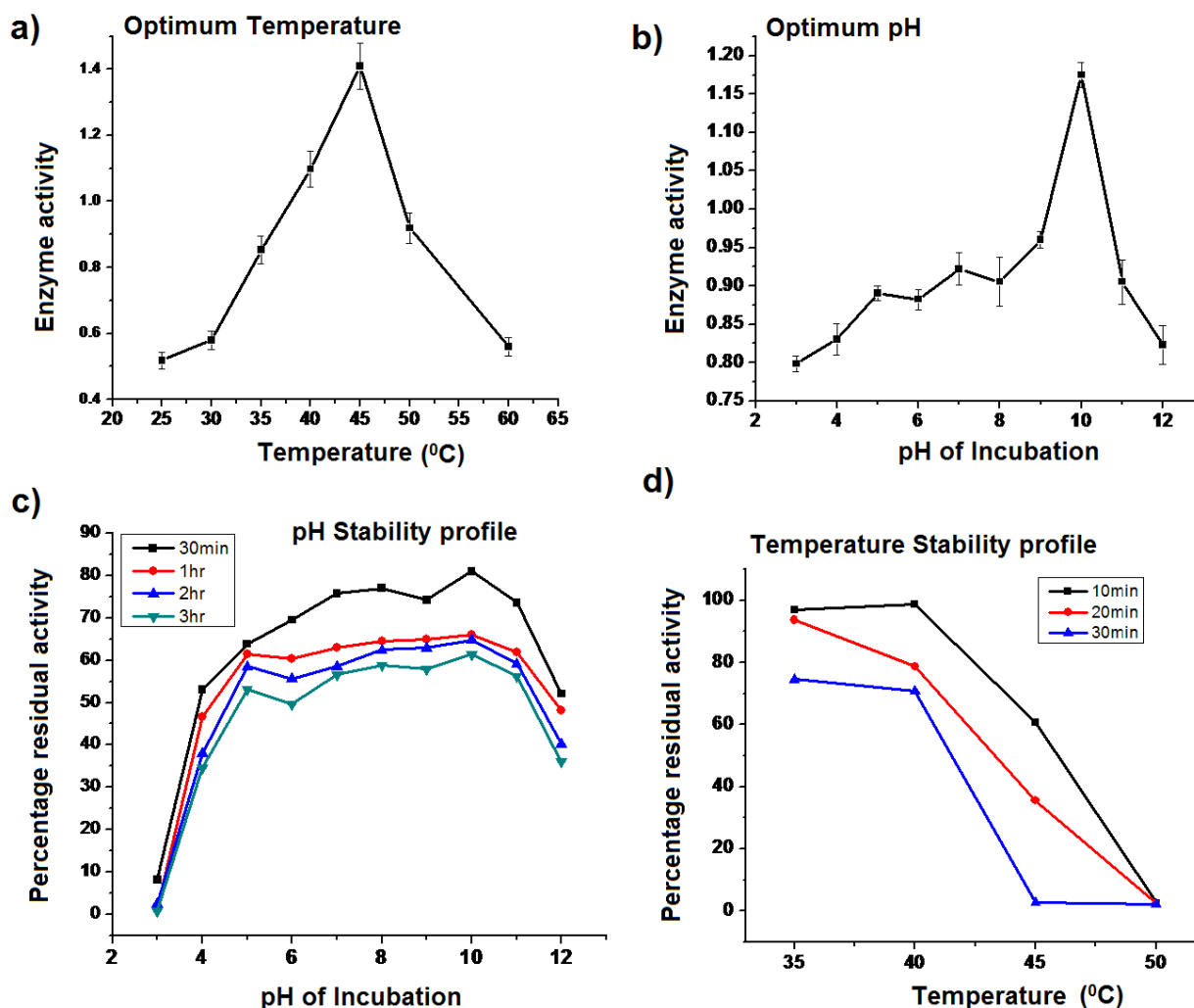


Figure 5.8: Panel a-b illustrates the optimum temperature and pH for PGA hydrolysis activity of *PdPGA* while panel c-d illustrates the stability profile of *PdPGA* at various pH and temperatures.

5.3.5 Temperature stability profile of *PdPGA*

The optimum temperature for substrate hydrolysis by *PdPGA* was observed to be 45 °C (Fig. 5.8a). The temperature stability profile of *PdPGA* was monitored up to 50 °C (Fig. 5.8d). At 50 °C upon 10 min of incubation, complete loss of enzyme activity was observed. Comparison of enzyme activity at 40 °C to that of 45 °C for 10 min of incubation revealed a 40% loss of enzyme activity at 45 °C. The enzyme was found to be stable for maximum of 20 min at 45 °C. Overall the enzyme was found to exhibit higher stability compared to *EcPGA*, while it was observed to be lower than that of both *AfPGA* and *AxPGA*. The loss of activity beyond 45 °C could be either due to global structural changes of the enzyme or due to local structural

changes near the active site. To understand this, thermal unfolding studies using both steady-state fluorescence and CD spectroscopy were carried out. The addition of modulators such as ions, detergents and organic solvents didn't show any significant effect on *Pd*PGA thermostability.

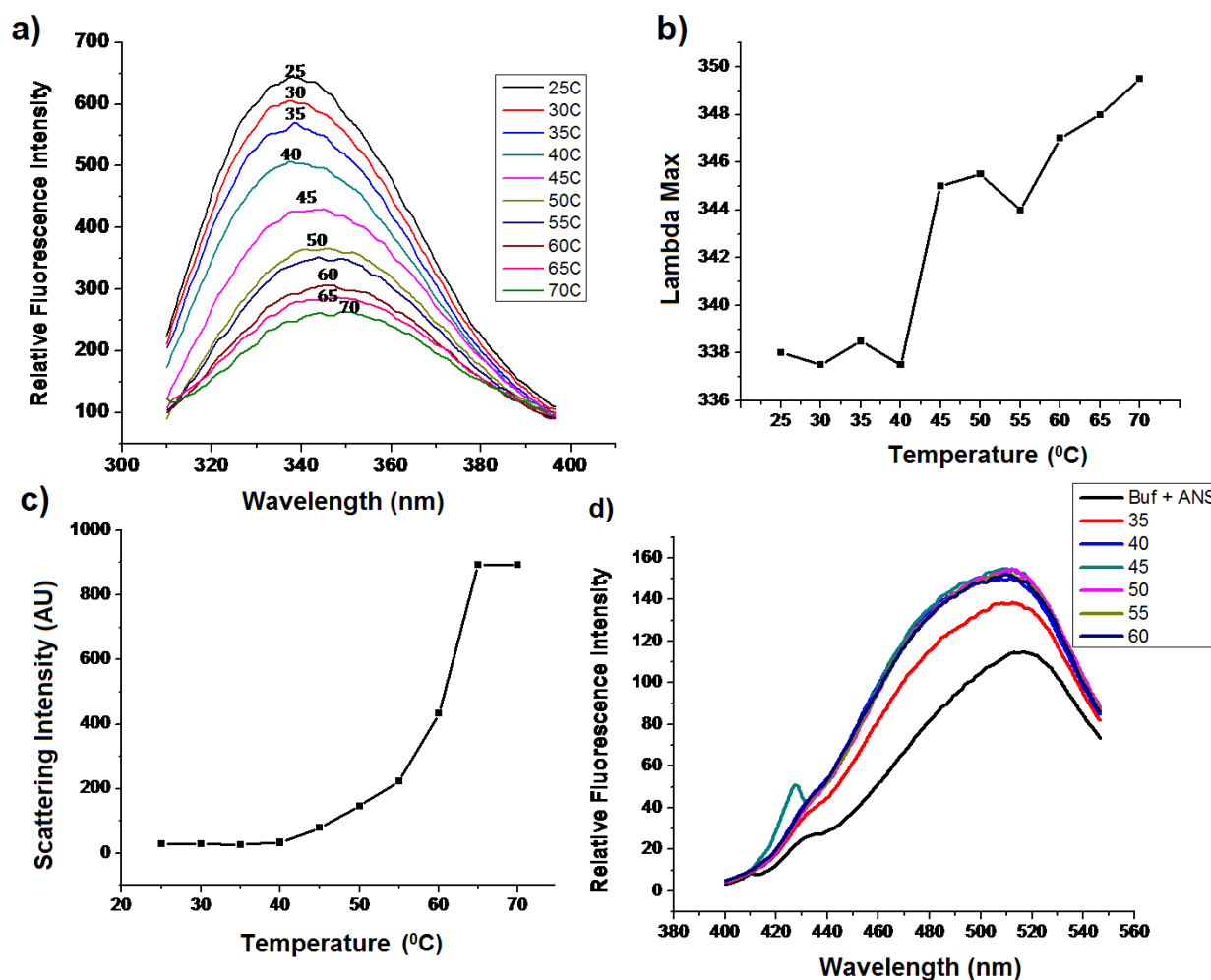


Figure 5.9: a) Trp fluorescence spectra of *Pd*PGA subjected to various temperatures from 25 to 70 °C. b) Line plot showing the increase of λ_{max} of fluorescence spectra (red shift) of *Pd*PGA at higher temperatures. c) Line plot depicting the scattering intensities of *Pd*PGA at various temperatures. d) ANS binding spectra of *Pd*PGA under various temperatures. The black line corresponds to the spectrum of buffer with ANS.

5.3.6 Probing structural changes using fluorescence

Compared to tyrosine and phenylalanine residues, tryptophan (Trp) residues exhibit a much stronger fluorescence on excitation. Since Trp fluorescence is dependent on its microenvironment even minor changes in this microenvironment can lead to changes in the Trp fluorescence spectra. *PdPGA* has 20 Trp residues of which 5 residues were found to be within 10 Å of its N-terminal catalytic residue β Ser1 based on modeled structure. Thus heat-induced conformational changes in *PdPGA* could be easily monitored by recording Trp fluorescence spectra at various temperatures. The λ_{max} in intrinsic fluorescence spectrum of *PdPGA* was near 338 nm between 25 to 40 °C (Fig. 5.9). However, a red-shift to 345 nm at 45 °C indicated a possible alteration of Trp microenvironment to a partially hydrophilic nature due to conformational change in the protein upon heat treatment. These observations could be correlated with the observed 40% decrease in enzyme activity at 45 °C as well as the complete loss of activity at 50 °C. Beyond 50 °C further red-shift was observed reaching 350 nm at 70 °C. Rayleigh light scattering experiment carried out using the above experimental setup was used to study protein aggregation due to heat-treatment (Fig. 5.9). A gradual increase in scattering intensity after 45 °C indicated protein aggregation due to heat-induced denaturation resulting in loss of enzyme activity. ANS hydrophobic dye binding study did not show binding of ANS either to native enzyme or heat-denatured enzyme, indicating not much exposure of hydrophobic patches during thermal denaturation (Fig. 5.9).

5.3.7 Circular Dichroism study

The change in secondary structure of *PdPGA* in the course of thermal denaturation was observed in Far UV CD analysis. Figure 5.10 shows the CD-spectra of *PdPGA* from 25 to 70 °C. Gradual reduction in ellipticity above 45 °C correlated with the loss of activity as well as altered Trp environment. The CD-pro analysis showed a slight decrease in α -helical content but a significant decrease in β -sheet content of enzyme between 45 and 50 °C compared to 40 °C (Table 5.1). The decrease in ordered secondary structure content led to an increase in turns and unordered regions. This observation can be linked with the possible unfolding of the $\alpha\beta\beta\alpha$ core Ntn-hydrolase fold which contains the catalytic site, thus leading to loss of enzyme activity.

Consensus based analysis identified 24 different sites on *Pd*PGA where conserved stabilizing residues of *Af*PGA and *Ax*PGA can be introduced for improving its thermostability (Table 4.5). Bidirectional stability prediction analysis further filtered 6 substitutions ($\alpha 80\text{Ser}\rightarrow\text{Arg}$, $\beta 218\text{Gln}\rightarrow\text{Leu}$, $\beta 280\text{Asp}\rightarrow\text{Gln}$, $\beta 308\text{Glu}\rightarrow\text{Gln}$, $\beta 436\text{Asp}\rightarrow\text{Gln}$ and $\beta 457\text{Ser}\rightarrow\text{Arg}$; residue numbering as per *Ec*PGA) which might show higher potential of thermostabilization when introduced in *Pd*PGA.

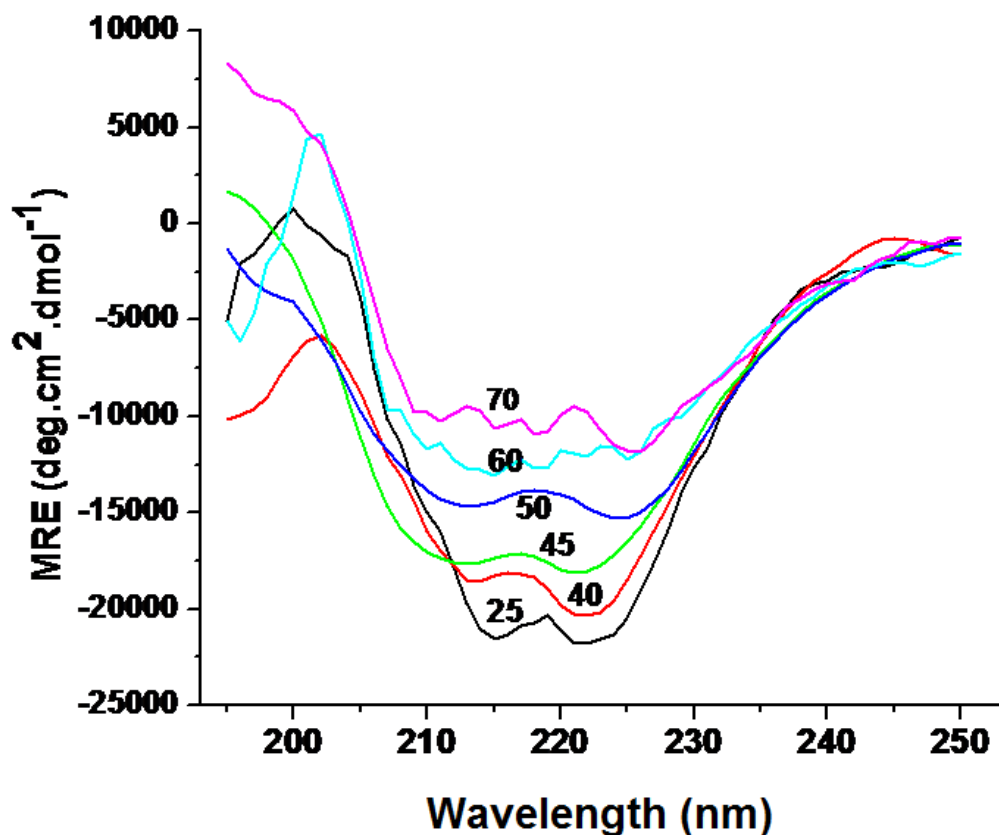


Figure 5.10: Far UV CD spectra of *Pd*PGA subjected to various heat treatments.

Table 5.1: Percentages of various secondary structure elements of *Pd*PGA estimated at various temperatures using CDPro software.

Temperature (°C)	%Helix	%Sheet	%Turn	%Unordered
40	47	23	13	15
45	44	12	17	24
50	43	11	19	24

5.4 Summary

The current study highlights the complex sequence-structure-function relationship in the PGA family of enzymes. Though several mechanisms for protein structure stabilization have already been reported, the lack of any dominant or universal strategy which could be applied to all enzymes makes it difficult to screen novel candidates based on a single strategy. A multiple strategy based analysis has been carried out in case of PGAs (Chapter 4) to identify a possible thermostable candidate. From the MEROPS database, three sources of putative thermostable PGAs (ptPGAs: *Sw*PGA, *Pd*PGA and *Ao*PGA) were identified having possibility to have a disulfide bond similar to thermostable *Af*PGA enzyme. The *consensus sequence-based* analysis (Chapter 4) classified these three identified ptPGAs under consideration to a mesostable group. The analysis also highlighted that ptPGAs could potentially be more thermostable as compared to the widely used *Ec*PGA but lesser stable than both *Ax*PGA and *Af*PGA. Of the 24 sites identified, three specific sites could be short-listed as potential targets for mutation-based thermostabilization studies for these mesostable enzymes. Of the three, the site corresponding to $\alpha 80$ of *Ec*PGA was found to be the most promising site which was also in accordance with available reports. Extensive *structural analysis* carried out on the ptPGAs and their comparison with the other PGAs singled out *Pd*PGA as the most potent candidate for the exploration of its thermostable nature (Chapter 4).

Presence of disulfide bonds which was considered as a screening criterion was indeed found to be a positive contributor to *Pd*PGA stability. *Pd*PGA on treatment with 50 and 100 mM reducing agent DTT showed 40% and 80% decrease in enzyme activity at 40 °C. The temperature stability profile of *Pd*PGA was observed to be intermediate to that of *Ec*PGA and *Af*PGA/*Ax*PGA.

*Pd*PGA displayed a broad range of pH stability and was found to be particularly unique in terms of its optimum pH which was found to be at pH 10, showing its distinct alkaline stability. The broad range of pH stability of *Pd*PGA could be attributed to the presence of extensive ion-pair networks. Stability at broader pH range has vital industrial advantages especially during processes like purification, immobilization as well as chemical modifications. In their recent attempts towards making *Ec*PGA more industrially useful, Suplatov *et al.*, 2014, have attempted to make *Ec*PGA alkaline stable by carrying out mutation of β Asp484 residue to

Asn, which resulted in 9-fold increase of its stability at pH 10 (Suplatov *et al.*, 2014). In case of *Pd*PGA, it was observed that at the position corresponding to *Ec*PGA β Asp484, an Asn residue is naturally present, which could act as a major contributing factor towards its unique alkaline stability.

In our approach towards prediction of thermostability, consensus site-specific sequence-based approach gave a more realistic estimate of *Pd*PGA thermostability although the known thermostability factors were overwhelmingly in its favour. Unfortunately, the results from experimental characterization of *Pd*PGA thermostability limits its potent application as an industrial biocatalyst, however, the sustained effort towards increasing its thermal stability through substitution mutation of the identified sites (Chapter 4), could transform this uniquely alkaline stable PGA into an extremely viable industrial biocatalyst.

Chapter 6

Conclusions

Two pharmaceutically and medicinally important families of enzymes (CGH and PGA) belonging to Ntn-hydrolase superfamily have been studied in this thesis. The work involved the use of computational approaches along with experimental analysis to study the structure-function relationship among these enzymes.

Substrate specificity annotation remains a challenging problem for members of CGH family due to high degree of sequence similarity among themselves. The CGH family enzymes such as BSH and PVA had been grouped together under a single family and there was a need for development of sequence-based method for their distinction. In this thesis, we have combined phylogenetic information along with binding site and substrate specificity information of CGH enzymes in order to develop Binding Site Similarity (BSS) based annotation method which allowed annotation of CGH family members accurately into BSH or PVA. This new method was validated with all experimentally characterized BSH/PVA enzymes as well as earlier functional annotations given by Lambert *et al.*, 2008. A total of 198 representative family members were considered as local dataset for annotation by BSS based scoring system. Solely based on sequence, this method could be used for functional annotation of BSH and PVA enzymes. The source code of computer program developed and written in Perl is freely available upon request. The study also deciphered the mode of substrate binding among BSH and PVA enzymes. Polar complementarity was observed as the basis for bile salt binding affinity while for penicillin V binding, aromatic interaction in the binding site was observed as influencing factor. Detailed understanding of the enzyme-substrate interactions in these enzymes would further help to tailor the substrate specificities to desired levels and design novel enzymes for pharmaceutical applications. The phylogenetic analysis identified an important evolutionary characteristic amongst BSH/PVA members of Gram-positive and Gram-negative members that they form two distinct sub-families of enzymes which diverge not only with respect to their structural chemistry near the substrate binding site but also with respect to their quaternary structure assembly. An important feature identified was, the absence of tetramer assembly motif among members of Cluster2 (dominated by Gram-negative bacteria) members. Theoretical analysis showed the importance of the tetramer assembly in attaining thermodynamic stabilities of the quaternary assemblies. The evolution of CGH family members could be related to antibiotics-selection pressure hypothesis of Gupta *et al.*, 2011. While analyzing the distributions of BSH and PVA enzymes, it was observed that PVA enzymes were distributed among many environmental

microbes involved in bioremediation and pathogenesis. Currently the entire focus of research in PVA enzymes is towards their industrial application for synthesis of semi-synthetic penicillins. However, the current findings have opened up new dimension towards investigation to decipher the true physiological role of these enzymes and understand the structural and functional role of tetramer assembly motif.

The iRDP web server developed as part of this research is freely available at <http://irdp.ncl.res.in>. It includes three independent modules namely iCAPS, iStability and iMutants, which can either be used separately or can be used as a pipeline to solve any protein engineering problem. **iCAPS** can be used to compare large number of protein structures simultaneously. Currently maximum 100 structures can be compared with respect to 20 different structural stabilization mechanisms and 288 different structural parameters that are known to contribute to protein stability and function. Users can not only analyze the known PDB files but can upload any valid PDB formatted files. For example, users can upload molecular dynamics simulation trajectory frames and study the dynamics of various non-bonded interaction networks in a protein. Similarly NMR ensemble models and protein family members can also be analyzed. The detection of non-covalent interaction networks is one of the most important features of iCAPS, which is currently not available in any public domain structural analysis tools. Another unique advantage about iCAPS module is the introduction of features such as analysis of helix dipole stabilization, conformational strain release, classification of residues in terms of their solvent accessibility and few others, which is available for the first time to users. Users can download all the results in a single zip file. The files are tab-delimited so that users can further analyze the results using other statistical softwares such as Excel and R. Since iCAPS generates 288 quantitative features for every input protein, these 288 parameters can be considered as input feature vector in many machine learning tools such as support vector machine, support vector regression and neural networks, for prediction of any function.

iStability incorporates four stability prediction tools and provides an interface in which user can identify potential stabilizing sites in the input protein structure. In near future we are planning to integrate more protein design strategies and integrate other stability prediction tools. The unique feature of iStability is that, if a user does not have prior information of mutations to carry out, the user can identify potential thermostabilizing mutations based on selected strategy.

The third module **iMutants** is a novel tool incorporated in iRDP server using which user can get an in depth knowledge about the structural and interaction changes that might happen at the mutation site due to mutation. The local interaction profile gives a quick summary of the loss/gain of interactions due to mutations. Users can also download the mutants in PDB format for further downstream analysis. The evolutionary conservation score given for the mutation site is an important parameter which could help users to determine whether a particular site should be considered for mutation or not. Finally iRDP also provides **iATMs** information resource, in which iMutants analysis was extended for all experimentally characterized single/double/multiple mutants present in ProTherm database. The information available in iATMs database will help researchers to correlate the experimental parameters with structural parameters, and thus could help in efficient protein designing.

PGA enzymes are NtSn-hydrolases that have wide applications in the antibiotics industry. With the objective of identifying novel sources of thermostable PGA enzymes, a computational approach of sequence and structure analysis was used through which we could identify three putative thermostable PGA enzymes of which *PdPGA* (PGA from *Paracoccus denitrificans*) was found to be the most potential thermostable PGA, possessing many features of thermostabilization. The consensus site-specific sequence-based approach predicted *PdPGA* to be more thermostable than *EcPGA* but not as thermostable as *AxPGA*. Experimental verification proved this correct, although several thermostability factors favoured a much higher thermostability for the enzyme. *PdPGA* has an advantage of being active at alkaline pH. Another positive outcome of the analysis was the identification of mutations which could increase the thermostability of *PdPGA*. Site-specific analysis identified 24 sites of thermostabilization which can be considered for site-directed mutagenesis among the PGA enzymes to enhance their stability. Stability prediction analysis helped to further filter down these sites to identify most potent sites. Amongst all sites, $\alpha 80$ site was found to be the most-potent site for which experimental validation is also available. *PdPGA* enzyme could be useful in industry for its comparatively higher pH stability behavior. The future direction could be towards making this enzyme industrially more useful by incorporating the identified stabilizing mutations to make *PdPGA* both pH stable as well as thermostable. The study also identified *SwPGA* and *AoPGA* as other two enzymes, which could be considered for experimental characterization. Since the computational analysis carried out for *PdPGA* enzyme is based on homology model,

determination of the actual three-dimensional structure of this enzyme could prove to be more informative in correlating structure with the experimental findings.

In summary, stability and substrate specificities of members belonging to two groups of Ntn-hydrolase family have been studied using various computational, biochemical and biophysical techniques. These studies have provided useful insights into the complex structure-function relationship of these enzymes which would further help to improve their application potential and design better enzymes for industrial and therapeutic purpose. The BSS method introduced for the CGH family explores an additional dimension in the field of enzyme annotation by taking into consideration the micro-environment of the binding sites. The implementation of the iRDP web server represents a major effort to introduce *in silico*, multiple aspects considered in experimental characterization of thermostability in enzymes. While the iCAPS and iStability are primarily analytical in nature, iMutants uniquely investigates mutation sites through molecular interactions thereby adding a new perspective to the field of designing of proteins by means of mutations. Finally the site-specific consensus analysis combined with bi-directional stability prediction approach among PGA enzymes could be extended to other enzyme families towards identification of potential thermostabilization sites.

BIBLIOGRAPHY

- Abraham, E. P. (1981). The beta-lactam antibiotics. *Sci Am* 244, 76-86.
- Agah, S., Larson, J. D. & Henzl, M. T. (2003). Impact of proline residues on parvalbumin stability. *Biochemistry* 42, 10886-10895.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-3402.
- Ambedkar, S. S., Deshpande, B. S., Sudhakaran, V. K. & Shewale, J. G. (1991). Beijerinckia indica var. penicillanicum penicillin V acylase: enhanced enzyme production by catabolite repression-resistant mutant and effect of solvents on enzyme activity. *J Ind Microbiol* 7, 209-214.
- Anderson, D. E., Becktel, W. J. & Dahlquist, F. W. (1990). pH-induced denaturation of proteins: a single salt bridge contributes 3-5 kcal/mol to the free energy of folding of T4 lysozyme. *Biochemistry* 29, 2403-2408.
- Andreini, C., Cavallaro, G. & Lorenzini, S. (2012). FindGeo: a tool for determining metal coordination geometry. *Bioinformatics* 28, 1658-1660.
- Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science* 181, 223-230.
- Antikainen, N. M. & Martin, S. F. (2005). Altering protein specificity: techniques and applications. *Bioorganic & medicinal chemistry* 13, 2701-2716.
- Aplin, R. T., Baldwin, J. E., Cole, S. C., Sutherland, J. D. & Tobin, M. B. (1993). On the production of alpha, beta-heterodimeric acyl-coenzyme A: isopenicillin N-acyltransferase of Penicillium chrysogenum. Studies using a recombinant source. *FEBS Lett* 319, 166-170.
- Arroyo, M., de la Mata, I., Acebal, C. & Castillon, M. P. (2003). Biotechnological applications of penicillin acylases: state-of-the-art. *Appl Microbiol Biotechnol* 60, 507-514.
- Artymiuk, P. J. (1995). A sting in the (N-terminal) tail. *Nat Struct Biol* 2, 1035-1037.
- Auerbach, G., Huber, R., Grättinger, M., Zaiss, K., Schurig, H., Jaenicke, R. & Jacob, U. (1997). Closed structure of phosphoglycerate kinase from Thermotoga maritima reveals the catalytic mechanism and determinants of thermal stability. *Structure* 5, 1475-1483.
- Auerbach, G., Ostendorp, R., Prade, L., Korndörfer, I., Dams, T., Huber, R. & Jaenicke, R. (1998). Lactate dehydrogenase from the hyperthermophilic bacterium Thermotoga maritima: the crystal structure at 2.1 Å resolution reveals strategies for intrinsic protein stabilization. *Structure* 6, 769-781.
- Avinash, V. S., Panigrahi, P., Suresh, C. G., Pundle, A. V. & Ramasamy, S. (2013). Structural modelling of substrate binding and inhibition in penicillin V acylase from Pectobacterium atrosepticum. *Biochem Biophys Res Commun* 437, 538-543.
- Baker, E. N. & Hubbard, R. E. (1984). Hydrogen bonding in globular proteins. *Progress in biophysics and molecular biology* 44, 97-179.

Baker, S. C., Ferguson, S. J., Ludwig, B., Page, M. D., Richter, O. M. & van Spanning, R. J. (1998). Molecular genetics of the genus *Paracoccus*: metabolically versatile bacteria with bioenergetic flexibility. *Microbiol Mol Biol Rev* 62, 1046-1078.

Barends, T. R., Yoshida, H. & Dijkstra, B. W. (2004). Three-dimensional structures of enzymes useful for beta-lactam antibiotic production. *Curr Opin Biotechnol* 15, 356-363.

Batchelor, F. R., Doyle, F. P., Nayler, J. H. C. & Rolinson, G. N. (1959). Synthesis of Penicillin: 6-Aminopenicillanic Acid in Penicillin Fermentations. *Nature* 183, 257-258.

Batta, A. K., Salen, G. & Shefer, S. (1984). Substrate specificity of cholyglycine hydrolase for the hydrolysis of bile acid conjugates. *J Biol Chem* 259, 15035-15039.

Baumann, B., Snozzi, M., Zehnder, A. J. & Van Der Meer, J. R. (1996). Dynamics of denitrification activity of *Paracoccus denitrificans* in continuous culture during aerobic-anaerobic changes. *J Bacteriol* 178, 4367-4374.

Begley, M., Sleator, R. D., Gahan, C. G. & Hill, C. (2005). Contribution of three bile-associated loci, *bsh*, *pva*, and *btlB*, to gastrointestinal persistence and bile tolerance of *Listeria monocytogenes*. *Infect Immun* 73, 894-904.

Begley, M., Hill, C. & Gahan, C. G. (2006). Bile salt hydrolase activity in probiotics. *Appl Environ Microbiol* 72, 1729-1738.

Béguin, P. (1999). Hybrid enzymes. *Current opinion in biotechnology* 10, 336-340.

Ben-Bassat, A., Bauer, K., Chang, S. Y., Myambo, K., Boosman, A. & Chang, S. (1987). Processing of the initiation methionine from proteins: properties of the *Escherichia coli* methionine aminopeptidase and its gene structure. *J Bacteriol* 169, 751-757.

Bennett, T. P. & Frieden, E. (1969). *Modern Topics in Biochemistry*, pg 43-45, Macmillan, London.

Berg, J., Tymoczko, J. & Stryer, L. (2002). *Enzymes Accelerate Reactions by Facilitating the Formation of the Transition State*. New York: Biochemistry. 5th edition. New York: W H Freeman; 2002., Available from: <http://www.ncbi.nlm.nih.gov/books/NBK22431/>.

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

Betz, S. F. (1993). Disulfide bonds and the stability of globular proteins. *Protein Science* 2, 1551-1558.

Bezsonova, I., Singer, A., Choy, W.-Y., Tollinger, M. & Forman-Kay, J. D. (2005). Structural comparison of the unstable drkN SH3 domain and a stable mutant. *Biochemistry* 44, 15550-15560.

Bjørk, A., Dalhus, B., Mantzilas, D., Sirevåg, R. & Eijsink, V. G. (2004). Large improvement in the thermal stability of a tetrameric malate dehydrogenase by single point mutations at the dimer-dimer interface. *Journal of molecular biology* 341, 1215-1226.

Blundell, T. L., Elliott, G., Gardner, S. P. & other authors (1989). Protein Engineering and Design. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 324, 447-460.

- Bogin, O., Peretz, M., Hacham, Y., Burstein, Y., Korkhin, Y. & Frolow, F. (1998). Enhanced thermal stability of *Clostridium beijerinckii* alcohol dehydrogenase after strategic substitution of amino acid residues with prolines from the homologous thermophilic *Thermoanaerobacter brockii* alcohol dehydrogenase. *Protein science* 7, 1156-1163.
- Bokhove, M., Nadal Jimenez, P., Quax, W. J. & Dijkstra, B. W. (2010a). The quorum-quenching N-acyl homoserine lactone acylase PvdQ is an Ntn-hydrolase with an unusual substrate-binding pocket. *Proc Natl Acad Sci U S A* 107, 686-691.
- Bokhove, M., Yoshida, H., Hensgens, C. M., van der Laan, J. M., Sutherland, J. D. & Dijkstra, B. W. (2010b). Structures of an isopenicillin N converting Ntn-hydrolase reveal different catalytic roles for the active site residues of precursor and mature enzyme. *Structure* 18, 301-308.
- Bomstein, J. & Evans, W. G. (1965). Automated Colorimetric Determination of 6-Aminopenicillanic Acid in Fermentation Media. *Analytical Chemistry* 37, 576-578.
- Borgo, B. & Havranek, J. J. (2012). Automated selection of stabilizing mutations in designed and natural proteins. *Proceedings of the National Academy of Sciences* 109, 1494-1499.
- Bornhorst, J. A. & Falke, J. J. (2000). Purification of proteins using polyhistidine affinity tags. *Methods Enzymol* 326, 245-254.
- Brannigan, J. A., Dodson, G., Duggleby, H. J., Moody, P. C. E., Smith, J. L., Tomchick, D. R. & Murzin, A. G. (1995). A protein catalytic framework with an N-terminal nucleophile is capable of self-activation. *Nature* 378, 416-419.
- Bruggink, A. & Roy, P. (2001). Industrial Synthesis of Semisynthetic Antibiotics. In *Synthesis of β -Lactam Antibiotics*, pp. 12-54: Springer Netherlands.
- Burley, S. & Petsko, G. (1985). Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science* 229, 23-28.
- Burlingame, R. & Chapman, P. J. (1983). Catabolism of phenylpropionic acid and its 3-hydroxy derivative by *Escherichia coli*. *J Bacteriol* 155, 113-121.
- Buse, M. G. (2006). Hexosamines, insulin resistance, and the complications of diabetes: current status. *Am J Physiol Endocrinol Metab* 290, E1-E8.
- Cai, G., Zhu, S., Yang, S., Zhao, G. & Jiang, W. (2004). Cloning, overexpression, and characterization of a novel thermostable penicillin G acylase from *Achromobacter xylosoxidans*: probing the molecular basis for its high thermostability. *Appl Environ Microbiol* 70, 2764-2770.
- Cao, L., van Rantwijk, F. & Sheldon, R. A. (2000). Cross-linked enzyme aggregates: a simple and effective method for the immobilization of penicillin acylase. *Org Lett* 2, 1361-1364.
- Capriotti, E., Fariselli, P. & Casadio, R. (2005). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic acids research* 33, W306-W310.
- Carlsen, F. & Emborg, C. (1982). *Bacillus sphaericus* V-penicillin acylase. II. Isolation and characterisation. *Journal of Chemical Technology and Biotechnology* 32, 808-811.

- Chakravarty, S. & Varadarajan, R. (2000). Elucidation of determinants of protein stability through genome sequence analysis. *Febs Letters* 470, 65-69.
- Chandra, P. M., Brannigan, J. A., Prabhune, A., Pundle, A., Turkenburg, J. P., Dodson, G. G. & Suresh, C. G. (2005). Cloning, preparation and preliminary crystallographic studies of penicillin V acylase autoproteolytic processing mutants. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 61, 124-127.
- Cheng, J., Randall, A. & Baldi, P. (2006). Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins: Structure, Function, and Bioinformatics* 62, 1125-1132.
- Christiaens, H., Leer, R. J., Pouwels, P. H. & Verstraete, W. (1992). Cloning and expression of a conjugated bile acid hydrolase gene from *Lactobacillus plantarum* by using a direct plate assay. *Appl Environ Microbiol* 58, 3792-3798.
- Cole, M. (1964). Properties of the Penicillin Deacylase Enzyme of *Escherichia coli*. *Nature* 203, 519-520.
- Coleman, J. P. & Hudson, L. L. (1995). Cloning and characterization of a conjugated bile acid hydrolase gene from *Clostridium perfringens*. *Appl Environ Microbiol* 61, 2514-2520.
- Colovos, C. & Yeates, T. O. (1993). Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci* 2, 1511-1519.
- Costantini, S., Colonna, G. & Facchiano, A. M. (2008). ESBRI: a web server for evaluating salt bridges in proteins. *Bioinformation* 3, 137-138.
- Csardi, G. & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems* 1695.
- Daumy, G. O., Danley, D. & McColl, A. S. (1985). Role of protein subunits in *Proteus rettgeri* penicillin G acylase. *J Bacteriol* 163, 1279-1281.
- Davoodi, J., Wakarchuk, W. W., Carey, P. R. & Surewicz, W. K. (2007). Mechanism of stabilization of *Bacillus circulans* xylanase upon the introduction of disulfide bonds. *Biophysical chemistry* 125, 453-461.
- Deckert, G., Warren, P. V., Gaasterland, T. & other authors (1998). The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* 392, 353-358.
- Dehouck, Y., Kwasigroch, J. M., Gilis, D. & Rooman, M. (2011). PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC bioinformatics* 12, 151.
- Delpino, M. V., Marchesini, M. I., Estein, S. M., Comerci, D. J., Cassataro, J., Fossati, C. A. & Baldi, P. C. (2007). A bile salt hydrolase of *Brucella abortus* contributes to the establishment of a successful infection through the oral route in mice. *Infect Immun* 75, 299-305.
- Demain, A., Solomon, N. & Abraham, E. P. (1983). History of beta-Lactam Antibiotics. In *Antibiotics*, pp. 1-14: Springer Berlin Heidelberg.
- Deuschl, F., Kollmann, K., von Figura, K. & Lubke, T. (2006). Molecular characterization of the hypothetical 66.3-kDa protein in mouse: lysosomal targeting, glycosylation, processing and tissue distribution. *FEBS Lett* 580, 5747-5752.

- Dijkstra, B. W., Smith, J. L. & Salvesen, N. D. R. (2013). Chapter 809 - Self-Processing Cysteine-Dependent N-terminal Nucleophile Hydrolases. In *Handbook of Proteolytic Enzymes*, pp. 3653-3657: Academic Press.
- Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry* 29, 7133-7155.
- Ding, H., Gao, F., Liu, D., Li, Z., Xu, X., Wu, M. & Zhao, Y. (2013). Significant improvement of thermal stability of glucose 1-dehydrogenase by introducing disulfide bonds at the tetramer interface. *Enzyme and microbial technology* 53, 365-372.
- Dougherty, D. A. (1996). Cation- π interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp. *Science* 271, 163-168.
- Drasar, B. S., Hill, M. J. & Shiner, M. (1966). The deconjugation of bile salts by human intestinal bacteria. *Lancet* 1, 1237-1238.
- Dudley, E. G., Husgen, A. C., He, W. & Steele, J. L. (1996). Sequencing, distribution, and inactivation of the dipeptidase A gene (pepDA) from *Lactobacillus helveticus* CNRZ32. *J Bacteriol* 178, 701-704.
- Duggleby, H. J., Tolley, S. P., Hill, C. P., Dodson, E. J., Dodson, G. & Moody, P. C. (1995). Penicillin acylase has a single-amino-acid catalytic centre. *Nature* 373, 264-268.
- Dussurget, O., Cabanes, D., Dehoux, P., Lecuit, M., Buchrieser, C., Glaser, P. & Cossart, P. (2002). *Listeria monocytogenes* bile salt hydrolase is a PrfA-regulated virulence factor involved in the intestinal and hepatic phases of listeriosis. *Mol Microbiol* 45, 1095-1106.
- Eisenberg, D., Luthy, R. & Bowie, J. U. (1997). VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol* 277, 396-404.
- Elkins, C. A. & Savage, D. C. (1998). Identification of genes encoding conjugated bile salt hydrolase and transport in *Lactobacillus johnsonii* 100-100. *J Bacteriol* 180, 4344-4349.
- Erarslan, A. & Kocer, H. (1992). Thermal inactivation kinetics of penicillin G acylase obtained from a mutant derivative of *Escherichia coli* ATCC 11105. *J Chem Technol Biotechnol* 55, 79-84.
- Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M., Eramian, D., Shen, M. y., Pieper, U. & Sali, A. (2006). Comparative protein structure modeling using Modeller. *Current protocols in bioinformatics*, 5.6. 1-5.6. 30.
- Facchiano, A. M., Colonna, G. & Ragone, R. (1998). Helix stabilizing factors and stabilization of thermophilic proteins: an X-ray based study. *Protein engineering* 11, 753-760.
- Friedrich, C. G. & Mitrenga, G. (1981). Oxidation of thiosulfate by *Paracoccus denitrificans* and other hydrogen bacteria. *FEMS Microbiology Letters* 10, 209-212.
- Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., Sanschagrin, P. C. & Mainz, D. T. (2006). Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem* 49, 6177-6196.

Fu, H., Grimsley, G. R., Razvi, A., Scholtz, J. M. & Pace, C. N. (2009). Increasing protein stability by improving beta-turns. *Proteins: Structure, Function, and Bioinformatics* 77, 491-498.

Fuganti, C., Grasselli, P. & Casati, P. (1986a). Immobilized penicillin acylase: application to the synthesis of the dipeptide aspartame. *Tetrahedron Letters* 27, 3191-3194.

Fuganti, C., Grasselli, P., Seneci, P. F., Servi, S. & Casati, P. (1986b). Immobilized benzylpenicillin acylase: Application to the synthesis of optically active forms of carnitin and propranolol. *Tetrahedron Letters* 27, 2061-2062.

Gallastegui, N. & Groll, M. (2010). The 26S proteasome: assembly and function of a destructive machine. *Trends Biochem Sci* 35, 634-642.

Gallivan, J. P. & Dougherty, D. A. (1999). Cation- π interactions in structural biology. *Proceedings of the National Academy of Sciences* 96, 9459-9464.

Giollo, M., Martin, A. J., Walsh, I., Ferrari, C. & Tosatto, S. C. (2014). NeEMO: a method using residue interaction networks to improve prediction of protein stability upon mutation. *BMC Genomics* 15, S7.

Goldenberg, O., Erez, E., Nimrod, G. & Ben-Tal, N. (2009). The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. *Nucleic acids research* 37, D323-D327.

Gough, J., Karplus, K., Hughey, R. & Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* 313, 903-919.

Grant, B. J., Rodrigues, A. P., ElSawy, K. M., McCammon, J. A. & Caves, L. S. (2006). Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* 22, 2695-2696.

Groll, M., Brandstetter, H., Bartunik, H., Bourenkow, G. & Huber, R. (2003). Investigations on the maturation and regulation of archaeobacterial proteasomes. *J Mol Biol* 327, 75-83.

Guan, C., Cui, T., Rao, V., Liao, W., Benner, J., Lin, C. L. & Comb, D. (1996). Activation of glycosylasparaginase. Formation of active N-terminal threonine by intramolecular autoproteolysis. *J Biol Chem* 271, 1732-1737.

Guerois, R., Nielsen, J. E. & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of molecular biology* 320, 369-387.

Guex, N. & Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 18, 2714-2723.

Gupta, R. S. (2011). Origin of diderm (Gram-negative) bacteria: antibiotic selection pressure rather than endosymbiosis likely led to the evolution of bacterial cells with two membranes. *Antonie Van Leeuwenhoek* 100, 171-182.

Gutteridge, A. & Thornton, J. (2004). Conformational change in substrate binding, catalysis and product release: an open and shut case? *FEBS Letters* 567, 67-73.

Halgren, T. (2007). New method for fast and accurate binding-site identification and analysis. *Chem Biol Drug Des* 69, 146-148.

- Han, Z.-l., Han, S.-y., Zheng, S.-p. & Lin, Y. (2009). Enhancing thermostability of a *Rhizomucor miehei* lipase by engineering a disulfide bond and displaying on the yeast cell surface. *Applied microbiology and biotechnology* 85, 117-126.
- Haupt, V. J., Daminelli, S. & Schroeder, M. (2013). Drug Promiscuity in PDB: Protein Binding Site Similarity Is Key. *PLoS ONE* 8, e65894.
- Hay, J. W., Yu, W. M. & Ashraf, T. (1999). Pharmacoeconomics of lipid-lowering agents for primary and secondary prevention of coronary artery disease. *Pharmacoeconomics* 15, 47-74.
- Hazes, B. & Dijkstra, B. W. (1988). Model building of disulfide bonds in proteins with known three-dimensional structure. *Protein engineering* 2, 119-125.
- Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 89, 10915-10919.
- Hennig, M., Darimont, B., Sterner, R., Kirschner, K. & Jansonius, J. N. (1995). 2.0 Å structure of indole-3-glycerol phosphate synthase from the hyperthermophile *Sulfolobus solfataricus*: possible determinants of protein stability. *Structure* 3, 1295-1306.
- Herning, T., Yutani, K., Inaka, K., Kuroki, R., Matsushima, M. & Kikuchi, M. (1992). Role of proline residues in human lysozyme stability: a scanning calorimetric study combined with X-ray structure analysis of proline mutants. *Biochemistry* 31, 7077-7085.
- Herschlag, D. (1988). The role of induced fit and conformational changes of enzymes in specificity and catalysis. *Bioorganic Chemistry* 16, 62-96.
- Hershko, A. & Ciechanover, A. (1992). The ubiquitin system for protein degradation. *Annu Rev Biochem* 61, 761-807.
- Hubbard, S. J. & Thornton, J. M. (1993). Naccess. *Computer Program, Department of Biochemistry and Molecular Biology, University College London 2*.
- Hutchinson, E. G. & Thornton, J. M. (1996). PROMOTIF—a program to identify and analyze structural motifs in proteins. *Protein Science* 5, 212-220.
- Iqbal, A., Clifton, I. J., Chowdhury, R., Ivison, D., Domene, C. & Schofield, C. J. (2011). Structural and biochemical analyses reveal how ornithine acetyl transferase binds acidic and basic amino acid substrates. *Org Biomol Chem* 9, 6219-6225.
- Jäger, M., Dendle, M. & Kelly, J. W. (2009). Sequence determinants of thermodynamic stability in a WW domain—An all- β -sheet protein. *Protein Science* 18, 1806-1813.
- John, P. & Whatley, F. R. (1975). *Paracoccus denitrificans* and the evolutionary origin of the mitochondrion. *Nature* 254, 495-498.
- Johnson, L. R. (2003). *Bile secretion and gall bladder function, Essential Medical Physiology, p520-528*: Elsevier Science.

Johnson, M. R., Barnes, S., Sweeny, D. J. & Diasio, R. B. (1990). 2-Fluoro-beta-alanine, a previously unrecognized substrate for bile acid coenzyme A:amino acid:N-acyltransferase from human liver. *Biochem Pharmacol* 40, 1241-1246.

Jones, B. V., Begley, M. i., Hill, C., Gahan, C. G. M. & Marchesi, J. R. (2008). Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. *Proceedings of the National Academy of Sciences* 105, 13580-13585.

Jones, M. L., Chen, H., Ouyang, W., Metz, T. & Prakash, S. (2004). Microencapsulated Genetically Engineered *Lactobacillus plantarum* 80 (pCBH1) for Bile Acid Deconjugation and Its Implication in Lowering Cholesterol. *J Biomed Biotechnol* 2004, 61-69.

Joon Cho, K., Hyun Kim, K. & Salvesen, N. D. R. (2013). Chapter 811 - Cephalosporin Acylase Precursor, Glutaryl-7-aminocephalosporanic Acid Acylase Precursor. In *Handbook of Proteolytic Enzymes*, pp. 3659-3663: Academic Press.

Jordan, S. L., McDonald, I. R., Kraczkiewicz-Dowjat, A. J., Kelly, D. P., Rainey, F. A., Murrell, J. C. & Wood, A. P. (1997). Autotrophic growth on carbon disulfide is a property of novel strains of *Paracoccus denitrificans*. *Arch Microbiol* 168, 225-236.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577-2637.

Kasumi, T., Hayashi, K. & Tsumura, N. (1982). Roles of magnesium and cobalt in the reaction of glucose isomerase from *Streptomyces griseofuscus* S-41. *Agricultural and Biological Chemistry* 46, 21-30.

Katayama, Y., Hiraishi, A. & Kuraishi, H. (1995). *Paracoccus thiocyanatus* sp. nov., a new species of thiocyanate-utilizing facultative chemolithotroph, and transfer of *Thiobacillus versutus* to the genus *Paracoccus* as *Paracoccus versutus* comb. nov. with emendation of the genus. *Microbiology* 141 (Pt 6), 1469-1477.

Katchalski-Katzir, E. & Kraemer, D. M. (2000). Eupergit® C, a carrier for immobilization of enzymes of industrial potential. *Journal of Molecular Catalysis B: Enzymatic* 10, 157-176.

Kawamoto, K., Horibe, I. & Uchida, K. (1989). Purification and characterization of a new hydrolase for conjugated bile acids, chenodeoxycholytaurine hydrolase, from *Bacteroides vulgatus*. *J Biochem* 106, 1049-1053.

Kawamura, S., Kakuta, Y., Tanaka, I., Hikichi, K., Kuhara, S., Yamasaki, N. & Kimura, M. (1996). Glycine-15 in the bend between two α -helices can explain the thermostability of DNA binding protein HU from *Bacillus stearothermophilus*. *Biochemistry* 35, 1195-1200.

Khan, J. A., Dunn, B. M. & Tong, L. (2005). Crystal structure of human Taspase1, a crucial protease regulating the function of MLL. *Structure* 13, 1443-1452.

Kim, G. B., Miyamoto, C. M., Meighen, E. A. & Lee, B. H. (2004). Cloning and characterization of the bile salt hydrolase genes (bsh) from *Bifidobacterium bifidum* strains. *Appl Environ Microbiol* 70, 5603-5612.

Kim, G. B., Brochet, M. & Lee, B. H. (2005). Cloning and characterization of a bile salt hydrolase (bsh) from *Bifidobacterium adolescentis*. *Biotechnol Lett* 27, 817-822.

- Kim, J. K., Yang, I. S., Shin, H. J., Cho, K. J., Ryu, E. K., Kim, S. H., Park, S. S. & Kim, K. H. (2006). Insight into autoproteolytic activation from the structure of cephalosporin acylase: a protein with two proteolytic chemistries. *Proc Natl Acad Sci U S A* 103, 1732-1737.
- Kimura, S., Kanaya, S. & Nakamura, H. (1992). Thermostabilization of Escherichia coli ribonuclease HI by replacing left-handed helical Lys95 with Gly or Asn. *Journal of Biological Chemistry* 267, 22014-22017.
- Klock, H. E. & Lesley, S. A. (2009). The Polymerase Incomplete Primer Extension (PIPE) method applied to high-throughput cloning and site-directed mutagenesis. *Methods Mol Biol* 498, 91-103.
- Knapp, S., Kardinahl, S., Hellgren, N., Tibbelin, G., Schäfer, G. & Ladenstein, R. (1999). Refined crystal structure of a superoxide dismutase from the hyperthermophilic archaeon Sulfolobus acidocaldarius at 2.2 Å resolution. *Journal of molecular biology* 285, 689-702.
- Kolata, G. & Andrews, E. L. (2001). Anticholesterol drug pulled after link to 31 deaths. New York Times Online.
- Kovacikova, G., Lin, W. & Skorupski, K. (2003). The virulence activator AphA links quorum sensing to pathogenesis and physiology in Vibrio cholerae by repressing the expression of a penicillin amidase gene on the small chromosome. *J Bacteriol* 185, 4825-4836.
- Krissinel, E. & Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *J Mol Biol* 372, 774-797.
- Kumar, M. S., Bava, K. A., Gromiha, M. M., Prabakaran, P., Kitajima, K., Uedaira, H. & Sarai, A. (2006a). ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic acids research* 34, D204-D206.
- Kumar, R. S., Brannigan, J. A., Prabhune, A. A., Pundle, A. V., Dodson, G. G., Dodson, E. J. & Suresh, C. G. (2006b). Structural and functional analysis of a conjugated bile salt hydrolase from Bifidobacterium longum reveals an evolutionary relationship with penicillin V acylase. *J Biol Chem* 281, 32516-32525.
- Kumar, S., Tsai, C.-J. & Nussinov, R. (2000). Factors enhancing protein thermostability. *Protein engineering* 13, 179-191.
- Lambert, J. M., Bongers, R. S. & Kleerebezem, M. (2007). Cre-lox-Based System for Multiple Gene Deletions and Selectable-Marker Removal in Lactobacillus plantarum. *Applied and Environmental Microbiology* 73, 1126-1135.
- Lambert, J. M., Siezen, R. J., de Vos, W. M. & Kleerebezem, M. (2008). Improved annotation of conjugated bile acid hydrolase superfamily members in Gram-positive bacteria. *Microbiology* 154, 2492-2500.
- Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Cryst* 26, 283-291.
- Le, Q. A., Joo, J. C., Yoo, Y. J. & Kim, Y. H. (2012). Development of thermostable Candida antarctica lipase B through novel in silico design of disulfide bridge. *Biotechnology and bioengineering* 109, 867-876.

- Lee, H., Park, O. K. & Kang, H. S. (2000). Identification of a new active site for autocatalytic processing of penicillin acylase precursor in *Escherichia coli* ATCC11105. *Biochem Biophys Res Commun* 272, 199-204.
- Lee, Y. S. & Park, S. S. (1998). Two-step autocatalytic processing of the glutaryl 7-aminocephalosporanic acid acylase from *Pseudomonas* sp. strain GK16. *J Bacteriol* 180, 4576-4582.
- Lehmann, M., Loch, C., Middendorf, A., Studer, D., Lassen, S. F., Pasamontes, L., van Loon, A. P. & Wyss, M. (2002). The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng* 15, 403-411.
- Levine, G. N., Keaney, J. F., Jr. & Vita, J. A. (1995). Cholesterol reduction in cardiovascular disease. Clinical benefits and possible mechanisms. *N Engl J Med* 332, 512-521.
- Li, C., Heatwole, J., Soelaiman, S. & Shoham, M. (1999a). Crystal structure of a thermophilic alcohol dehydrogenase substrate complex suggests determinants of substrate specificity and thermostability. *Proteins* 37, 619-627.
- Li, S., Smith, J. L. & Zalkin, H. (1999b). Mutational analysis of *Bacillus subtilis* glutamine phosphoribosylpyrophosphate amidotransferase propeptide processing. *J Bacteriol* 181, 1403-1408.
- Lim, H. J., Kim, S. Y. & Lee, W. K. (2004). Isolation of cholesterol-lowering lactic acid bacteria from human intestine for probiotic use. *J Vet Sci* 5, 391-395.
- Lin, G., Hu G Fau - Tsu, C., Tsu C Fau - Kunes, Y. Z. & other authors (2006). Mycobacterium tuberculosis prcBA genes encode a gated proteasome with broad oligopeptide specificity. *Mol Microbiol* 59(5), 1405-1416.
- Liu, Y. & Kuhlman, B. (2006). RosettaDesign server for protein design. *Nucleic acids research* 34, W235-W238.
- Lodola, A., Branduardi, D., De Vivo, M., Capoferri, L., Mor, M., Piomelli, D. & Cavalli, A. (2012). A Catalytic Mechanism for Cysteine N-Terminal Nucleophile Hydrolases, as Revealed by Free Energy Simulations. *PLoS ONE* 7, e32397.
- Lowe, J., Stock, D., Jap, B., Zwickl, P., Baumeister, W. & Huber, R. (1995). Crystal structure of the 20S proteasome from the archaeon *T. acidophilum* at 3.4 Å resolution. *Science* 268, 533-539.
- Macedo-Ribeiro, S., Darimont, B., Sterner, R. & Huber, R. (1996). Small structural changes account for the high thermostability of [4Fe-4S] ferredoxin from the hyperthermophilic bacterium *Thermotoga maritima*. *Structure* 4, 1291-1301.
- Mahmood, T. & Yang, P. C. (2012). Western blot: technique, theory, and trouble shooting. *N Am J Med Sci* 4, 429-434.
- Mainfroid, V., Mande, S. C., Hol, W. G., Martial, J. A. & Goraj, K. (1996). Stabilization of human triosephosphate isomerase by improvement of the stability of individual α -helices in dimeric as well as monomeric forms of the protein. *Biochemistry* 35, 4110-4117.

- Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R. & other authors (2013). CDD: NCBI's conserved domain database. *Nucleic Acids Res* 43, D222-226.
- Marg, G. A. & Clark, D. S. (1990). Activation of glucose isomerase by divalent cations: evidence for two distinct metal-binding sites. *Enzyme and microbial technology* 12, 367-373.
- Margolin, A. L. (1996). Novel crystalline catalysts. *Trends in Biotechnology* 14, 223-230.
- Marshall, S. A., Morgan, C. S. & Mayo, S. L. (2002). Electrostatics significantly affect the stability of designed homeodomain variants. *Journal of molecular biology* 316, 189-199.
- Martin, J., Slade, A., Aitken, A., Arche, R. & Virden, R. (1991). Chemical modification of serine at the active site of penicillin acylase from *Kluyvera citrophila*. *Biochem J* 280 (Pt 3), 659-662.
- Martin, L., Prieto, M. A., Cortes, E. & Garcia, J. L. (1995). Cloning and sequencing of the pac gene encoding the penicillin G acylase of *Bacillus megaterium* ATCC 14945. *FEMS Microbiol Lett* 125, 287-292.
- Masso, M. & Vaisman, I. I. (2011). Structure-based prediction of protein activity changes: Assessing the impact of single residue replacements. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pp. 3221-3224: IEEE.
- Mata, L., Gripon, J.-C. & Mistou, M.-Y. (1999). Deletion of the four C-terminal residues of PepC converts an aminopeptidase into an oligopeptidase. *Protein engineering* 12, 681-686.
- Mateo, C., Abian, O., Fernandez-Lorente, G., Pedroche, J., Fernandez-Lafuente, R., Guisan, J. M., Tam, A. & Daminati, M. (2002). Epoxy sepabeads: a novel epoxy support for stabilization of industrial enzymes via very intense multipoint covalent attachment. *Biotechnol Prog* 18, 629-634.
- Matsumura, M., Signor, G. & Matthews, B. W. (1989). Substantial increase of protein stability by multiple disulphide bonds. *Nature* 342, 291-293.
- Matthews, C. R. (1993). Pathways of protein folding. *Annu Rev Biochem* 62, 653-683.
- McAuliffe, O., Cano, R. J. & Klaenhammer, T. R. (2005). Genetic Analysis of Two Bile Salt Hydrolase Activities in *Lactobacillus acidophilus* NCFM. *Applied and Environmental Microbiology* 71, 4925-4929.
- McDonald, I. K. & Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J Mol Biol* 238, 777-793.
- McDonough, M. A., Klei, H. E. & Kelly, J. A. (1999). Crystal structure of penicillin G acylase from the Bro1 mutant strain of *Providencia rettgeri*. *Protein Sci* 8, 1971-1981.
- Michalska, K., Borek, D., Hernandez-Santoyo, A. & Jaskolski, M. (2008). Crystal packing of plant-type L-asparaginase from *Escherichia coli*. *Acta Crystallogr D Biol Crystallogr* 64, 309-320.
- Milla, M. E., Brown, B. M. & Sauer, R. T. (1994). Protein stability effects of a complete set of alanine substitutions in Arc repressor. *Nature structural biology* 1, 518-523.
- Moser, S. A. & Savage, D. C. (2001). Bile salt hydrolase activity and resistance to toxicity of conjugated bile salts are unrelated properties in lactobacilli. *Appl Environ Microbiol* 67, 3476-3480.

- Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology* 247, 536-540.
- Nakaishi, Y., Bando, M., Shimizu, H., Watanabe, K., Goto, F., Tsuge, H., Kondo, K. & Komatsu, M. (2009). Structural analysis of human glutamine:fructose-6-phosphate amidotransferase, a key regulator in type 2 diabetes. *FEBS Lett* 583, 163-167.
- Nicholson, H., Becktel, W. & Matthews, B. (1988). Enhanced protein thermostability from designed mutations that interact with alpha-helix dipoles. *Nature* 336, 651-656.
- Nicholson, H., Tronrud, D., Becktel, W. & Matthews, B. (1992). Analysis of the effectiveness of proline substitutions and glycine replacements in increasing the stability of phage T4 lysozyme. *Biopolymers* 32, 1431-1441.
- Nielsen, H. & Krogh, A. (1998). Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc Int Conf Intell Syst Mol Biol* 6, 122-130.
- Oh, B., Kim, K., Park, J., Yoon, J., Han, D. & Kim, Y. (2004). Modifying the substrate specificity of penicillin G acylase to cephalosporin acylase by mutating active-site residues. *Biochem Biophys Res Commun* 319, 486-492.
- Ohashi, H., Katsuta, Y., Hashizume, T., Abe, S. N., Kajiura, H., Hattori, H., Kamei, T. & Yano, M. (1988). Molecular cloning of the penicillin G acylase gene from *Arthrobacter viscosus*. *Appl Environ Microbiol* 54, 2603-2607.
- Oinonen, C. & Rouvinen, J. (2000). Structural comparison of Ntn-hydrolases. *Protein Sci* 9, 2329-2337.
- Okada, T., Suzuki, H., Wada, K., Kumagai, H. & Fukuyama, K. (2006). Crystal structures of gamma-glutamyltranspeptidase from *Escherichia coli*, a key enzyme in glutathione metabolism, and its reaction intermediate. *Proc Natl Acad Sci U S A* 103, 6471-6476.
- Olsson, A. & Uhlen, M. (1986). Sequencing and heterologous expression of the gene encoding penicillin V amidase from *Bacillus sphaericus*. *Gene* 45, 175-181.
- Pace, C. N. (1992). Contribution of the hydrophobic effect to globular protein stability. *J Mol Biol* 226, 29-35.
- Pantoliano, M. W., Ladner, R. C., Bryan, P. N., Rollence, M. L., Wood, J. F. & Poulos, T. L. (1987). Protein engineering of subtilisin BPN': enhanced stabilization through the introduction of two cysteines to form a disulfide bond. *Biochemistry* 26, 2077-2082.
- Park, J. H. & Schuchman, E. H. (2006). Acid ceramidase and human disease. *Biochim Biophys Acta* 1758, 2133-2138.
- Parthiban, V., Gromiha, M. M. & Schomburg, D. (2006). CUPSAT: prediction of protein stability upon point mutations. *Nucleic acids research* 34, W239-W242.
- Perry, L. J. & Wetzel, R. (1984). Disulfide bond engineered into T4 lysozyme: stabilization of the protein toward thermal inactivation. *Science* 226, 555-557.

- Perutz, M. F. (1978). Electrostatic effects in proteins. *Science* 201, 1187-1191.
- Phadtare, S., Parekh, P., Gole, A., Patil, M., Pundle, A., Prabhune, A. & Sastry, M. (2002). Penicillin G acylase-fatty lipid biocomposite films show excellent catalytic activity and long term stability/reusability. *Biotechnol Prog* 18, 483-488.
- Polizzi, K. M., Chaparro-Riggers, J. F., Vazquez-Figueroa, E. & Bommarius, A. S. (2006). Structure-guided consensus approach to create a more thermostable penicillin G acylase. *Biotechnology Journal* 1, 531-536.
- Prieto, M. A., Perez-Aranda, A. & Garcia, J. L. (1993). Characterization of an Escherichia coli aromatic hydroxylase with a broad substrate range. *J Bacteriol* 175, 2162-2167.
- Prieto, M. A., Diaz, E. & Garcia, J. L. (1996). Molecular characterization of the 4-hydroxyphenylacetate catabolic pathway of Escherichia coli W: engineering a mobile aromatic degradative cluster. *J Bacteriol* 178, 111-120.
- Pronk, S., Pall, S., Schulz, R. & other authors (2013). GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 29, 845-854.
- Punta, M., Coggill, P. C., Eberhardt, R. Y. & other authors (2012). The Pfam protein families database. *Nucleic Acids Res* 40, D290-301.
- Rathinaswamy, P., Gaikwad, S. M., Suresh, C. G., Prabhune, A. A., Brannigan, J. A., Dodson, G. G. & Pundle, A. V. (2012). Purification and characterization of YxeI, a penicillin acylase from Bacillus subtilis. *Int J Biol Macromol* 50, 25-30.
- Rawlings, N. D., Barrett, A. J. & Bateman, A. (2012). MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res* 40, D343-350.
- Rawlings, N. D., Barrett, A. J. & Salvesen, N. D. R. (2013). Chapter 808 - Introduction: Clan PB Containing N-terminal Nucleophile Peptidases. In *Handbook of Proteolytic Enzymes*, pp. 3648-3653: Academic Press.
- Reddy, G. S., Prakash, J. S., Vairamani, M., Prabhakar, S., Matsumoto, G. I. & Shivaji, S. (2002). Planococcus antarcticus and Planococcus psychrophilus spp. nov. isolated from cyanobacterial mat samples collected from ponds in Antarctica. *Extremophiles* 6, 253-261.
- Reid, K. S. C., Lindley, P. F. & Thornton, J. M. (1985). Sulphur-aromatic interactions in proteins. *FEBS Letters* 190, 209-213.
- Ringer, A. L., Senenko, A. & Sherrill, C. D. (2007). Models of S/pi interactions in protein structures: comparison of the H2S benzene complex with PDB data. *Protein Sci* 16, 2216-2223.
- Robinson, N. E. (2002). Protein deamidation. *Proc Natl Acad Sci U S A* 99, 5283-5288.
- Rocchietti, S., Urrutia, A. S. V., Pregnotato, M., Tagliani, A., Guisãn, J. M., Fernãndez-Lafuente, R. & Terreni, M. (2002). Influence of the enzyme derivative preparation and substrate structure on the enantioselectivity of penicillin G acylase. *Enzyme and Microbial Technology* 31, 88-93.

Rossocha, M., Schultz-Heienbrok, R., von Moeller, H., Coleman, J. P. & Saenger, W. (2005). Conjugated bile acid hydrolase is a tetrameric N-terminal thiol hydrolase with specific recognition of its cholyl but not of its tauryl product. *Biochemistry* 44, 5739-5748.

Russell, R. J. M., Ferguson, J. M. C., Hough, D. W., Danson, M. J. & Taylor, G. L. (1997). The Crystal Structure of Citrate Synthase from the Hyperthermophilic Archaeon *Pyrococcus furiosus* at 1.9 Å Resolution. *Biochemistry* 36, 9983-9994.

Ruzzo, E. K., Capo-Chichi, J. M., Ben-Zeev, B. & other authors (2013). Deficiency of asparagine synthetase causes congenital microcephaly and a progressive form of encephalopathy. *Neuron* 80, 429-441.

Sadowski, M. I. & Jones, D. T. (2009). The sequence-structure relationship and protein function prediction. *Current Opinion in Structural Biology* 19, 357-362.

Sathyapriya, R. & Vishveshwara, S. (2004). Interaction of DNA with clusters of amino acids in proteins. *Nucleic acids research* 32, 4109-4118.

Schwarz, K., Walther, M., Anton, M., Gerth, C., Feussner, I. & Kuhn, H. (2001). Structural basis for lipoyxygenase specificity conversion of the human leukocyte 5-lipoyxygenase to a 15-lipoyxygenating enzyme species by site-directed mutagenesis. *Journal of Biological Chemistry* 276, 773-779.

Seemuller, E., Lupas, A. & Baumeister, W. (1996). Autocatalytic processing of the 20S proteasome. *Nature* 382, 468-471.

Serrano, L., Bycroft, M. & Fersht, A. R. (1991). Aromatic-aromatic interactions and protein stability. Investigation by double-mutant cycles. *J Mol Biol* 218, 465-475.

Serrano, L., Kellis Jr, J. T., Cann, P., Matouschek, A. & Fersht, A. R. (1992). The folding of an enzyme: II. Substructure of barnase and the contribution of different interactions to protein stability. *Journal of molecular biology* 224, 783-804.

Shewale, J., Kumar, K. & Ambekar, G. (1987). Evaluation of determination of 6-aminopenicillanic acid by p-dimethyl aminobenzaldehyde. *Biotechnology Techniques* 1, 69-72.

Shewale, J. G. & Sudhakaran, V. K. (1997). Penicillin V acylase: Its potential in the production of 6-aminopenicillanic acid. *Enzyme and Microbial Technology* 20, 402-410.

Shirley, B. A., Stanssens, P., Hahn, U. & Pace, C. N. (1992). Contribution of hydrogen bonding to the conformational stability of ribonuclease T1. *Biochemistry* 31, 725-732.

Shtraizent, N., Eliyahu, E., Park, J. H., He, X., Shalgi, R. & Schuchman, E. H. (2008). Autoproteolytic cleavage and activation of human acid ceramidase. *J Biol Chem* 283, 11253-11259.

Singh, J., Hamid, R. & Reddy, B. S. (1997). Dietary fat and colon cancer: modulating effect of types and amount of dietary fat on ras-p21 function during promotion and progression stages of colon cancer. *Cancer Res* 57, 253-258.

Sio, C. F. & Quax, W. J. (2004). Improved beta-lactam acylases and their use as industrial biocatalysts. *Curr Opin Biotechnol* 15, 349-355.

- Smith, C. A., Toogood, H. S., Baker, H. M., Daniel, R. M. & Baker, E. N. (1999). Calcium-mediated thermostability in the subtilisin superfamily: the crystal structure of Bacillus Ak. 1 protease at 1.8 Å resolution. *Journal of molecular biology* 294, 1027-1040.
- Souciet, J. L., Hermodson, M. A. & Zalkin, H. (1988). Mutational analysis of the glutamine phosphoribosylpyrophosphate amidotransferase pro-peptide. *J Biol Chem* 263, 3323-3327.
- Sousa da Silva, A. W. & Vranken, W. F. (2012). ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Res Notes* 5, 367.
- Srirangan, K., Orr, V., Akawi, L., Westbrook, A., Moo-Young, M. & Chou, C. P. (2013). Biotechnological advances on Penicillin G acylase: Pharmaceutical implications, unique expression mechanism and production strategies. *Biotechnology Advances* 31, 1319-1332.
- Stellwag, E. J. & Hylemon, P. B. (1976). Purification and characterization of bile salt hydrolase from *Bacteroides fragilis* subsp. *fragilis*. *Biochim Biophys Acta* 452, 165-176.
- Stites, W. E., Meeker, A. K. & Shortle, D. (1994). Evidence for strained interactions between side-chains and the polypeptide backbone. *Journal of molecular biology* 235, 27-32.
- Stouthamer, A. H. (1991). Metabolic regulation including anaerobic metabolism in *Paracoccus denitrificans*. *J Bioenerg Biomembr* 23, 163-185.
- Suemori, A. (2013). Conserved and non-conserved residues and their role in the structure and function of p-hydroxybenzoate hydroxylase. *Protein engineering, design & selection : PEDS* 26, 479-488.
- Suplatov, D., Panin, N., Kirilin, E., Shcherbakova, T., Kudryavtsev, P. & Svedas, V. (2014). Computational design of a pH stable enzyme: understanding molecular mechanism of penicillin acylase's adaptation to alkaline conditions. *PLoS One* 9, e100643.
- Suzuki, Y., Oishi, K., Nakano, H. & Nagayama, T. (1987). A strong correlation between the increase in number of proline residues and the rise in thermostability of five Bacillus oligo-1, 6-glucosidases. *Applied microbiology and biotechnology* 26, 546-551.
- Suzuki, Y. (1989). A general principle of increasing protein thermostability. *Proceedings of the Japan Academy Ser B: Physical and Biological Sciences* 65, 146-148.
- Takano, K., Ota, M., Ogasahara, K., Yamagata, Y., Nishikawa, K. & Yutani, K. (1999). Experimental verification of the stability profile of mutant protein (SPMP) data using mutant human lysozymes. *Protein engineering* 12, 663-672.
- Takano, K., Yamagata, Y. & Yutani, K. (2001). Role of non-glycine residues in left-handed helical conformation for the conformational stability of human lysozyme. *Proteins: Structure, Function, and Bioinformatics* 44, 233-243.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28, 2731-2739.

Tamura, T., Nagy, I. n., Lupas, A., Lottspeich, F., Cejka, Z., Schoofs, G., Tanaka, K., De Mot, R. & Baumeister, W. (1995). The first characterization of a eubacterial proteasome: the 20S complex of *Rhodococcus*. *Current Biology* 5, 766-774.

Tanner, J. J., Hecht, R. M. & Krause, K. L. (1996). Determinants of enzyme thermostability observed in the molecular structure of *Thermus aquaticus* D-glyceraldehyde-3-phosphate dehydrogenase at 2.5 Angstroms Resolution. *Biochemistry* 35, 2597-2609.

Tepljakov, A. V., Kuranova, I. P., Harutyunyan, E. H., Vainshtein, B. K., Froëmmel, C., Hoëhne, W. E. & Wilson, K. S. (1990). Crystal structure of thermitase at 1.4Å resolution. *Journal of Molecular Biology* 214, 261-279.

Theillet, F.-X., Kalmar, L., Tompa, P., Han, K.-H., Selenko, P., Dunker, A. K., Daughdrill, G. W. & Uversky, V. N. (2013). The alphabet of intrinsic disorder. *Intrinsically Disordered Proteins* 1, e24360.

Thomas, L. A., Veysey, M. J., Bathgate, T., King, A., French, G., Smeeton, N. C., Murphy, G. M. & Dowling, R. H. (2000). Mechanism for the transit-induced increase in colonic deoxycholic acid formation in cholesterol cholelithiasis. *Gastroenterology* 119, 806-815.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 25, 4876-4882.

Thompson, M. J. & Eisenberg, D. (1999). Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability. *Journal of molecular biology* 290, 595-604.

Tikkanen, R., Riikonen, A., Oinonen, C., Rouvinen, R. & Peltonen, L. (1996). Functional analyses of active site residues of human lysosomal aspartylglucosaminidase: implications for catalytic mechanism and autocatalytic activation. *EMBO J* 15, 2954-2960.

Tina, K., Bhadra, R. & Srinivasan, N. (2007). PIC: protein interactions calculator. *Nucleic acids research* 35, W473-W476.

Tobin, M. B., Cole, S. C., Miller, J. R., Baldwin, J. E. & Sutherland, J. D. (1995). Amino-acid substitutions in the cleavage site of acyl-coenzyme A:isopenicillin N acyltransferase from *Penicillium chrysogenum*: effect on proenzyme cleavage and activity. *Gene* 162, 29-35.

Topgi, R. S., Ng, J. S., Landis, B., Wang, P. & Behling, J. R. (1999). Use of enzyme penicillin acylase in selective amidation/amide hydrolysis to resolve ethyl 3-amino-4-pentynoate isomers. *Bioorg Med Chem* 7, 2221-2229.

Torres-Bacete, J. s., Arroyo, M., Torres-Guzmán, R., de la Mata, I., Castillãñn, M. a. & Acebal, C. (2000). Covalent immobilization of penicillin acylase from *Streptomyces lavendulae*. *Biotechnology Letters* 22, 1011-1014.

Torres, L., Ferreras, E., Cantero, A., Hidalgo, A. & Berenguer, J. (2012). Functional expression of a penicillin acylase from the extreme thermophile *Thermus thermophilus* HB27 in *Escherichia coli*. *Microbial Cell Factories* 11, 105.

Usher, K. C., De La Cruz, A. F. A., Dahlquist, F. W., James Remington, S., Swanson, R. V. & Simon, M. I. (1998). Crystal structures of CheY from *Thermotoga maritima* do not support conventional explanations for the structural basis of enhanced thermostability. *Protein science* 7, 403-412.

Valas, R. E. & Bourne, P. E. (2011). The origin of a derived superkingdom: how a gram-positive bacterium crossed the desert to become an archaeon. *Biol Direct* 6, 16.

Valle, F., Balbas, P., Merino, E. & Bolivar, F. (1991). The role of penicillin amidases in nature and in industry. *Trends Biochem Sci* 16, 36-40.

Van Heeke, G. & Schuster, S. M. (1989). The N-terminal cysteine of human asparagine synthetase is essential for glutamine-dependent activity. *J Biol Chem* 264, 19475-19477.

van Langen, L. M., van Rantwijk, F., Å vedas, V. K. & Sheldon, R. A. (2000). Penicillin acylase-catalyzed peptide synthesis: a chemo-enzymatic route to stereoisomers of 3,6-diphenylpiperazine-2,5-dione. *Tetrahedron: Asymmetry* 11, 1077-1083.

Vandamme, E. J. & Voets, J. P. (1974). Microbial penicillin acylases. *Adv Appl Microbiol* 17, 311-369.

Vandamme, E. J. & Voets, J. P. (1975). Properties of the purified penicillin V-acylase of *Erwinia aroideae*. *Experientia* 31, 140-143.

Vanoni, M. A. & Curti, B. (1999). Glutamate synthase: a complex iron-sulfur flavoprotein. *Cell Mol Life Sci* 55, 617-638.

Varshney, N. K., Suresh Kumar, R., Ignatova, Z., Prabhune, A., Pundle, A., Dodson, E. & Suresh, C. G. (2012). Crystallization and X-ray structure analysis of a thermostable penicillin G acylase from *Alcaligenes faecalis*. *Acta Crystallographica Section F* 68, 273-277.

Verhaert, R. M., Riemens, A. M., van der Laan, J. M., van Duin, J. & Quax, W. J. (1997). Molecular cloning and analysis of the gene encoding the thermostable penicillin G acylase from *Alcaligenes faecalis*. *Appl Environ Microbiol* 63, 3412-3418.

Vieille, C. & Zeikus, G. J. (2001). Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol Mol Biol Rev* 65, 1-43.

Vogel, H. & Weljie, A. (2002). Steady-State Fluorescence Spectroscopy. In *Calcium-Binding Protein Protocols: Volume 2: Methods and Techniques*, pp. 75-87: Springer New York.

Vogt, G. & Argos, P. (1997). Protein thermal stability: hydrogen bonds or internal packing? *Folding and Design* 2, S40-S46.

Vogt, G., Woell, S. & Argos, P. (1997). Protein thermal stability, hydrogen bonds, and ion pairs. *Journal of molecular biology* 269, 631-643.

Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. *Journal of molecular graphics* 8, 52-56.

Watanabe, K., Masuda, T., Ohashi, H., Mihara, H. & Suzuki, Y. (1994). Multiple Proline Substitutions Cumulatively Thermostabilize *Bacillus Cereus* ATCC7064 Oligo-1, 6-Glucosidase. *European Journal of Biochemistry* 226, 277-283.

Watanabe, K., Hata, Y., Kizaki, H., Katsube, Y. & Suzuki, Y. (1997). The refined crystal structure of *Bacillus cereus* oligo-1, 6-glucosidase at 2.0 Å resolution: structural characterization of proline-substitution sites for protein thermostabilization. *Journal of molecular biology* 269, 142-153.

Wiederstein, M. & Sippl, M. J. (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res* 35, W407-410.

Wijaya, A., Hermann, A., Abriouel, H., Specht, I., Yousif, N. M., Holzapfel, W. H. & Franz, C. M. (2004). Cloning of the bile salt hydrolase (bsh) gene from *Enterococcus faecium* FAIR-E 345 and chromosomal location of bsh genes in food enterococci. *J Food Prot* 67, 2772-2778.

Williams, K., Cullati, S., Sand, A., Biterova, E. I. & Barycki, J. J. (2009). Crystal structure of acivicin-inhibited gamma-glutamyltranspeptidase reveals critical roles for its C-terminus in autoprocessing and catalysis. *Biochemistry* 48, 2459-2467.

Wilson, C. A., Kreychman, J. & Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores1. *Journal of Molecular Biology* 297, 233-249.

Worth, C. L., Preissner, R. & Blundell, T. L. (2011). SDM—a server for predicting effects of mutations on protein stability and malfunction. *Nucleic acids research* 39, W215-W222.

Xu, Q., Buckley, D., Guan, C. & Guo, H. C. (1999). Structural insights into the mechanism of intramolecular proteolysis. *Cell* 98, 651-661.

Ye, Y. & Godzik, A. (2003). Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* 19 Suppl 2, ii246-255.

Yip, K. S., Stillman, T. J., Britton, K. L. & other authors (1995). The structure of *Pyrococcus furiosus* glutamate dehydrogenase reveals a key role for ion-pair networks in maintaining enzyme stability at extreme temperatures. *Structure (London, England : 1993)* 3, 1147-1158.

Zalkin, H. & Smith, J. L. (1998). Enzymes utilizing glutamine as an amide donor. *Adv Enzymol Relat Areas Mol Biol* 72, 87-144.

Zamost, B., Nielsen, H. & Starnes, R. (1991). Thermostable enzymes for industrial applications. *Journal of Industrial Microbiology* 8, 71-81.

Zhang, T., Bertelsen, E. & Alber, T. (1994). Entropic effects of disulphide bonds on protein stability. *Nature Structural & Molecular Biology* 1, 434-438.

Zmijewski Jr, M. J., Briggs, B. S., Thompson, A. R. & Wright, I. G. (1991). Enantioselective acylation of a beta-lactam intermediate in the synthesis of loracarbef using penicillin G amidase. *Tetrahedron Letters* 32, 1621-1622.

Zwickl, P., Kleinz, J. & Baumeister, W. (1994). Critical elements in proteasome assembly. *Nat Struct Biol* 1, 765-770.

List of Publications

1. **Panigrahi, P.**, Sule, M., Sharma, R., Ramasamy, S. and Suresh, C.G. An improved method for specificity annotation shows a distinct evolutionary divergence among the microbial enzymes of the cholyglycine hydrolase family, *Microbiology*, 160, 1162-1174 (2014).
2. **Priyabrata Panigrahi**, Manas Sule, Avinash Ghanate, Sureshkumar Ramasamy and C.G. Suresh. Engineering proteins for thermostability with iRDP web server. (Communicated).
3. **Priyabrata Panigrahi**, Deepak Chand, Ruchira Mukherji, Sureshkumar Ramasamy and C. G. Suresh. Developing a proof of concept approach towards selecting thermostable penicillin acylases for industrial applications. (Communicated).
4. Avinash, V.S.*, **Panigrahi, P.***, Suresh, C.G., Pundle, A.V. and Ramasamy, S. Structural modelling of substrate binding and inhibition in penicillin V acylase from *Pectobacterium atrosepticum*, *Biochemical and biophysical research communications*, 437, 538-543 (2013). *Equal contribution.
5. Joshi, R.R., **Panigrahi, P.R.** and Patil, R.N. Dimensionality reduction in computational demarcation of protein tertiary structures, *Journal of molecular modeling*, 18, 2741-2754 (2012)
6. Sharma, R., **Panigrahi, P.** and Suresh, C.G. In-Silico Analysis of Binding Site Features and Substrate Selectivity in Plant Flavonoid-3-O Glycosyltransferases (F3GT) through Molecular Modeling, Docking and Dynamics Simulation Studies, *PloS one*, 9, e92636 (2014)
7. Mukherji, R., Varshney, N.K., **Panigrahi, P.**, Suresh, C.G. and Prabhune, A. A new role for penicillin acylases: Degradation of acyl homoserine lactone quorum sensing signals by *Kluyvera citrophila* penicillin G acylase, *Enzyme and Microbial Technology*, 56, 1-7 (2013)
8. Yashwant Kumar, Bhushan B Dholakia, **Priyabrata Panigrahi**, Narendra Y Kadoo, Ashok P Giri, Vidya S Gupta. Metabolic profiling of chickpea-Fusarium interaction identifies differential modulation of disease resistance pathways, *Phytochemistry* (2015).