

Large-scale structural analysis of enzymes to understand the basis of enzyme promiscuity

A Thesis

Submitted by

Preeti Choudhary

PH12108

For the award of the degree of
Doctor of Philosophy



Department of Biological Sciences
Indian Institute of Science Education and Research (IISER) Mohali
Sector-81, Mohali, 140306, Punjab, India

January, 2019

Dedicated to my parents
Two most brightest stars of my universe

Declaration

The work presented in this thesis entitled “Large-scale structural analysis of enzymes to understand the basis of enzyme promiscuity” has been carried out by me under the supervision of Dr. Shashi Bhushan Pandit at the Indian Institute of Science Education and Research (IISER) Mohali. This work has not been submitted in part or full for a degree, a diploma, or a fellowship to any university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgement of collaborative research and discussions. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in the references.

Preeti Choudhary
(Candidate)

In my capacity as the supervisor of the candidate’s thesis work, I certify that the above statements by the candidate are true to the best of my knowledge.

Date:

Place:

Dr. Shashi Bhushan Pandit
(Supervisor)

Acknowledgements

Bioinformatics was one of the most least liked subjects when I was an undergraduate. Infact, my initial few months as PhD scholar at IISER Mohali were spent in “fear of coding”. But flipside to this, today I love coding and it has become a useful skillset which I exploit both professionally and personally. This could not have been possible without the support and help from people who provided me courage and confidence and most importantly believed in me. Personally, my supporting pillar, my inspiration and my source of energy- my mother, she is the sole reason I am here today (finishing PhD which starting as a big question mark), and she made sure that I caterpillar in the cocoon didn't become a burnt pancake! She makes my life beautiful.

Professionally, first and foremost, I express my deepest gratitude to my Ph.D. supervisor Dr. Shashi Bhushan Pandit. Without his support, insight and encouragement, this thesis would not have been possible. I still remember the initial days in his lab, when I was still in the cocoon, with struggling my computational biology phobia, he told me “You know how to click a mouse, right! That's all you need”. With that level of support and understanding, he made me comfortable with field to computational biology. His rock-solid patience level has been a source of inspiration. He always says “devil lies in details”, data should be carefully analyzed which is the most crucial part of science. I admire him for being a person who always makes sure that everything should be done perfectly. His cordial nature and guidance has immensely helped me in a better understanding of the work. I particularly appreciate his constructive feedback, which (usually) manifests in the form of red strike out words (in Microsoft review mode) all over the doc file I have ever mailed him. His patience for me is amazing, despite my endless questions (Many times knocking his office door again minutes after the discussion is over!). He just seems to have a knack of setting me straight, and telling me not to sway when in doubt. As they say, the first PhD student is usually a tricky one! Even though a many time, I might have been a guinea pig, throughout this journey, but surely I have learnt lifelong lessons during this time.

Although, we did not have privilege of exploiting the knowledge and experience of seniors PhD labmates (as we didn't have any), I was lucky to that Prof. Somdatta Sinha shifted to IISER Mohali from CCMB, Hyderabad. With her came my only two seniors, the two most beautiful and kind souls I have ever met- Ashutosh and Priya. Ashutosh immensely helped me in coming out of my comfort zone and overcome by programming fears during my formative years of PhD. I always enjoyed scintillating scientific discussions with him (certainly miss discussions with him sometimes!). I personally appreciate his daily critical reviews and comments which certain made me a both a better researcher and a better person personally too. Priya is again a great friend and my all-time coffee buddy.

I would like to thank Dr. Kuljeet Singh Sandhu, a great teacher (fan of his teaching) who ignited my interest in Intrinsically disordered proteins. He has played crucial role in scintillating general interest in science during my formative months of PhD. I express my sincere thanks to my doctoral committee members, Prof. Purnananda Guptasarma and Dr. Rachna Chaba for their time, support, and valuable suggestions.

I profoundly thank our collaborators- Dr. Kuljeet Singh Sandhu, Nitish Tayal, Prof. Anand Kumar Bachhawat, Dr. Shaiesh Kumar, Dr. Kausik Chattopadhyay and Amritha Sreekumar for giving me opportunity to work with them on interesting questions.

Many thanks to Prof. Somdatta Sinha, she is my inspiration and it is my privilege that I got the opportunity to interact with such an astonishing inter-disciplinary scientist. She has taught me many professional and personal life-long lessons. She is like my scientific mother who has always taken care of me and I know I can always rely on her for keeping my back. Also many thanks to her for organizing house parties (always a great stress busters) and for bringing delicious chocolates and sweets from her abroad trips. The most memorable lab trip of "Morni Hills" could never have been possible if it wasn't for her.

I would like to thank Prof. N. Sathyamurthy, former Director, IISER Mohali, for giving me the opportunity to work at this premier research institute. I am also thankful to Prof. Debi Prasad Sarkar, the Director, IISER Mohali for kindly permitting me to use the

excellent infrastructure for carrying out my research work. I express my immense gratitude to all faculties and students of Department of Biological Sciences, IISER Mohali for providing me access to departmental facilities, for the exciting discussions and encouraging interactions during seminars and presentations. I want to take a moment and thank IISER Mohali library facilities which immensely helped me concentrating more and was fast getaway from distractions while writing my thesis.

I am thankful to all my B.tech teachers Prof. Meenu Kapoor, Prof. Raghu Ram, Prof. Kannan and Prof. Rita Singh for their lessons and encouragement to pursue science as a research carrier (especially RaghuRam Sir- for those although very long but highly intellectual talks which really motivated me).

I would like to thank Department of Biotechnology (DBT), Government of India, for the fellowship during my PhD tenure, IISER Mohali, Centre for Protein Science, Design and Engineering (CPSDE) (special thanks for funding me towards the end of my PhD), Ministry of Human Resource and Development (MHRD), EMBL and EMBL-EBI for providing international travel fellowship and bursary award and providing an excellent opportunity to visit and present my work at Heidelberg, Germany and Hinxton, UK. The rigorous scientific discussions really helped me a lot in better understanding of work. Not to forget the exquisite cuisine serve at both the places were mesmerizing too!

I am extremely thankful to all my lab members for their incredible support, fruitful scientific discussions, homely environment in the lab, and dinner parties we enjoyed together. Special thanks to Paras and Deeksha for their immense support during the end of my PhD, at times they literally have fed me when I was going hay-wire and jaded while consolidating my thesis (Occasions when I used to hit a writer's block!). I am grateful to other lab mates- Dr. Nidhi, Shashi and Rivi for their love and support. Special thanks to Rivi for the unique bond we share owed to our epiphany moments. Sincere thanks to entire computational biology group Arash, Meenakshi, Ken, Yachna, Harpeet and Mohan for always keeping the environmental of lab cool, friendly, pleasant with lots of happy vibes. I am really going to miss you guys. I would like to thank former members of lab- Srishti,

Nitika, Amrinder, Swapnil and Aditya, for thanks guys for always being there for me professionally and for all the instant party plans- made PhD survivable. Special thanks to Meenakshi and Nitika, my all-time buddies for chili's and mocha. I thank Sandeep and Meenakshi for our late night movie shows or Viswajeet's canteen chit chats, these were great mood lifters. I am grateful to Sandeep for always pulling my leg, being a critic and always giving an honest opinion which really helped me to grow as a person.

I am extremely lucky to have a bunch of astonishing friends and would like to acknowledge them for the scientific and non-scientific discussions, their support, and for the birthday celebration parties. I am thankful to Poonam, Saikat, Sandeep, Prince, Rohan, Soumitro, Krishna, Bhupinder, Shashank and Shiv. Many thanks to Poonam and Saikat for the countless happiest moments and unforgettable memories which I am going to cherish throughout my life. Thank you for always being with me, giving me confidence whenever I was shattered and for our stress buster trips to Himachal (Saikat especially to you as you never let your mountain sickness come on the way). I would also like to thank Rajyavardhan, for his critical reviews during our scientific discussion has surely enlightened my scientific perceptiveness. I always learn a new thing, every time we talk. Thanks for being my scientific preacher. Also thanks for being a part of my one of the few precious moments of PhD. Although you are my biggest critic (Yes, more than my supervisor!), you have always been an amazing friend and I will always cherish the things you made me do which otherwise I wouldn't have explored. Thanks for making me a better version of me.

Most importantly, I owe my deepest gratitude to my family for their constant support and understanding Thank you for believing in me, and my dreams. Your sheer love is unparalleled. I am short of words to say thanks to you guys for your blessings, unconditional love, prayers, support, sacrifices, and faith. Thanks a lot for being the driving force of my life, helping me in everything and making my life so simple.

Lastly, thanks to all the hard-working crystallographers and experimentalists who have done the work to generate the data from which I've built this thesis.

P.S: Special thanks to Nescafe Coffee machines in Academic block which was my quick getaway when I was stuck in some serious code-cracking!

Abbreviations

RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuations
MACiE	Mechanism and Catalytic Site Atlas
CSA	Catalytic Site Atlas
CSmetaPred	Catalytic Site meta Predictor
EC number	Enzyme Classification Number
PDB	Protein Data Bank
LPC	Ligand Protein Contact
URL	Uniform resource locator
CATH	Class Architecture Topology Homologous
CRpred	Catalytic Residue Predictor
DNA	Deoxy Ribonucleic Acid
RNA	Ribonucleic Acid
HAD superfamily	Haloalkanoate dehalogenase superfamily
AP superfamily	Alkaline Phosphatase superfamily
AADS	Absorbance-Activated droplet sorting
NMR	Nuclear Magnetic Resonance
GP	Gaussian Process
PROPER	Promiscuity Predictor
K_M	Michaelis constant
GST	glutathione S-transferase
CYP	cytochrome P450
GCL	γ -glutamylcysteine ligase
PROMISE	Promiscuity Indices Estimator
CLASP	Catalytic Active Site Prediction
K_{cat}	Catalytic Efficiency/ Turnover number
V_{max}	Maximum enzyme velocity
WT	Wild type
MT	Mutant type
L-CYS	L-Cysteine
L-GLU	L-Glutamate
ATP	Adenosine Triphosphate
NAD	Nicotinamide adenine dinucleotide
PS-Score	Pocket Similarity Score
VMD	Visual molecular Dynamics
NAMD	Nanoscale Molecular Dynamics
MD simulation	Molecular Dynamics simulation
RCSB	Research Collaboratory for Structural Bioinformatics
MSA	Multiple Sequence Alignment
BUB dataset	Bound-unbound dataset

EXIA	Enzyme Catalytic residue side-chain arrangement
ROC curve	Receiver operating characteristic
PR curve	Precision Recall curve
AveS	Average Specificity
MAS	Mean Average Specificity
MAP	Mean Average Precision
AUCROC	Area Under Curve of ROC curve
AUCPR	Area Under Curve of PR curve
FPR	False Positive Rate
TPR	True Positive Rate
WCN	Weighted Contact Network
PScN	Protein side chain network
PSCDB	Protein Structural Change DataBase
GCLC	GCL catalytic subunit
GCLM	GCL modifier subunit
GSH	Glutathione
GSSG	Glutathione disulfide
γ -GC	γ -glutamylcysteine
P2S	(2s)-2-amino-4-[[[(2r)-2carboxybutyl](phosphono)sulfonimidoyl]butanoic acid
BSO	Buthionine Sulphoximine
APoc	Alignment of Pockets
CSG	Completely sequenced genomes
EcGCL	<i>Escherichia coli</i>
BjGCL	<i>Brassica juncea</i>
ScGCL	<i>Saccharomyces cerevisiae</i>
HMM	Hidden Markov Model
Sb1	Subgroup 1
Sb2	Subgroup 2
Sb3	Subgroup 3
Sb4	Subgroup 4
POOL	Partial Order Optimum Likelihood
CD-HIT	Cluster Database at High Identity with Tolerance
UniProt	Universal Protein Resource
iTOL	Interactive Tree of Life
MEGA	Molecular Evolutionary Genetics Analysis
PON1	serum paraoxonase 1

Synopsis

Most of the enzyme research has been focused on understanding specificity of enzymes as aptly put in Beadle and Tatum's "one gene-one enzyme-one reaction" hypothesis. Previous studies have shown that enzymes are remarkable efficient and specific to a particular substrate or reaction they catalyze. However, recent studies have shown that many enzymes also harbor inherent capability to catalyze alternate reaction/s or substrate/s apart from their physiologically evolved activities, usually involving same active site. Such secondary adventitious reaction/s are referred to as promiscuous reactions, which in general are catalytically less efficient compared to the native activity of the given enzyme. Despite being inefficient to catalyze certain substrates under normal conditions, promiscuous enzymes can become essential under gene deletions or environment perturbations. Thus, promiscuous activities can act as a repertoire of catalytic activities, which could be recruited to confer fitness benefit to an organism under selective pressure. Moreover, these enzymes could serve as a starting point for the emergence of new efficient enzyme functions by gene duplication and divergence. Promiscuous enzymes catalyzing naturally occurring metabolites forms a part of 'underground metabolism', which can impart resilience to metabolic network against genetic or environment perturbations.

The availability of experimental protein tertiary structures owing to the structural genomics efforts in the last decade and recent high-throughput experimental approaches developed to decipher new enzymatic (promiscuous) activities provide a unique opportunity to systematically investigate and derive general structural basis of enzyme promiscuity. Further, this enhanced understanding of mechanistic basis of promiscuity can facilitate protein engineering such as designing catalytically efficient enzymes for desired substrates. Moreover, it can provide insights into evolution of enzymes. The overall objective of the work reported in this thesis is to systematically explore the general structural principles of enzyme promiscuity especially with respect to the roles of binding site and catalytic site residues. The brief overview of work carried out during my doctoral research is described below.

Understanding residue conformational variability of binding and catalytic residues between specialist and generalist enzymes

Understanding mechanistic and structural aspects of enzyme promiscuity can facilitate rational enzyme engineering as well as provide insights into their functional divergence. Based on recent studies the proposed mechanisms of promiscuity are: conformational diversity of active sites, sub-sites in an active site, different protonation state of catalytic residues, and assistance through metal/cofactor/water. However, these mechanisms are mostly based on individual case studies. In an attempt to find general features that may facilitate promiscuity, we have performed systematic comparison of

structural features of binding/catalytic sites based on hypothesis that structural features of these sites vary between generalist (promiscuous) and specialist (non-promiscuous).

To compare differences in structural features between generalist and specialist, we used a curated list of *Escherichia coli* metabolic enzymes classified as generalist (404) and specialist (677) by Palsson's group. Subsequent to obtaining enzymes with known tertiary structures, the potential substrate binding sites were defined as predicted pocket/s (Fpocket program) harboring at least one known catalytic residues (MACiE or CSA database). This resulted in a total of 41 and 129 predicted binding sites in 35 generalist and 104 specialist enzymes respectively. Among various active site structural features (solvent accessible area, B-factor, hydrophobicity score, secondary structure content and residue propensities) analyzed, we observed that generalist tends to have greater hydrophobicity and non-polar solvent accessible surface area relative to specialist. This suggests that in general non-specific nature of hydrophobic interactions at active site might facilitate promiscuity. Even though the role of flexibility in enzyme promiscuity has been suggested before, we did not observe this as general distinguishing feature (as measured by B-factor) between generalist and specialist.

Next, we analyzed ligand induced conformational changes in enzymes and extended it to find any characteristic feature in this aspect among promiscuous enzymes. For this analysis, we constructed dataset of wild-type enzyme pairs having ligand unbound (apo) and bound (holo) structures with only one bound ligand, which is similar (score ≥ 0.8) to cognate substrate/product/cofactor of the enzyme using EC-PDB database. Subsequently, a subset of this dataset is generated with enzymes classified as specialist and generalist. On an average, global C α Root Mean Square Deviation (RMSD) and local C α RMSD (binding/catalytic residues) does not show large change (≤ 1 Å) upon ligand binding as has been reported previously. Further, local conformational changes in the binding and catalytic residues are analyzed using measures such as side-change torsional angle ($\Delta\chi_1$) and change in functional group angle (Δ C-C α -Fg). In general, only small set of binding/catalytic residues (~20%) show $\Delta\chi_1 > 20^\circ$ between apo and holo structures. Of these, a greater fraction binding residues has large ($>120^\circ$) $\Delta\chi_1$. Similarly, Δ C-C α -Fg is slightly more for binding site residues. This shows substrate binding is mostly facilitated by conformational changes involving small number of residues and catalytic residues are relatively conformationally restrained. Further, comparison of the same measures between generalist and specialist showed similar trend. Interestingly, in some specialist enzymes catalytic residues undergo greater structural change whereas little/no structural change is observed in generalist. This indicates a possibility that conformationally restrained catalytic residues in generalist may facilitate catalysis/binding of alternate reaction/substrate.

CSmetaPred: a meta-approach to predict catalytic residues

In our previous work discussed above, we observed that catalytic residues are not known for many enzymes. Moreover, elucidation of catalytic residues requires laborious experimental characterization. To facilitate experimental efforts to identify and characterize catalytic residues, we developed a consensus based method CSMetaPred (Catalytic Site **META Predictor**) to predict catalytic residues that combines prediction results from four well-known predictors *viz.* EXIA2, CATSID, DISCERN and CRpred. The residue scores obtained from these methods are normalized and then averaged (meta-score) to rank residues in CSMetaPred. Further, an improved variant of meta-approach CSMetaPred_poc was developed that combined predicted binding pockets information with CSMetaPred meta-score. The benchmarking and evaluation of methods were performed on five distinct datasets using binary classification measure such as ROC/PR (Receiver Operating Characteristic/Precision Recall) curves. Using these measures of assessment on all datasets, both meta-predictors outperform its constituent methods with CSMetaPred_poc as the best predictor among assessed methods. CSMetaPred_poc (CSmetaPred) attains highest Mean Average Specificity (ROC quantitative measure) of 0.97 (0.96) on CSAMAC dataset. The catalytic rank analysis suggested that most of the known catalytic residues have CSMetaPred_poc predicted rank ≤ 20 . On CSAMAC dataset, CSMetaPred_poc achieves average precision, recall and accuracy of 0.14, 0.87, and 0.94 respectively at rank 20. Moreover, at this rank, CSMetaPred_poc correctly predicts more than half of catalytic residues for ~93% of proteins and all catalytic residues for ~73% of proteins. Importantly, this consistent performance was observed in all datasets.

The comparison between predicted and experimentally characterized catalytic residues of enzyme structures determined subsequent to the development of our meta-predictors showed that in ~83% cases (30 out of 36 enzymes) of all known catalytic residues are within top 20 ranked positions as observed in benchmark studies. We have experimentally verified one of the predicted catalytic residues of *E. coli* GCL in collaboration with Dr. Anand Kumar Bachhawat and Shaliesh Kumar. Both meta-predictors are freely available for public use as webservers at <http://14.139.227.206/csmetapred/>.

Understanding substrate promiscuity in *E. coli* γ -glutamyl cysteine ligase (EcGCL)

γ -glutamyl cysteine ligase (GCL) enzyme catalyzes the first step in glutathione biosynthesis and forms a peptide bond between the γ -carboxylate group of L-glutamate (L-GLU) and α -amino group of L-cysteine (L-CYS) in the presence of ATP and Mg^{2+} . Dhillon and Copley, in their previous work classified GCL enzymes into Group 1 (mostly γ -proteobacteria), Group 2 (non-plant eukaryotes) and Group 3 (mostly plants and α -proteobacteria). Despite insignificant sequence similarity among these groups, tertiary structures are similar as assessed by TM-score (all-against-all TM-score ≥ 0.60). Previous studies have shown that members of group 1 and group 2 exhibit substrate promiscuity. For

instance, GCL of *E. coli* (group 1) accepts polyamines, amino acids and cysteine analogs as substrates and rat GCL (group 2) can catalyze cysteine analogs. We used docking and molecular dynamics simulations to understand the basis of substrate promiscuity in EcGCL. First we analyzed cysteine-binding pockets in all enzymes structure. The quantitative comparison of similarity in cysteine binding pockets using PS-score (pocket similarity measure calculated using APoc program) showed that these are variable across various EC class and EcGCL does not share similarity to other known cysteine binding sites.

Next, we used docking approach to elucidate binding modes of cysteine and alternate substrates of EcGCL. Here, we have docked two categories of alternate substrates: a) amino-acids (including cysteine as wild type) and b) polyamines. The comparison of docking binding energies and experimentally observed relative activities (from earlier studies) of EcGCL alternate substrates showed that α -amino-group and/or γ -carboxylate group of the substrate play a crucial role in substrate recognition. However, substrate positioning is most likely facilitated by hydrophobic interactions between hydrophobic constituent of substrate and hydrophobic residues F61, Y131, and L135 of binding site. The shallow cysteine binding site imposes another restraint on the size of substrates, which can be accommodated and oriented appropriately to facilitate catalysis. These suggest that substrate promiscuity in EcGCL is mostly likely governed by hydrophobic interactions and size of the binding site. To assess binding stability of EcGCL substrates, we docked L-CYS, L-GLU and ATP to EcGCL followed by 50 ns molecular dynamics simulation with explicit water at 300 K temperature, 1 atm pressure. Using RMSD as a measure of ligand stability, we found that L-CYS shows maximum fluctuation compared to other two ligands. Moreover, fraction of common substrate binding residues (taking initial set of residues as reference) during MD simulation is variable for L-CYS indicating that L-CYS is probably more loosely bound in EcGCL compared to other two substrates.

Enhanced function annotation and phylogenetic analysis of γ -glutamyl cysteine ligase (GCL) superfamily

We investigated sequence divergence of GCL superfamily to understand evolutionary origin of GCL and possible origin of γ -glutamyl cysteine (γ -GC) biosynthesis. In this study, we performed systematic study to: a) analyze sequence divergence of GCL families and further enhanced function annotation of GCL superfamily by classifying its families into subfamilies and b) derive possible ancestral relationship among 3 families using phylogenetic analysis. Sequence analysis showed that among three GCL families, group 3 is the most diverse and can be classified further into following seven putative subfamilies. Of these, 3 subfamilies *viz.* YbdK, Plant-like and EgtA have been characterized in earlier studies, whereas subgroup 1, subgroup 2, subgroup 3 and subgroup 4 are classified in the present study based on their sequence similarities. The substrate binding site motif analysis showed that binding residues of Mg/L-GLU are relatively more conserved than L-CYS. Especially, the binding site of cysteine is variable among group 3 subfamilies suggesting

that these subfamilies may show differing affinities possibly required for their function. Further, we annotated GCL from 5930 completely sequenced bacterial and eukaryotic genomes (including 295 plant draft genomes) by using curated HMM profile of each family and its respective subfamilies. This resulted in classification of 1083, 288 and 2325 sequences into groups 1, 2 and 3 respectively. Among group 3, YbdK has maximum members with 1225 sequences followed by Plant-like and EgtA having 460 and 181 sequences respectively. The genome analysis showed that group 3 members have undergone process of gene duplication and a possible horizontal gene transfer that is mostly observed in YbdK subfamily. Interestingly, we found members of GCL subfamily (mostly YbdK) present in organisms that are not known to produce glutathione (such as *Halobacteria* sp.). It is most likely that these organisms can produce γ -GC (reported in *Halobacteria*), which can serve similar role as glutathione. Hence, in GCL pathway evolution the γ -GC biosynthesis step might have evolved earlier than the step involving glutathione biosynthesis from γ -GC. Moreover, in few such organisms, γ -GC might act as a substitute for glutathione to combat cellular oxidative stress.

Since GCL consists of extensively diverged sequences, we manually curated structure guided sequence alignment of representative members from all three groups/families of GCL to generate a multiple sequence alignment, which is further used for phylogenetic analysis. This showed that group 1 has probably evolved independent of group 2 and group 3 while the latter groups show relatively more evolutionary relatedness amongst themselves. Further phylogeny for individual groups, showed that group 1 has distinct lineages of bifunctional GCL and traditional GCL. Group 2 is largely monophyletic which constitute non-plant eukaryotes including an exception of red and brown algae. The phylogeny of group 3 members showed that Plant-like, subgroup 3 and EgtA subfamilies are more closely related to each other with plants and green algae more evolutionary related to α -proteobacteria than to cyanobacteria. Further, subgroup 4 (which constitutes Euryarchaeota) forms a distinct group.

Appendix I: Study of conformational variability of Conserved Recognition Elements (CoREs) in long disordered regions using molecular dynamics simulations

This work was performed in collaboration with Dr. Kuljeet Singh Sandhu and Nitish Tayal. In this work, a repertoire of short evolutionarily Conserved Recognition Elements (CoREs) were identified in long disordered regions. Moreover, based on structural analysis, it was suggested that CoREs retain their three-dimensional conformation in comparison to their adjacent regions. Moreover, significantly lower median RMSD (0.37 Å) was observed for CoREs compared to its neighboring regions (2.16 Å) in multiple structural alignments of the CoRE motifs found in non-redundant PDB entries. Further, we investigated the conformational variability of these short peptides and their neighboring regions of four representative proteins using explicit water MD simulations done for 50 ns (at 298 K temperature and 1 atm pressure). The comparison of C α RMSD during MD simulations

showed that the CoRE regions were conformationally more restrained in comparison to their neighboring regions.

Appendix II: Understanding the low pH induced structural changes of *Helicobacter pylori* TlyA using molecular dynamics simulations

This work was performed in collaboration with Dr. Kausik Chattopadhyay and Amritha Sreekumar. The protein TlyA from *Helicobacter pylori* (HpTlyA) is a membrane-damaging toxin with amyloidogenic tendencies. To understand low pH induced structural change, we performed explicit water MD simulations for 200 ns (300K temperature and 1atm pressure) at low and neutral pH conditions. The analysis of MD simulations showed that at HpTlyA shows large structural changes at low pH as evident by high median C α RMSD (9.5 Å) compared to neutral pH (median C α RMSD of 6.4 Å). Moreover, compared to neutral pH conditions, structure at low pH has higher radius of gyration and reduced buried surface area between the domains. These suggest relative domain motion is primarily responsible for observed structural changes. This can be compared to the experimentally observed physiological properties of HpTlyA at low pH conditions.

List of publications arising from this work:

1. **Choudhary Preeti**, Kumar Shailesh, Bachhawat Anand Kumar, and Pandit Shashi Bhushan. 2017. “CSmetaPred: A Consensus Method for Prediction of Catalytic Residues.” *BMC bioinformatics* 18(1): 583.
2. Tayal Nitish, **Choudhary Preeti**, Pandit Shashi Bhushan, and Sandhu Kuljeet Singh. 2014. “Evolutionarily Conserved and Conformationally Constrained Short Peptides Might Serve as DNA Recognition Elements in Intrinsically Disordered Regions.” *Molecular bioSystems* 10(6): 1469–80.
3. Sreekumar Amritha, **Choudhary Preeti**, Pandit Shashi Bhushan, and Chattopadhyay Kausik “Understanding the role of acidic pH in the structural and functional mechanism of a non-conventional hemolysin *Helicobacter pylori* TlyA.” 2018. *Manuscript to be submitted*.
4. **Choudhary Preeti** and Pandit Shashi Bhushan. 2018. “Enhanced function annotation and phylogenetic analysis of γ -glutamyl cysteine ligase (GCL) superfamily.” *Manuscript in preparation*.
5. **Choudhary Preeti**, Kumar Shailesh, Bachhawat Anand Kumar, and Pandit Shashi Bhushan. 2018. “Resilience of cysteine binding pocket of *E. coli* γ -glutamyl cysteine ligase (EcGCL).” *Manuscript in preparation*.

Table of Contents

Chapter 1	<i>Review of literature</i>	1
1.1	Enzymes: essential molecular machineries to sustain life	1
1.1.1	Enzyme specificity	4
1.1.2	Enzyme binding sites	5
1.1.3	Catalytic site of enzymes	10
1.1.4	Large-scale analyses of binding sites	24
1.2	Enzyme Promiscuity	25
1.2.1	Significance of enzyme promiscuity	28
1.2.2	Dilemma in defining enzyme promiscuity	33
1.2.3	Levels of enzyme promiscuity	34
1.2.4	Types of enzyme promiscuity	37
1.2.5	Identifying enzyme promiscuity	41
1.2.6	Quantifying enzyme promiscuity	46
1.2.7	Prevalence of enzyme promiscuity at various levels	48
1.2.8	Energetics of promiscuous activities	51
1.2.9	Proposed mechanism facilitating enzyme promiscuity	51
1.2.10	Importance of inefficiency in promiscuous enzymes	60
1.2.11	Generalist and Specialists	62
1.2.12	Applications of enzyme promiscuity	63
Chapter 2	<i>Understanding residue conformational variability of binding site and catalytic residues between specialist and generalist enzymes</i>	69
2.1	Introduction	69
2.2	Comparison of structural features of active sites of generalist and specialist enzymes	72
2.2.1	Methods	72
2.2.2	Results	75
2.2.3	Conclusions	84
2.3	Investigation into global/local structural in enzymes upon substrate binding	84
2.3.1	Background	85
2.3.2	Methods	87
2.3.3	Results	94
2.3.4	Conclusions	111

Chapter 3 *CSmetaPred: a meta-approach for prediction of catalytic residues* __ 113

3.1	Introduction	113
3.1.1	Present status of catalytic residue prediction tools	115
3.1.2	Overview of the study	119
3.2	Methods	119
3.2.1	Dataset to study anatomy of active site	119
3.2.2	Overview of CSmetaPred/CSmetaPred_poc methodology	121
3.2.3	Processing of prediction server outputs and meta-score computation	123
3.2.4	Benchmark datasets for meta-predictor	126
	Generation of homology models	127
3.2.5	Metrics used in evaluation meta-predictor	128
3.3	Results	130
3.3.1	Analysis of distance geometries between catalytic residues and bound substrates/cofactors	130
3.3.2	Evaluation of meta-predictor prediction	135
3.4	Conclusions	163

Chapter 4 *Understanding substrate promiscuity in E. coli γ -glutamyl cysteine ligase (EcGCL)* __ 165

4.1	Introduction	165
4.2	Methods	170
4.2.1	Analysis of L-CYS binding enzymes	170
4.2.2	Overview of Docking procedure	172
4.2.3	MD simulations of bound and unbound states of EcGCL	177
4.3	Results	178
4.3.1	Analysis of L-CYS binding sites in enzymes	178
4.3.2	Docking studies to understand substrate promiscuity of EcGCL	184
4.3.3	MD simulations of substrate bound and unbound structures of EcGCL	191
4.4	Conclusions	197

Chapter 5 *Enhanced function annotation and phylogenetic analysis of γ -glutamyl cysteine ligase (GCL) superfamily* __ 201

5.1	Introduction	201
5.2	Methods	206
5.2.1	Construction of GCL sequence dataset	206
5.2.2	HMM profiles of GCL sub-family	207

5.2.3	Subfamily classification of GCL group 3	207
5.2.4	GCL family/subfamily annotation in completely sequenced genomes	209
5.2.5	Multiple sequence alignment of representative sequences of each GCL group	212
5.2.6	Phylogenetic reconstruction of GCL family/subfamily	214
5.3	Results	215
5.3.1	Classification of GCL into families and subfamilies	215
5.3.2	Analysis of conservation of substrate/metal binding residues	217
5.3.3	Function annotation of GCL homologues encoded in completely sequenced genomes	224
5.3.4	Analysis of sequence variation and insertion/deletion regions among GCL families	229
5.3.5	Phylogenetic analysis of GCL families/subfamilies	231
5.4	Conclusions	242
	References	253

Chapter 1

Review of literature

1.1 Enzymes: essential molecular machineries to sustain life

A popular aphorism defines life as “*a series of chemical reactions*”. Synthesis of bio-macromolecules like proteins, nucleic acids and lipids, all aspects of intermediate metabolism and intercellular communication are mostly mediated *via* a series of chemical reactions to maintain life’s critical function. However, most of these essential biochemical reactions are utterly slow, if uncatalyzed, it won’t be able to suffice life. For instance, the *de novo* biosynthesis of pyrimidines, one of building block of all nucleic acid, requires the formation of uridine monophosphate (UMP) *via* the decarboxylation of orotidine monophosphate (OMP). Measurements of the rate of OMP decarboxylation have estimated the half-life of this chemical reaction to be approximately 78 million years! Obviously a reaction this slow need very significant rate enhancement in order to sustain life. The enzyme orotidine 5'-phosphate decarboxylase is capable of accelerating the uncatalyzed reaction rate by a factor of 10^{17} enabling the completion of the catalysis of this reaction within 18 milliseconds (Radzicka and Wolfenden 1995). Thus, enzyme catalysis is essential for life with enzymes being the molecular machines that sustain life by catalyzing the most of chemical reactions associated with metabolism in all living organisms.

Enzymes represent most versatile group of all proteins and constitute significant fraction of the genome. Infact, enzymes are common products of the translation of genetic information and represent ~45% of the collective protein products of all the genomes catalogued by resources such as the UniProt Knowledge Base (UniProtKB) (The UniProt Consortium 2017). Enzymes act on substrate/s to generate product/s and during this process, they not either consumed or alter the equilibrium of catalytic

reaction. Rather, enzymes lower the activation energy barrier to enhance rate of catalytic reaction (Figure 1.1). Owing to their ability to accelerate the rate of a chemical reaction, enzymes are often the focal point for myriad of biotechnology and industrial application. Moreover, enzymes are principal target in pharmacological intervention, with a large number of approved drugs acting to modify the behavior of enzymes implicated in human disease as well as disease causing pathogens. Hopkins and Groom (2002) found that nearly half (47%) of all marketed small molecule drugs inhibit enzymes as their molecular target.

Enzyme Commission (E.C.) classifies enzymes based on their chemical reactions in a hierarchical classification system. The EC number of enzyme is a four-level descriptor, where first number indicates the reaction type, the second and third number indicates the occurring chemistry and the last number gives the substrate specificity. For example, enzyme tripeptide aminopeptidases have the EC number 3.4.11.4, where each level indicates the following:

1. EC **3** enzymes are hydrolases (enzymes that use water to break up some other molecule)
2. EC **3.4** are hydrolases that act on peptide bonds
3. EC **3.4.11** are those hydrolases that cleave off the amino-terminal amino acid from a polypeptide
4. EC **3.4.11.4** are those that cleave off the amino-terminal end from a tripeptide

There are six major classes of enzymes based on the top-level of EC number and are listed in Table 1.1. However, it has become apparent that none of these could describe an important group of enzymes, which catalyze movement of ions or molecules across membranes or their separation within membranes. Several of these involve the hydrolysis of ATP and had been previously classified as ATPases (EC 3.6.3.-), even though hydrolytic reaction is not their primary function. Such enzymes have now been classified under a new EC class of translocases (EC 7) (<http://www.enzyme-database.org/news.php>). The EC descriptors provide a classification scheme for all enzymatic reactions and also facilitate their comparison. The primary EC database is

hosted in ExplorEnz (A. G. McDonald, Boyce, and Tipton 2009) and replicated by IntEnz (Fleischmann et al. 2004) and ExPASy ENZYME (Bairoch 2000). To date (July, 2018), there are total 7269 different enzymes with defined EC number present in ExplorEnz database. Further, the similarity between enzymatic reactions can be calculated based on bond changes, reaction centers or substructure metrics using tools like EC-BLAST (Rahman et al. 2014) and SimCAL(Sivakumar et al. 2018).

Table 1.1 Six major classes of enzymes based on top-level EC number. *Recently a new enzyme level 7 has been defined to classify Translocases.*

EC top level	Enzyme class	Reaction catalyzed	Total enzyme count (ExplorEnz database)
1	Oxidoreductases	catalyze oxidation/reduction reactions; transfer of H and O atoms or electrons from one substance to another	1,832
2	Transferases	Transfer of a functional group from one substance to another. The group may be methyl-, acyl-, amino- or phosphate group	1,830
3	Hydrolases	Formation of two products from a substrate by hydrolysis	1,347
4	Lyases	Non-hydrolytic addition or removal of groups from substrates. C-C, C-N, C-O or C-S bonds may be cleaved	680
5	Isomerases	Intramolecular rearrangement, i.e. isomerization changes within a single molecule	287
6	Ligases	Join together two molecules by synthesis of new C-O, C-S, C-N or C-C bonds with simultaneous breakdown of ATP	205
7	Translocases	catalyze the movement of ions or molecules across membranes or their separation within membranes	Recently added, count unavailable yet

Three-dimensional (3-D) structures of enzymes can provide insights into mechanistic details of enzymatic reaction. Hence, there has been concerted effort to experimentally determine 3-D enzyme structures. Currently, there are 61,168 enzyme structures present in RCSB-PDB database(Berman et al. 2000). Table 1.2 summarizes enzyme information deposited in sequence/structure databases.

Table 1.2: The extent of enzyme data in various databases as in August 2018.

S.No.	Data type	Total count	Source database
1	Known enzyme reactions (unique EC numbers)	6,181	ExplorEnz: an open-access, manually curated and peer-reviewed enzyme database of the of the IUBMB Enzyme Nomenclature List
2	Enzymes in Uniprot	4,156,658 (56,635 reviewed in Swiss-Prot)	Uniprot: comprehensive resource for protein sequence and annotation data.
3	Known enzyme structures	72,727 PDB-enzyme entries in the PDB involving 61,168 separate PDB files (some files having more than one E.C. number associated with them)	EC-PDB/Enzyme-structure database: contains the known enzyme structures that have been deposited in the Protein Data Bank (PDB).
4	Known EC reactions in PDB	2,729	Protein Data Bank (PDB): archives information about the 3D structure of proteins, nucleic acids, and complex assemblies.
5	Known catalytic site and reaction mechanism for enzymes	964 hand-curated entries, 682 of them with detailed mechanistic description. covering 852 EC numbers	M-CSA: database of enzyme reaction mechanism and annotated catalytic residue for enzymes.

1.1.1 Enzyme specificity

Tracing the history of enzymology, almost a century of research can be found which focused on specificity of enzymes, which led to Beadle and Tatum's "one gene-one enzyme-one reaction" hypothesis (1941) (Beadle and Tatum 1941). These studies showed that enzymes are remarkably efficient and specific to a particular substrate or reaction they catalyze. For this, enzymes require binding of substrate/s or reactant/s in appropriate spatial orientation for efficient catalysis (Bartlett et al. 2002). However, optimization of binding pocket residues over evolutionary time has led to variability in the specificity of the binding of an enzyme to its reactants (Nobeli, Favia, and Thornton 2009). Due to various evolutionary pressures experienced by enzymes, their specificity is a continuum, which varies from highly specific to non-specific substrate/reaction

catalysis. For instance, highly specific enzymes such as glucokinase accepts only glucose for phosphorylation as opposed to non-specific enzymes like Cytochrome P450 family, which can catalyze multiple substrates. Although, grading degree of enzyme specificity is extremely challenging, in general, there are four major categories of enzyme specificity which are as follows (Freehold NJ 1972) and shown in Figure 1.1:

1. **High specificity:** an enzyme catalyzes only one reaction at a particular active site
2. **Group specificity:** an enzyme catalyzes a specific group of substrates by cleaving/ligating a specific type of bond in a particular molecular environment
3. **Bond specificity:** an enzyme acts on a specific type of bond irrespective of molecular environment
4. **Low specificity:** an enzyme acts on multiple substrates and may use same/multiple sites to bind and catalyze these substrates.


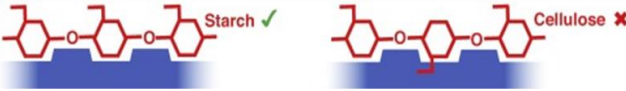


Type	Example	
1. Low	CYP450 3A4: can fit many substrates in many orientations in its binding pocket	
2. Bond	α-amylase: cleaves α -glycosidic bonds	
3. Group	Trypsin: cleaves amino group of basic amino -acids	
4. High	Glucokinase: phosphorylates only glucose	

Figure 1.1 Four major categories of enzyme specificity. Modified from (Tyzack et al. 2017).

1.1.2 Enzyme binding sites

Enzymes have “sticky” regions, few of which are, specially designed to interact with other molecules and are referred as enzyme binding site (Ringe 1995). An enzyme can have different types of binding sites, which differ in their function and bind to different molecules. Among these, the most important is active site of an enzyme, which is comprised of two components: a. catalytic site of an enzyme harboring the catalytic

toolkit of one/two to six residues, which perform catalytic reaction b substrate binding site, which recognizes substrate of an enzyme and provide binding energy to keep substrate bound to the enzyme during catalysis. Sometimes the activity of an enzyme can be regulated by another molecule that binds to a binding site distant from the enzyme active site. These molecules are called allosteric regulator and the phenomenon is referred as allosteric regulation.

The analysis of many 3-D structures of enzymes clearly showed that substrates and accessory molecules such cofactors and metals bind to specific regions on protein surface. This suggests that binding site harbors unique features that distinguish them from other regions of enzyme surface. Thus, enabling binding sites to specific substrate among myriad of other molecules present in the living cell. The substrate binding sites in 3-D structure of enzyme bound to substrate or its analogue can be identified by finding residues, which are within atomic distance cut-off (4.5 Å) of ligands and/or involved in hydrogen bonding interaction with ligands. The program such as Ligand Protein Contact (LPC) can be used to compute the atomic contacts and hydrogen bonds among ligand and binding site residues (Sobolev et al. 1999). It determines hydrogen bond using HBPLUS (I. K. McDonald and Thornton 1994) which purely uses geometrical criteria with simple atomic interaction criteria between ligand and binding site residues. We discuss some of the most important characteristics of enzyme binding sites (Figure 1.2):

1. Active sites are in the largest and deepest clefts: In 70-80 % of the enzymes, substrate or relevant cofactors bind to the largest pocket of the enzyme (R A Laskowski et al. 1996; Noyal and Honig 2006). For instance, the active site of ribosyl-transferase (PDB: 1og3) is present in the largest pocket as determined by SURFNET program (R A Laskowski 1995). The average volume of binding site depends on the type of ligand it binds and varies usually from 400-2000 Å³ (Kahraman et al. 2007). The active site of an enzyme is often found in the deepest cleft of an enzyme, which aids in maximizing the number of interactions with the substrate (Kraut et al. 2006). The average depth of the cleft that harbor binding site depends on the size of the protein and can be up to 30 Å (Coleman and Sharp 2006).

- 2. Shape of binding site is not truly complementary to ligand shape:** The specificity of enzyme are usually explained by invoking two models – the Lock and Key model by Fischer (Koshland et al. 2004) and the induced fit model by Koshland (Koshland 1958). Both of these models assume that the shape of enzyme binding pocket is complementary to the shape of the ligand. This can play crucial role in recognition process by allowing initial interaction between ligand and proteins. However, several recent studies have shown that the shape complementarity is rarely achieved between ligand and its binding site residues (Kahraman et al. 2007, 2010).

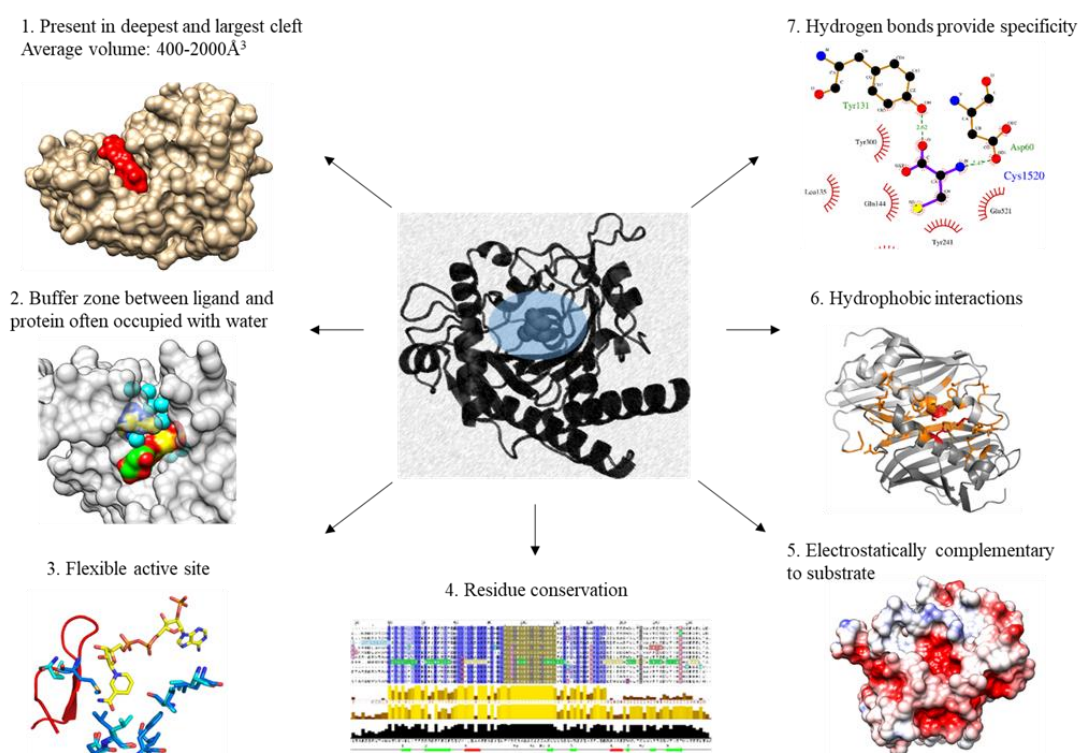


Figure 1.2 Characteristics of binding site.

Essentially, the complementary fit between ligand and binding is not perfect. In many cases, binding site residues partially enclose ligands with rest of it being exposed to the solvent. Even in the enclosed region between the ligand and its binding site residues there are empty spaces, which act as ‘buffer zone’ (Kahraman et al. 2007). This is filled with solvent water, which probably enhances the complementarity as well as provides space for motions of ligands. For instance, not every atom of substrate Adenosine monophosphate contacts with binding pocket in Adenine phosphoribosyltransferase from *Leishmania donovani* (pdbid 1qb8) and these empty spaces are filled with water. Further this buffer zone has been suggested to avoid

complete loss of ligand and binding site residue's entropy by providing the space for motions of enzyme, ligand or water molecules (Bohm, Klebe, and Böhm 1996). The geometrical shapes of binding sites can be visualized and analyzed using mathematical functions called spherical harmonics (R. J. Morris et al. 2005), which were implemented in CleftXplorer tool, which describes the shape and size of the binding sites and the ligand.

3. **Binding site residues are flexible:** The binding sites may undergo conformational change upon substrate binding. In enzymes having binding sites located in flexible loop show large conformation changes. However, fraction of such enzymes showing large conformational change is small, with majority of enzymes showing only little or no conformational changes upon ligand binding as assessed by average C α RMSD between the binding site residues of bound and unbound form of an enzyme, which is observed to be less than 1 Å (Gutteridge and Thornton 2005). Interestingly, on average, the backbone residues in the binding site are found to be more flexible compared to catalytic residues of the enzyme but both catalytic and binding residues show comparable side-chain flexibility (Gutteridge and Thornton 2005). This suggests that the active site adjusts its geometric shape according to the transition state (the conformation of the substrate at its highest reaction point). and further allow the completion of catalytic reaction. The typical method for measuring the flexibility of binding site of an enzyme is to calculate the Root Mean Square Deviation (RMSD) between different conformations of the binding site. The RMSD between two given binding site conformation is calculated using the equation:

$$\text{RMSD} = \sqrt{\frac{\sum_{i=0}^N [(x1 - x2)^2 + (y1 - y2)^2 + (z1 - z2)^2]}{N}}$$

where x , y , z are the Cartesian coordinates of the enzyme binding site atoms and N is the number of compared atoms. The qualitative analysis of the flexibility of binding site can be done using the STRuster webserver (Domingues, Rahnenführer, and Lengauer 2004). This webserver uses “all conformation versus all conformation” distance matrix to cluster each conformation according to its level of flexibility and further group similar conformations.

4. **Binding site residues are highly conserved:** It has been observed that within a given protein family, the binding site residues are usually highly conserved (Bartlett et al.

2002), which is usually to facilitate same biological function. As each protein family sequence evolves, it is subjected to natural variation under different environmental conditions and, hence, often face different evolutionary pressure and different mutation rates. Further, the tolerance level for a mutation across various residue positions is variable. For instance, mutation of functionally important residues may not be tolerated, as it will lead to loss of function in comparison to mutation of other residues. Since catalytic residues facilitate the enzymatic reaction, these are the most important residues in an enzyme and as a consequence are highly conserved. Most of the catalytic residues (~70 %) are either polar or charged (Bartlett et al. 2002). In order to calculate the conservation of each amino-acid in a given protein sequence, one can use ConSurf (Glaser et al. 2003). ConSurf uses evolutionary trace method (Lichtarge, Bourne, and Cohen 1996) in order to compute conservation and groups the residue of the query sequence into nine classes, with 1 representing the least conserved and 9 representing the most conserved residue.

- 5. Binding sites are electrostatically complementary to substrate:** Many DNA binding enzymes have a large pocket enriched in positively charged amino acid residues in order to attract their corresponding negatively charged binding partner (Tsuchiya, Kinoshita, and Nakamura 2004). Certain enzymes like copper zinc superoxide dismutase have catalysis rate close to diffusion limit. This enzyme family applies a positive electric field over the active site leading to attraction of negatively charged oxygen radicals towards the active site copper ion (Livesay et al. 2003). Prevalence of such enzyme families has led to the proposal of a theory regarding the electrostatic complementarity between the binding site and its bound substrate, which suggests that electrostatic potential forces are strong enough to attract the ligand from the solvent into the active site. The electrostatic potential can be calculated using electrostatic methodologies like APBS (Baker et al. 2001) and Delphi (L. Li et al. 2012). The pre-computed electrostatic potential surfaces for all PDB structures is present in eF-site database (Kinoshita, Furui, and Nakamura 2002). This property of complementary electrostatic potentials of binding site surface can be exploited to calculate estimated binding energies between ligand and binding site. Q-SiteFinder (Laurie and Jackson 2005) is one such tool which can be used to calculate the potential binding energies on the protein surface and detects energetically favorable surface patches that may present ligand binding site.

6. **Hydrophobic interactions essential for binding:** Many hydrophobic ligands like heme and steroids are often bound to binding sites that expose mainly hydrophobic residues. Previous experiments have shown that binding affinities of substrates can be increased by promoting hydrophobic interactions between the substrate and binding site (A. M. Davis and Teague 1999). The program HINT (Hydrophobic interactions) can be used to calculate the hydrophobicity of a molecule using experimental octanol/water partition coefficients and constructs complementarity map or hydrophathy field for a given enzyme binding site (Kellogg, Semus, and Abraham 1991).
7. **Hydrogen bonds provide binding specificity:** Since the orientation between hydrogen bond donor and acceptor atoms dictates the ability and strength of hydrogen bond, it confers directionality and specificity to the interaction between ligands and binding site residues (Martin and Derewenda 1999). The substrate must achieve appropriate orientation in the binding site to facilitate optimal interaction with binding sites. In fact, ~10-20 % of hydrogen bonds are formed with the ligand atoms (Bartlett et al. 2002). Remaining majority of hydrogen bonds in binding sites are formed among the atoms of binding residues in order to stabilize the positions of the catalytic residues. On an average, there are around 10 bonds present in protein-ligand complexes, of which one-third are hydrogen bond donors and two-thirds are hydrogen bond acceptors (Panigrahi and Desiraju 2007).

Having discussed the general features of enzyme binding sites, we describe unique features of catalytic site residues. As mentioned before, catalytic residues directly participate in enzyme catalyzed reactions. The catalytic residues are mostly the subset of binding residues (Kahraman and Thornton 2008; Tseng and Li 2011), on an average there is ~70 % overlap between the catalytic and binding site residues (Cilia and Passerini 2010).

1.1.3 Catalytic site of enzymes

1.1.3.1 Criteria to define catalytic residues

In order to study and understand the functional role of catalytic residues in an enzyme, it is necessary to define catalytic residues. However, there is no consistent way of defining

catalytic residues in literature. Different authors in the literature use different criteria in describing residues as “catalytic”. Bartlett and co-workers used the following rules in order to classify the active site residues as catalytic (Bartlett et al. 2002):

1. Residues directly involved in the catalytic mechanism- for example, a residue act as a nucleophile.
2. Exerting an effect on another residue or water molecule, which is directly involved in the catalytic mechanism and aids in catalysis. For example, by electrostatic or acid-base action.
3. Stabilization of a proposed transition-state intermediate
4. Exerting an effect on a substrate or cofactor, which aids catalysis. For example, by polarizing a bond which is to be broken, or steric and electrostatic effects

Subsequently, developer of MACiE database (discussed in details in section 1.1.3.3) database the residues defined as ‘catalytic’ based on the above definition is further divided into two categories as follows: (Holliday, Almonacid, Bartlett, et al. 2007)

1. **Reactant:** are catalytic residues, which are directly involved in catalytic reaction mechanism and their chemical structure is modified during the course of reaction. For example, the residue involved in covalent catalysis, electron shuttling or proton shuttling.
2. **Spectator:** residues plays an indirect but essential role in enzymatic reaction mechanism. The chemical structure of these residues remains unmodified during the reaction. For instance, residues that polarize or alter the pKa of a residue, a water molecule or part of the substrate directly involved in the reaction, affect the stereospecificity or regiospecificity of the reaction, or stabilize the reactive intermediates (either by stabilizing the transition states or the intermediates themselves, or destabilizing the ground states of the substrates).

1.1.3.2 Identification of catalytic residues

The catalytic residues of an enzyme are mostly determined using tedious and cumbersome experimental studies. The site-directed mutagenesis is the most common experimental approach to characterize catalytic residues that involves testing the loss of enzyme activity upon mutation of crucial residues. For example, mutation of active site

residue arginine to lysine of tyrosine phosphatase leads to 8,200-fold reduction in K_{cat} but K_m remained unchanged (Z.-Y. Zhang 2003). The unchanged K_m and reduced K_{cat} suggests that R to K mutation has probably disrupted catalysis without disturbing the substrate binding affinity. Even though both arginine and lysine are positively charged residues, lysine uses ϵ -amino group to provide positive charge, unlike arginine, which uses guanidinium group. Thus, it was deduced that the guanidinium group of arginine must have been making specific contacts in the transition state, which aids catalysis. Hence, arginine is probably a catalytic residue of tyrosine phosphatase enzyme. Another brute force and systematic technique is alanine scanning approach, which is often used in the absence of any prior knowledge enzyme function. In this approach, every residue is mutated to alanine and effect of the same is measured on the enzymatic reaction (Morrison and Weiss 2001). Such mutational analyses are very cumbersome and time consuming. Moreover, many times the removal of an essential catalytic residue doesn't lead to complete abolishment of catalysis and enzyme adopts an alternate mechanism which often lead to catalysis but at a comparatively slow rate (Peracchi 2001).

Chemical labelling (Aktories 1997) is an alternative approach to identify catalytic residues by forming covalent bonds with the residues directly involved in enzyme reaction mechanism. Use of inhibitors that form covalent links with the catalytic residues involved in catalysis can also be used to elucidate catalytic residues. For instance, the catalytic machinery of serine proteases was elucidated by the use of inhibitors diisopropylfluorophosphate (DFP) and tosyl-L-phenylalanine chloromethyl ketone (TPCK), which made covalent bond with catalytic serine and histidine residue respectively (Hedstrom 2002). Moreover, pH rate profiles and NMR experiments can also suggest the involvement of different chemical groups in the active site, by tracking protonation changes and subsequent changes in the enzymes activity. In pH rate profiling, measuring the enzyme activity at variable pH conditions may reveal the protonation states, which are the most conducive for catalysis.

Measurement of kinetic parameters like K_{cat} and K_m can also help in elucidating the residues important for binding or/and catalysis. These parameters can be calculated by determining the rate of formation of products or depletion of substrates using the techniques like spectrophotometry, spectrofluorimetry and radioactivity assays. Such techniques also aid in identifying the number and sequence of intermediate processes in

an enzyme mechanism. For instance, during the catalysis of the hydrolysis of p-nitrophenyl acetate by chymotrypsin, a sudden burst of product release is observed after which the breakdown occurs relatively slowly. Quick formation of an intermediate leads to this sudden burst during which p-nitrophenol is released. However, as the enzyme is saturated by intermediate molecules the turnover of the enzyme is slowed down subsequently (Fersht A 1999).

These traditional molecular biology techniques for finding catalytic sites, such as mutagenesis (Morrison and Weiss 2001), pH dependence (Zhou and Toney 1999) and chemical labelling (Aktories 1997) are generally time consuming, and rely on some prior knowledge of the function of the protein to allow it to be assayed.

Catalytic residues can also be identified by analysis their conservation in their related homologs. In the absence of above mentioned, direct evidences for identification of catalytic residues, one can also predict catalytic residues computationally. There is plethora of catalytic site prediction tools which use sequence or/and structure based information in order to predict catalytic residue and exploit properties like conserved residues, conserved surface patches, geometric or network based properties. The various catalytic site prediction tools available are further discussed in details in chapter 3.

1.1.3.3 Catalytic residue annotation in various databases

As discussed in previous section, the experimental characterization and identification of catalytic residues is not a trivial task. To create a compendium of known catalytic residues for various enzymes requires collating data from research articles that needs both enormous efforts and careful curation. This is essentially because catalytic residue information is usually embedded in the text of various research publications and although can be automatically extracted, it requires manual curation for any reliability. Further, to weigh evidences for each residue in any proposed catalytic mechanism require expert annotators. For instance, evidence includes analyzing the data from number of sources: mutagenesis, pH profiling and chemical labelling. An important way to identify catalytic residue is experimentally determining substrate or its analogue or transition state analogue enzyme bound structure and then use proposed reaction mechanism to find catalytic residues. Hence, most catalytic site information is always linked with

information derived from 3-D structure, especially analyzing the ligand bound enzyme structure which facilitates the identification of appropriately oriented residues relative to the substrate and essential for catalysis. Below, we discuss publicly important catalytic site databases *viz.* MACiE and CSA.

Catalytic Site Atlas (CSA) is a database which documents catalytic site of enzymes with known 3-D structure in RCSB PDB database (Porter, Bartlett, and Thornton 2004; Furnham et al. 2014). There are two types of entries in CSA database and are as follows:

1. **CSA literature:** includes hand-annotated which are primarily derived from various research publications and are manually curated.
2. **CSA homology:** includes homologous entries with catalytic residues defined as equivalent sites in related proteins derived subsequently using various sequence comparison methods to one of entries original entries.

The CSA webserver is available at: <http://www.ebi.ac.uk/thornton-srv/databases/CSA/>. A user can access CSA using pdbid, Uniprot ID or EC number. Last updated version of CSA; CSA 2.0 lists 968 and 32,216 enzyme structures as CSA literature and CSA homology entries (Furnham et al. 2014). This covered ~70 % (34,096/49,049) of enzyme structures deposited in PDB at that time (year 2014) and have 1,189 EC numbers covering the enzymes from all classes and subclasses and most sub-subclasses. This update extended the data in CSA enormously compared to just 177 CSA literature and 2,608 CSA homology in CSA 1.0 (Porter, Bartlett, and Thornton 2004). Later in the year 2017, the information contained in CSA database is no longer updated and a newer form of this database M-CSA(Ribeiro et al. 2018) is made available and is discussed later in this section.

MACiE (Mechanism, Annotation and Classification in Enzymes) database provides catalytic residues of various enzymes and apart from this, it also describes their respective putative role in enzyme reaction mechanism (Holliday et al. 2012; Holliday, Almonacid, Bartlett, et al. 2007). This is the only database to provide the role of catalytic residue at various reaction steps. An important motivation to create MACiE database was to understand role of amino acids in catalyzing a vast array of chemical reactions and also enlist various mechanism of catalytic reactions. Further, the detailed catalytic steps of reaction mechanism present in this database helps us in understanding the evolution of

enzymes as well as their reaction mechanisms (Holliday et al. 2011; Glasner, Gerlt, and Babbitt 2006). Moreover, reaction mechanism of two related can be compared to understand common a set of steps in a reactions (Akiva et al. 2014).

This database is evolved from CSA database and following criteria is followed while adding an entry in MACiE database:

1. Enzyme structure deposited in PDB database (Berman et al. 2000)
2. Well understood mechanism of enzyme catalytic reaction, acquired from various research publications and deciphered from numerous methods such as chemical and biochemical studies, structural biology reports and quantum mechanical calculations.
3. Each enzyme is unique at a hierarchical classification system of protein domain structures of CATH classification(Sillitoe et al. 2015), unless its homologue has a significantly different catalytic mechanism
4. Preference given to wild type PDB structure in case where multiple 3D structures are available of a given enzyme. Mutated or engineered structures are ignored unless an alternate wild type structure is absent.

The MACiE database can be accessed at the following url <http://www.ebi.ac.uk/thornton-srv/databases/MACiE/index.html>. The latest version of MACiE (version 3.0) contains 335 fully annotated enzyme reaction mechanism, comprising 372 distinct CATH codes and 321 EC numbers with 182 EC sub-subclasses.

This database was further extended to include metalloenzymes and a separate database Metal MACiE was made available for catalytic metal ions. This database aids in understanding the functional role metal ions involved in catalysis and is available at the following url http://www.ebi.ac.uk/thornton-srv/databases/Metal_MACiE/home.html. This database currently covers 75 % of the metal-dependent EC sub-sub classes(Andreini et al. 2009).

M-CSA (Mechanism and Catalytic Site Atlas) is a database of enzyme active sites and reaction mechanisms (Ribeiro et al. 2018) and is mostly obtained by merging MACiE and CSA database with updated information. M-CSA can be accessed at: <https://www.ebi.ac.uk/thornton-srv/m-csa/>. This unified resource facilitates the searching of active sites and remove redundancy between the resources. M-CSA

currently has 961 entries, among which 423 entries have detailed mechanism information and remaining 538 entries have information about catalytic residues only. Among total 6,028 enzyme PDB structures with associated EC number, M-CSA database covers 81 % (195/241) of third level EC number and 30 % (840/2793) of the fourth level EC number. Further, the coverage of M-CSA is extended to 51,993 enzyme structures and over five million sequences by searching close homologues in PDB and UniprotKB databases respectively. It should be noted that the above mentioned resources does not include residue solely involved in binding of the substrate and hence differs from other resources like UniProtKB annotations.

The growth of these catalytic site databases with respect to time is shown in Figure 1.3.

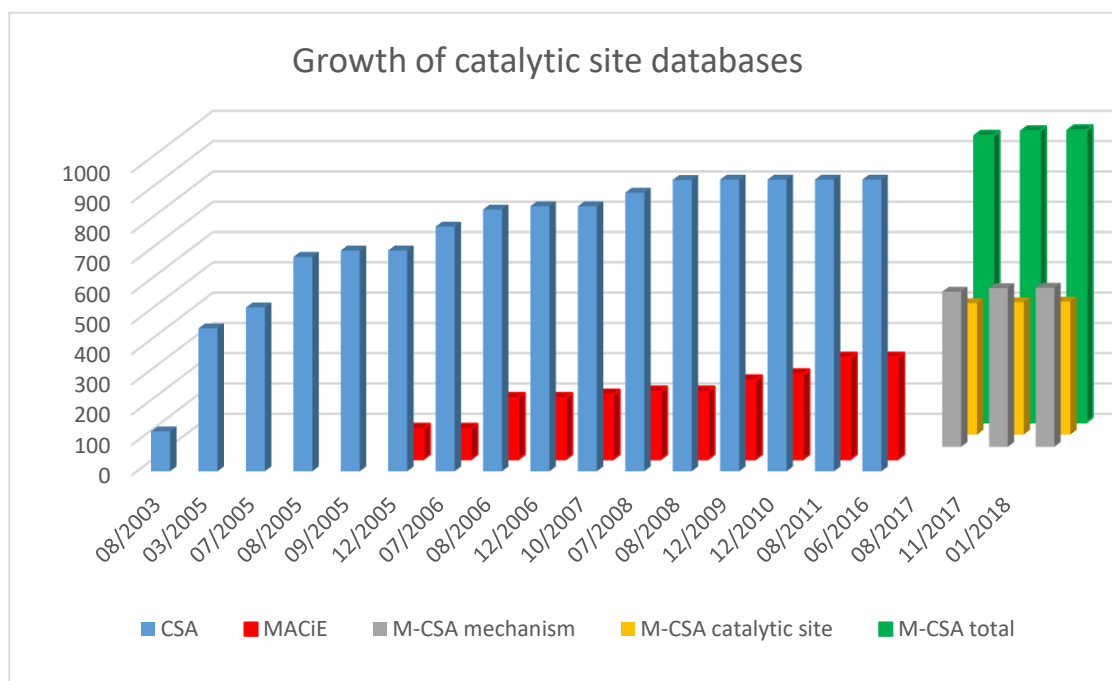


Figure 1.3 Summary of growth of catalytic site databases with time. (Data source: M-CSA database)

1.1.3.4 Characteristics of catalytic site residues

Bartlett and co-workers analyzed the catalytic residues in 178 enzyme active sites in order to gain insight into general active site environment (Bartlett et al. 2002). Except two NMR-derived enzyme structures, all other were X-ray crystal structure with resolution varying from 1.5 to 3.2 Å. All the enzyme in their dataset have well defined active site and known reaction mechanism available in literature. The criteria for

defining catalytic site are already discussed in section 1.1.3.1. Following are the summarized general characteristics of catalytic site identified in their study.

Frequency of amino acid type and their role in catalysis: Since enzyme catalysis majorly involve electrostatic interactions and are driven by such forces, which facilitates common steps in catalysis such as movement of electrons/protons and charge stabilization, catalytic residues are mostly (~92%) composed of charged and polar residues (Bartlett et al. 2002). Of these, most (65%) catalytic residues charged amino acids such as Histidine, Arginine, Lysine, Glutamate and Aspartate. The polar amino acids constitute 27% of catalytic residues that includes Glutamine, Threonine, Serine, Asparagine, Cysteine, Tyrosine and Tryptophan. Rest 8% of catalytic residues were hydrophobic amino acids. The hydrophobic residues are usually involved in the non-specific interactions between the ligand and the enzyme but they can also exert steric strain on the substrates and help in lowering the energy of transition state of the reactant or can provide neutral environment to increase the catalytic power of the charged moiety (Bartlett et al. 2002). Owing to its unique property to be either charged or neutral at physiological pH, Histidine is the most common (18%) catalytic amino acid residue despite of its low (2.7%) overall abundance. Histidine can act as nucleophile, acid, base or stabilize the transition state of the reaction. Amino acids Aspartate and Glutamate contribute 15% and 11% to catalytic residues respectively. Despite their similar natural abundance (5.7% for D and 5.9% of E), Aspartate is more preferred over Glutamate due to shorter side-chain making the side-chain less flexible and aid in catalysis by holding the residues in correct orientation. Arginine and Lysine comprise 11% and 9% of catalytic residues respectively. Despite of its lower abundance (4.9%) compared to lysine (5.9%), arginine is more favored as all nitrogen groups in its side-chain can be involved in electrostatic interactions in contrast to only one nitrogen group in the lysine side-chain. Arginine also harbors appropriate geometry to stabilize the oxygen atoms on the phosphate group, which is commonly found biological moiety. 5.6% of catalytic residues are constituted by rarely occurring amino acid Cysteine (1.2% natural abundance). This prevalence of cysteine in catalytic residues indicates the importance of thiol group and disulphide bridges involvement in catalysis by certain enzymes. These include enzymes like glutathione reductase (Karplus and Schulz 1989; Pai and Schulz 1983) and protein

disulphide isomerase (Wilkinson and Gilbert 2004; Gruber et al. 2006; Galligan and Petersen 2012).

Catalytic propensity/catalycity of residue types: The catalytic propensity of an amino acid indicates how often a given amino acid is involved catalysis compared to it being involved randomly and is calculated using the equation:

$$\text{Catalytic propensity} = \frac{\text{percentage of a given residue in catalytic residues}}{\text{percentage of given residue in all residues}}$$

Catalytic propensity is >1 indicates propensity for that residue to be catalytic is more than expected by chance, and if it is <1, then the residue is less catalytic than might be expected (Holliday et al. 2011; Ribeiro et al. 2018). The residues Histidine and Cysteine have the highest catalytic propensity (Figure 1.4). Further, in the list of residues with high catalycity, charged residues are followed by polar and hydrophobic residues have least catalytic propensity. The high catalycity of cysteine and histidine can be due to their side-chain pKa values, which is closest to the biological pH. The side-chain pKa of free histidine is 6.0 and only a small fraction of its side chain will be protonated at the cellular pH of 7.4. Moreover, histidine can readily alter its protonation state depending on its surrounding environment (S. Li and Hong 2011). Similarly, the –SH group of cysteine can easily be deprotonated as its pKa value is between 8 and 9 (Dudev and Lim 2002; Bombarda et al. 2001). Hence, ability of these amino acids to protonate and deprotonate at physiological pH is advantageous in acid-base reactions during catalysis where these states can act as acid/base/nucleophile in catalytic reaction steps or aid in stabilization of transition state (Bartlett et al. 2002). This catalytic propensity computed is in consensus with similar prior analyses (Zvelebil and Sternberg 1988), (Ribeiro et al. 2018) and Bartlett's work (Bartlett et al. 2002). The number enzymes used in previous work was far smaller (17) compared to recent work, which included 961 enzymes suggesting the catalycity may be indeed general characteristics of enzymes.

Although histidine and cysteine are observed to be catalytic residues among all EC classes, the catalycity for other residues differs in six EC classes (Holliday et al. 2011). Infact, only 10 catalytic residues (Arg, Asp, Cys, Glu, His, Lys, Ser, Thr, Trp and Tyr) can perform all the functions associated with all EC classes (Holliday, Almonacid, Mitchell, et al. 2007; Holliday, Mitchell, and Thornton 2009).

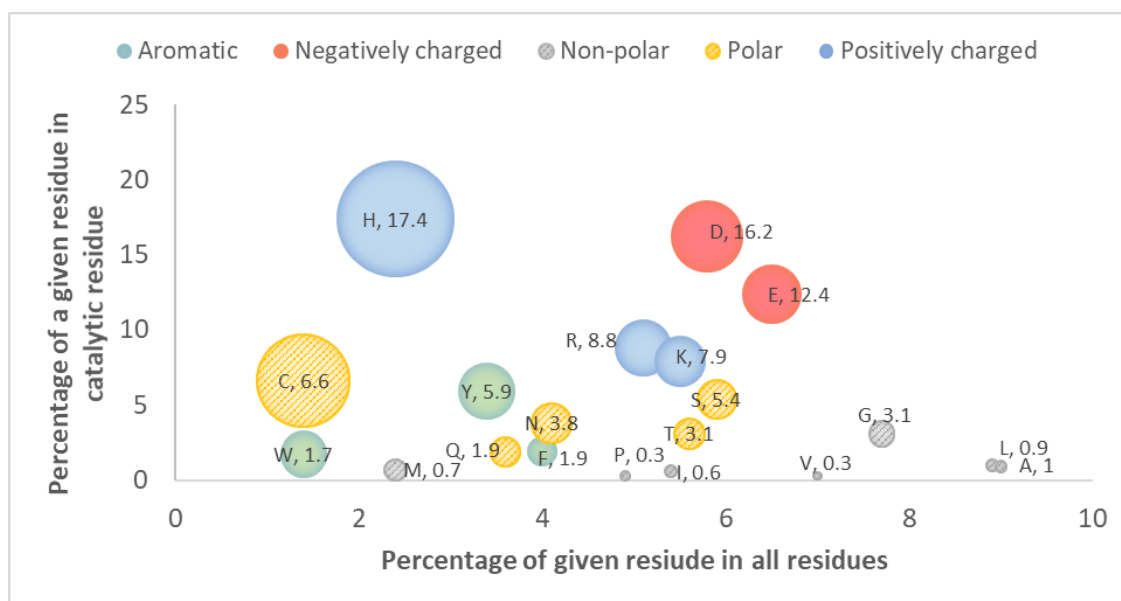


Figure 1.4 Percentage of a given residues among catalytic and all residue in enzyme. The name of the concerned amino acid and its catalytic propensity is labelled on the circle. The size of the circle is proportional to the catalytic propensity of a given residue, which is colored according to nature of amino-acid. (Data source: M-CSA database)

The difference in the catalytic propensities of these residues is further reflected in the type of functions these residues can perform in different EC class of enzymes (Holliday, Mitchell, and Thornton 2009; Holliday et al. 2011). Broadly, the functions performed by amino acid residues during catalysis can be split into seven categories as follows:

1. Activation – residues that are responsible for activating other species
2. Steric role – residues that affect the outcome of the reaction through steric considerations
3. Stabilization – residues that (de)stabilize other species
4. Proton shuttling – residues that donate, accept or relay protons
5. Hydrogen radical shuttling – residues that donate, accept or relay hydrogen atoms
6. Electron shuttling – residues that donate, accept or relay electrons, either singly or in pairs
7. Covalent catalysis – residues that become covalently attached to a reaction intermediate

The function profiling using these categories revealed that stabilisation seems to be the most important and dominating role of residues during catalysis (Holliday, Mitchell, and

Thornton 2009). All functional category except hydrogen radical shuttling and covalent catalysis can be performed by all catalytic residues.

Similar to catalytic propensities, functional profiles of the catalytic residues varies in different EC class (Holliday, Mitchell, and Thornton 2009; Holliday et al. 2011). For instance, serine is mostly involved in the covalent catalysis in hydrolase class but not seen in this role in any other EC class (Holliday, Mitchell, and Thornton 2009). Although, difference in catalytic propensity can be related to different functional role a residue, it is still unclear “why an amino acid residue prefers certain catalytic function in one EC class in comparison to other EC classes” (Holliday et al. 2011).

Side chain interactions are more common in catalytic residues: Among catalytic residues, 92% of these use their side chain atoms for interaction, while only 8% of them interact *via* their main chain atoms (Bartlett et al. 2002). Among main chain atoms, N-H group interactions are more predominant (82%), compared to C=O group. The main chain interactions are usually involved in stabilization of the transition state intermediates. Glycine is the most common catalytic residue, which uses main-chain interactions as it can fit in any gap in active site given its small size. For instance, G30 in phospholipase A2 is used to stabilize oxyanion holes (D. L. Scott et al. 1990). Other non-polar catalytic residues also mostly act through their main chain atoms (Holliday et al. 2011).

Prevalence of catalytic residues in coil secondary structure: The analysis of secondary structure propensity of catalytic residues showed that most (50%) catalytic residues lie in coil regions (Bartlett et al. 2002). The catalytic residues show similar prevalence in both regular secondary structures α -helices and β -strands with each having 28% and 22% of catalytic residues respectively. In α -helices, catalytic residues are often present on internal face of the helix, whereas in β -sheets, these are either present in an edge strand or at the end of the strand. Thus, unlike α -helices, catalytic residues in β -sheets are available for interaction with substrate (Bartlett et al. 2002). A recent large-scale analysis of catalytic residues derived from 379 non-redundant CATH enzyme superfamilies revealed that large number of catalytic residues are present in the loops of large (~77 %) proportion of enzyme superfamilies and are functionally diverse and usually belong to alpha/beta fold (Furnham et al. 2016).

Lower solvent accessibility of catalytic residues: In comparison to fully exposed residues, 89% of catalytic residues have a relative solvent accessibility of less than 30% (Bartlett et al. 2002). In fact, 5% catalytic residues are totally buried with 0% RSA, 50% of catalytic residues are found in 0-10% RSA and 25% in 10-20% RSA. Thus, majority of catalytic residues are less exposed to solvent, which might restrict the motion of catalytic residue and orient them in appropriate orientation. As mentioned before, majority of the catalytic residues are present in the largest and deepest cleft of the enzyme (R A Laskowski et al. 1996), allowing the ligand to bind in solvent-free environment.

Limited residue flexibility in catalytic residues: Catalytic residues tend to be more rigid as indicated by their relative low B-factor compared to non-catalytic residues (Bartlett et al. 2002; Yuan, Zhao, and Wang 2003). The average residue B-factor in the crystal structure can be correlated with residue flexibility. It has been shown that the restricted motions at the catalytic sites are not due to the presence of substrates at those sites but are intrinsic mechanical properties of the enzymes themselves irrespective of bound molecules (L.-W. Yang and Bahar 2005). Interestingly residues having long side chains such as arginine, lysine, aspartate, and glutamate generally have flexible side-chains, however, these tend to have lower B-factor when they play catalytic role (Bartlett et al. 2002). Gerek et al. in 2013, proposed a novel dynamic flexibility index (*dfi*) to quantify dynamic properties of individual residues when a perturbation is introduced in a given protein (Nevin Gerek, Kumar, and Banu Ozkan 2013). They showed that catalytic residues usually exhibit low *dfi* profiles and are dynamically robust residues. Such robustness can be beneficial for catalytic residues as mostly they are buried within the core of the protein (Bartlett et al. 2002) or co-localized with inflexible hinge regions (Yang and Bahar 2005). In comparison to catalytic residues, binding residues tend to have high dynamic flexibility aiding them to accommodate binding induced conformational changes. Further, in Gerek's dataset, majority of catalytic and binding residues were present in loops regions. Although, the loop regions usually have high *dfi* values, these functionally critical residues showed less dynamic flexibility than other positions (Nevin Gerek, Kumar, and Banu Ozkan 2013).

High degree of conservation in catalytic residues: As discussed earlier, catalytic residues are highly conserved as these play crucial role in enzymatic function (Capra and Singh 2007) suggesting strong selection pressure against mutating these

catalytic residues in comparison to residues lying in the binding site cavity important for substrate recognition. In order to understand the evolution of catalytic machinery with different functions, Furnham et al. studied the changes in catalytic site among related proteins and analyzed 101 enzyme superfamilies with at least two families with different functions (Furnham et al. 2016). Among these, ~40 % superfamilies have at least one catalytic residue completely conserved among all their functional families indicating the residues essential for a common catalytic step conserved across the superfamilies.

Enriched hydrogen bond interaction by catalytic residues: In a study on properties of catalytic residues showed that majority (93%) of catalytic residues are involved in at least one hydrogen bond interaction either as hydrogen donor or acceptor (Bartlett et al. 2002). Among these, most hydrogen bonds are with another residue suggesting role of these residues in orienting catalytic residues in the binding pocket. The hydrogen bonds help in maintaining the structural integrity of active site. The spectator catalytic residues often forms hydrogen bond interactions with other species (Holliday, Almonacid, Mitchell, et al. 2007).

Preference of residue type for performing specific functions during catalysis: The most common function of catalytic residue is stabilization of the transition state formed during catalysis (Bartlett et al. 2002; Holliday et al. 2011). Among catalytic residues, Arginine is the most common residue involved in stabilization of transition state. Further, it has been shown that hydrophobic residues like G, F, L, M, A, I, P and V are also involved in stabilization of a proposed transition state intermediate (Bartlett et al. 2002). Generally, cysteine and serine act as a nucleophile and negatively charged amino acids like aspartate and glutamate are typically acid/bases during catalysis.

As discussed earlier, the binding of ligand to an enzyme is primarily due to complementary physiochemical properties between the ligand and its binding site. However, this is not always true, there are many examples which show binding despite of non-complementarity. These include enzymes like DNase I (pdbid:2dnj), sulphate binding proteins (pdbid:1sbp), phosphate receptors (pdbid: 1pbp) and flavodoxin structures (pdbid: 2fox). Although, all these structures bind to a negatively charged substrate, still a negative electrostatic field is exerted in their respective binding site. The protein stabilizes the anion charges by Van der Waals interactions and an extensive local

hydrogen bonding network. For instance, as shown in Figure 1.5, in DNase I, the divalent cations bound to its structure compensates for the otherwise non-complementary (negative electrostatic potential) surface (Gueroult et al. 2010). Nakamura *et al.* gave following reasons to justify the absence of electrostatic complementarity in few protein-ligand complexes (Nakamura et al. 1985):

1. The substrate interacts mainly with solvent molecules.
2. Interactions are primarily driven by hydrophobic interactions.
3. Incorrect assignment of dissociation of the ionizable protein residues
4. Effect of additional ionic substrates on the electrostatic energy experienced by the ligand

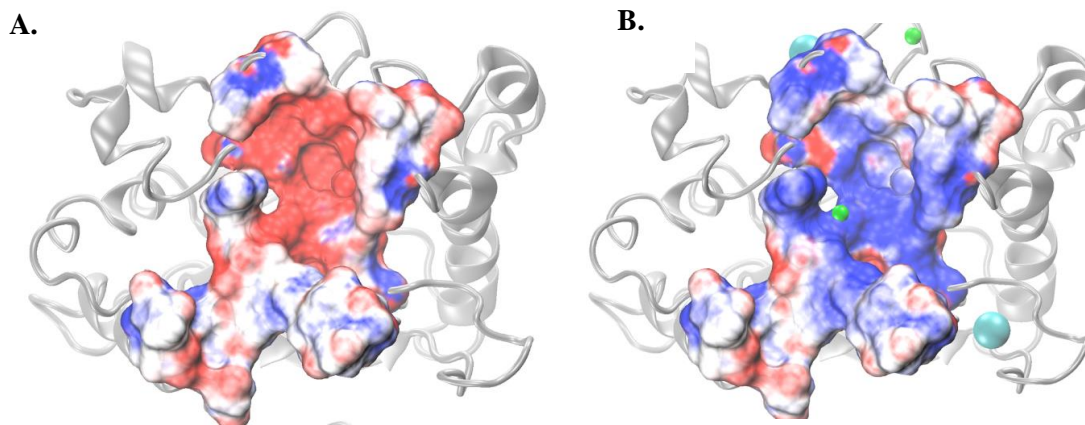


Figure 1.5 Non complementary of substrate binding site assisted by metal ions. *Electrostatic potential surface of DNaseI structure (3dni) from MD simulations A) negatively charged in the absence of counter-ions, thereby inhibiting DNA binding B) highly positively charged in presence of Mg²⁺ and Ca²⁺ facilitating DNA binding. Here the electrostatic potentials are colored from -5 kT/e in red to 5 kT/e in blue. Adapted from (Gueroult et al. 2010).*

The absence of complementarity among the ligand and binding site arose numerous questions which still need to be answered such as if their binding site possessed complementarity in ancestral enzymes which was lost subsequently during the course of evolution. Some of the previous large-scale analyses of binding sites are discussed below which shed more light on the general behavior of binding sites.

1.1.4 Large-scale analyses of binding sites

Considering the geometric and physicochemical complementarity exhibited between the binding site of enzyme and its bound substrate, binding site that bind the same cognate ligand should share very similar properties. However, this assumption is not always as expected (Gueroult et al. 2010; Kahraman et al. 2010). While analyzing 100 non-homologous protein binding pockets, Kahraman *et al.* found that same substrate can bind to pockets with distinct physiochemical environments (Kahraman et al. 2010). The higher variation in binding site was often associated with large energy fluctuations necessary for function of the protein. Thus, the degree of complementarity between the substrate and the binding site is often dictated by the functional role of the ligand rather than its chemical composition. Interestingly, the hydrophobicity show comparatively less variations compared to electrostatic potentials. Thus, nature has evolved multiple binding solutions for the same substrate. However, this study was limited to the proteins binding to any of the nine ligands- AMP, ATP, FAD, FMN, glucose, heme, NAD, phosphate and steroid molecules (dehydroepiandrosterone and estradiol). Further Khazanov and Carlson did much more large-scale analysis to explore the generic composition of protein-ligand binding site and analyzed 3,295 non-redundant proteins with 9,114 non-redundant binding sites (Khazanov and Carlson 2013). More importantly, this study distinguishes between valid (biologically functionally relevant) and invalid (crystallographic additives) ligand contacts using annotations from binding MOAD database (Hu et al. 2005; Ahmed et al. 2015). They found that all the residues have more contacts to valid ligands in comparison to invalid ones apart from cysteine and histidine, which showed more and equal contacts in invalid-ligand bound sites respectively. However, this might be an artifact of the dataset as valid ligand usually included bulky substrates. Further, Cys, Ala, Thr, Asp, Ser, and Gly were found have most (≥ 0.78) contacts/heavy atom while Leu, Ile, Val, Phe, and Pro had the least contacts/heavy atom (≤ 0.62). Total raw contacts for all residues interacting with valid ligands. This study suggested that Trp and Cys are “sticky” for all small molecules irrespective of valid or invalid ligands. Moreover, the valid binding site were found to be more biased for Ala, Ile, Met, and Val while His, Pro, Gln, Glu, Lys, and Arg showed biasness for invalid binding site. These observations can further be exploited to improve scoring of the

predicted binding site with higher weight given to residues with higher propensities in the valid ligand binding sites.

Another large-scale analysis done by Gao and Skolnick showed degeneracy in structural space of pockets (Gao and Skolnick 2013). In this study, they curated and analyzed a non-redundant set of 20,414 substrate-bound pockets which contained 9,485 unique ligands and found only 1,000 pocket shapes to be sufficient to represent the structural space of protein pockets. Such degeneracy suggests that the same ligand can bind to similar shaped pockets. Further, they investigated the relation between ligand similarity and pocket similarity using Tanimoto coefficient (Tc) and P-values of the PS-scores (a quantitative measure for pocket similarity). In this study, it was observed that ~13% of ligand pairs share significant similarity (Tc >0.4), which are binding to similar pockets with $0.01 < P\text{-value} < 0.05$. Further, on increasing the binding pocket similarity level to P-value of 1×10^{-3} and 1×10^{-5} , the similar ligands population was further increased to 31% and 37% respectively. Surprisingly, on further increasing the P-value $< 1 \times 10^{-5}$, the similarity among ligand pairs drops to 18%. In such prodigious cases, same pockets were recognized by chemically different substrates and are referred as “promiscuous pockets”. Moreover, these pockets are essentially from the same protein structure solved with different substrates and constitute 34% (6,913 unique pockets) of all pockets in their dataset. This clearly indicates that promiscuity is not an exception. The earlier predominant view of enzyme specificity is shifted to enzyme promiscuity and many enzymes are now discovered which apart from their specific native active, perform other side activities for which a given enzyme wasn't evolved and are referred as promiscuous activities. Further, enzyme promiscuity is discussed in details below.

1.2 Enzyme Promiscuity

Tracing back the history of enzymology, almost a century of research can be found which focused on specified of enzymes as aptly put in Beadle and Tatum's “one gene-one enzyme-one reaction” hypothesis given in 1941 (Beadle and Tatum 1941). Owing to this, biochemistry textbooks states that enzymes are remarkably efficient and specific to a particular substrate or reaction they catalyze. However, recent studies have shown that many enzymes harbor inherent capability to catalyze alternate reaction/s or substrate/s

apart from their physiologically evolved activities, usually assumed through same active site. These enzymes are called as promiscuous (meaning “mixed-up” in Latin) enzymes and their secondary adventitious reaction/s are referred to as promiscuous reactions. Thus, traditional view of enzymes as highly specific biocatalysts is now being replaced by “avant-garde view” in which enzymes may exhibit promiscuity.

In general, promiscuous reactions are catalytically inefficient as values of K_{cat} and K_{cat}/K_m are often very low compared to their physiologically relevant activity of enzyme. For example, the native activity of a decarboxylase, Malonate semialdehyde decarboxylase (MSAD) is 3.5 fold more efficient than its promiscuous hydratase activity (G J Poelarends, Veetil, and Whitman 2008). Such low level of promiscuous activity is usually undetectable under normal conditions, however, it may become important if metabolites or concentration of metabolic enzymes changes due to environmental/genetic factors. For instance, transaldolase (talA/talB) deleted strain of *E. coli* ($\Delta talA/\Delta talB$) is not expected to grow in media having xylose as a sole carbon source, however, experimentally no growth defect was observed (Nakahigashi et al. 2009). Under xylose as carbon source, transaldolase becomes an essential gene as it is involved in gluconeogenesis as well as biosynthesis of precursor erythrose-4-phosphate. Here, *E. coli* ($\Delta talA/\Delta talB$) is able to grow due to promiscuous activity of two enzymes, phosphofructokinase (pfkA) and fructose 1,6-bisphosphate aldolase (FbaA). The accumulation of Sedoheptulose-7 phosphate (S7P) triggers the promiscuous activity of phosphofructokinase (PfkA) which phosphorylates S7P into Sedoheptulose-1,7-bisphosphate (S1,7P) (Figure 1.6). Further, S1,7P cleaves into erythrose 4-phosphate and dihydroxyacetone phosphate by a promiscuous activity of fructose 1,6-bisphosphate aldolase (FbaA). Thus, promiscuous activities can act as a repertoire of catalytic activities, which could be recruited to confer fitness benefit to the organism under selective pressure. Moreover, these enzymes could serve as a starting point for the emergence of new efficient enzyme functions by gene duplication and divergence.

Further, promiscuous activities of enzyme, usually low levels, can increase due to mutations in the process of adaptation of organism in the new environment. For instance, *Pseudomonas sp.* gained ability to degrade atrazine, a herbicide, subsequent to its extensive usage and the first enzyme chlorohydrolyase in atrazine degradation pathway evolved from melamine deaminase as both enzymes are 98% identical (Copley 2009).

The promiscuous activity was increased to more than 500-fold by introducing single mutation in case of 4-oxalocrotonate tautomerase and γ -glutamyl phosphate reductase (Rahimi et al. 2016; Khanal et al. 2015).

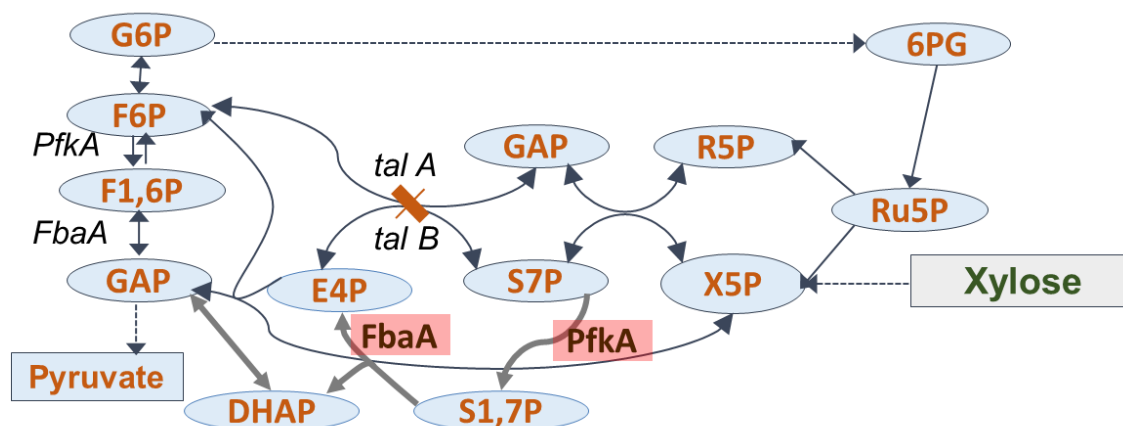


Figure 1.6 Alternate metabolic routes involving promiscuity in *E. coli* (Δ talA/ Δ talB) strain. Solid lines show normal xylose utilization pathway. Thick lines show alternate metabolic route under talA/tabB deletions involving promiscuous activity of FbaA and PfkA highlighted in red box.

In some cases, such as muconate lactonizing enzyme II from *Pseudomonas sp. P51* (MLE II), mutating E323 to G323 resulted in more than 10^6 fold increase in its promiscuous *o*-succinylbenzoate synthase activity (Schmidt et al. 2003). Conversely, mutations can also decrease promiscuity. For instance, γ -humulene synthase from *Abies grandis* exhibits product promiscuity (discussed in details in section 1.2.4) and can catalyze farnesyl diphosphate into more than 50 different sesquiterpenes. Residues responsible for its promiscuity were identified and mutated in order to create novel terpene synthases which catalyzed synthesis of only one/few different sesquiterpenes (Yoshikuni, Ferrin, and Keasling 2006).

Promiscuous enzymes catalyzing naturally occurring metabolites forms a part of ‘underground metabolism’, which can impart resilience to metabolic network against genetic or environment perturbations. More than 260 underground reactions are known in *E. coli* which covers approximately the 10% of the metabolic reactions (Notebaart et al. 2014). In order to indicate the wide prevalence of promiscuity, recently Copley estimated with each enzyme exhibiting on an average 10 promiscuous activities, even in simplest bacteria which has 1000-2000 enzymes, $\sim 10,000$ - $20,000$ promiscuous activities

are present (Copley 2017). Thus, the space of enzyme promiscuity is vast and its significance is further discussed below.

1.2.1 Significance of enzyme promiscuity

Enzyme promiscuity confers various advantages to an organism and is of evolutionary and physiological importance. In late 1970's, Jensen and Ycas independently proposed a general model of enzyme evolution in which ancestral enzymes possessed broad specificity (Jensen 1976; Ycas 1974). Owing to such "substrate ambiguity", enzymes were able to catalyze diverse set of substrates but with low catalytic efficiency. During evolution, the process of gene duplication, mutations and functional divergence leads to diversification of enzyme families and depending of specific requirements some enzymes can become highly specific for one of the substrates. Thus, it can be assumed that some modern day enzymes are "specialist", which have evolved to accept specific substrate/s or catalyze one reaction with high catalytic efficiency. However, there are some enzymes, which still harbor secondary promiscuous reactions and are referred to as "generalist". A genome-scale metabolic network model analysis of *E. coli* revealed that ~37% of its metabolic enzymes are generalist which catalyze majority (~65%) of the catalytic reactions (Nam et al. 2012). Prevalence of such promiscuous activities confers variety of physiological and evolutionary benefits to the cell discussed below.

1.2.1.1 Physiological benefits of enzyme promiscuity

1. **Robustness to adverse conditions:** Promiscuous enzymes serve as reservoir of novel catalytic activities utilizing which an organism can show physiological adaptation to genetic or environmental perturbations (Copley 2014, 2017). Promiscuous activities can provide starting base for evolution of novel catalytic activities such as new nutrient utilization for example, evolution of atrazine degradation pathway. (C. Scott et al. 2009). Promiscuous activities can also compensate for the deletion of essential enzymes (Patrick et al. 2007; Oberhardt et al. 2016; Notebaart et al. 2014). Promiscuous enzymes from various metabolic pathways can further be stitched together to generate "Serendipitous" pathways that occur only under certain (adverse) conditions and perform other functions in normal conditions. Three of such serendipitous pathways were unveiled when multicopy

suppression approach (discussed in details in section 1.2.5) was used to identify seven genes that restored the growth of *E. coli* mutant lacking 4-phosphoerythronate dehydrogenase (PdxB) on M9/glucose media (Kim et al. 2010). PdxB is an essential enzyme as it synthesizes pyridoxal phosphate (PLP), a cofactor which is utilized by at least 60 enzymes in *E. coli*. Surprisingly none of these genes encoded for enzyme which have promiscuous activity for PdxB. Rather, overexpression of these genes induced three serendipitous pathways that takes metabolites from other metabolic pathways and feed into the normal PLP synthesis pathway downstream of the blocked step. Detailed characterization revealed that the reactions in one such serendipitous pathway are being catalyzed by i) promiscuous activity of the enzyme (ThrB) whose native activity is a part of another pathway, ii) a broad –specificity enzyme (LtaE) and iii) a protein of unknown function (YeaB). Thus, promiscuous activities enable *E. coli* to generate novel pathways required for its survival under adverse conditions. These impart resilience against genetic or environmental perturbation resulting in metabolic robustness of the cell. Promiscuous activities play a significant role in higher eukaryotes as well. For instance, in plants, which are known to produce numerous secondary metabolites, often promiscuous activities of enzymes involved in primary metabolism are partitioned and recruited into these specialized metabolic pathways (Moghe and Last 2015).

2. **Balancing metabolite pools:** Promiscuous activities towards diverse set of substrates can also help in balancing of various metabolites in the cell. For example, thioesterase YciA native activity is to catalyze hydrolysis of cellular Acyl-CoA thioesters. The broad substrate specificity of YciA and its orthologues is suggested to facilitate recycling of CoA and maintaining the cellular levels of fatty acyl-CoA and involved in membrane biogenesis (Zhuang et al. 2008). Thus, this enzyme is tightly regulated by strong CoA feedback inhibition. Structural elucidation of YciA revealed the role of only thioester pantetheine moiety in substrate recognition.
3. **Xenobiotic metabolism:** Many mammalian enzymes have broad substrate specificities which enable them to bind, oxidize, and eliminate the putative foreign compounds like drugs, plant alkaloids or other toxic compounds (Jakoby and Ziegler 1990). These promiscuous enzymes may also interfere with drugs used for treatment of various diseases or infections. For example, Human cytosolic 5'-nucleotidase II (cN-II) is a member of promiscuous HAD superfamily. The native activity of this

enzyme is dephosphorylation of 5'-nucleotide monophosphates and thereby regulates the cellular level of purine nucleotides. This enzyme shows broad in vivo activity and also interferes with phosphorylation-dependent activation of nucleoside analogues used in the treatment of cancer and viral diseases by dephosphorylating them as well (Wallden and Nordlund 2011).

4. **Removal of anti-metabolites:** Apart from destroying the foreign compounds, promiscuous enzymes can also remove anti-metabolites synthesized within the cell. For example, a member of HAD superfamily, cytosolic 5'-nucleotidase III-like protein (cN-IIIB) can catalyze diverse range of substrates (Monecke et al. 2014). 7-methylguanosine (m⁷GMP) nucleotide are produced during eukaryotic mRNA degradation, which on accumulation in cytosol may be incorporated in the nucleic acid. Owing to its ability to recognize wide range of substrates, m⁷GMP is hydrolyzed preventing its undesirable accumulation in the cell.
5. **Production of toxic compounds and their repair system:** Promiscuous activities may not be always beneficial and can also result in production of toxic compounds (Kim and Copley 2012). To circumvent these, metabolite damage repair system has evolved. For example, two key enzymes of glycolysis: mammalian glyceraldehyde 3-phosphate dehydrogenase and pyruvate kinase produce 4-phosphoerythronate and 2-phospho-L-lactate respectively. These products are toxic to the cell as they inhibit an enzyme of the pentose phosphate pathway (PEP) and the enzyme producing the glycolytic activator fructose 2,6-bisphosphate respectively. Further, a single conserved repair enzyme phosphoglycolate phosphatase (PGP) was identified in mammals which dephosphorylates both of these toxic compounds and maintains the constant flux of metabolites through PEP and glycolysis (Collard et al. 2016). More such metabolite repair systems for toxic compounds generated because of promiscuous activities are also known (Linster, Van Schaftingen, and Hanson 2013).

1.2.1.2 Evolutionary aspect of enzyme promiscuity

Considering only native substrates from KEGG database, detailed analysis of kinetic parameters for all the enzymes from BRENDA database showed that an “average enzyme exhibits moderate efficiency” with a k_{cat} of $\sim 10 \text{ s}^{-1}$ and a k_{cat}/K_M of $\sim 10^5 \text{ s}^{-1} \text{ M}^{-1}$ (Bar-

Even et al. 2011). Another recent study commented “real world enzymes are sloppy and mediocre” (M. S. Newton et al. 2018). As mentioned earlier, one of the key concept of evolution of enzyme as hinted independently by Jensen and Ycas in late 1970’s, proposes that the modern day “specialist” enzymes are evolved from ancestral “generalist” which harbors with broad specificity(Jensen 1976; Ycas 1974) during the course of functional divergence due to gene duplications and mutations. In this process, some enzymes may retain low levels of their ancestral (promiscuous) activities and some enzymes may acquire new activities, which were not even present in their ancestor (Figure 1.7A). This model indicates the importance of understanding promiscuous activities of existing enzymes as today’s enzymes are tomorrow’s ancestor and serve as a starting point of evolution. One of the suggested possible routes of divergence of a new function is indicated in Figure 1.7B. This model assumes that there is trade-off between the native activity and the promiscuous activity of the enzyme selected for evolution of new function (O. K. and D. S. Tawfik 2010; Khersonsky, Roodveldt, and Tawfik 2006). Further, the dynamics of divergence may vary according to the possible route adapted for the evolution of new function. If convex route is followed, there a strong trade-off between the native and promiscuous activity and gene duplication is essential at early stage of evolution prior to acquisition of new functions. Early gene duplication is indispensable as even low levels of new function is accompanied by large loss of native activity, or else adaptation will be constrained till the native function is essential. On the other hand, if concave route is followed, there is weak trade-off among the native and promiscuous activities, enzyme can acquire new function without losing the native activity and can adapt prior to gene duplication through an intermediate, which are “generalists”. In various directed evolution studies, promiscuous activities of an enzyme tend to increase under various selection pressure while there is small/no effect on its native activity (O. K. and D. S. Tawfik 2010). This supports the adoption of convex route during divergence of new functions. Based on these observations, Twafik and co-workers have suggested that there is weak trade-off between native and promiscuous activity while emergence of new function. Further, they proposed that the evolution of a new function is driven by mutations that only affect promiscuous activity without compensating their native activity at all (D. S. Tawfik 2010; Aharoni et al. 2005; O. K. and D. S. Tawfik 2010). Thus, robustness (tolerance) in native activity towards mutations is a prerequisite during evolution. However, a very recent study highlighted

the existence of bias in selection pressure in usual laboratory directed evolution methods (Kaltenbach et al. 2016). While there is no selection pressure against original activity in these experiments, it usually increases new (targeted) activity. This results in sampling and isolation of only those mutations, which strongly increase the target activity. Thus, it does not necessarily mean that native activity is more tolerant to the mutations compared to promiscuous activity during evolution. In this study, they analyzed mutations that results in complete functional switch of native to new activity in an enzyme. Using two such case studies, they showed that lack of robustness (tolerance) of the native activity towards mutations. Moreover, weak trade-off was observed due to opposite epistatic effects in the native (antagonistic) and new (synergistic) activities (Kaltenbach et al. 2016).

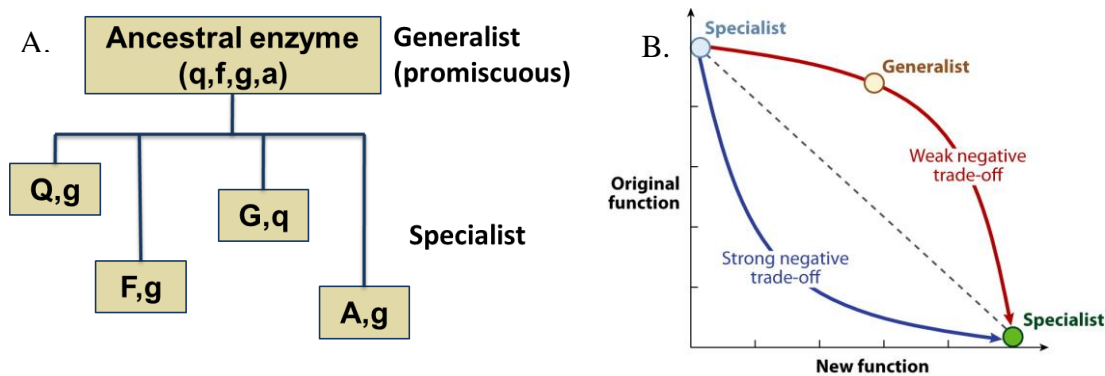


Figure 7 Schematic representation of A) Jensen and Ycas hypothesis for enzyme evolution. Modified from (Khersonsky, Roodveldt, and Tawfik 2006). B) Proposed routes of functional divergence of enzymes assuming trade-off between the native activity and the promiscuous activity of the enzyme. Adapted from (O. K. and D. S. Tawfik 2010).

Thus, understanding mechanistic aspect of promiscuity will aid us in understanding enzyme evolution. Moreover, promiscuous enzymes have already been used for direction evolution of a novel function in various studies (Renata, Wang, and Arnold 2015). Further, based on the observation that level of promiscuity varies among orthologues, another study has proposed use of more than one scaffolds for directed evolution experiments (Khanal et al. 2015). In line with this approach, “Scaffold sampling” is also adapted where beneficial mutations are transferred among series of homologous protein structures and attaining a new scaffold with superior activity in comparison to initial donor scaffold (Dunn et al. 2016).

1.2.2 Dilemma in defining enzyme promiscuity

The term ‘enzyme promiscuity’ was imprinted through a review by O'Brien and Herschlag (O'Brien and Herschlag 1999). Since then the term ‘enzyme promiscuity’ is loosely defined and broadly used to define enzymes, which can catalyze more than more substrate/reaction. However, this term is often perceived differently by different people. Some of these definitions are highlighted below and are differentiated with respect to the terms with which enzyme promiscuity is interchangeably used.

1.2.2.1 Promiscuity versus Broad Specificity

Most of the biochemists and molecular biologists, usually, categorize broad specificity enzymes as promiscuous. While this definition is reasonable, evolutionary biologist restricts the definition of promiscuity only to those enzymes where these are secondary activities, which are not physiologically relevant. (O. K. and D. S. Tawfik 2010; Copley 2003). In reviews both by Copley (Copley 2017) and (O. K. and D. S. Tawfik 2010) Tawfik groups, have highlighted broad specificity enzymes *viz.* glutathione S-transferases and cytochrome P450 and suggested that these detoxification enzymes harbor broad substrate specificity in order to protect the organism from numerous toxic compounds. Thus, giving clearly an evolutionary beneficial trait required for their native function. It is further emphasized that occurrence of secondary activities in these detoxification enzymes has no physiological prevalence and usually exhibit low efficiency. Thus, promiscuous enzymes are defined as the enzymes, which harbor secondary adventitious activity apart from their native one for which they physiologically specialized or evolved. The same definition of enzyme promiscuity is followed in this thesis as well.

1.2.2.2 Promiscuity versus Moonlighting

Further, promiscuity is also confused with moonlighting. Moonlighting is the ability of any protein to perform more than one physiological function (Copley 2003). Often these functions exhibit spatial or temporal separation and occur at different places or times. In most cases, they are acquired as additional non-enzymatic, regulatory or structural functions, which usually does not involve active site of protein. The classic example is

of crystallins whereby metabolic enzymes were recruited as structural component of eye later in evolution. Albaflavenone synthase and mARC (mitochondrial Amidoxime Reducing Component) are some other examples of moonlighting enzymes (Llamas et al. 2017; Zhao et al. 2009). According to definition state above, moonlighting is clearly distinct from promiscuity, which is restricted to only physiologically irrelevant activities.

1.2.2.3 Promiscuity versus alternate-site promiscuity

While defining promiscuity, it is commonly assumed that both native and promiscuous activities are occurring within the same active site. Further, it has been shown in various studies that native and promiscuous activity occur within the same active site (O. K. and D. S. Tawfik 2010; Tokuriki et al. 2012; Pandya et al. 2014). However, there are few reports where native catalytic site is not involved in promiscuous reactions and is referred as **alternate-site enzyme promiscuity**. The first example of alternate-site enzyme promiscuity was reported in thermostable enzyme tHisF from *Thermotoga maritima* which is primarily cyclization reactions involved in histidine biosynthesis (Taglieber et al. 2007). Further the residues responsible for esterase-like promiscuous activity are distant from the native active site.

1.2.3 Levels of enzyme promiscuity

Enzyme promiscuity can be exhibited at various molecular levels varying from single protein to superfamily level (Figure 1.8). Few examples of enzyme promiscuity occurring at various levels are listed below.

1. **Single protein (enzyme) level:** There are many enzymes which catalyze other promiscuous activity at very low level. For example, the native activity of Aspartate aminotransferase is transamination of dicarboxylic substrates with $k_{cat}/K_M = 9.1 \text{ M}^{-1}\text{s}^{-1}$ (Rothman and Kirsch 2003). It also harbors promiscuous activity and can catalyze transamination of tyrosine and phenylalanine with $k_{cat}/K_M = 0.055 \text{ M}^{-1}\text{s}^{-1}$ and $0.012 \text{ M}^{-1}\text{s}^{-1}$ respectively. Further, using directed evolution, mutations in this enzyme lead to 130 and 270-fold higher promiscuous activity in the evolved enzyme.
2. **Family level:** Promiscuity is exhibited by enzymes and their close homologs. Enzymes and their homologs with sequence similarity >40%, often have substrate

promiscuity acting on variety of substrates while sharing same reaction mechanism (Nobeli, Favia, and Thornton 2009). Promiscuity is a typical property of enzyme families (Baier, Copp, and Tokuriki 2016). For example, systematic functional profiling revealed the prevalence of substrate promiscuity in “ β -keto acid cleavage enzyme” or BKACE family (Bastard et al. 2013). It was observed that >60% of the enzymes that were tested (50 of 80) catalyzed more than one β -keto acid substrate, whereas remaining 40% (30 of 80) enzymes were found to be specific to only one substrate. It should be noted here that promiscuity can also appear in one family member and not in the others. For example, among three families of guanidino-modifying enzyme superfamily, no promiscuity is observed in arginine deiminase (PaAgDI) family while the remaining two family members exhibit promiscuity. Similarly, in mammalian paraoxonases family, promiscuity is prevalent in PON1 but barely detected in PON2 and PON3 (Khersonsky and Tawfik 2010).

- 3. Superfamily level:** promiscuity is observed in enzymes and their remote homologs. At superfamily level catalytic promiscuity is more prevalent, thereby enzymes with sequence identity < 30%, usually differ in both substrate and reaction chemistry (Nobeli, Favia, and Thornton 2009). For instance, members of Alkaline phosphatase superfamily possess catalytic promiscuity and also known to show cross-promiscuity where product of one member is often catalyzed by other member. Similarly, members of pectin superfamily exhibit both substrate and catalytic promiscuity (Linsky and Fast 2010). Another example includes enolase superfamily which is highly mechanically diverse although the formation of an enolate anion intermediate is conserved in this superfamily. Thus, many related enzymes harboring promiscuity may conserve only a part of catalytic reaction.

Functional Promiscuity in related proteins (family or superfamily level) is very common often originated by gene duplication followed by divergence to acquire different specific function. Moreover, promiscuity is characteristic of many enzyme families and superfamilies (Baier, Copp, and Tokuriki 2016). Enzyme families/superfamilies tend to harbor similar binding pockets, despite large sequence divergence among its individual members and hence are prone to show substrate ambiguity.

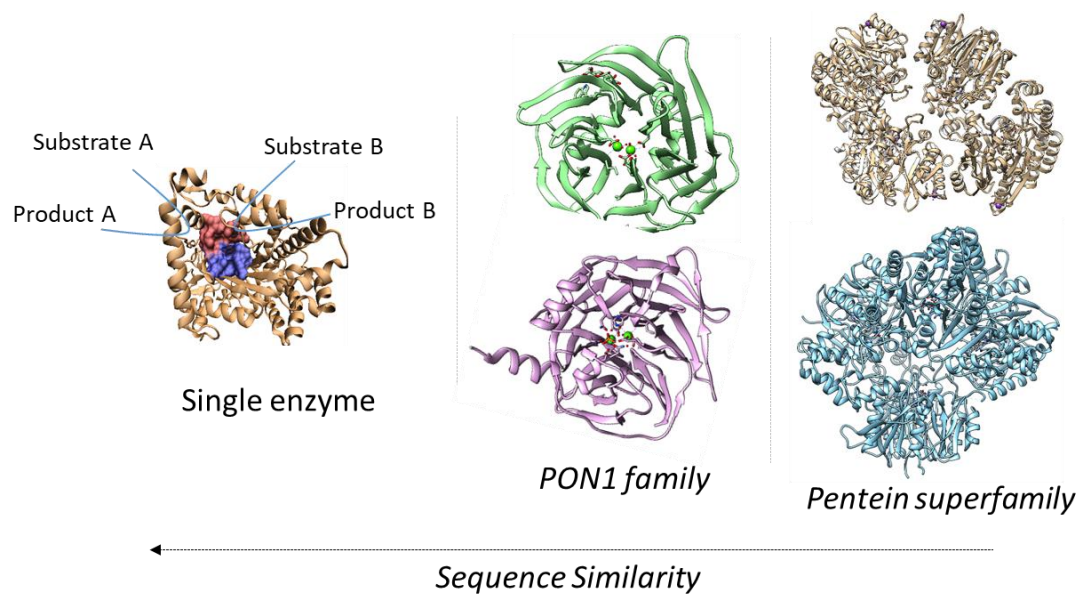


Figure 1.8 Various level of promiscuity.

1.2.3.1 Reciprocal promiscuity

Homologs often show promiscuity towards each other's native reactions and thus tend to have reciprocal promiscuity. However, a recent study demonstrated a novel example of reciprocal promiscuity in two pyridoxal 5'-phosphate dependent non-homologous enzymes (Soo et al. 2016). The enzymes belonging to at least five different folds are known to use pyridoxal 5'-phosphate as a cofactor. The representative of fold type III, *E. coli* alanine racemase (ALR), is a promiscuous cystathionine β -lyase (CBL) a representative from fold-type I. Similarly, *E. coli* CBL is promiscuous ALR. Further single mutation can improve these promiscuous activities in both the cases. ALR variant with mutation Y274F harbor CBL activity near native $K_m \sim 3.3$ mM (but poor $K_{cat} \sim 10h^{-1}$) due to enlarged active site allowing better access to cystathionine. In case of CBL, its variant with single mutation P113S has ALR activity with higher $K_{cat} 22 s^{-1}$ (but poor K_m 58 mM). Detailed analysis revealed that improved in ALR activity in this mutant was due to optimization of pKa of Y111, which act as a catalytic acid while racemizing alanine. Thus, short mutational routes may have been sufficient to evolve families of PLP with specific function regardless of fold similarity. Further such reciprocal promiscuity may be evolved from common multifunctional ancestor.

1.2.4 Types of enzyme promiscuity

Hult and Berglund majority classified enzymatic promiscuity in three types which are as mentioned below (Hult and Berglund 2007). An additional fourth type is also introduced in another review and mentioned later (Gatti-Lafranconi and Hollfelder 2013). A schematic overview of the classification of promiscuity is shown in the Figure 1.9 and are discussed below in details.

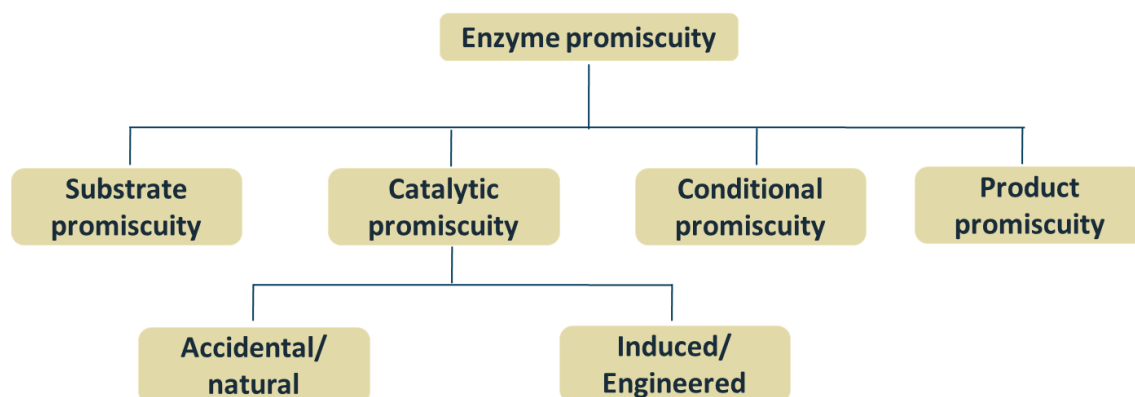


Figure 1.9 Different types of enzyme promiscuity.

1. **Substrate promiscuity:** This is defined as catalysis of alternate substrate apart from the native substrate using the same active site and through the same reaction mechanism as well as same transition state. For example, in hyperthermophilic Archaea *Sulfolobus solfataricus*, both D-2-keto-3-deoxygluconate (KDG) and D-2-keto-3-deoxygalactonate (KDGal) are catalyzed by the metabolic enzyme KDG aldolase, leading to production of pyruvate and D-glyceraldehyde (Theodossis et al. 2004). Substrate promiscuity can also be introduced in an enzyme using methods like random mutagenesis and directed evolution. For instance, the native activity of aspartate aminotransferase is to catalyze the interconversion of aspartate and α -ketoglutarate to oxaloacetate and glutamate (Figure 1.10A). However, the evolved variant of this enzyme generated via directed evolution can catalyze both L-Aspartate and L-Tyrosine (Rothman and Kirsch 2003).

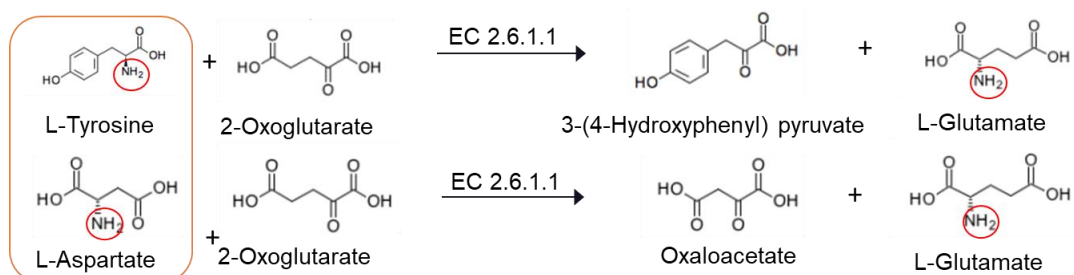


Figure 1.10 A) Example of substrate promiscuity. *Aspartate aminotransferase* can catalyze both *L-Tyrosine* and *L-Aspartate* and transfers the same chemical moiety (amine group) in order to produce *L-Glutamate*.

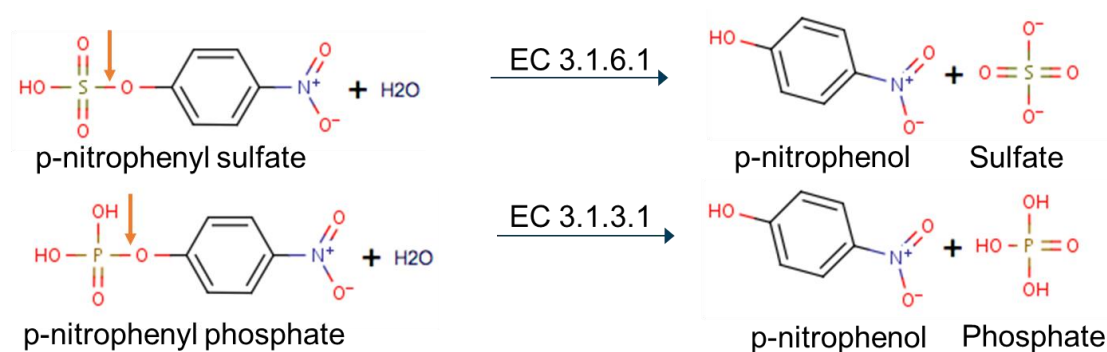


Figure 1.10 B) Example of catalytic promiscuity. *Pseudomonas aeruginosa* arylsulfatase (PAS) show both sulfate ester hydrolysis (native activity) and phosphomonoesters hydrolysis (promiscuous activity). While catalyzing the native and promiscuous activity, PAS cleaves different types of bond- S-O and P-O (as shown by orange colored arrow). The enzyme has two different EC numbers for two reactions.

- Catalytic promiscuity:** The catalysis of chemically distinct reactions with possible different transition states is defined as catalytic promiscuity. For example, *Pseudomonas aeruginosa* arylsulfatase (PAS) show both sulfate ester hydrolysis (native activity) and phosphomonoesters hydrolysis, however latter reaction is much less efficient in comparison to former (Olguin et al. 2008) (Figure 1.10B). Interestingly, this enzyme also possess promiscuous diesterase catalytic proficiency of which is compared to its native activity and is of the order of 4×10^{18} (A. C. Babtje et al. 2009). This enzyme along with other (Kim et al. 2010; Steinmetz et al. 2010; P. Wang, Jin, and Zhu 2012; van Loo et al. 2010; F. Zhang et al. 2018) cases, where the catalytic efficiency of native and promiscuous activities is comparable has challenged the idea of requirement of specialization for efficient catalysis. Although individual members of an enzyme superfamily are specific with each catalyzing a different

reaction, many times they can exhibit catalytic/substrate promiscuity for one another's native reaction/substrate. This phenomenon is referred as "**cross-promiscuity**". Alkaline Phosphatase is one of the well-studied superfamily in this aspect (Sunden et al. 2017; Duarte, Amrein, and Kamerlin 2013; Pabis and Kamerlin 2016; López-Canut et al. 2011). The members of this superfamily shows catalytic promiscuity and can catalyze many chemically distinct substrates such as sulfo-carbohydrate, phosphonocarbohydrate and phosphocarbohydrate thereby cleaving S-O, P-O and P-C bonds respectively (Jonas and Hollfelder 2009). Few of the promiscuous activities towards each other's native substrate are shown in Figure 1.11. Catalytic promiscuity can further be classified into two subtypes based on the nature of type of enzyme performing the promiscuous activity:

- a. **Accidental/natural promiscuity:** catalysis of secondary activity by wild-type enzyme
 - b. **Induced/engineered promiscuity:** when promiscuous reaction induced by one/more mutation/s.
3. **Condition promiscuity:** The promiscuity dependent on reaction conditions such as extreme pH/temperature/salinity conditions, presence of organic solvent or anhydrous media. For example, the native activity of malate dehydrogenase catalyzes the interconversion between malate and oxaloacetate. However, it also produces L-2-hydroxyglutarate in the presence of high amount of α -ketoglutarate in the cell (Rzem et al. 2007). Many lipases use alternate substrates in the presence of organic solvents and are exploited in several industrial applications (Schmid et al. 2001). Many enzymes perform different function in different oligomer states. For instance, in monomeric state pyruvate kinase acts as thyroid hormone bind while it acts as metabolic enzyme in tetrameric state (Parkison et al. 1991). Recently, a bacteriophage DNA was reported to rescue auxotrophic *E. coli* *ilvA* mutant in by triggering overexpression of promiscuous activity of bacterial enzyme MetB (Jerlström Hultqvist et al. 2018). Such an increase in promiscuous activity was induced by reduced level of S-adenosylmethionine SAM due to hydrolysis by bacteriophage-encoded- SAM hydrolases.

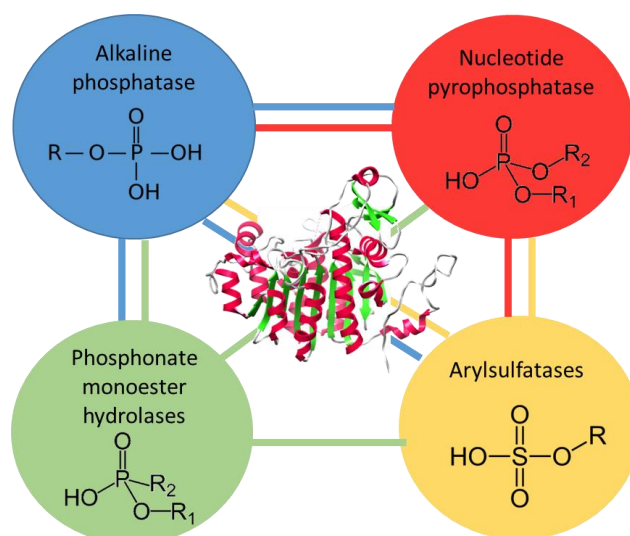


Figure 1.11 Cross-promiscuity shown by various family members of Alkaline Phosphatase. Four circle represents the native substrate for four family members. The colored lines indicate the cross-promiscuous activities among these family members i.e. these members are able to catalyze each other's native substrate as well apart from their own native substrate. Modified from (Mohamed and Hollfelder 2013)

4. **Product promiscuity:** It is defined as same reaction generating alternative products. For instance, catalysis of the same substrate (Phenetole) by enzyme Naphthalene 1,2-dioxygenase (NDO) results in formation of different products in multiple ratio (Vinyloxybenzene and S-Ethyl phenyl sulfoxide in 1:13 ratio)(Ferraro et al. 2017). This is facilitated by substrate flexibility i.e. ligand bound in different conformations yielding different products. Sometimes, first the bound ligand is catalyzed to produce initial product, which in turn binds again and is catalyzed into different product. The multiple ratios of the different products formed from the same substrate suggests that in the large pocket of NDO multiple cycles of reaction can take place when smaller substrates are involved: for example, indan could first be converted to indene and then to a monohydroxylated or dihydroxylated product of indene. Similar behavior where the product formed by first bound ligand is catalyzed to yield second product is also shown during catalysis by biphenyl dioxygenase using docking studies (Pham and Sylvestre 2013; Pham, Tu, and Sylvestre 2012). Many detoxification enzymes also possess product promiscuity. Infact, it has been proposed that product promiscuity may facilitate detoxification function(Cook and Atkins 1997; Atkins, Lu, and Cook 2002). Another example includes sesquiterpene synthases, Cop4 which

catalyzes cyclization of all-trans-farnesyl pyrophosphate (FPP) into multiple products (Lopez-Gallego et al. 2010).

1.2.5 Identifying enzyme promiscuity

Initial discovery of promiscuous enzyme was mostly by serendipity during studies involving mutations of ligand binding residues or in genetic perturbations for functional characterization of enzymes. However, today we know numerous promiscuous activities identified in various routes. Some of these are discussed below and summarized in Figure 1.12.

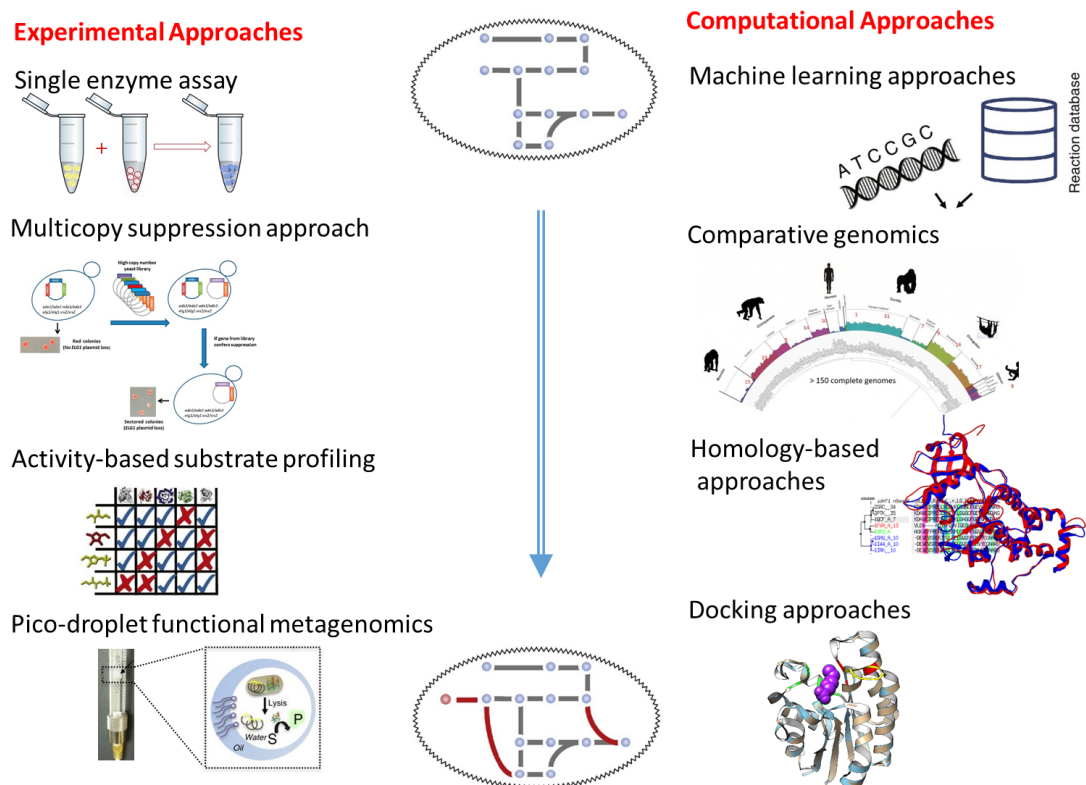


Figure 1.12 Various experimental and computational approaches used for identification of promiscuity. Modified from (Notebaart et al. 2018)

1.2.5.1 High-throughput experimental approaches

1. Multicopy suppression approach: This simple approach was one of the first initial high-throughput experimental approaches to identify promiscuity at genome level. Here a library of genes or genomic DNA cloned into overexpression plasmids that replace an essential enzyme are used to identify promiscuous activities.

Overproduction of a promiscuous enzyme induces sufficient (promiscuous) activity is which is capable of replacing the native activity of the deleted (essential) enzyme. For instance, prevalence of promiscuous enzymes in *E. coli* proteome was revealed by complementation of each of 3,985 single-gene knockout strains of Keio collection by selection from ASKA (A complete Set of *E. coli* K-12 open reading frame Archive) library (Patrick et al. 2007). ASKA library contains every *E. coli* open reading frame (ORF), each individually cloned in expression vector pCA24N. In this study, 21 out of 104 single-gene knockout strains were rescued by overexpression of at least one non-cognate *E. coli* gene. Many other studies used similar approach to identify enzyme promiscuity (Patrick and Matsumura 2008; Soo et al. 2016; Miller and Raines 2004). However, not every time the survival is due replacement of deleted enzyme and may involve other mechanisms like presence of alternate metabolic routes. This approach will also be unable to detect very low levels of promiscuity activity due to production of undetectable enzyme or inhibition by a cellular metabolite.

2. Activity-based metabolic profiling or large-scale functional/substrate profiling:

This technique is simply scaled-up version of classical *in vitro* enzyme assays which allows assaying hundreds of enzymes family members for a wide range of substrates. For example, a screen of 217 members of haloalkanoate dehalogenase superfamily (HADSf) from 86 different species against 169 phosphorylated compounds revealed that among these 204 members catalyze hydrolysis of a median of ~16 substrates (H. Huang et al. 2015). Surprisingly, this screen identified few members can catalyze up to 143 substrates indicated the enormous extent of substrate promiscuity in HADSf. More importantly, such large-scale metabolic profiling has revealed that promiscuity is intrinsic property of many enzyme families where usually enzymes can catalyze 2 to 5 additional reactions apart from their native functions (Baier, Copp, and Tokuriki 2016).

3. Picodroplet functional metagenomics: Metagenomic libraries constitute several bacterial communities, which cannot be cultured (grown) in laboratory conditions, providing a rich source of biocatalysts and reservoir of many promiscuous enzymes with potential novel catalytic activities. One of the initial functional screening of metagenomics library using microfluidic picolitre droplets was demonstrated by Colin and coworkers (Colin et al. 2015). In this ultra-high throughput technique,

DNA (eDNA/Environmental DNA) of each member of metagenomic library is cloned into high-copy plasmid and transformed into *E. coli*. Single transformed bacteria are then encapsulated into “water-in-oil” droplets together with fluorogenic ‘bait’ substrates and lysing agents. After single cell lysis, cytoplasmically expressed enzymes are able to catalyze substrate and hence this system is analogous to “miniaturize cell lysate assays”. Since fluorogenic substrates are synthetic, any detected activity is certainly promiscuous. Further, droplets showing high fluorescence are selected using fluorescence-activated sorting. Finally, plasmids containing eDNA coding for active catalyst is isolated and sequenced. Using such a combination of small-insert shotgun metagenomics and droplet microfluidics, Colin and co-workers identified enzymes with promiscuous activities from a metagenomics library containing 1.25 million sequences (Colin et al. 2015). Owing to its high sensitivity, this method identified new hydrolases with promiscuous (weak) activities catalyzing sulfate monoesters and phosphotriesters. Speed and low screening cost are one of the major advantages of this method. Using this method upto ~108 biochemical reactions can be performed per day, typically in pico- to femtolitre volume. However, requirement of fluorogenic substrate can often be a limitation for high-throughput screening of enzyme activity. This limitation was overcome by another extension of the method absorbance-activated droplet sorting (AADS) given by Gielen and co-workers (Gielen et al. 2016). This method enabled the absorbance-based detection of enzymatic activity in lysates of single bacterial cells.

- 4. Metabolite-responsive biosensors:** Promiscuous activities are difficult to detect if they result in only a small change in absorbance or fluorescence. In such instance, the sensitivity of the method to detect promiscuous activity can be increased using metabolite-responsive biosensor. In such biosensors, metabolite or substrate detection is coupled by alteration in gene expression that allows screening by production of GFP or selection against an antibiotic. Although currently this method is used only in metabolic engineering, it certainly has a potential in detecting in enzyme promiscuity as well (Copley 2017).

1.2.5.2 Computational approaches to detect enzyme promiscuity

Apart from experimental techniques, many computational approaches have been developed to identify enzyme promiscuity. However, the level to which these predict promiscuity varies. For instance, there are few methods which only determine whether an enzyme is promiscuous or non-promiscuous such as Promis server (Carbonell and Faulon 2010) or only predict novel reactions that can potentially be acting promiscuously in a metabolic network without linking them to specific enzymes such as GEM-Path algorithm (Campodonico et al. 2014). Further, many specific level prediction tools are also available, which predict enzyme-reaction pairs possessing promiscuity and are discussed below in details.

1. **Machine learning approaches:** These exploit and learn from experimentally verified enzyme-reaction pairs data. One such method, Gaussian process (GP) predictor was developed by Mellor and co-workers where, given a reaction and an enzyme, it provides a probability estimate that the enzyme can catalyze the reaction (Mellor et al. 2016). This method uses information from both enzyme features and the reaction signatures in order to provide a probability estimate. Further, this study demonstrated an approach to predict the Michaelis constant K_M for a given reaction-enzyme pair using Gaussian process regression. K_M for a reaction is an indicator of binding affinity between an enzyme and the substrate it catalyzes where lower the K_M value, higher is the binding affinity. Further, GP predictor was used to predict putative promiscuous reactions catalyzed by *E. coli* enzymes among which reaction producing novel metabolite N-Acetyl-L-Leucine was shown experimentally.
2. **Comparative genomics:** exploits the observation that usually native activity of one enzyme may be promiscuous activity of its homologue (O. K. and D. S. Tawfik 2010; Kuznetsova et al. 2006). The first genome-wide method to predict enzyme promiscuity, PROPER (**PRO**miscuity **PrE**dicto**R**) relies on this observation and uses a permissive PSI-BLAST based phylogenetic tree to predict promiscuous reactions (Oberhardt et al. 2016). In this phylogenetic tree, query enzyme is the root and the reactions catalyzed by all other enzymes, which are different than root enzyme reactions are predicted to be promiscuous. Another modified version, GEM-PROPER was developed integrating PROPER with genome-scale metabolic modelling to predict promiscuous replacements via alternate metabolic pathways. In total, using

PROPER and GEM-PROPER, the group predicted 2811 direct and 98 indirect target-replacer pairs in *E. coli* respectively. Out of these 4 novel target-replacer pairs were experimentally validated. One of these was the novel promiscuous pathway for synthesis of pyridoxal 5'-phosphate (the active form of Vitamin B6) where the essential enzyme *pdxB* can be replaced by a new indirect replacer *thiG* as predicted by GEM-PROPER. Elucidation and experimental confirmation of putative active site residues responsible for promiscuous activity of *thiG* was performed.

- 3. Homology-based approaches:** Prediction performance of comparative genomics can further be enhanced by incorporating information of active site conformations. Steinkellner and co-workers demonstrated an approach to predict promiscuous activity by screening the PDB database using specific 3D templates, which mimic a 'minimal' catalytic active site constellation ('catalophore') (Steinkellner et al. 2014). Using this motif based catalophore approaches, two enzymes PhENR and TtENR were predicted to harbor ene-reductase promiscuous activity which was further validated experimentally to be NADPH-dependent quinone reductases. Moreover, despite high sequence and structural divergence, both these enzymes showed significant OYE-like side activities (Old Yellow Enzyme).
- 4. Structure-based prediction/modelling of alternate substrate binding site/s:** Structure-based docking with high-energy forms of potential substrates can also be used to predict the binding site of a given (alternate) substrates. Although classical molecular docking approaches were limited in predicting enzyme activity reliably, recent improvements made in this area such as development of covalent docking approach can facilitate the docking of alternate substrates in their respective binding site more accurately (Kolb et al. 2009). In a recent work, London and co-workers exploited covalent docking for substrate profiling for various enzymes from haloalkanoate dehalogenase superfamily (London et al. 2015). They also identified an enzyme in *Bacteroides* that can catalyzes the orphan phosphatase reaction in the riboflavin biosynthetic pathway. Thus this approach can be used to predict unknown promiscuous activities of an enzyme.

Owing to these currently available experimental and computational approaches, we know promiscuity is not an exception anymore. However, the existing computational approaches still need to be systematically evaluated on large-scale benchmark datasets

in order to judge their prediction performance. The field of promiscuity prediction is still at an early stage and slowly many reliable methods are being developed. The main limitation in the development of predictors is the availability of experimentally verified promiscuous enzymes.

1.2.6 Quantifying enzyme promiscuity

Apart from identification of enzyme promiscuity, it is important to quantitate it in order to systematically study the extent of promiscuity in a given class of enzymes. It will be helpful in understanding the mechanistic aspect of promiscuity, relationship between promiscuity and enzyme activity, stability or protein dynamics. For instance, it is unclear that till what extent an enzyme needs to be promiscuous in order to it evolve new catalytic function. Further, such quantitative metrics can also be used to determine if high degree of promiscuity have any effect on the stability of the protein. Such information will aid in improving protein engineering.

1. **Promiscuity Index:** Nath and Atkins defined an entropy based metric to predict enzyme promiscuity (Nath and Atkins 2008). This metric is modified version of Shannon entropy which is typically used to measure degree of uncertainty or randomness in information theory and is calculated by the following formula:

$$H = - \sum_{i=1}^N P_i \log P_i$$

where H is Shannon entropy for a defined set of N possible outcomes, with each outcome i having a probability of occurrence of P_i . Using the same analogy, Nath and Atkins defined the entropy of the enzyme in order to quantitate the substrate promiscuity of a given enzyme and defined promiscuity index I as follows:

$$I = - \frac{1}{\log N} \sum_{i=1}^N \frac{e_i}{\sum_{j=1}^N e_j} \log \frac{e_i}{\sum_{j=1}^N e_j}$$

where I is the promiscuity index of a promiscuous enzyme with N substrates, each associated with catalytic efficiency $e_i = \frac{k_{cat}}{K_M}$ where k_{cat} is rate at which product is generated by an enzyme in saturating substrate concentrations and K_M (Michaelis constant) is the concentration of substrate that yields a half-maximal rate.

$I=1$ indicates that a given enzyme is perfectly promiscuous and all the substrates are equally well-metabolized. On the other hand, $I=0$ implies that the given enzyme is highly specific and catalyzes only one substrate. Since I is a functional parameter of a given set of substrates, its value can be only compared among two different enzymes if it is calculated using the same set of substrates. Ideally, enzymes that catalyzes two chemically distinct substrates with equal rates are more promiscuous compared to the one which catalyzes two chemically similar substrates. From this perspective the definition of Promiscuity Index is incomplete. Thus, another modified version of I called weighted Promiscuity Index J was defined this study which accounts for substrate similarity aswell. This metric was computed for three different enzyme classes: serine/cysteine proteases, glutathione S-transferase (GST) isoforms, and cytochrome P450 (CYP) isoforms in order to understand the extent of promiscuity in each of these enzyme class (Figure 1.13). Further for each enzyme class, correlation between substrate promiscuity and an enzyme's activity toward its most-favored substrate was also drawn.

2. **PromIndex:** Chakraborty and Rao demonstrated a computational method **Promiscuity Indices Estimator (PROMISE)** to predict and quantitate relative promiscuity of set of enzymes with known active sites based on the signatures derived from spatial and electrostatic properties of the catalytic residues (Chakraborty and Rao 2012). This method relies on previously developed method CLASP, which identifies active site in a query protein by searching its spatially congruent matches among pre-defined motif library derived from catalytic sites collated in CSA database (Chakraborty et al. 2011). In order to reduce false positives, this method exploits Potential Difference (PD) as PD of equivalent catalytic residues from different enzymes of same family is usually similar. CLASP also unveiled promiscuous protease activity in shrimp Alkaline Phosphatases, which were validated experimental in *vitro studies*. Extending this work, in PROMISE, a CLASP score is calculated for active site matched among the pre-defined library of catalytic residue motifs and additionally a quantitative metric PromIndex was calculated for a set of 305 non-redundant proteins. Among these proteins, Carboxypeptidase A and Ribonuclease A were ranked most promiscuous. PromIndex is calculated by comparing three main features among the predicted promiscuous activity and the

native activity: the difference in EC number, quality of the spatial and electrostatic congruence and occurrence in same vicinity. PromIndex varies from 0 to 1, with 1 being most promiscuous protein. Enzymes with mean PromIndex > 0.28 were regarded highly promiscuous. Moreover, a weak correlation indicating high percentage of polar and charged residues in vicinity (3Å) of catalytic site of highly promiscuous proteins was observed. Although, this was not statistically significant, the reciprocal relationship among them was statistically significant and enzymes with less percentage of polar and charged residues in the neighborhood of catalytic residues usually implied less promiscuous protein.

3. **Difference in EC number:** Khersonsky and Twafik proposed using difference in EC numbers to assess the degree of enzyme promiscuity (O. K. and D. S. Tawfik 2010). Here, the enzymes possessing substrate promiscuity would differ only in fourth digit of EC number. On the other hand, difference in first, second or third digit of EC number would indicate different reaction category and hence would imply catalytic promiscuity.

1.2.7 Prevalence of enzyme promiscuity at various levels

1.2.7.1 Promiscuity within *E. coli* metabolome

It can be well appreciated that not every enzyme in a cell is specific for their evolved substrate/reaction. Numerous naturally occurring metabolites can act as alternate substrate to the promiscuous enzymes present in a cell. As mentioned earlier, during normal conditions, these promiscuous reactions are usually carried out at very low levels owed to their characteristic low catalytic efficiency (low K_{cat}/K_m) compared to their native activity and hence are difficult to detect. D'Ari and Casadesus coined the term “**underground metabolism**” to describe these naturally occurring low level reactions catalyzed by promiscuous enzymes (wild-type) using endogenous metabolites as alternate substrate (D'Ari and Casadesus 1998).

Moreover, numerous cross-reactions occur within living cells. These are often revealed by various knockout studies where loss of an essential enzyme is complemented by alternate activities of enzymes from different pathway. The level of such cross-reactivity was indicated by the first quantitative survey of enzyme promiscuity done in *E.coli*. In

this study, about 20% of the total (104) single-gene knockout strains screened were rescued mostly by complementation of promiscuous action of at least one non-cognate overexpressed *E.coli* gene or an alternate metabolic pathways (Patrick et al. 2007). Moreover, this study suggested that promiscuity is often a “product of contingency”. This terminology was influenced from Gould’s book on evolution where he used the term “contingency” to describe “an instance when a feature evolved long ago for a different use has fortuitously permitted survival during a sudden and unpredictable change in rule (Gould SJ 1989). Further, exploiting the information from BRENDA database and literature mining, *E. coli* metabolic network was extended by adding 262 weak underground promiscuous activities which covers ~11% of its metabolic reactions (Notebaart et al. 2014). Further, the contribution of these underground activities in adapting to novel nutrient environment was predicted using FBA and these *in-silico* predictions showed this ~11% of increase in network space via underground reactions expanded the scope of utilizable nutrient by 3%. These predictions showed significant agreement with in vivo genome-wide gene overexpression screen across numerous carbon sources indicating that these weak underground promiscuous activities confer fitness benefit to *E. coli* under specific environments. Recently, the collection of underground activities or underground metabolism is referred as “**enzyme promiscuome**” (Notebaart et al. 2018).

1.2.7.2 Prevalence of enzyme promiscuity in archaea, bacteria, fungi and plants

Around 10% of total enzymatic repertoire in bacterial and archaeal organisms are promiscuous enzymes (M. A. Martínez-Núñez et al. 2013). Another systemic analysis of 89 and 705 non-redundant archaeal and bacterial genomes showed 8.31% and 8.76% of average promiscuous enzymes in these genomes respectively (M. Martínez-Núñez et al. 2017). Further, on an average 82% and 73% of unique structural domains were found to be associated with the promiscuous enzymes in archaeal and bacterial genomes in contrast to specialist which were associated with 42% and 36% unique structural domains in these genomes respectively. Thus, generalist tend to have more structural diversity which might facilitate plasticity in them. Interesting, in archaeal genomes substrate promiscuity was found to be 2.5 times more prevailing compared to catalytic promiscuity

suggesting that it is difficult to evolve new catalytic mechanism compared to new substrate-binding modes.

Many fungal enzymes involved in secondary metabolism have broad specificity and exhibit promiscuous activities. For instance, aromatic prenyl transferase (aPTases) from *Aspergillus terreus* can catalyze 72 different aromatic substrates (Chen et al. 2017). Many enzymes involved in secondary metabolism in plants also harbor promiscuous activities (Moghe and Last 2015). These often facilitate synthesis of pigments, flavors or defense molecule synthesis and helps in bioremediation of anthropogenic chemicals.

As discussed earlier in section 1.2.3, enzyme promiscuity is common in many enzyme families. For instance, while identification of physiological activities for 217 haloacid dehalogenase family members from 86 species using metabolite- profiling against 169 phosphorylated compounds, promiscuous activities in various families were also unveiled. Around 101 members could catalyze 6 to 40 substrates and these screened members could catalyze median of 15 substrates.

1.2.7.3 Prevalence of enzyme promiscuity influenced by life-style of an organism

Analysis of 761 bacterial species showed that prevalence of promiscuous enzymes is influenced by the life-style of bacteria. These bacterial species were categorized into free-living, extremophiles, pathogens and intracellular organisms where majority of promiscuous enzymes were found in free-living bacteria (Martinez-Nunez, Rodriguez-Vazquez, and Perez-Rueda 2015). This study suggested that these organisms show enrichment in promiscuous enzymes as an adaptive measure to combat the effects of the fluctuating environment experienced by these free-living bacteria. Additionally, they found about one-third of promiscuous enzymes arose by duplication events in these organisms which allows such functional divergence. On contrary, the organism which usually live in stable environment such as intra-cellular organisms harbor lower proportion of duplicated and promiscuous enzymes (Martinez-Nunez, Rodriguez-Vazquez, and Perez-Rueda 2015).

1.2.8 Energetics of promiscuous activities

As discussed in section 1.1, enzymes accelerate the rate of catalysis by lowering the activation energy such that a small amount of energy can facilitate the conversion of reactants to products *via* a transition state. Few ensemble approaches used to study the basis of catalytic promiscuity have suggested that promiscuous enzyme lower energy barrier to enable conformational rearrangements required for the binding to diverse range of substrates (Honaker et al. 2011; Colletier et al. 2012; F. Zhang et al. 2018). For instance, comparison of free energy landscape of two isoforms of glutathione S-transferase (GST): one specific (GSTA4-4) and other promiscuous (GSTA1-1) revealed thermodynamically smooth active site landscape, with barrier less transitions between enthalpy states for GSTA1-1 (Honaker et al. 2011). The regions around the active site of GSTA1-1 samples large number of conformational sub-states without significant change in free energy barriers among them enabling its active site to be more “fluid-like”. Moreover, this active site C terminal is tethered to a more stable core scaffold of GSTA1-1 (than core of GSTA4-4), which may be required for its heterogeneity. Similarly, basis of catalytic promiscuous activity was revealed for a sesquiterpene/farnesyl pyrophosphate (FPP) cyclases, TEAS enzyme which catalyzes both native (trans, trans)-FPP and promiscuous (cis, trans)-FPP substrates efficiently in order to generate diverse range of products. While native substrates were found to catalyze major 1,10-closure pathway, the promiscuous substrates were found to be catalyzed by a major 1,6-closure (lower energy barrier) pathway (F. Zhang et al. 2018). Moreover, the promiscuity of this enzyme is dependent on three major factors, folding mode of substrate, flexibility of intermediate and plasticity of active site pocket. Similar to this enzyme, many a times more than one factor is involved in the facilitating enzyme promiscuity and are discussed in details in next sections.

1.2.9 Proposed mechanism facilitating enzyme promiscuity

It has been shown that native and promiscuous activity occur within the same active site (O. K. and D. S. Tawfik 2010; Tokuriki et al. 2012; Pandya et al. 2014). For instance, numerous enzyme exhibit phosphatase and sulfatase activity within the same active site (Pabis, Duarte, and Kamerlin 2016). While comprehensive analysis of the non-redundant

20,414 ligand binding pockets bound to 9,485 unique ligands, Gao and Skolnick, found around more than 34 % of binding pockets to be promiscuous and able to bind more than one chemically distinct ligands (Tanimoto coefficient <0.3). Analysis of conserved atomic contacts at ligand binding site revealed that ~58% of atomic contacts are conserved between pair of ligand-bound complexes, thereby sharing similar type of interactions. For example, hydrophobic interactions, hydrogen bonding or aromatic. Further, flexibility in active site or different binding modes may also facilitated promiscuity. Based on several case studies many molecular mechanisms which facilitate enzyme promiscuity have been reviewed and proposed (O. K. and D. S. Tawfik 2010; Nobeli, Favia, and Thornton 2009). These mechanisms may involve enzyme (receptor macromolecule), reactant/substrates or even involve both. Some of these are discussed below in details and summarized in Figure 1.13.

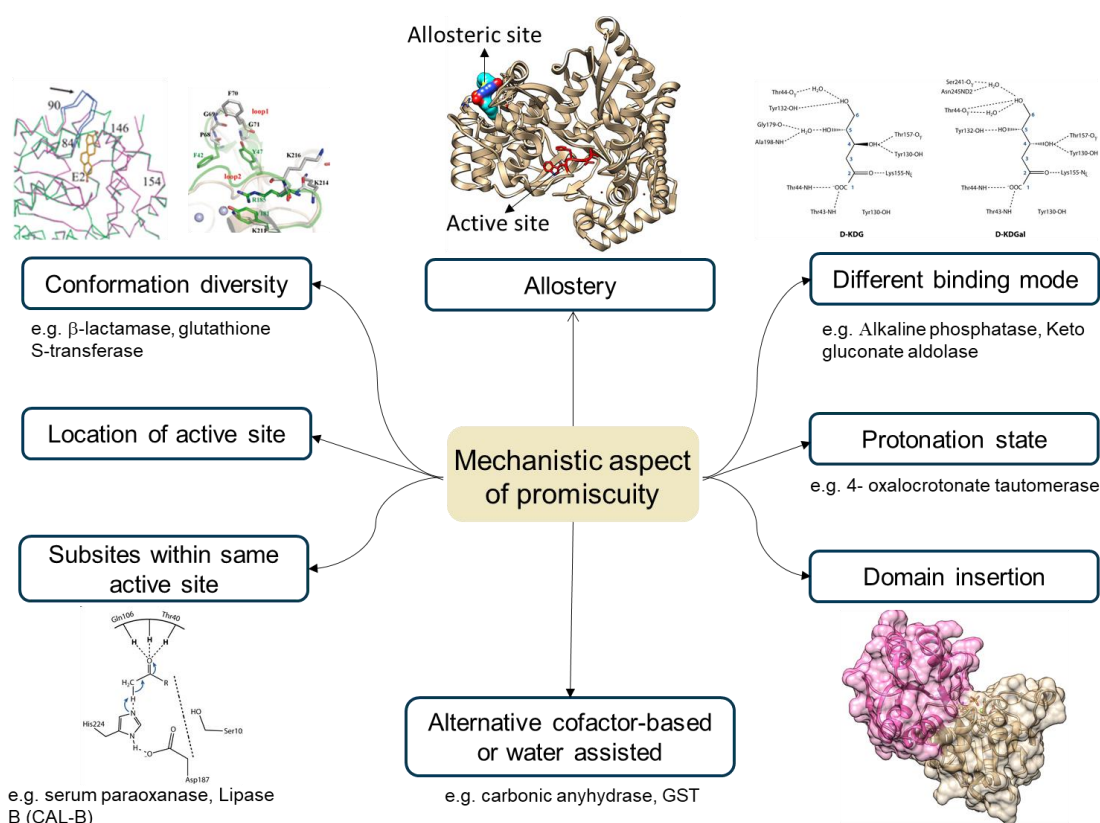


Figure 1.13 Proposed mechanisms of enzyme promiscuity

- 1. Conformational Diversity:** Often the native and promiscuous activities are mediated by different active site conformations. For example, Human SULT1A1 binds to different substrates in different active-site conformations (Gamage et al. 2005). It has

been suggested that flexibility in active site especially loops leads to differential conformational sampling and active site reshaping which enables it to bind and catalyze diverse range of substrates. For instance, promiscuous phosphotriesterase activity in lactonase SsoPox (member of PLL- **Phosphotriesterase-Like Lactonases**) from *Sulfolobus solfataricus* is governed by conformational diversity of active site loop region (Hiblot et al. 2013). Mutation in only single residue W263 (present in active site loop and also a part of interface while enzyme dimerization) reshapes the active site, re-orientes enzyme homodimer and increase in the flexibility of the loop, thereby improving promiscuous activities. Comparison of activity profiles of mutants with increased native and promiscuous activities showed that promiscuous activity is facilitates with distinct loop conformation and different subset of conformational loop landscape. Another similar case is enzyme mammalian serum paraoxonase 1 (PON1) where owing to its flexible active site loop, this enzyme binds to diverse set of substrates with alternative active-site configurations (Ben-David et al. 2012). Such plasticity of active site mediating promiscuity is known in other enzymes as well (Yu Zhang et al. 2015; McMillan et al. 2014; Lopez-Gallego, Wawrzyn, and Schmidt-Dannert 2010).

2. **Different modes of interactions:** Same active site configuration can catalyze both native and promiscuous reactions. Often same catalytic machinery is utilized for native and promiscuous activities, however interacting residues involved in recognition of their respective substrate might differ. As mentioned before (section 1.2.4), KDG alolase enzyme from hyperthermophilic Archaea *Sulfolobus solfataricus* harbor substrate promiscuity and can catalyze both KDG and D-2-keto-3-deoxygalactonate (KDGal) to produce pyruvate and D-glyceraldehyde (Theodossis et al. 2004). The catalysis of both these substrates involves formation of Schiff's base by K155, followed by hydration and cleavage. However, different set of residues are involved in hydrogen bonding with KDG and KDGal. Another such example is of penten superfamily which include guanidine modifying enzymes and harbor hydrolase, dihydrolase and amidinotransferase activities (Linsky and Fast 2010). Although their catalytic core (constituting Cys, His and two polar guanidine binding residues) remains conserved, they use different set of residues to recognized diverse set of substrates. Interestingly, amidinotransferase reorients its substrate (arginine) by a shift of $\sim 120^\circ$ pivoted around guanidine carbon in contrast to hydrolases and

able to utilize same catalytic machinery to cleave different C-N bonds. Thus, apart from different set of interacting residues from enzyme, different substrate/ligand binding modes/orientation may also facilitate promiscuity. Such substrate assisted catalysis is also reported in salicylic acid binding protein 2 (SABP2) which possess promiscuous esterase activity apart from catalyzing its natural substrate methyl salicylate (MeSA) (Yao et al. 2015). Further, using combination of computational and experimental analysis, it was demonstrated that the hydroxyl group of MeSA is responsible for substrate discrimination among natural and other promiscuous substrates.

- 3. Location of active site:** In many cases, location and secondary structural content of active site facilitates enzymes to bind to diverse set of substrates. The active site in TIM barrels-the most reoccurring structural fold in proteins, is located at the one end of the barrel comprising of many loops which enable them to bind to diverse substrates in various proteomes (David, Joshua, and Laura 2016). Dellus-Gur, Tawfik and co-workers introduced the concept of “Polarity” according to which active site is composed of flexible loops juxtaposed and well-separated from rest of the highly ordered core scaffold (Dellus-Gur et al. 2013). Experimentally studying variants TEM-1 β -lactamase, they proposed such a polarity leads to active site flexibility while stabilizing the scaffold (core) and is a key feature of enzymes amenable to evolution and gaining new functions. Similar behavior was observed during Differential Scanning Calorimetry (DSC) analyses of two structurally similar glutathione S- transferase (GST) isoforms – one highly specific GSTA4-4 and other highly promiscuous GSTA1-1 (Honaker et al. 2011). The catalytic promiscuity was facilitated in latter isoform by existence of “fluid active site” capable of sample many conformations while being attached to its rigid scaffold which usually remains in strictly folded state. Further, lowering of the chemical transition state barrier by GSTA1-1 is more in comparison to GSTA4-4, which brings about the conformational rearrangement necessary for binding to diverse range of substrates.
- 4. Different protonation states:** Protonation state of a catalytic residue may differ while performing its respective native or promiscuous activity. This is commonly observed among the members of catalytically promiscuous tautomerase superfamily, which has characteristic β - α - β structural fold and also possess conserved catalytic amino-terminal Proline residue in different protonation states. Depending on the

protonation state, the mechanism of catalysis is altered where catalytic proline act as general acid at high pKa (~9.2) or as general base at low pKa (~6.4). In this aspect, two out of the five known families of this superfamily: 4-Oxalocrotonate tautomerase (4-OT) and malonate semialdehyde decarboxylase (MSAD) are discussed further. The native activity of enzyme 4-OT is to catalyze the isomerization of unconjugated α -keto acids such as 2-oxo-4-hexenedioate to its conjugated isomer 2-oxo-3-hexenedioate through a dienolate intermediate. The catalytic Pro1 residue of this enzyme is usually uncharged at cellular pH and with pKa of ~6.4, it acts as a general base, abstracting 2 hydroxyl proton of α -keto acid and transferring it yield its conjugated isomer. The crystal structure of 4-OT showed that the residues within 9 Å of Pro1 are predominantly hydrophobic creating a site with low dielectric constant (Taylor et al. 1998). However, during experimental studies at pH 7.3, the secondary amine of Pro1 act as a nucleophile and forms an imine/enamine with various aldehyde and ketone compounds (Zandvoort et al. 2011). 4-OT also possess weak trans-3-chloroacrylic acid dehalogenase (CaaD) activity (S. C. Wang, Johnson, and Whitman 2003). It should be noted that cis-3-chloroacrylic acid dehalogenase (cis-CaaD) is one among the five families of this superfamily and difference in the catalytic efficiency of CaaD activity of 4-OT with typical CaaD might be a consequence of low pKa value of Pro1 (6.4 for 4-OT vs. 9.2 for CaaD). Another member of tautomerase superfamily, Malonate semialdehyde decarboxylase (MSAD), in addition to its native decarboxylase activity also exhibit promiscuous hydratase activity primarily because of protonated (cationic) catalytic Pro1 (Gerrit J Poelarends et al. 2004). Both MSAD and CaaD have hydratase activity and ~9.2 pKa of catalytic Pro1 suggesting that they both may have diverged from common ancestral protein. It has been suggested that this hydratase activity in MSAD removes covalent adducts between Pro1 and reactive aldehydes. However, in some MSAD homologs from *Burkholderia phymatum strain STM815* exhibit significant hydratase activity but lack decarboxylase activity suggesting that hydratase activity might be a new activity for an unknown substrate (Huddleston, Burks, and Whitman 2014).

5. **Different subsites within the same active sites:** Many a times, the main feature of the catalytic machinery of a promiscuous enzyme is shared while catalyzing either native or promiscuous activity but the rest of the catalytic machinery may differ while catalyzing these activities. The classical example of this mechanism is enzymes

which form “oxyanion hole” while catalysis in order to stabilize the negative charge built up on transition state analogue. For instance, the native lipid hydrolysis activity of CAL-B (*Candida antarctica* lipase B) is mediated by Ser-His-Asp catalytic triad (Uppenberg et al. 1994). CAL-B is of particular interest owed to its strong stereospecificity on chiral substrates while hydrolysis, which in turn is facilitated by narrow and deep channel leading to its open active site containing oxyanion hole. Using this oxyanion hole, CAL-B can also catalyzes other promiscuous reactions like aldol condensations and Michael addition of secondary amines (Branneby et al. 2003; Carlqvist et al. 2005; Torre, Alfonso, and Gotor 2004). However, catalytic SER does not participate while catalysis of these promiscuous activities. Another such example is of PON1 (serum paraoxonase) which physiologically act as lactonases but also possess promiscuous esterase and phosphotriesterase activities. The characteristic active-site feature of this enzyme is catalytic calcium ion, which is coordinated to phosphoryl/carbonyl oxygen and deeply seated in the hydrophobic active site and forms oxyanion hole. Both the native lactonase and arylesterase promiscuous activity is mediated by His115-His134 dyad (Khersonsky and Tawfik 2006). However, promiscuous phosphotriesterase activity involves other residues which can act as nucleophile or base (Blum et al. 2006). Further this promiscuous activity is independent of His115-His134 dyad, mutating them leads to drastic decrease in native activity but upto 300-fold increase in promiscuous phosphotriesterase activity (Gabriel Amitai et al. 2006; Yeung, Lenz, and Cerasoli 2005).

- 6. Domain insertion:** In certain superfamilies like HAD (Haloacid dehydrogenase), domain insertion near active site has enabled the binding of other non-native substrates. The active site of HAD is present in its conserved core domain which is constituted by Rossmann fold. Often core domain is supplement by an inserted domain referred as cap domain which enable them to recognize diverse range of substrates and regulates the access to the active site (Park et al. 2015). The native activity of HAD is to hydrolyze phosphate monoesters and involves cleavage of P-O bond through covalent catalysis by nucleophilic Asp residue. Insertion of cap domain extends the active site and provides an additional Lys residue which provides the electron sink for catalysis of C-P bond-cleavage occurring while phosphonate hydrolase activity (Morais et al. 2000). For example, NagD– a member of HAD superfamily from *E. coli* can dephosphorylate wide range of substrates (Tremblay,

Dunaway-Mariano, and Allen 2006) facilitated by extension of binding site by insertion of cap domain shown in pink color in Figure 1.14.

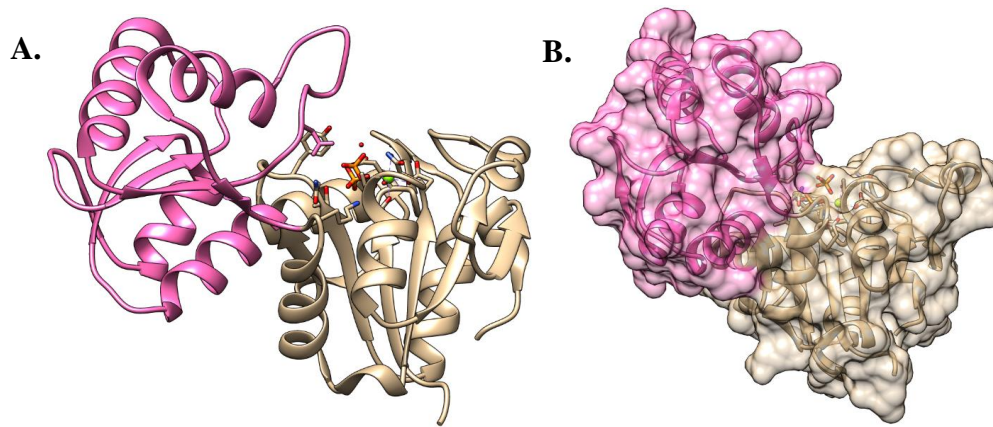


Figure 1.14 Role of domain insertion in enzyme promiscuity in *E. coli* NagD. Structure of *E. coli* NagD (*2c4n*) shown in A) new cartoon representation B) surface representation. The active site shown in licorice representation is present in conserved core domain shown in pale brown color. Inserted cap domain is shown in pink color, which enable them to recognize diverse range of substrates and regulates the access to the active site.

Further, HAD superfamily which shows catalytic promiscuity is further classified into four subfamilies I, IIA, IIB, and III based on the topology and insertion site of cap domain. Another such example was revealed during high-throughput enzymatic activity screening of 124 representative members from DUF849 Pfam family (Bastard et al. 2013). 20% of these screened members were found to catalyze five or more substrates and members of two subfamilies showed substrate ambiguity with a capability to bind to range of aliphatic and polar substrates. G1, one of these subfamily harbor a cap domain which covers the active site and increase its size enable the enzyme to bind to hydrophobic substrates with an ease.

7. **Allostery:** Allosteric interactions can also facilitate promiscuity in certain cases. A member of HAD superfamily, Cytosolic 5'-nucleotidase II (cN-II) catalyzes the dephosphorylation of 6-hydroxypurine nucleoside 5'-monophosphates (NMP). It also possesses phosphotransferase activity by which is transfers phosphate from donor NMP to an acceptor nucleoside. cN-II also dephosphorylates other allosterically activated nucleoside analogues that are commonly used in treatment of cancer and viral diseases (Wallden and Nordlund 2011). These allosteric effectors molecules increase both nucleotidase and phosphotransferase activity. Binding of effector

molecule induces disorder-to-order transition of helix A of cN-II leading to coordinated active site Mg^{2+} and enables substrate to bind by anchoring its phosphate group. Further, different interacting modes in cap domain enable it to catalyze wide range of substrates. For instance, while catalyzing UMP, Y210 makes vander Waals interaction with ribose moiety and R202, D206 and H209 are involved in hydrogen bonding. However, to accommodate larger substrate dGMP, Y210 moves out of active site and H209 is no more involved in hydrogen bonding. A recent study showed that promiscuity of ADP-glucose pyrophosphorylase in *E. coli* is controlled by allosteric activator fructose-1,6-bisphosphate (FBP) (Ebrecht et al. 2017). The catalytic efficiency for ATP is nearly ~600 higher than other nucleotides in the presence of FBP while in its absence this value is just ~3 fold higher.

- 8. Other mediators like alternative cofactors/amino-acids or water:** In many enzymes, cofactors such as metal ions induce promiscuous activities. For instance, di- Cu^{2+} substituted aminopeptidase from *Streptomyces griseus* exhibit promiscuous oxidase activity comparable to native catechol oxidase (da Silva and Ming 2005). Incorporation of tungstate (IV) in neutral zinc dependent protease thermolysin from *Bacillus proteolyticus rokko* lead to peroxidase activity (Bakker, Rantwijk, and Sheldon 2002). Another such example is substitution of native Zn^{2+} by Mn^{2+} in carbonic anhydrase which induces enantioselective epoxidation of styrene, whose native activity is reversible hydration of carbon dioxide (Fernández-Gacio et al. 2006). In another study, carbonic anhydrase was converted into first cofactor-independent reductase by substituting rhodium (I) into it thereby enabling it to catalyze hydrogenation of olefins (Jing, Okrasa, and Kazlauskas 2008). Water may also mediate promiscuity by forming hydrogen bonds with the substrates. Water molecules can also act as general acid, base or nucleophile during catalysis of promiscuous reactions. Further, they also act a buffer to opposing charges or dipoles among the active site residues and the substrate. A recent study showed that PON1 discriminates among its native lactonase and promiscuous organophosphatase activity by selective hydration of the active site (Blaha-Nelson et al. 2017). As usually PON1's substrate (native/promiscuous) are hydrophobic, substrate binding triggers the closing of the flexible loop covering the substrate bound active site and creates a hydrophobic cage with only nucleophilic water in the vicinity of reactants. In order to study the role of functionally critical residue Y71 present in this flexible loop,

various Y71 mutants were analyzed using detailed computational and experimental studies. These mutations apart from alteration in PON1's active site volume, flexibility, shape, resulted in increase in solvent accessibility of active site. This increase decreased the promiscuous organophosphatase activity but native lactonase activity was unaffected. The increase in active site size, leads to more influx of water and increase the electrostatic environment of the active site making the environment unfavorable for binding of preferential organophosphatase hydrolases's substrates such as lipolactones or neutral phosphate esters which are usually large and hydrophobic in nature. Thus, exclusion of solvent molecules from active site is clearly advantageous for promiscuous organophosphatase activity. This feature is in contrast to other promiscuous phosphatase such as members of alkaline phosphatase superfamily where active site polarity and charge facilitates binding of different anionic substrates.

The discussed mechanisms of promiscuity have native and promiscuous activities catalyzed within the same active site. However, as discussed earlier in the section 1.2.2, there are different definitions of promiscuity in this area of research. In much broader terms, enzyme promiscuity is referred as the ability of the enzyme to catalyze distinctly different chemical substrates/reactions, which are physiologically irrelevant. Most of the studies by assume promiscuous enzymes utilizing their evolved catalytic site for these side-reactions based on the fact that an enzyme will not evolve or harbor many catalytic center, rather it will use already evolved active sites and alternate substrates/reactions are usually related to each other. In few reports, native catalytic sites are not involved in promiscuous reactions and these are referred as **alternate-site enzyme promiscuity**. The first example of alternate-site enzyme promiscuity was reported in thermostable enzyme tHisF from *Thermotoga maritima* which is primarily cyclization reactions involved in histidine biosynthesis (Taglieber et al. 2007). This enzyme was shown to catalyze p-nitro-phenyl esters and thereby exhibits esterase-like promiscuous activity. However, mutations at active site residues required for its native activity do not influence its promiscuous activity and the catalytic efficiency of this side reaction essential remained constant in these mutants. Such a phenomenon is of general importance as its promiscuity originates solely from mutation in native active center and role of such alternate-site promiscuity in conferring fitness benefit to an organism during the course of evolution.

Further a recent large-scale analysis of evolution of catalytic machinery of all experimentally annotated enzyme superfamilies in CATH database revealed that at least 87% of all enzyme functions seems to evolved from another function or from a more generic ancestor (Furnham et al. 2016). Thus, remaining 13% of enzyme function may have generated using sites other than native active center, and could be alternate-site promiscuous enzymes (Furnham et al. 2016).

1.2.10 Importance of inefficiency in promiscuous enzymes

As mentioned earlier, the promiscuous activities are physiological irrelevant and usually occur at so low levels that they almost undetectable under normal conditions. Skolnick and Gao referred these low-level adventitious reactions as “biochemical noise” which cannot (or difficult) be eliminated and are inherent to enzymes (Skolnick and Gao 2013). In lines in Ycas and Jensen model of evolution, Twakif and Khersonsky suggests that promiscuity is inherent to the enzymes and modern day enzymes have evolved from primordial enzymes catalyzing numerous low-level reactions (Khersonsky, Roodveldt, and Tawfik 2006; O. K. and D. S. Tawfik 2010). Further, in various directed evolutionary experimental studies often low-level desired function is achieved after only few generations suggesting that this low-level promiscuity is inherent property of the protein (Jürgens et al. 2000; Pande, Szewczyk, and Grover 2010; Khersonsky et al. 2011; Khare et al. 2012; Soo et al. 2016; Rahimi et al. 2016; Khanal et al. 2015). The promiscuous enzymes are usually inefficient so that these secondary side-reactions do not interfere with the enzyme’s primary function. Below we discuss the mechanisms, which facilitate such inefficiency of promiscuous reaction within the enzymes native active site.

1.2.10.1 Mechanisms facilitating the inefficiency of promiscuous reactions

1. **Sub-optimal orientation of the promiscuous substrates:** Often these non-native substrates do not bind in an optimal position and hence are unable to take full advantage of the catalytic toolkit. For instance, AiiA, a metallo- β -lactamase from *Bacillus thuringiensis* exhibit very low-level promiscuous phosphotriesterase activity. MD simulations of paraoxon docked AiiA structure showed that this promiscuous substrate is mis-positioned with respect to water attacking during

catalysis (G. Yang et al. 2016). Moreover, mutations in peripheral regions lead to displacement of non-mutated active site residue F68 by ~ 3 Å which facilitate stacking interaction of F68 with p-nitrophenyl group of paraoxon (Figure 1.15). This leads to favorable orientation of the promiscuous substrate (paraoxon) towards catalytic machinery and lead to 1000-fold improvement in phosphotriesterase activity.

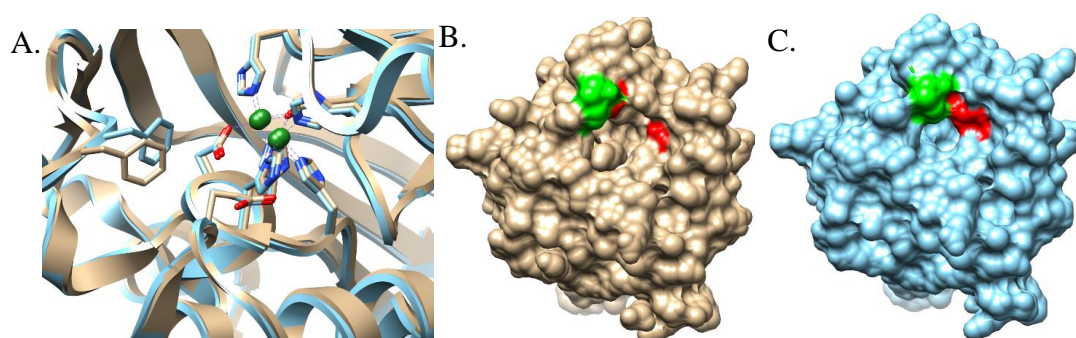


Figure 1.15 Comparison of wild-type and mutant-type AiiA, a metallo- β -lactamase from *Bacillus thuringiensis*. A) Superimposed wt (*pdbid 5eh9*) and mt (*pdbid 5eht*) AiiA with catalytic residues highlighted in pale brown and blue licorice representation respectively, F68 is displaced by ~ 3 Å in mt AiiA leading to 1000-fold increase of its phosphodiesterase promiscuous activity in mt. Structural change in F68 (green color) and two active site mutated residues S20F and V69G (red) shown in B) wt and C) mt AiiA.

2. **Binding of promiscuous substrate to sparsely populated conformations:** Various studies have shown that enzymes exhibit as an ensemble for various conformations among which some are better binding scaffold for some substrates enabling their efficient catalysis. The promiscuous substrates may bind to the conformations, which are not that frequent in the conformational ensemble and are inefficient binders. One such example in which this is observed is TER-1 β -lactamase. The catalytic efficiency for hydrolysis of ampicillin is much higher (2600-fold) than for cefotaxime, a much larger substrate (Dellus-Gur et al. 2015). However, this difference drastically drops to only 11-fold in case of G238S mutant of this enzyme. This mutated residue lies in the loop (238-loop), which flanks active site. In contrast to wild-type enzyme where 238-loop exists in closed conformations, this loop in mutant exits in major open conformation. This clearly indicates that in wild-type the promiscuous activity towards cefotaxime occurs in sparsely populated conformer whose population increases in G238S mutant.

Understanding how such low-level promiscuous activities are mediated will be beneficial in understanding how enzyme attains specificity and further these principles can be used in enzyme engineering.

1.2.11 Generalist and Specialists

As mentioned earlier, in 2012, in genome-scale metabolic network of *E. coli*, Nam et al found that 37% of generalists which catalyzed majority (~65%) of the cellular reactions. Further, stark differences were found among generalists and specialist (Nam et al. 2012b). For instance, in contrast to specialists, generalists showed the following features:

- 1) **Lower essentiality:** The genes encoding generalists are more frequently non-essential in-vivo. On the other hand, specialists encoding genes are more essential and involved in central metabolic pathways. In fact, simulated growth conditions (174) showed that most (56%) essential reactions were catalyzed by specialists.
- 2) **Smaller flux:** The fluxes through reactions catalyzed by generalists were comparatively smaller compared to flux through specialized catalyzed reactions. However, the flux through specialists was found to be more sensitive to environmental changes compared to generalists. The change in flux through generalists during various environmental shifts simulated was negligible.
- 3) **Reduced regulatory requirements:** Regulation of metabolic flux mediated by metabolite/substrate binding to enzyme or posttranslational modifications of enzyme. These regulatory interactions were observed more predominantly in specialists compared to generalists. Thus, during the course of evolution, during duplication enzymes are specialized with more regulated fluxes.

Further, prevalence of numerous generalists was found in other genomes including archaeon *Methanosarcina barkeri* and the eukaryotes *Saccharomyces cerevisiae* and *Chlamydomonas reinhardtii* (Nam et al. 2012b). Based on, it was suggested that these prevailing generalists are not specialized yet during evolution as these events may not provide adequate fitness advantages to counteract the cost of gene duplication and maintenance.

1.2.12 Applications of enzyme promiscuity

There are numerous applications of promiscuous enzymes. This “dark side of enzyme specificity” is extensively used in various biotechnological applications (Arora, Mukherjee, and Gupta 2014). Some of these are highlighted below and summarized in Figure 1.16.

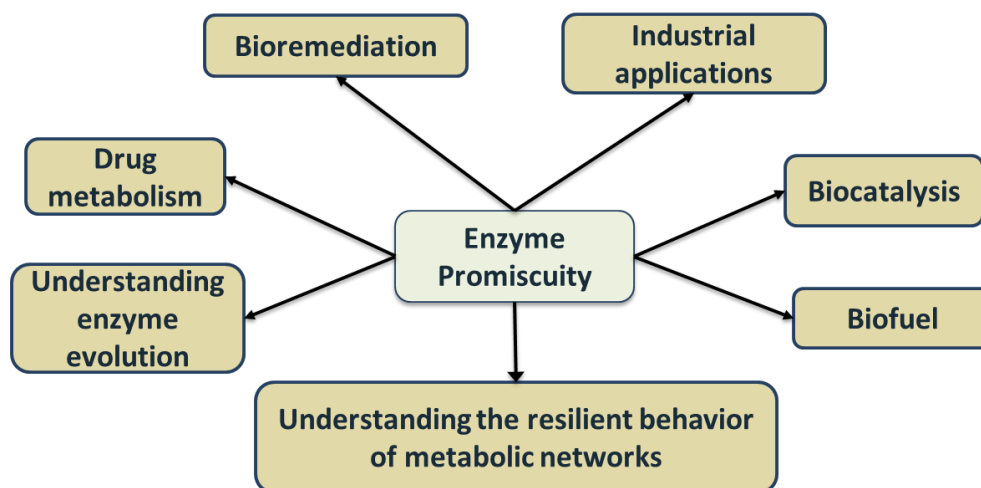


Figure 1.16 Applications of enzyme promiscuity

Bioremediation: Naturally occurring promiscuous enzymes can be used to degrade environmental pollutants, toxic compounds such as insecticides or pesticides. For instance, although Organophosphorus acid anhydrolases (OPAA) from *Alteromonas sp. strain JD6.5* exhibits native proline dipeptidase activity, it also has a high level of promiscuous activity towards degrading insecticide class Organophosphates (DeFrank and Cheng 1991; Gabriel Amitai et al. 2006). Further, a recent metagenomics study identified new esterase enzymes capable of degrading carbamates from the bovine rumen microbiome. Carbamates have a variety of uses such as pesticides or elastomers. This study clearly highlights the potential of using diverse microbiomes such as the bovine rumen for mining of promiscuous enzymes which can be exploited for commercial, biotechnological or industrial uses (Ufarte et al. 2017). Recently, a whole new class of promiscuous enzymes, P450 aryl-O-demethylase, has been reported which can convert plant waste into sustainable products (Mallinson et al. 2018).

Industrial applications: Many industries have exploited the capability of promiscuous enzymes to act as a reservoir of novel catalytic activities. For instance, pyruvate decarboxylase exhibits both substrate and catalytic promiscuity and is extensively used

in industries (BEUBERG 1921). Owing to their substrate, catalytic and conditional promiscuity lipases are widely used in various industries such as food, detergent, pharmaceutical, paper and textile industries (Kapoor and Gupta 2012). The lipases promiscuity is further enhanced in the presence of various organic solvents which is exploited by various industries (Schmid et al. 2001). Such conditional promiscuity of lipases is also exploited in solid-gas bioreactors for commercial scale production of esters (Lamare et al. 2004). Further, the naturally occurring low levels promiscuous reactions are often raised to the levels as per desire using rational designing and protein engineering.

Drug metabolism: Human carboxylesterase 1 (hce1) are promiscuous enzymes which are involved in drug metabolism apart from other biological processes (M R Redinbo, Bencharit, and Potter 2003). For instance, these are known to hydrolyze narcotics like heroin, cocaine and also involved in cholesterol metabolism. The structural elucidation of hce1 revealed that the active site of this enzyme harbor both specific and flexible pockets enabling it to act both specifically and promiscuously (Bencharit et al. 2003). Exploiting the promiscuity of metabolic enzymes of human gut microbiota, recently a “DRUGBUG” tool has been developed which predicts the gut bacterial species-specific metabolic enzyme capable of biotransformation of xenobiotic/drug molecule (A. K. Sharma et al. 2017).

Biocatalysis: Some reactions sought in biocatalysis are absent in nature. After the identification of promiscuous enzymes catalyzing the required reaction, directed evolution of rational designing approaches can be used to evolve as its specialized form, which exhibit high catalytic efficiency. For example, ω -transaminase is a commonly evolved enzyme which can produce chiral amines and is being used for rapid biomining by Codexis company on commercial scale (Shin and Kim 2001). Another catalytically promiscuous enzyme nuclease p1 from *Penicillium citrinum* is shown to catalyze asymmetric aldol reactions between isatin derivatives and cyclic ketones and is extensively used in biosynthesis of various pharmaceutically active compounds (Z.-Q. Liu et al. 2014).

Biofuel: Commercial glucose oxidase (GOx) exhibits substrate promiscuity and used in enzyme fuel cell which generates electrical energy by oxidizing variety of sugars (Milton

et al. 2015). Recently an efficient enzyme biofuel cell was designed using combination of two enzymes which harbor substrate promiscuity: pyranose dehydrogenase (PDH) and a broad glucose oxidase (bGOx) (Holade et al. 2017).

Understanding the resilient behavior of metabolic network: Microorganisms often show physiological adaptation to genetic and environmental perturbations. For instance, ~35% (80 out of 227) of metabolic enzymes are not essential for the growth the *E. coli* on glucose media (Kim and Copley 2007). This robustness is mostly attributed by dynamic interactions among its various components, which maintains of constant flow of essential metabolites allowing to bypass the defective step in disrupted pathway/s. These alternate metabolic pathways often involved promiscuous enzymes, which can catalyze analogous substrates/reactions. Thus, promiscuous enzymes can impart resilience towards various genetic or environmentally induced alterations and maintains plasticity of metabolic networks.

Understanding enzyme evolution: As discussed in section 1.2.1.2, low-level promiscuous activities can serve as repertoire of novel catalytic activities and could be important starting point in evolution of novel enzyme functions. Based on this, it can be started that present day enzymes are ancestors for future enzymes. Enormous space of promiscuity in extant enzymes gives us the tool-kit to understand enzymatic evolution. Such a vast prevalence of promiscuous activities in current enzymes has improved our knowledge of enzyme evolution and it is clear that evolution has not converged to a point where all enzymes are specific. A recent review highlighted an emerging view of enzyme evolution citing several examples indicating that while novel innovation during enzyme evolution is easy, it is their optimization which is complicated (M. S. Newton et al. 2018). Further, it emphasized that “most enzymes are far from perfect catalysts — evolution is not the pursuit of perfect enzymes. 'Real world' enzymes are sloppy and mediocre.” This emerging view provides a more realist perception of modern-day enzymes, which should always be kept in mind while protein engineering.

Protein designing and retrosynthesis:

Promiscuous enzymes are used to design or evolve desired catalysts via mutational studies or directed evolution experiments. It was found that mutations may have differential effect on promiscuous activities of different orthologues (Khanal et al. 2015).

Motivated by this observation, Khanal *et al.*, proposed the use of more than one scaffold in directed evolution experiments (Khanal *et al.* 2015). This strategy was further modified in “scaffold sampling” where beneficial mutations identified in one enzyme are transferred to a series of homologous enzymes (Dunn *et al.* 2016). Further, keeping in mind the vast space of promiscuity in extant enzymes, a recent study formulated generalized rules for chemical transformation and modelling of enzymatic promiscuity which were incorporated in the tool RetroPath 2.0 which can be used for retrosynthesis (Delépine *et al.* 2018). This tool take set of metabolites as input and target compounds and generates reactions through which these target compounds can be produced given the set of metabolites. It can identify various promiscuous reactions as shown for transaminase EC number 2.6.1.1 in this study.

Specificity being the cornerstone for such a long time in enzymology, the promiscuous activities in extant enzymes have just started to emerge by exploitation of the recent advancement in various experimental tools and techniques for their identification. We are still only looking at the tip of the ice-berg. In this thesis, we focused to decipher the general rules of promiscuity. Using large-scale structural analyses, we systematically explored the general structural principles of enzyme promiscuity especially with respect to the roles of binding site and catalytic site residues.

The overall objective of the work reported in this thesis is to systematically explore the general structural principles of enzyme promiscuity especially with respect to the roles of binding site and catalytic site residues. Since catalytic residues are mostly identified using laborious experimental methods, we have implemented a meta-approach for prediction of catalytic residues to improve prediction accuracy. We have studied structural basis of promiscuity in a sequence diverse γ -glutamyl cysteine ligase (GCL) superfamily using both traditional docking and molecular dynamics approaches. Further, we have classified families of GCL superfamily into subfamilies to improve their function annotation and understand evolutionary perspective of such a diverse GCL sequence diversity. In Chapter 2, we have used several structural parameters of binding/catalytic site residue to identify possible distinguishing factor/s between known promiscuous (generalist) and non-promiscuous (specialist) enzymes in *E. coli*. Subsequently, we have characterized ligand induced conformational changes of binding/catalytic residues by comparing differences in various properties in bound and

unbound state of an enzyme. The major properties analyzed include side chain torsional angles, angle defined to capture conformational changes involving functional groups, centroid distance of all-against-all pairwise catalytic/binding site residues. Chapter 3 describes development of meta-approaches (CSmetaPred and CSmetaPred_poc) for catalytic residue prediction to improve their prediction accuracy. In meta-approach, we have used a consensus approach, which involves computing residue average scores obtained from results from four well-known catalytic residue prediction methods. The evaluation of meta-predictor showed it improved catalytic prediction method over the best prediction methods among its constituent methods. In Chapter 4, we have investigated structural basis of γ -glutamyl cysteine ligase (GCL) enzyme promiscuity and used docking studies to understand the binding modes of various alternate substrates for *E. coli* GCL. Further, classification of GCL families into sub-families, their function annotation, and phylogenetic analysis has been performed to understand evolution of GCL function, which is discussed in Chapter 5. The two appendices provide brief overview and results of additional work carried out in collaboration during PhD.

Chapter 2

Understanding residue conformational variability of binding site and catalytic residues between specialist and generalist enzymes

2.1 Introduction

Ancestral enzymes are proposed to be generalists (having broad specificity) which have been suggested to undergo functional specialization to become specialists (having specific function) during the course of evolution mostly because of selection pressure (Jensen 1976). Thus, enzymes are expected to evolve towards performing specific catalytic activity with high catalytic efficiency. In a recent study, *E. coli*, has been shown to have 63% of enzymes as specialist (non-promiscuous) and 37% of enzymes as generalist (promiscuous) (Nam et al. 2012). Surprisingly, in this study generalists were found to perform most of the metabolic enzymatic reactions (65%). Moreover, promiscuous enzymes were shown to be involved in peripheral metabolic pathways and usually carry low metabolic fluxes. Subsequently, there have been concerted experimental efforts to decipher promiscuous enzymes using high throughput techniques. For example, promiscuous activities are now revealed by multiple knockout mutant analysis (Nakahigashi et al. 2009), multicopy suppression (Oberhardt et al. 2016), in which the lethality caused due to inactivation of a ‘target gene’ is rescued by over-expression of promiscuous ‘replacer gene’ and activity-based metabolic profiling (Prosser, Larrouy-Maumus, and de Carvalho 2014; Sevin et al. 2017). Recent systems biology approaches have enabled the mapping of these “underground activities” and prediction and analysis of their contribution in new metabolic functions which further facilitate resilience towards various new environments (Notebaart et al. 2018). It has now been recognized that promiscuity is an important driving force, which confers robustness

to metabolic networks and acts a reservoir of novel catalytic activities (Copley 2017). These promiscuous enzymes are still prevalent in extant enzymes and is a typical property of enzyme families (Baier, Copp, and Tokuriki 2016). In fact, 10 % of total bacterial and archaeal enzymes are promiscuous enzymes (M. A. Martínez-Núñez et al. 2013).

Numerous studies have shown that enzymatic catalysis often requires precise optimization of their active site, which allows favorable binding of substrate and spatial proximity and orientation of functional groups (Bartlett et al. 2002; Gutteridge and Thornton 2004, 2005). It is of considerable interest to understand how promiscuous enzymes can harbor both native and side activity within the same active site. It has been suggested that specificity of an enzyme is regulated both by protein-substrate/s interactions such as by enthalpy-driven interactions like hydrogen bonds and appropriate positioning of substrate/s relative to catalytic machinery. Whereas, promiscuous substrates usually involve non-specific interactions like hydrophobic interactions (Nobeli, Favia, and Thornton 2009; James and Tawfik 2005). Based on previous studies, it has been suggested that mechanism of promiscuous activity can completely or partly overlapped with the native activity of enzyme, however, it can differ in some cases (Pandya et al. 2014; O. K. and D. S. Tawfik 2010; Tokuriki et al. 2012). The studies on individual enzyme families have proposed following mechanisms of promiscuity: conformational diversity of active sites, sub-sites in an active site, different protonation state of catalytic residues, and assistance through metal/cofactor/water (also summarized in section 1.2.9). Of these, it has been suggested that conformational dynamics and flexibility of active site plays important role in evolution of new enzymatic functions as well as promiscuous behavior of enzymes (Gatti-Lafranconi and Hollfelder 2013; Zou et al. 2015). The details of these mechanism has been discussed previously in Chapter 1 (see section 1.2.9). These insights into mechanistic and structural basis of enzyme promiscuity can facilitate in rational design of enzyme as well as help in understanding the evolution of enzyme function and their divergence.

Previous study on *E. coli* enzymes has identified three main distinguishing characteristic of generalist as: a) mostly being non-essential, b) maintains lower metabolic flux and, c) possesses less regulatory mechanism compared to specialist (Nam et al. 2012). Moreover, lifestyle of prokaryotes has an impact on the repertoire of promiscuous enzymes. Among, free-living, extremophiles, pathogens, and intracellular organisms, free-living organisms

were found to have larger genomes and more promiscuous enzymes, where fluctuating environments aid their emergence (Martinez-Nunez, Rodriguez-Vazquez and Perez-Rueda 2015). A recent large-scale study combined classical analysis of relationship between sequence and structure with new qualitative measures of similarity of function to investigate the change in chemistry or substrate specificity in various families of 379 CATH superfamilies (Furnham et al. 2016). 91% of these 379 CATH superfamilies have diverse substrate specificity. Some superfamilies have changed the reactions they perform without changing catalytic machinery. In others, large changes of enzyme function, in terms of both overall chemistry and substrate specificity, have been brought about by significant changes in catalytic machinery. Interestingly, in some superfamilies, relatives perform similar functions but with different catalytic machineries (Furnham et al. 2016).

In the present work, we are interested in identifying general structural principles governing enzyme promiscuity. It has been observed that changes in substrate specificity, due to incremental binding site mutations is more likely to occur compared to changes in chemistry which may require complementary mutations to key catalytic residues without disrupting enzyme activity (Furnham et al. 2016; Tyzack et al. 2017). Thus, in the present work, we have explored among general structural features of active sites, which could be distinguish promiscuous from non-promiscuous enzymes. Here, we formulated two hypotheses: a. substrate/cofactor binding sites have attributes to distinguish promiscuous (generalist) and specialist enzymes, and b. catalytic sites of promiscuous enzymes are pre-primed to catalyze enzymatic reactions. Such pre-priming of catalytic residues in generalist may to facilitate catalysis/binding of alternate reaction/substrate. Deciphering such structural features in generalist could possibly facilitate development of accurate prediction methods to identify promiscuous enzymes. Moreover, this knowledge has numerous applications in drug designing, bioremediation, directed evolution and many industrial applications. It will help us in understanding the resilient behavior of metabolic networks.

2.2 Comparison of structural features of active sites of generalist and specialist enzymes

We have extracted active sites of generalist and specialist enzymes and compared their general structure features in order to identify any characteristic property of promiscuous enzymes.

2.2.1 Methods

Construction of generalist and specialist enzymes dataset

In order to compare structural difference between generalist and specialist, we have used manually curated classification of *E.coli* enzymes into these two categories (Nam et al. 2012) and used them to construct ‘eco-spec-gen’ dataset for further analysis. In their work on reconstruction of genome scale metabolic model of *E.coli*, 1081 enzymes were classified into 404 generalist and 677 specialist enzymes based on experimental and literature evidences (Nam et al. 2012). The schematic flowchart of ‘eco-spec-gen’ dataset generation is illustrated in Figure 2.1.

Briefly, first we excluded enzymes, which are part of multi-enzyme complexes or are involved in transport reactions. Then, rest of the 764 enzymes (236 generalists and 528 specialists) were searched in the PDB database (this work was performed in the period of 2013 to 2014) to find proteins with known tertiary structures. Thus, we identified 100 generalist and 235 specialist enzymes with known tertiary structure. In this work, we are concerned with analysis of active sites of enzymes. Ideally, we need substrate bound structures, to identify substrate/s binding sites, however, determination of these is not a trivial task and we relied on predicted binding sites. Although, while mapping the enzyme to their tertiary structure, we usually found more than one structure associated with a given enzyme, among these we always preferred the structure bound to a ligand, still not all representative structures of enzymes (8 structures) were bound to their cognate ligand. Hence, we resorted this caveat by using ligand binding pockets prediction method to first identify potential binding sites, then, use a simple criteria of spatial proximity between

catalytic site and binding sites as these need to be closer to catalyze enzymatic reaction (Cilia and Passerini 2010). The advantage of this generalized scheme to identify predicted substrate binding site is that it can be extended to construct a larger dataset in later studies. For this, first we used Catalytic Site Atlas (CSA) database (Porter, Bartlett, and Thornton 2004) and found 178 pdbidentifiers having defined catalytic residues. Then, we used Fpocket (Le Guilloux, Schmidtke, and Tuffery 2009) method to predict ligand binding sites and among these predicted pocket/s having at least one residue overlap with the known catalytic residues (obtained from CSA) is defined as potential substrate/s binding sites. In case there are more than one site, we retain both sites since either of them can be potential binding sites. This procedure resulted in a total 180 sites in 132 specialists and 65 sites in 46 generalists.

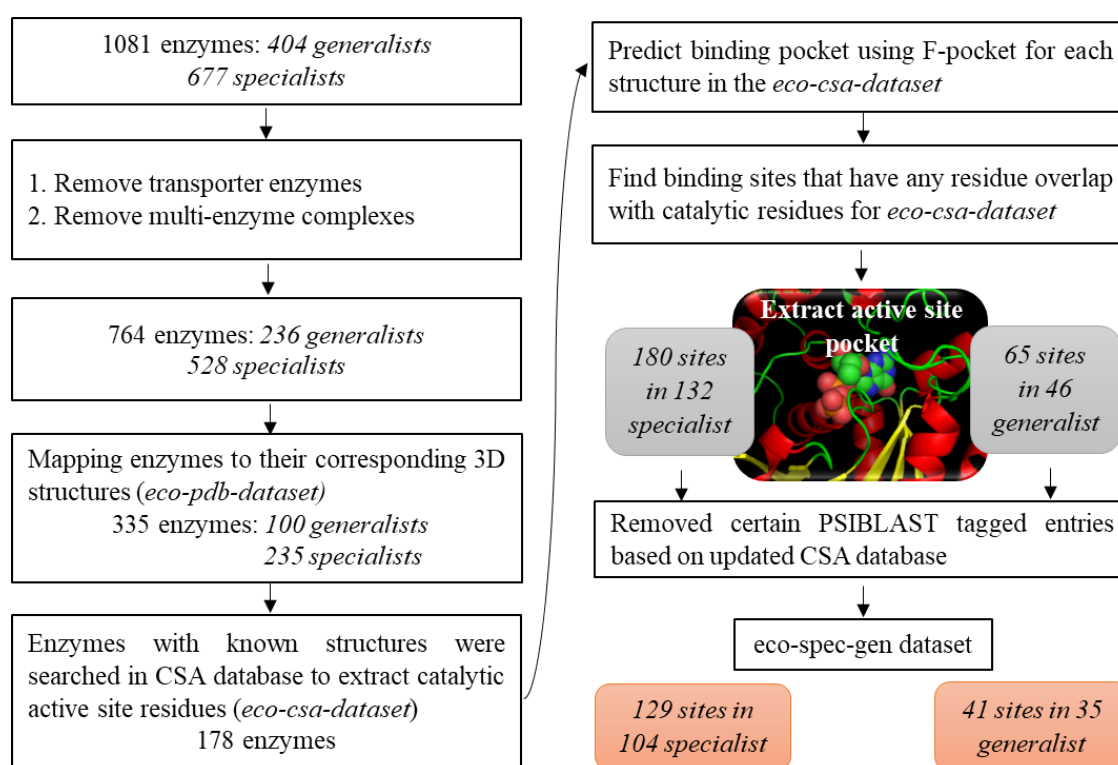


Figure 2.1 Flowchart showing the construction of *eco-spec-gen* dataset.

The CSA database consists of two sets of catalytic residues: 1) manually curated residues primarily derived from literature and 2) additional homologous catalytic residues inferred using PSI-BLAST and sequence alignment to the literature entries of CSA database. The catalytic residue annotations of the latter type were given a tag ‘PSIBLAST’ in the CSA database files by database developers of CSA. We performed analysis in 2012, when this tag was provided in the files. However, in 2014 with a major update of CSA database

(file: CSA_2_0_121113.txt) and many catalytic residues defined with the tag ‘PSIBLAST’ were removed from prior database (file: CSA_2_2_12.dat). Hence, in order to remove these false positives/obsolete entries from our dataset, we reassessed mapping between tertiary structure and CSA database and eliminated some enzymes and/or binding sites (which were absent in updated CSA) from the ‘eco-spec-gen’ dataset. Thus, we have a total 129 sites in 104 specialists and 41 sites in 35 generalists in the final dataset on which we have performed our analysis. This dataset is provided in supplementary material (c2.1_s1_eco_spec_gen_dataset.xlsx) for chapter 2 provided in a CD along with this thesis.

It should be noted that the above mentioned selection of active site from predicted set of binding pockets is motivated from the observation that the catalytic residues are mostly the subset of binding residues (Kahraman and Thornton 2008; Tseng and Li 2011) and on an average there is ~70 % overlap between the catalytic and binding site residues (Cilia and Passerini 2010; for details see section 3.3.1 of chapter 3). As preference was given to ligand bound structures during mapping of 3D structures for enzymes, out of total 139 pdbchains in our final dataset we have known binding residues for 134 pdb chains. For these enzymes, the mean overlap between known binding residues and catalytic residues is 67.59% with a 40.16 as standard deviation. Here, the number of overlapping residues is defined as number of common residues among binding and catalytic residues/number of catalytic residues.

Structural features

We analyzed following general structural features of active sites:

1. **Active site residue properties:** The **residue propensity** was calculated for residues lying at the active site, and **secondary structure content** was obtained from DSSP (Joosten et al. 2011; Kabsch and Sander 1983). The residue propensity ($RProp_i$) for a given amino-acid residue type i , is calculated using equation:

$$RProp_i = \frac{\text{percentage of residue type } i \text{ that is catalytic}}{\text{percentage of residue type } i \text{ in whole protein}}$$

Apart from computing residue propensities, we also calculated propensities for a group of residues, which are grouped based on their physiochemical properties.

2. Residue hydrophobicity index was obtained from the Fpocket output.
3. Residue flexibility was assessed by **B-factor**. Since we compared B-factor across proteins solved at various resolution, we z-score normalized B-factor using following equation:

$$z_{Bi} = \frac{B_i - \langle B \rangle}{\sigma_B}$$

where, Z_{Bi} is z-score of B-factor of an atom, $\langle B \rangle$ is mean B-factor and σ_B is standard deviation of B-factors.

We compared residue B-factor, which is average of normalized atom B-factor.

4. Residue solvent exposure as measured by **total accessible surface area, polar accessible surface area, and non-polar accessible surface area**. NACCESS was used to compute accessible surface area of the proteins (S. J. Hubbard 1992). We computed the fraction of polar and non-polar solvent accessible surface area of a pocket using the following equation:

$$Frac_x = \frac{\text{Absolute } x \text{ surface area of the pocket}}{\text{Absolute total surface area of the pocket}}$$

where x can be either polar or non-polar

5. Active site cavity size measured by **volume**, which is obtained from Fpocket output.

2.2.2 Results

The main objective of this study was to explore and possibly identify distinguishing structural features of active sites in promiscuous enzymes that can provide structural aspects of promiscuity. Moreover, such features can be potentially employed

in prediction of enzyme promiscuity. In this analysis, we have compared following structural features such as, flexibility, solvent accessible surface area of residues, and binding pocket size. Apart from this, we also analyzed residue hydrophobicity index, residues propensity and secondary structure content of the active site residues.

Before, analysis of active sites, we assessed heterogeneity as well as representation of various enzyme functional classes in both generalist and specialist using simple EC number variation at level 1 and 2 (Figure 2.2). This showed in our dataset both specialist and generalist have enzymes from all six classes of enzymes except generalist where no representative from E.C.6.-.- Ligases class was present.

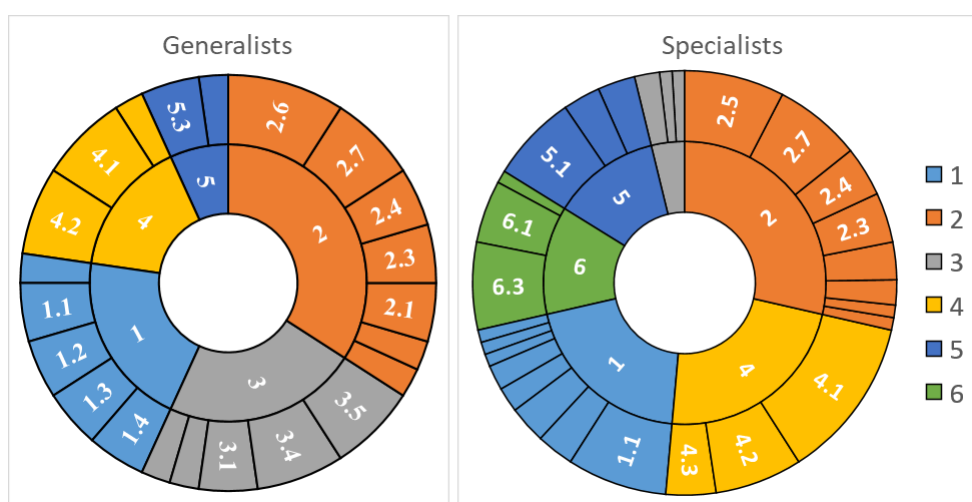


Figure 2.2 EC wheel functional classification for generalists and specialists representing the span of various enzyme classes.

First, we compared difference in sequence properties of the active sites between generalist and specialist. As shown in Figure 2.3A, most hydrophobic residues (F, Y, M, I, and L) have slightly higher propensity in generalist compared to specialist. Phenylalanine shows maximum difference in propensity. The residue Asparagine also shows higher propensity in generalist. In specialist, slight higher propensities are observed for residues such as C, P and Q. The grouping of residues based on their physiochemical properties showed (Figure 2.3B) predominant occurrence of amino acids having aromatic residues (H, Y, F and W). These suggest that active sites of generalist have a preponderance of hydrophobic residues. However, the difference in propensity between specialist and generalist is not statistically significant. Next, we looked at difference in hydrophobicity index (Figure 2.4A) obtained from F-pocket output. It is evident from cumulative distribution of hydrophobicity index, that generalist tends to

have higher hydrophobicity compared to specialist. Moreover, density distribution of hydrophobicity values (inset of Figure 2.4A) shows there are generalist pockets having more hydrophobicity than specialist. However, observed difference was found to be not statistically significant. This could essentially be because there is far less number of pockets in generalist compared to specialist and this may influence the significance.

Next, analysis of various structural features was performed. Distribution of average residue B-factor, fraction of polar ASA, fraction of non-polar ASA and pocket volume are shown in Figures 2.4 B, C, D and E respectively. As seen in these Figures, there does not seem to be much difference in any of the structural properties. Interestingly, the residue flexibility has been suggested to play an important role in conferring promiscuity (O. K. and D. S. Tawfik 2010; A. Babbitt, Tokuriki, and Hollfelder 2010; Pandya et al. 2014). However, we could not find this in distribution of average residue B-factor. The residue B-factor can be influenced by the ligand bound in tertiary structure at the active site that usually results in masking ligand binding residues from solvent and reducing their B-factor. In general, ~50% of residues are involved interaction with the ligand for both specialist and generalist enzymes. Since B-factor is a fitted parameter in x-ray structure determination that depends on many factors such as resolution of crystallographic data, the comparison of B-factor across structure of variable resolutions may not be an appropriate measure. Alternatively, it is quite possible that there is indeed not much difference in the active site residue flexibility between specialist and generalist. This indicates active site residues are somewhat equally flexible in both groups of enzymes suggesting that residue flexibility could potentially important in evolution of new functions of an enzyme irrespective it is specialist or generalist.

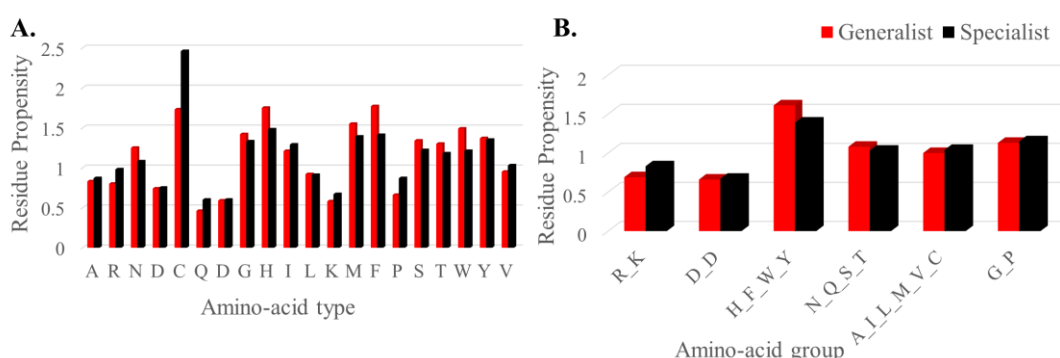


Figure 2.3 Histogram showing propensities of active site A) amino-acids and B) similar physiochemically grouped amino acid lying in the active sites of generalist and specialist.

The polar and non-polar accessible surface area did not show variation contrary to previous observation that hydrophobicity is predominant feature in case of generalist (not shown). In order to find contribution of non-polar ASA to the total ASA, we recalculated the fraction of non-polar ASA with respect to total ASA. As shown in Figure 2.4 D, non-polar ASA is slightly more in case of generalist than specialist that is in concurrence with sequence based features.

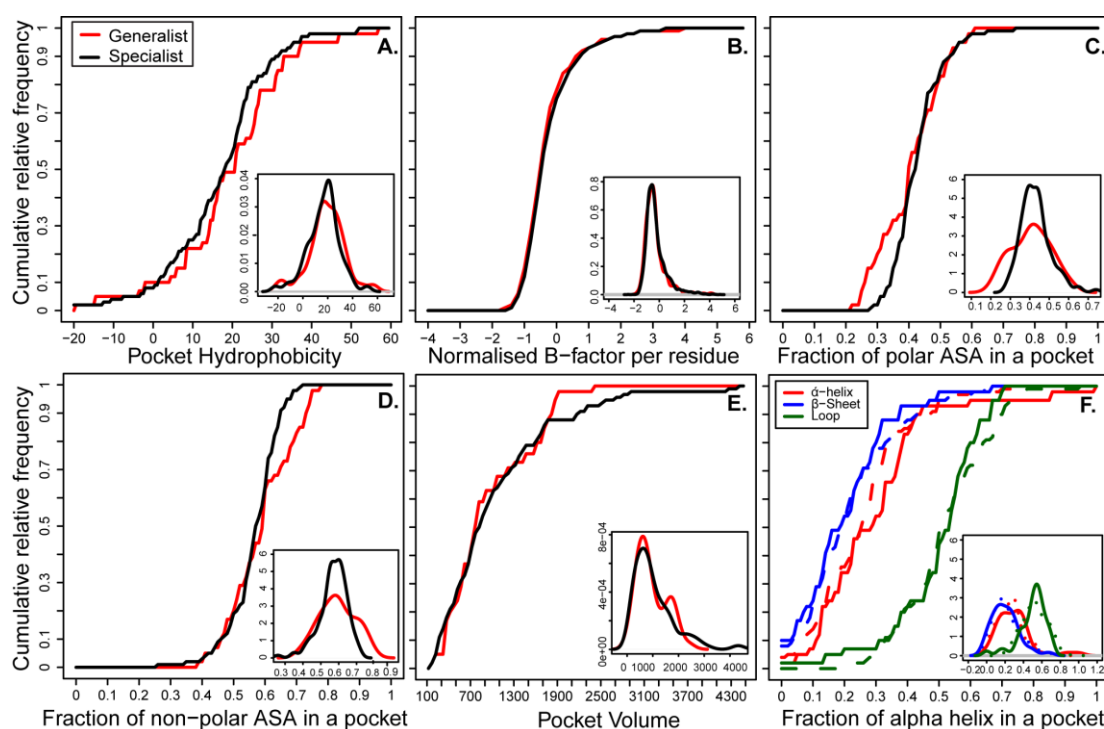


Figure 2.4 Cumulative distribution of active site features. A) *pocket hydrophobicity index* B) *Normalized residue B-factor* C) *polar ASA* D) *non-polar ASA* E) *pocket volume* and F) *secondary structure content of pocket residues*. Inset shows the density plot of the respective structural feature. In F) the solid line represents generalists while the dotted line represents dashed lines.

Even though we did not find large differences in various measures of hydrophobicity such as Hydrophobicity index, amino acid propensity, and non-polar ASA, the predominance of hydrophobic residues at active site residues suggests role of hydrophobicity in conferring promiscuity to enzymes. The hydrophobicity can allow non-specific molecular recognition at the binding site that can eventually evolve as a substrate under selective evolutionary pressure. Similar role of hydrophobicity has been suggested previously as well (A. Babbie, Tokuriki, and Hollfelder 2010). In general, hydrophobic interactions are thought to contribute the most in substrate-recognition by an enzyme (Wilfried and N. 2018). Unlike electrostatic or hydrogen bonding, which

requires specific arrangement of functional groups, hydrophobic interactions are non-specific in nature. Previous studies have shown that the catalytic efficiency of many promiscuous enzymes is dependent on substrate hydrophobicity and its interaction with deeply seated hydrophobic active sites (M. and Florian 2009; Khersonsky and Tawfik 2005; Afriat et al. 2006). Below we discuss examples of hydrophobicity observed in specialist and generalist.

2.2.2.1 Role of hydrophobic active site pocket in enzyme promiscuity

Here we discuss case studies enzymes having very high or low hydrophobicity index and their prospective role in enzyme promiscuity or influence specificity. The enzyme β -hydroxydecanoyl thiol ester dehydrase (herein it will be referred as β -dehydrase) pdbid: 1mkaA shows high hydrophobicity index of 56.56. This enzyme is essential in biosynthesis of unsaturated fatty acids. β -dehydrase is a catalytically promiscuous enzyme as it catalyzes two reactions on fatty acid thiol esters of acyl carrier protein (ACP): a. dehydration of (R)-3hydroxydecanoyl-ACP (ACP: Acyl carrier protein) to (E)-2-decenoyl-ACP (elongation step) and, b. isomerization of (E)-2-decenoyl-ACP to (Z)-3-decenoylACP. Importantly, both reactions involve same active site with any role of metals or cofactors unlike other enzymes catalyzing similar reactions (Leesong et al. 1996). The substrate binding site of β -dehydrase is a tunnel-shaped pocket resembling a worm hole isolated from rest of the solvent (Figure 2.5) with polar residues lying the entrance of this pocket and innermost half of tunnel harbors hydrophobic residues, which facilitates binding of long tail of fatty acid.

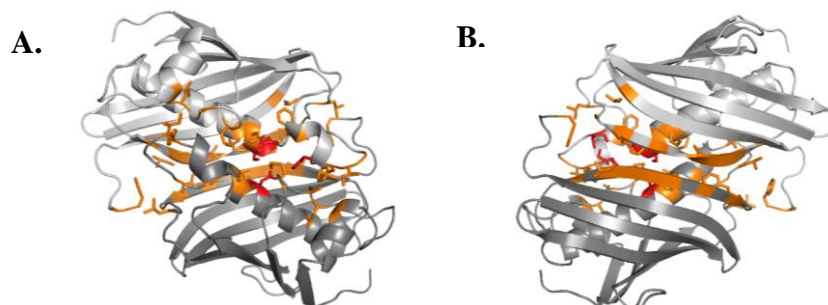


Figure 2.5 Substrate binding site of β -dehydratase. A) Front view and B) rear view of β -dehydratase (pdbid 1mkaA) in cartoon representation with the chain A and chain B colored in light grey and dark grey. Hydrophobic interface residues are represented in licorice with orange color. The catalytic residues H70 A, Gly 79 A, Cys 80 A, Val 76 A and Asp 84 B are colored red.

Among specialist, the substrate binding site of enzyme Lysyl-tRNA synthetase (pdbid:1bbuA) (LTS) has low hydrophobicity index of 2.8. As this enzyme is listed as specialist, it is expected to exhibit substrate specificity towards its cognate substrate L-Lysine and t-RNA. It is known that L-Lysine is recognized mostly by charged and polar side chains of amino-acid residues (Y280, E428, E240, N424, R262 and E278) (Onesti et al. 2000) as shown in Figure 2.6 that is probably responsible for specificity. In addition to this, upon substrate binding their concerted reorganization of the active site, which involves conformation change of residues 393-409; ordering of residues 215-217; 444-455, and rotation of a 4-helix bundle domain by 10° . Based on these, extent of this conformational change has been suggested to dictate specificity of tRNA synthetases. Here polar and charged interaction play a crucial role in catalysis and suggested as trigger in conformational change.

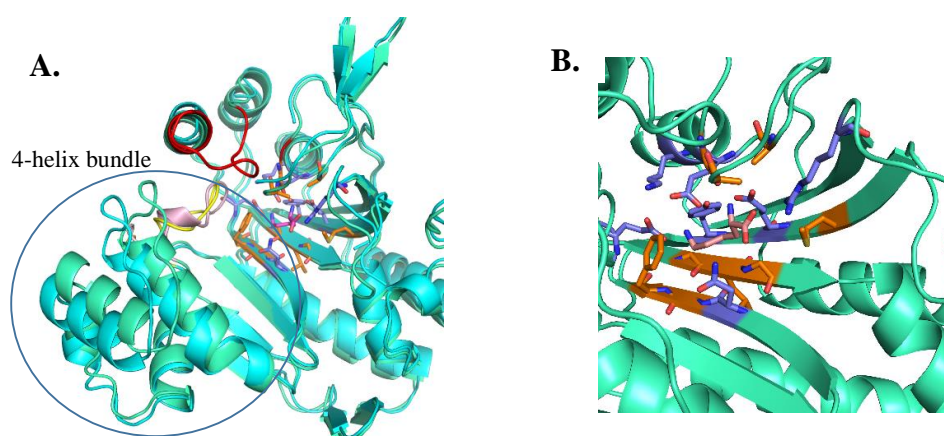


Figure 2.6 Conformational changes of Lysine binding in Lysyl-tRNA synthetase (LTS). A) Cartoon representation of superposed structures of LTS bound (pbbid: 1bbuA) and unbound (pbbid: 1bbwA) state shown in aqua-green and cyan respectively. Encircled 4 helix bundle rotates by 10° and region (215-217 and 444-455) highlighted in red becomes ordered on substrate binding B) Superimposed substrate binding where substrate LYS, polar and hydrophobic binding residues are shown in licorice representation colored magenta, blue, and orange respectively.

Increase in hydrophobicity of substrate binding site due to switch loop movement governing substrate specificity in *E. coli* enzyme TAP

Detailed analysis of the dataset we found that the B-factor for the binding residue in *E. coli* thioesterase I/protease I/lysophospholipase L 1 (TAP) (pdbid:1u8uA) is as high as 104.17 \AA^2 . TAP harbors diverse set of enzymatic activities including the thioesterase,

lysophospholipase, esterase, arylesterase, and protease activities, and stereoselectivity for amino acid derivatives. The substrate specificity of this promiscuous enzyme is dictated by its switch loop movement (Lo et al. 2005). Substrate binding initiates the conformational change in TAP's switch loop (residue 75 to 80), which is shown in red color in Figure 2.7A.

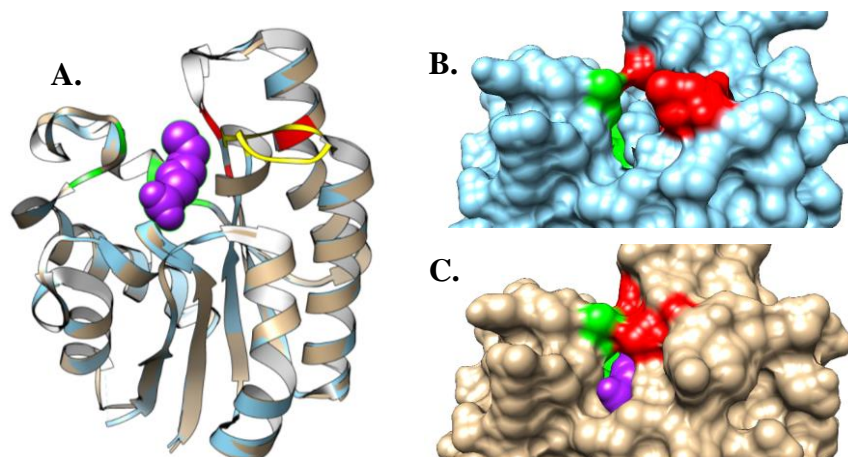


Figure 2.7 Substrate induced change in hydrophobicity of binding site on substrate binding *E. coli* thioesterase I/protease I/lysophospholipase L1(TAP). A) Cartoon representation of superimposed bound and unbound enzyme showing conformational change in the switch loop (colored in yellow) on substrate binding. The unbound (*1ivnA*) and bound (*1u8uA*) states are shown in cyan and pale brown respectively. The substrate octanoic acid is shown in purple colored space-fill representation. B) The unbound state have two distinct hydrophobic clusters colored in green and red colors which become continuous in C) bound state. The green colored cluster is formed by L11, F139, F140, M141, I156, and P158 while the red colored cluster is formed by G72, G75, L76, I107, L109, P110.

Studies have shown that short acyl chain substrates could not trigger TAP's switch loop movement and suggest that switch loop movement is acyl chain length dependent. In the unbound state, the substrate binding site has two separate hydrophobic clusters, with cluster 1 constituting Leu11, Phe139, Phe140, Met141, Ile156, and Pro158 and cluster 2 constituting Gly72, Gly75, Leu76, Ile107, Leu109, Pro110, and Phe121. Upon OCA binding, the C 4 atoms of OCA are in close contact with residues in the cluster 1, while the C 5 -C 8 atoms of OCA are in close contact with those in cluster 2. Thus, they form a continuous hydrophobic surface as shown in Figure 2.7C, thus increasing the hydrophobicity around the substrate-binding site of TAP. It has been suggested that this increase in hydrophobicity of substrate-binding site is a trigger for switch loop movement. Such examples motivated us to investigate the extent of variation in ligand-

induced conformational changes among generalist and specialist enzymes and this is further discussed in section 2.3.

In our previous analysis of cumulative distribution of active site features (Figure 2.4), except hydrophobicity and non-polar solvent accessible surface area of active sites, none of other properties showed any difference between specialist and generalist. We explored possibility of combining these features together to predict catalytic residues. It has been suggested previously that promiscuous proteins use a combination of hydrogen bond, hydrophobic interactions and flexibility to bind range of substrates (Matthew R Redinbo 2004). For this, we performed PCA of eco-gen-spec dataset to analyze the contribution of various active site features in distinguishing generalist and specialist. Approximately ~70% of variability is accounted within first 3 eigenvectors. Figure 2.8A shows PCA bi-plot where the individual enzymes in the dataset are colored by two different types of enzymes. In this biplot, an individual on the same side of a given variable (active site feature) has a high value for this variable. An individual that is on the opposite side of a given variable has a low value for this variable. In general, contribution of hydrophobicity and α -helix is more in case of generalist. For specialists, polar surface area, non-polar surface area and volume contribute more. Although, as seen in Figure 2.8A, there is a significant overlap between the two class of enzymes. Thus, current set of active site features are not sufficient enough to distinguish generalist and specialist. It should be noted that in a biplot the coordinates of individual and variables are not constructed on the same space. Therefore, one should only focus on the direction of the variable but not on their absolute position on the plot. Next, we plotted correlation circle/variable correlation plots (Figure 2.8B) for specialist and generalist separately in order to find different correlating feature among them if any. In this plot, the correlation between the variable and a principle component is used as the coordinates of the variable on the PC. The representation of variables differs from the plot of the observations: The observations are represented by their projections, but the variables are represented by their correlations. In this correlation circle plots, positively correlated variables are grouped together, negatively correlated variables are positioned on opposite sides/quadrants of the plots. The distance of the variable (in this case, the property of active site) from origin represents the estimate of the quality of representation on the principal component. If the variable is positioned close to the circumference of

correlation, then it has good representation on the principal component (PC) (Abdi and Williams 2010).

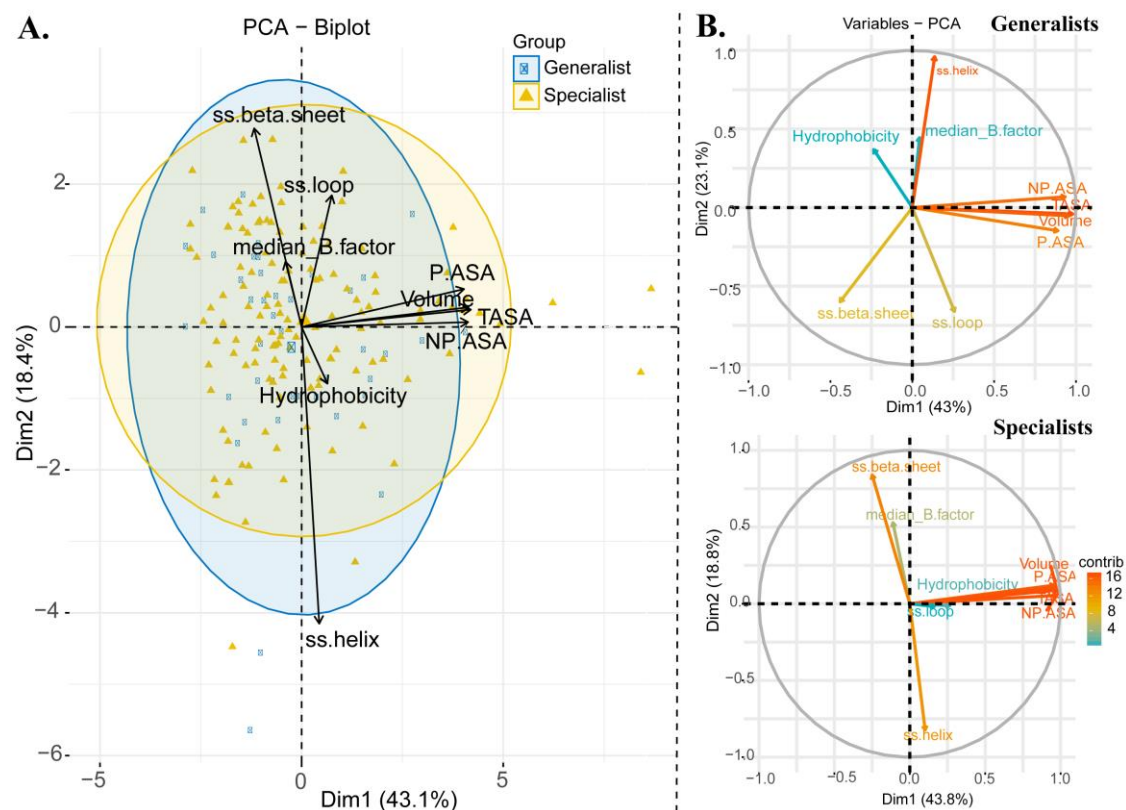


Figure 2.8 Biplots for PCA analysis. Biplot showing A) individuals (enzymes) of the eco-gen-spec dataset and various active site features (variables) analyzed. B) Biplot showing the variable correlation plots of active site properties after principal component analysis (PCA) for a) generalist and b) specialist.

As shown in Figure 2.4B, total ASA, non-polar ASA, polar ASA and volume of active site are correlated in both generalist and specialist. Hydrophobicity and loop are negatively correlated in generalist but positively correlated in specialist. Higher distance from origin indicates that, the contribution of hydrophobicity and loop content in accounting for the variability in PC is more in case of generalist compared to specialist. Further, α -helices content of active site residues and B-factor are positively correlated in generalist with β -sheet content negatively correlated to them in contrast to positively correlated B-factor and β -sheet content of the active site residues in specialists which in turn correlated negatively to α -helices content. These observations suggest that B-factor does have a distinct contribution in controlling the substrate specificity at least in case of *E. coli* metabolic enzymes. Further, hydrophobicity and loop content also seem to play a significant role in influencing promiscuity.

2.2.3 Conclusions

The present analysis suggests that there is no general set of features to distinguish between generalist and specialist. We observed that only hydrophobicity of binding sites is slightly higher in case of generalist suggesting promiscuous enzymes tends to harbor more hydrophobic residues compared to specialist. This indicates that specialist may use more sets of specific interactions such as hydrogen bonds for the recognition of its cognate substrates. Importantly, from PCA analysis, we found that the hydrophobicity is important to demarcate generalist from specialist. It is quite likely that many of structural features combine to provide appropriate local environment in promiscuous enzymes. PCA analysis showed that it is possible to combine these features using methods such as Support Vector Machine (SVM) to develop promiscuous enzyme predictor tool with availability of large training datasets.

The limitation of the present work is the size of datasets and variability of enzymes from various source organisms (we relied only on *E. coli* metabolic enzymes). However, advantage of using enzymes from same (source) organism is that enzymes have roughly similar evolutionary time scale. The difficulty in generation of dataset arises from the limited knowledge on promiscuous enzymes as the experimental endeavors to identify and to characterize them is very tedious and expensive process. With the availability of high throughput techniques, these are going to improve in coming years. A more robust analysis can then be performed to identify structural aspects of generalist.

2.3 Investigation into global/local structural in enzymes upon substrate binding

As stated earlier in section 2.1, we formulated two hypotheses in this chapter: a. substrate/cofactor binding sites have attributes to distinguish promiscuous (generalist) and specialist enzymes, and b. catalytic sites of promiscuous enzymes are pre-primed to catalyze enzymatic reactions. After testing our first hypothesis in previous section, this section of the chapter describes analysis of the second hypothesis. In this work, first we have analyzed global/local changes in binding/catalytic residues of enzymes upon

substrate binding using various measures. Subsequently, we analyzed local changes of catalytic residues between promiscuous and non-promiscuous enzymes.

2.3.1 Background

Enzymes are dynamic molecules with inherent flexibility and their movements are often essential for catalysis (Hammes 2002; Yon, Perahia, and Ghéllis 1998). Many enzymes harbor flexible loop regions, which enable correct positioning of catalytic residues. The classical example of this case is Triosephosphate isomerase, which is characterized by prominent closure of the phosphate gripper loop 6 over the ligand phosphodianion group, upon binding of substrate DHAP (Jogl et al. 2003) or intermediate analogs such as 2-phosphoglycolate (PGA) (Lolis and Petsko 1990) and 2-phosphoglycolohydroxamate (PGH) (Davenport et al. 1991). Although such dynamic behavior is prevalent in enzymes, the extent to this dynamic nature is variable from enzyme to enzyme. While some enzymes undergo limited or no conformational change through local fluctuations in side chains, there are enzymes which undergo large conformational change involving whole domain motions (Villali and Kern 2010). Even though we are aware of such a conformationally dynamic nature of enzymes, till date inclusion of the features to account for flexibility in function prediction or enzyme-substrate docking is still not a trivial task. Currently many structural template-based methods are available for the prediction of function of enzymes (Whisstock and Lesk 2003; Barker and Thornton 2003). However, if the enzyme undergoes significant conformational change upon substrate binding or catalysis, such template-based methods may fail to identify the correct function of the enzyme. Knowledge of ligand induced conformational changes in enzymes, can be used to enrich the conformational space of the template structures and hence increase the chances of correct prediction especially in enzymes undergoing large conformational change. Homology modelling is currently one of the popular tools used in drug designing (Wieman et al. 2004; Franca 2015). Accounting for enzyme (target) conformational flexibility in homology modelling will lead to further help in designing more specific and efficient drugs and less clinical failures. In fact, recently, the conformational targeting of proteins is adapted for neurodegenerative diseases (Krishnan et al. 2017). Further, many docking methods allow full flexibility of ligand/substrate but treat the receptor molecule either rigid or allow motion of very limited number of residues. Incorporation of protein flexibility in docking algorithms is shown to improve ligand discovery (M. Fischer et al.

2014). Thus, understanding the conformational changes occurring in enzymes upon ligand binding will help in developing such flexible docking methods with more accuracy and success rate. Further, knowing the extent of ligand-induced conformational changes will help in understanding the mechanism and motion, which facilitate catalysis and can further be used in enzyme-ligand engineering.

The extent of conformational changes upon substrate binding in proteins has been explored before in many studies in various contexts. In general, the only a small number of residues in binding pocket undergo change upon ligand binding and it is observed that polar residues are more flexible compared to aromatic residues (Najmanovich et al. 2000). In another study, analysis of bound and unbound forms of 98 proteins was performed into to tease out the distinguishing sequence and structural features in rigid and flexible binding site (Gunasekaran and Nussinov 2007). This study showed high preference of polar-polar residue pair interactions and hydrogen bonding interactions in proteins that do not undergo conformational change compared to proteins, which undergo large conformational change where aromatic-aromatic, hydrophobic-hydrophobic and hydrophobic-polar interactions the binding site were preferred. Previous study has shown that most of the proteins undergo relatively small conformational rearrangements in tertiary structure upon substrate binding (Brylinski and Skolnick 2008). It was shown that there is a clear difference in the extent of structural change observed upon substrate binding in case of single and multi-domain proteins. While in case of single-domain proteins, the mean global RMSD between bound and unbound state was found to be $< 1\text{\AA}$, multi-domain proteins showed global RMSD was $> 1\text{\AA}$. Most of these high RMSD cases were associated with large-scale hinge-bending movements of entire domains. For 839 non-redundant set of proteins, another study annotated and classified the structural changes upon ligand binding into seven categories based on location of motion (Amemiya et al. 2011). Protein Structural Change DataBase (PSCDB) enlists the pdbentries in each of these seven categories of ligand-induced protein structural change (Amemiya et al. 2012). It should be noted that although, this study focused more on the proteins showing significant ligand-induced motion, large number of protein (~37% or 311/839) in their dataset showed no significant motion upon ligand binding as well. Above mentioned studies were based on proteins and did not distinguish between enzymes and non-enzymes.

Gutteridge and Thornton reviewed conformational changes in enzymes during each part of catalytic reaction in a small set of 11 enzymes for which crystal structures were known for the apo, substrate-bound and product-bound forms (Gutteridge and Thornton 2004). Interestingly, in this study, they observed that most conformational change occurred during the substrate binding and product release steps of catalysis rather than during the transformation of substrate into product. Later they analyzed ligand-induced conformational changes in the active site of a larger dataset with 60 different enzymes and observed that most of the enzymes undergo less than 1 Å RMSD between the apo and substrate-bound forms across the whole protein (Gutteridge and Thornton 2005). While the extent of side-chain flexibility in binding and catalytic was similar, they also observed significant differences in the motion of binding site and catalytic site residues with backbone flexibility only shown by binding residues.

In the present study, we revisited this question of ligand-induced conformational changes in enzymes with more enriched dataset and analyzing the dataset in more details. In this study, first we investigated if there are any statistically significant differences in enzymes and non-enzymes observed in extent of conformational changes upon substrate binding. Thereafter, we particularly focus on structural changes observed in enzymes. We investigated local structural changes in the binding and catalytic site in more details especially in terms of the side-chain conformation of the binding/catalytic residue. Here we also study the extent conformational change separately in catalytic and binding site to explore if the two sets of residues undergo conformational changes to different extents using three different quantitative metrics. Subsequently, we categorized the enzymes into generalist and specialist and analyzed the differences in the extent of structural changes observed in them upon ligand binding. In our dataset, we have considered only one cognate substrate/cofactor bound to enzyme and also ensured that both bound and unbound structures are for wild type sequences in order to alleviate the effect of putative mutation on structure as well as to avoid comparison of mutant and wild type structure.

2.3.2 Methods

As mentioned before, in order to study ligand induced conformational changes in proteins, we have constructed dataset of protein tertiary structure pairs wherein one structure has no ligand bound (unbound or apo) and other is bound (holo) to only one ligand. For holo structure, we have considered only one ligand bound structure in order

to delineate effect of bound ligand and not compound effect of any other ligand. Here, ligands are defined as chemical compounds, which have a minimum of 6 heavy atoms.

2.3.2.1 Construction of enzyme bound-unbound (Enz-BUB) dataset

In order to construct enzyme bound-unbound (Enz-BUB) dataset, we took bound structure form of enzymes, which are either bound to cognate substrate/cofactor or their close analogs defined as having at least 80% similarity to enzyme cognate substrate/cofactor. For this, we have relied on EC-PDB database (<https://www.ebi.ac.uk/thornton-srv/databases/enzymes/>), which is comprehensive compilation of known enzyme structures with EC numbers and has mapping of bound ligands to the structure with substrate or cofactor. For every ligand bound to an enzyme structure in EC-PDB, it is assigned a similarity score its cognate substrate/cofactor and this score varies from 0 to 100 (exact)% where 0 is no match and 100 is identical to substrate/cofactor.

The overview of Enz-BUB dataset construction is outlined in Figure 2.9. Briefly, we parsed all the structures from EC-PDB and aligned their atom record sequences to full length sequence (Uniprot) using locally written script, which used Needleman-Wunsch algorithm (Needleman and Wunsch 1970) for sequence alignment. Any pdb structure having less than 95% sequence identity was removed from the list. Subsequent to this, the structures were renumbered with Uniprot sequence numbers based on the sequence alignment obtained in the previous step. This constituted ec-pdb-seq dataset was used subsequently to select bound and unbound representatives for enzymes. Next, we computed ligand binding sites for each ligand bound structure in ec-pdb-seq dataset using LPC program (Sobolev et al. 1999) to identify substrate/cofactor enzyme bound structures. Here, a holo structure is defined as having at least three interacting residues with the ligand having a minimum of 6 heavy atoms. Further, we also mapped the catalytic site for these enzymes using MACiE (Holliday, Almonacid, Bartlett, et al. 2007) or CSA-Literature (Furnham et al. 2014) database.

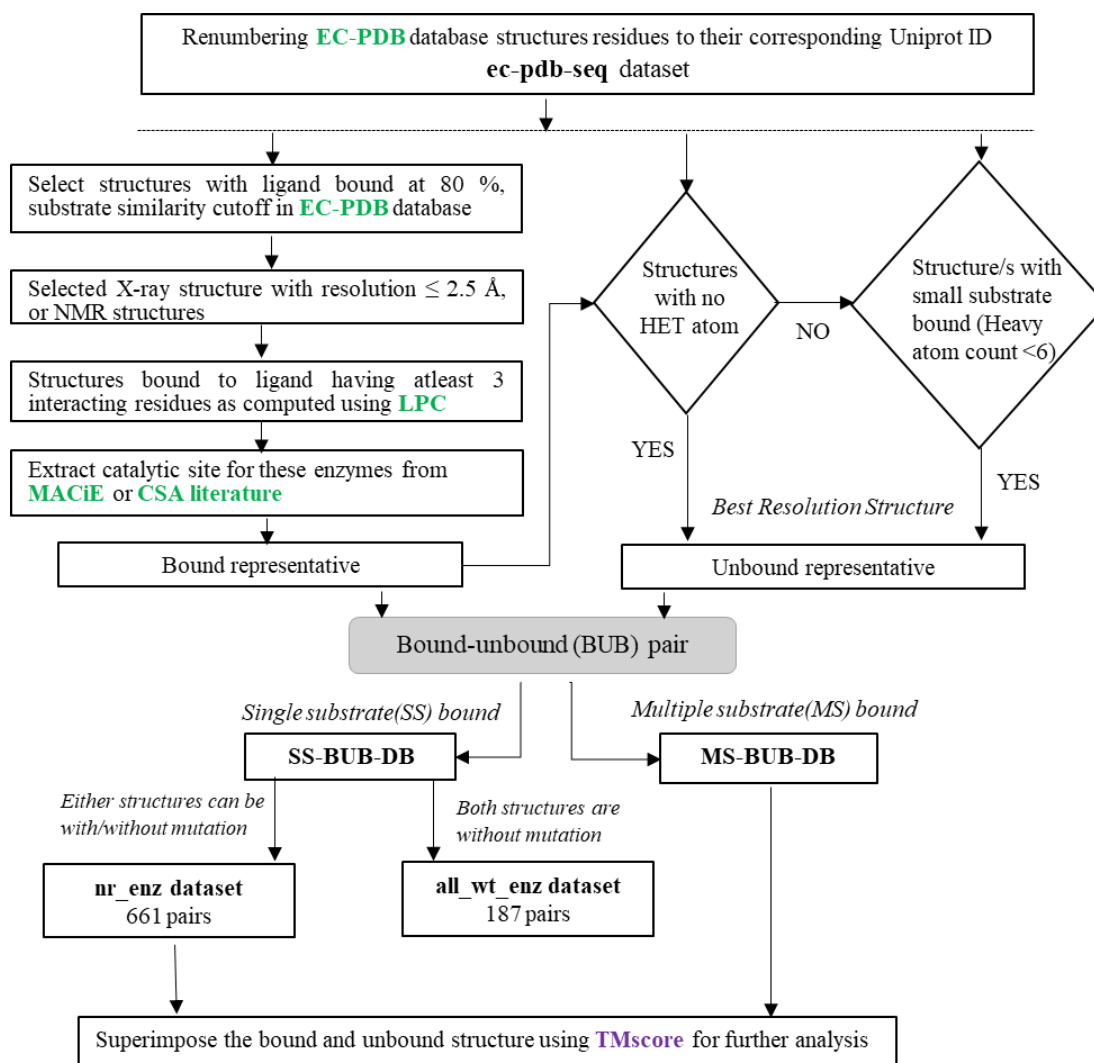


Figure 2.9 Overview of Enz-BUB dataset generation. Flowchart for i) single substrate bound (SS-BUB-DB) and ii) multiple substrate bound (MS-BUB-DB) are shown.

These mapped structures constitute holo (bound) enzyme structures. We identified unbound form of these holo enzyme structures using Uniprot identifier of bound form and searching in ec-pdb-seq dataset (also has Uniprot identifiers). The ligand unbound form (apo) structures could be either have no ligand bound to it or has a ligand having heavy atom less than 6 (based on our definition is not a ligand). In selecting apo form, the former is preferred, however, in absence of any such available structure to pair with holo form the latter is used. Subsequently, pairs of holo and apo structures are prepared. In case, there is more than one structure of a given sequence and are bound to same ligand, the structure with the best resolution is selected as representative. The same is followed for unbound structure. Following this procedure, we obtained pairs in holo and apo representative enzyme structures. Further, on the basis number of ligand bounds, we

categorized each bound-unbound pair into single –substrate bound dataset (SS-BUB-DB) and multiple-substrate bound dataset (MS-BUB-DB). These datasets are made non-redundant at 60% sequence identity using CD-HIT (W. Li and Godzik 2006) with respect to their bound ligand. Thus, we obtained 661 Enz-BUB or nr-enz dataset. This dataset is provided in supplementary material (*c2.2_s2_bub_dataset.xlsx*) for chapter 2 provided in a CD along with this thesis.

While preparing Enz-BUB, we did not consider mutation/s in either holo/apo form of structure. The limitation in such a dataset is that if we have pair of wild type and mutant enzyme structures it will be difficult to segregate conformational changes in enzyme structure because of ligand binding or mutational effects. This led us to construct holo-apo enzyme pairs having only wild type enzymes. This resulted in non-redundant dataset consisting of 187 pairs of enzymes that we refer to as all-wt-enz dataset.

For a given dataset, we aligned the bound and the unbound structure of each BUB pair using TMscore program and used these superimposed structures for further analysis.

2.3.2.2 Construction of non-enzymes bound unbound (nEnz-BUB) dataset

We extracted structures bound to biologically relevant ligands from MOAD database (Ahmed et al. 2015; Hu et al. 2005). We pruned the list of structures to remove enzymes and filtered 5545 non-enzymes. Subsequently, we identified ligand bound structure using same definition of ligand and its interaction with protein used in previous dataset generation. This provides dataset of holo structures. Next we mapped their corresponding unbound structures and collated a list of total 1,219 non-enzyme bound unbound pairs. Following the enz-BUB dataset generation procedure, we constructed non-redundant dataset of 780 non-Enzyme bound-unbound structures. This dataset is provided in supplementary material (*c2.2_s2_bub_dataset.xlsx*) for chapter 2 provided in a CD along with this thesis.

In order to analyses extent of conformational change upon substrate/cofactor binding in generalist versus specialist, we have constructed two datasets *viz.* Generalist – Specialist (GS-1) and GS-2. These two datasets are subset of all-wt-enz dataset.

2.3.2.3 Construction of GS-1 and GS-2 dataset

The construction of GS-1 dataset is straightforward. We extracted subset of generalist and specialist enzymes as defined in previous work (section 2.2) from the all-wt-enz dataset. We could identify bound-unbound pairs of enzymes for 7 generalists and 34 specialists. This dataset is referred as GS-1 dataset. This dataset is provided in supplementary material (*c2.2_s2_bub_dataset.xlsx*) for chapter 2 provided in a CD along with this thesis.

Since we did not have large numbers of generalist/specialist enzymes (GS-1 dataset) for analysis, we used BRENDA (Chang et al. 2015) database to expand list of promiscuous and non-promiscuous enzymes. For this first we take list of enzymes from all-wt-enz dataset that is also listed in BRENDA database. Then, according to the number of reactions catalyzed by a given enzyme, it is classified as generalist and specialist, where any enzyme involved in more than one distinct reaction are classified as promiscuous enzymes. This procedure resulted in 45 generalists and 75 specialists.

2.3.2.4 Quantitative metrics to measure ligand induced conformational change

Measuring global structural changes upon substrate binding

We have calculated C α Root Mean square deviation (RMSD) between holo and apo enzymes structures to analyze ligand induced conformational change. The bound and unbound structures are superposed using TM-score (Yang Zhang and Skolnick 2004) program to obtain global RMSD (gRMSD) between these two structures. We also calculate all-atom RMSD using C α superposed coordinates.

Measuring local changes in catalytic/binding site upon substrate binding

To measure local changes, we have used TM-score program to optimally superposed C α coordinates of catalytic/substrate binding residues of holo and apo structures and calculate local RMSD (lRMSD). The superposed coordinates can be used to compute all-atom RMSD. Any catalytic/substrate binding site has less than 3 residues are not used for superposition.

Since IRMSD gives deviation for C α atoms and all-atoms RMSD gives average deviation across all atoms types, such measures may not capture small changes involving residue side-chain of 1-2 residues. Moreover, at least three residues are required in order to align and subsequently calculate RMSD. To capture the small variations between holo and apo form of the structures, we have calculated three different parameters schematically shown in Figure 2.10 to quantify side-chain conformational changes upon ligand binding.

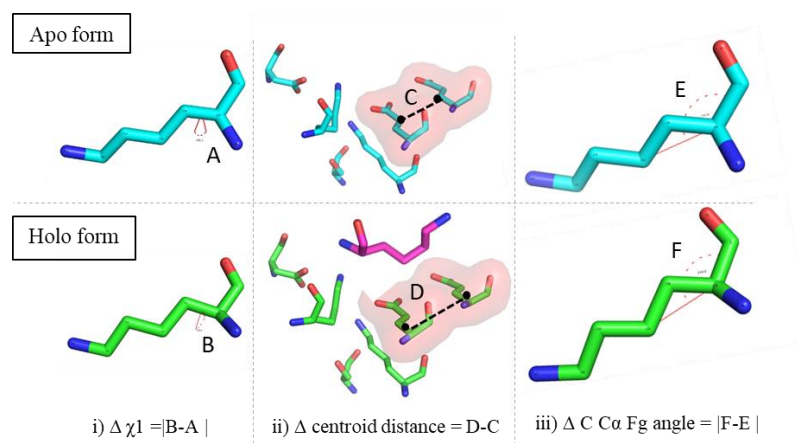


Figure 2.10 Schematic representation of the three metrics used to measure the change in side chain conformation of catalytic/binding residue. Green and Cyan colors represent bound and unbound forms respectively.

These metrics are as follows:

- I. Change in side chain torsional angle ($\Delta \chi_1$) between bound and unbound structures of enzyme.
- II. We computed difference of the angle C-C α -Fg between bound and unbound forms of enzyme. The angle is defined as angle between vectors of backbone carbonyl C-C α and C α -Fg where Fg is functional group, using definition of Gutteridge and Thornton (Gutteridge and Thornton 2005)
- III. Change in centroid distances calculated for catalytic/binding site residues between bound and unbound state of enzyme.

Quantifying structural changes at remote sites (sites distant from catalytic/binding site) upon substrate binding

Apart from conformational changes of catalytic/binding site residues, which are proximal to substrate/cofactor binding sites, we also analyzed any structural effects on residues located distantly from ligand binding sites.

In this analysis, we have constructed protein side-chain network (PScN) of an enzyme in its apo and holo form separately using PSN-Ensemble program (Bhattacharyya, Bhat, and Vishveshwara 2013). In PScN, each amino acid is represented as a node and the two nodes are connected to each other by an edge based on the strength of the non-covalent interaction (I_{ij}) between their side-chains atoms. I_{ij} between the two nodes (amino acid side chains) is calculated in percentage using the following equation:

$$I_{ij} = \left(\frac{n_{ij}}{\sqrt{N_i \times N_j}} \right) \times 100$$

where, n_{ij} is the number of distinct atom pairs between the side chains of amino acid residues i and j , which come within a distance of 4.5 Å, and N_i and N_j are the normalization factors for residue types i and j as defined by (Brinda and Vishveshwara 2005). Further, based on the user-defined cutoff value, I_{\min} , any two residue pair ij are connected if the $I_{ij} > I_{\min}$. Further it has been shown that PScN exhibit complex topological network behavior where the size of the largest cluster (total number of amino-acids in a cluster) is a function of the interaction strength cutoff. It was found that in majority of the proteins, irrespective of their globular fold or their size/length, the size of the largest cluster drops drastically at a certain I_{\min} value and is termed I_{critical} . In this study, we used I_{critical} of a given protein as a measure of the strength of non-covalent interaction (I_{\min}). I_{critical} is defined as I_{\min} at which size of the largest cluster is half the size of I_{\min} at 0% (Brinda and Vishveshwara 2005). We analyzed side-chain network property to find differences between two states (apo and holo) of enzyme structures. For this, we calculated change in the degree (Δ degree) of the equivalent residue between holo and apo structures and took this as a measure of structural change upon ligand binding (Figure 2.11). The degree of a node is defined as number of connections with other nodes *i.e.* number of contacting residues of a given residue. To analyze change in degree of residues as a function of distance from the binding site, we constructed concentric spherical shells of radii incremented by 5 Å for every sphere and the center of sphere is always centroid of binding sites (Figure 2.11C). Thus, first sphere will have all the residues within 5 Å from the centroid of the binding site, next sphere will have all the residues within 5-10 Å and so on. In this manner, a given residue will be a part of only one sphere and the sphere to which it belongs will indicate its distance from the binding site. Higher the sphere number a residue belongs, greater is its distance from the binding site.

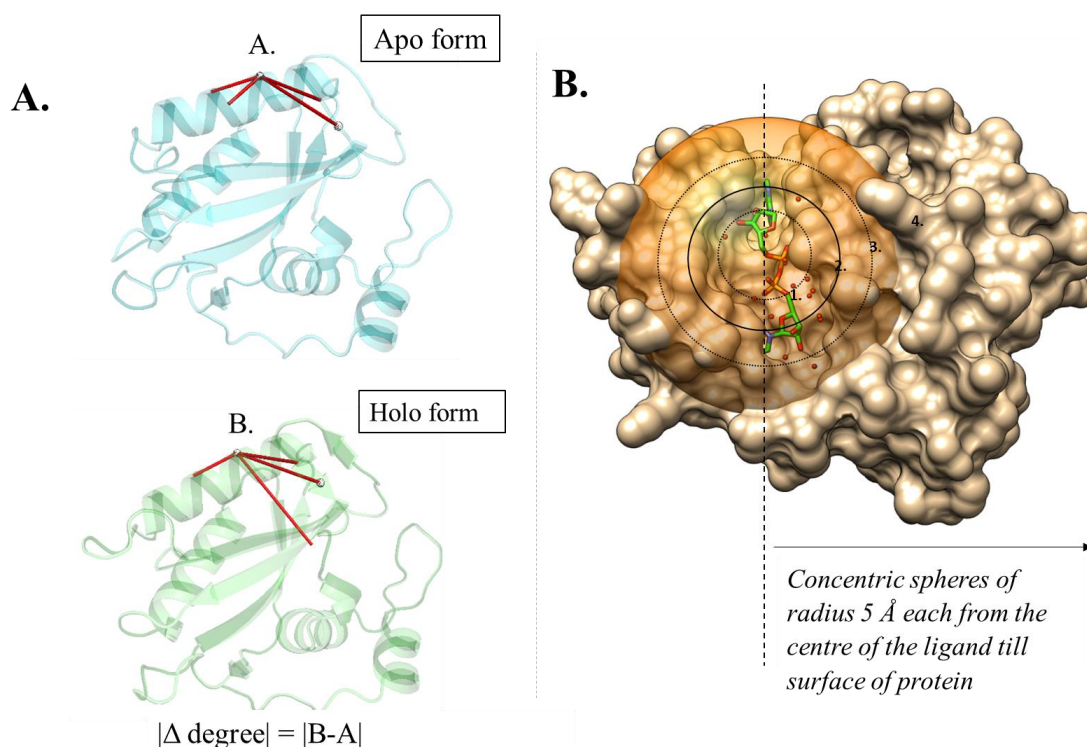


Figure 2.11 Schematic representation of the metric used to measure the structural change in remote sites catalytic/binding residue. A) Change in the degree of equivalent residue distant from catalytic site. B) Concentric spherical shells of radii incremented by 5 Å for every sphere with the center of sphere as centroid of binding sites in order to investigate the change in the degree from the sites distant from binding site.

Further, we also analyzed cases where degree remains unchanged while underlying nature of side-chain interaction undergoes a large change. For this, first contacts for a given residue that is defined as any heavy atom within 4.5 Å of it is calculated. Then, we find residues with no change in degree and zero overlap of contacting residues between apo and holo form of the enzyme. Essentially, such residues have changed their contacting residues while maintaining the degree of the residue (node). Thus, we analyzed at the fraction of overlap in contacts of each equivalent residue in apo and holo form.

2.3.3 Results

Previous studies on ligand induced conformational changes have shown that proteins undergo relatively small conformational change upon ligand binding (Brylinski and Skolnick 2008; Gutteridge and Thornton 2005). In the initial part of the work, we have also performed similar analysis but asked whether enzymes or non-enzymes show

more conformational change with respect to the other. Subsequently, we moved from global structural changes to local structural changes and restricted analysis to measure minute local changes and understand their role in enzymatic function.

2.3.3.1 Comparison of ligand induced conformational change between enzymes and non-enzymes

We compared global as well as local structural changes between enzymes and non-enzymes upon ligand binding to understand whether conformational changes are predominant in either of two groups. The global structural change is assessed using gRMSD calculated between apo and holo form of proteins (see Methodology section 2.3.2). On an average, there is not much gRMSD difference between enzymes and non-enzymes (mean (sd)/median gRMSD = 1.12 (1.64)/0.65 and 1.3 (1.57)/0.7 respectively). As is also evident from Figure 2.12 that there is not much difference in the distribution of gRMSD of Enz-BUB and nEnz-BUB groups. Moreover, the extent of conformational change in tertiary structure of protein is similar in both categories. Thus, for both enzymes and non-enzymes average ligand induced conformation changes are ~ 1 Å. During our study, we observed that there are structures of mutant proteins in our dataset. Such cases are difficult to study, as the observed conformational change upon substrate binding can either be a result of mutation or substrate binding. Such cases cannot be handled in an automated manner and need to be dealt manually. To overcome this issue of nr-enz dataset, we constructed all-wt-enz dataset as mentioned in methods section. Even though the number of enzymes is reduced from 661 into 187 in all-wt-enz, it is worth to look into more detailed manner, as it is more pristine than nr-enz dataset. As seen in Figure 2.12A and 2.13A, there was not much of change in the overall distribution of gRMSD in nr-enz and all-wt-enz dataset. Thus, we have shown results only for all-wt-enz dataset from here onwards until mentioned otherwise.

Next, we evaluated influence of domains on the structural changes in enzymes. As seen in Figure 2.13B the mean (standard deviation) gRMSD is 0.95 (1.36) Å and median gRMSD is 0.71 Å. There was only one enzyme structure having 4 domains, having gRMSD of 2 Å. It should also be noted that usually the cases where enzymes have higher gRMSD, were bound to the inhibitors/ analogs/covalent intermediates. It should be noted

that this observation for enzymes and non-enzymes is similar to the previously observed ligand induced conformational change in proteins (Brylinski and Skolnick 2008).

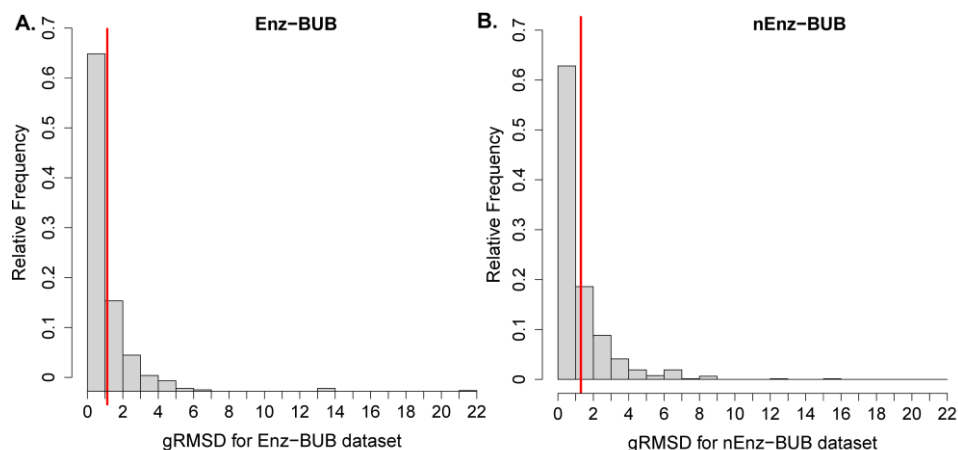


Figure 2.12 Histogram showing the distribution of $C\alpha$ gRMSD between bound and unbound forms of A) Enzymes B) Non-enzymes using *Enz-BUB* dataset and *nEnz-BUB* dataset respectively.

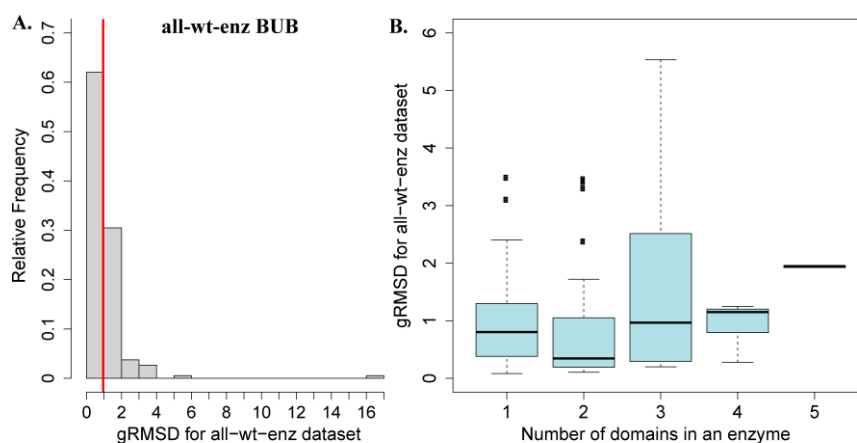


Figure 2.13 $C\alpha$ gRMSD between bound and unbound forms of all-wt-enz dataset. A) Histogram showing the distribution of $C\alpha$ gRMSD between bound and unbound forms of all-wt-enz dataset B) Box plot showing the distribution of domain-wise $C\alpha$ gRMSD between bound and unbound forms of all-wt-enz dataset.

2.3.3.2 Conformational variability of binding/catalytic site residue of enzymes upon substrate binding

Next, we performed detailed analysis of local structural of residues responsible for catalysis or lying at binding site/s to understand conformation variability of these residues.

Local structural changes occurring in the enzymes upon ligand binding

We investigated the local structural changes specifically at enzyme's catalytic site and binding site upon substrate binding. In general, binding site and catalytic site overlap with each other (Cilia and Passerini 2010). Despite this large overlap, catalytic and binding residues have their own role during catalysis. While catalytic residues are directly involved in catalysis, binding residues usually assist binding of substrates and orientating the substrate to facilitate reaction. First, we computed the local all-heavy atom RMSD between bound and unbound state of a) catalytic and b) binding site residues after local alignment of only catalytic/binding residues. A high RMSD values suggest high deviation/difference between the two structures. The distribution of local all-heavy atom RMSD (IRMSD) between bound and unbound forms of all-wt-enz dataset for catalytic/binding site residues is shown in Figure 2.14.

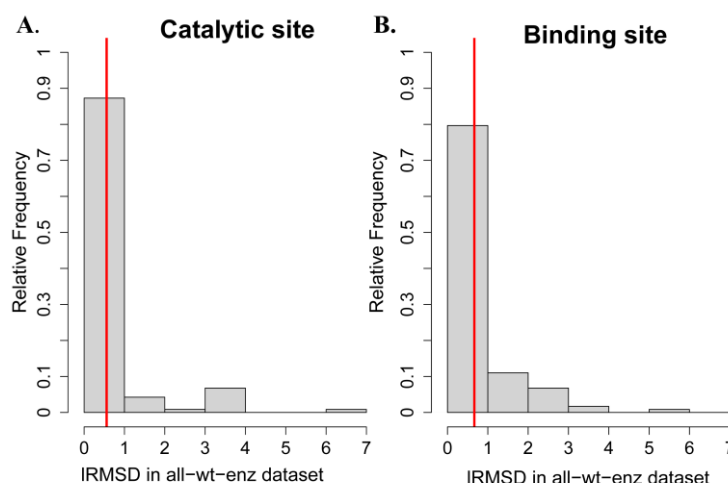


Figure 2.14 Histogram showing the distribution of local C α atom local RMSD (C α -IRMSD) between bound and unbound forms of all-wt-enz dataset for A) catalytic site residues B) binding site residues.

In our dataset, the local conformational changes are more apparent in the residue side-chain conformation of structure of the enzyme upon substrate binding compared to main-chain (C α atoms) conformation for binding/catalytic residues as indicated by mean (standard deviation) C α -IRMSD of 0.67 (0.85) / 0.56 (1.07) Å or median C α -IRMSD of 0.33/0.19 Å compared to mean (standard deviation) all-IRMSD \sim 1.23(1.03)/1.50(2.23) Å or median all-IRMSD \sim 0.97/0.75 Å), with few proteins showing even >6 Å all-IRMSD. As shown in Figure 2.14B, most of the residues either catalytic or binding have all-IRMSD within 1-2 Å. However, there are few proteins where catalytic residues show

higher all-IRMSD ($> 6 \text{ \AA}$). Thus, even though they do not occur in majority, there are enzymes, which show large conformational change in catalytic residues as well. Some of these cases are discussed later when we discuss the detailed changes in the side chain conformation of catalytic/binding residues.

Side-chain conformational variability of catalytic/binding residue upon substrate binding

Next, we investigated extent of conformational change in side-chain of binding/catalytic residues between apo and holo enzymes. As mentioned in Methodology section 2.3.2, we assessed this using following three properties: 1) $\Delta\chi_1$ angle, 2) $\Delta C-C\alpha-Fg$ angle and 3) Δ centroid distances between centroids of all pairwise catalytic/binding residues. Here, Δ is obtained by subtracting unbound (initial) state from bound (final) state values. Below we discuss changes observed in these parameters:

1. Change in χ_1 angle of catalytic/binding residue between apo and holo enzymes

The side-chain torsion angle χ_1 is defined as dihedral angle between atoms N-C α -C β -C γ /C γ 1/O γ /O γ 1, where identity of fourth atom could be any of these depending on the amino acid. Due to steric hindrance between γ side-chain and backbone atoms, χ_1 angle has distribution restricted around ± 60 and ± 180 . We calculated $\Delta\chi_1$ of catalytic as well ligand binding residues (see methodology). The relative cumulative distribution of $\Delta\chi_1$ for all catalytic and binding site residues across proteins is shown in Figure 2.15A. As shown in inset of Figure, there are 3 peaks, corresponding 0, ± 120 and ± 240 . As is evident from Figure, most (68/71%) of binding/catalytic site residues have change in torsion angle within 15° . In general, comparatively larger changes are observed mostly for catalytic site residues in comparison to binding residue. About 80/83% of binding/catalytic residues show changes $\sim 100^\circ$.

An example of large change in torsion angle is *E. coli* dihydrofolate reductase (ecDHFR). The detailed experimental and computational studies have suggested that during catalysis M20 loop oscillates between closed and occluded form, which involves conformational change in central portion of loop from β -sheet to 3_{10} helix as shown by pink color in Figure 2.15B. Residues M20 and L28 undergo major change in torsion angle $|\Delta\chi_1|$ of 65.39° and 48.35° respectively as seen in Figure 2.15C.

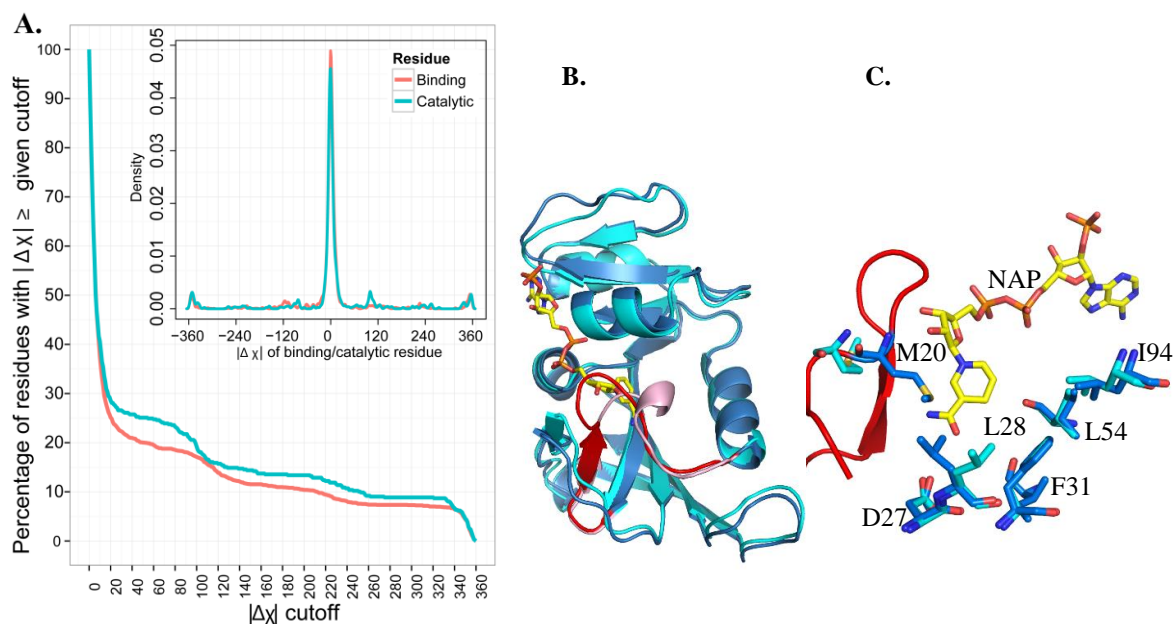


Figure 2.15 Change in side chain torsion angle $|\Delta\chi_1|$. A) Fraction of catalytic/binding residues with $|\Delta\chi_1| \geq$ given cutoff, with varying cutoff from 0° to 360° . Inset shows the density distribution of $|\Delta\chi_1|$ for catalytic/binding residues B) ecDHFR bound to NADP (pdbid:1rx1A) superimposed to its unbound state (5dfra). The closed conformation of M-loop shown in red color shields the reactants from the solvent. C) upon substrate binding, catalytic residue – M20 and L28 undergo major change $|\Delta\chi_1|$ of 65.39° and 48.35° respectively.

2. Change in C-C α -Fg angle of catalytic/binding residue between apo and holo forms

Usually, the functional group of a catalytic residue side-chain directly participates in a given enzymatic reaction. In order to understand the change in orientation of these functional groups upon substrate binding, we defined angle between functional group vector (C α -Fg) and backbone vector (C-C α) and analyzed change in this angle (see methods). The definition of functional groups for various amino acids (Table 2.2) were obtained from previous work (Porter, Bartlett, and Thornton 2004). The Δ C-C α -Fg angle is computed as difference between unbound to bound form (see section 2.3.2.4). The distribution of Δ C-C α -Fg for both catalytic and substrate binding residues is summarized in Figure 2.16. As is evident in Figure, for both catalytic and binding site residues, most residues ($\sim 80\%$ residues with $< 10^\circ$ Δ C-C α -Fg angle) do not show large change in functional group angle. However, rest 20% of residues shows change as large as 100° . One such example is Y335 residue of SET7/9 histone methyltransferase enzyme (HMTase), which has Δ C-C α -Fg of 65.95° (Figure 2.17). SET7/9 is involved transfer of

methyl group from S-adenosyl-L-methionine (AdoMet) to K4 of Histone H3. The active site of SET forms a knot-like structure involving β strands 19,20 and 22 (Figure 2.17). The structural change of residue Y335 is only major change between apo (1mufA) and holo (1n6a) form the enzyme (Jacobs et al. 2002). The motion of Y335 enables AdoMet to adopt compact conformation, which fits into narrow binding cavity.

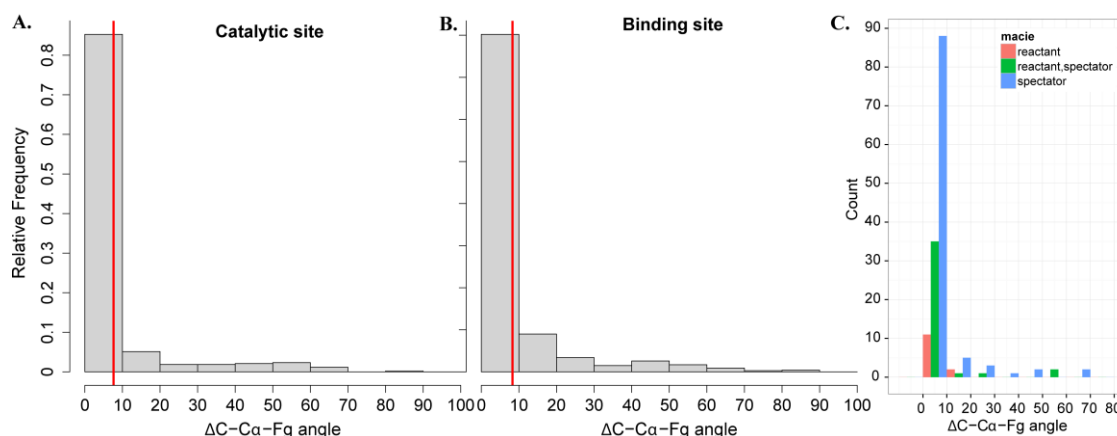


Figure 2.16 Histogram showing the distribution of $\Delta C-C\alpha-Fg$ angle upon ligand binding in A) catalytic site B) binding site C) shows the distribution of $\Delta C-C\alpha-Fg$ angle upon ligand binding only for those catalytic residue which have a functional role specified in MACiE database.

Table 2.2 List of amino-acid usually having functional atom located on the side-chain and their respective functional atom used to calculate $\Delta C-C\alpha-Fg$ angle

Amino acid	R	N	D	C	Q	D	H	K	S	T	W	Y
Functional atom	CZ	CG	CG	SG	CD	CD	NE	NZ	OG	OG	CZ	OH

In MACiE database, depending on the role of catalytic residues in a reaction, a residue is categorized into reactant, spectators, and reactant_spectators. Reactant catalytic residues are directly involved in chemical reaction and their structure is usually modified during the reaction such as residue involved in electron shuttling from substrate (Holliday et al. 2011). On the other hand, spectator catalytic residues are essential for reaction but are not actively involved in a reaction such as residues are involved in stabilization of reaction intermediates. Some catalytic residues can act as both reactant as well as spectator in different stages of a reaction and those are referred to as reactant_spectator.

To find out which of these classes of catalytic residues show unusual change in functional group angle, we analyzed distribution of $\Delta C-C\alpha-Fg$ shown in Figure 2.17C. This showed that reactant catalytic residues do not undergo much change ($<20^\circ$) in functional group angle. Spectator residues show large conformation changes that suggest these may get oriented subsequent to binding of substrate.

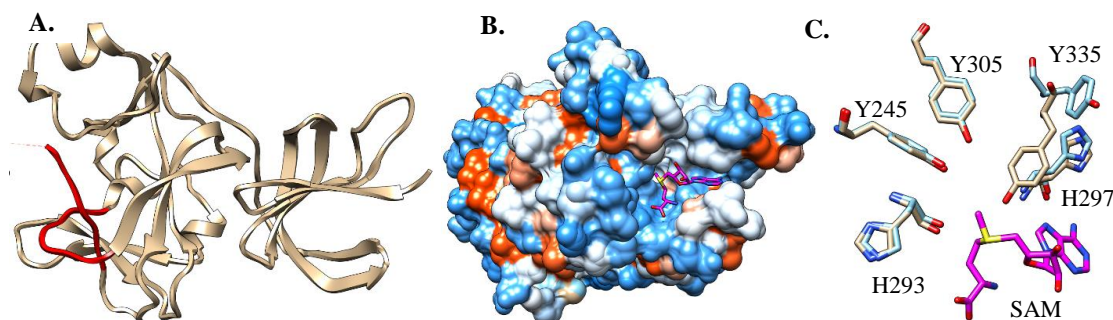


Figure 2.17 $\Delta C-C\alpha-Fg$ angle upon ligand binding in SET enzyme A) Cartoon representation of SET (pdbid 1n6aA) with knot-like structured active site shown in red color B) Adomet bound in compact conformation (shown by magenta licorice representation) bound in compact conformation to 1n6aA shown in surface representation colored by hydrophobicity. Here orange color indicates hydrophobic while blue represents polar residues C) Catalytic residues shown in superimposed bound 1n6aA (pale brown color) and unbound pdbid 1mufA (cyan color) form. The movement of catalytic residue Y335 restrain the adomet conformation.

Change in centroid distances of catalytic/binding residue between ligand bound and unbound form of enzymes

In order to estimate the change in mutual orientations of catalytic/binding site residues between apo and holo forms, for both forms we calculated all possible pairwise distances between centroid of catalytic/binding site residues and computed differences between equivalent residues. Here, large Δ centroid distance indicates that pair of residues has moved away from each other. The distribution of centroid distances for all possible pairs of catalytic/binding site residues is shown in Figure 2.18. The mean (standard deviation) centroid distance of catalytic residues and binding site residues are 0.57 (2.14) and 0.43 (1.67) respectively.

As has been observed before most residue pairs 85 do not show much change (within 1 Å) upon ligand binding. However, 5/6% of binding/catalytic residues show >3 Å difference in centroid distances. For instance, in enzyme isocitrate lyase (ICL) many

residue pairs such as C191:L348, D108:H193, 191C:T347 and Y93:H193 show changes varying between 13-20 Å between bound (1f8mA) and unbound form of enzyme (1f61A). The unbound form has active site loop harboring Cys191 far from the main catalytic site, which adopts closed conformation in the holo enzyme (1f8mA), which is bound to inhibitor 3-bromopyruvate, an is analogue of the substrate isocitrate and traps the active site making it inaccessible to the solvent (V. Sharma et al. 2000). Further, we analyzed any possible association between Δ centroid distances with reactant, spectator, or reactant_spectator role of catalytic residues as defined in MACiE. Figure 2.18C summarizes the distribution for each of these 3 categories of catalytic residues. As has been observed in functional group angle, spectators are mostly reoriented subsequent to ligand/substrate binding.

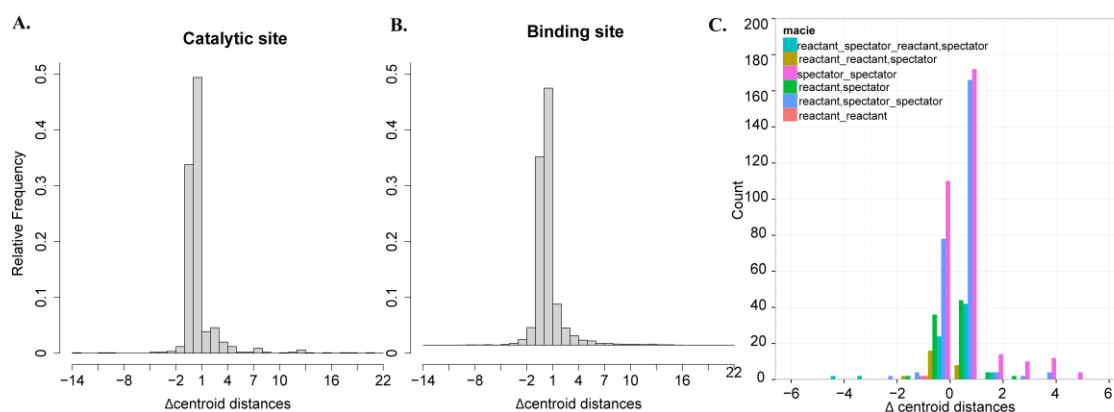


Figure 2.18 Histogram showing the distribution of Δ centroids distances of all-against all pairwise a) catalytic and b) binding residue in bound and unbound state, c) shows the distribution Δ centroid distances between the centroids of only those catalytic residues which have a functional role specified in MACiE database.

2.3.3.3 Generalist and specialist binding/catalytic site residue conformation variability upon substrate binding

The analysis of global/local conformational changes due to ligand binding showed that on average most enzymes have RMSD less than 1 Å. Moreover, local conformational changes of catalytic/binding site residues side-chain as assessed by differences in their centroid, functional group and χ_1 angles suggested that small number of residues (<20%) change their orientations upon ligand binding. Even though we observe marginal differences in local conformational changes, we compared such differences exists between generalist and specialist. Here, the basic presumption is that specialist enzymes would need re-orientation of catalytic residues to facilitate reaction in

comparison to generalist, which probably have functional residues pre-primed or oriented in a manner to catalyze reaction with alternate substrates. Based on this hypothesis, we have compared the ligand induced conformational changes between specialist and generalist using the previously described (Nam et al. 2012b) dataset GS-1 and BRENDA dataset GS-2. The comparison of global RMSD is summarized in Figure 2.19 shows specialist tends to have larger conformational change.

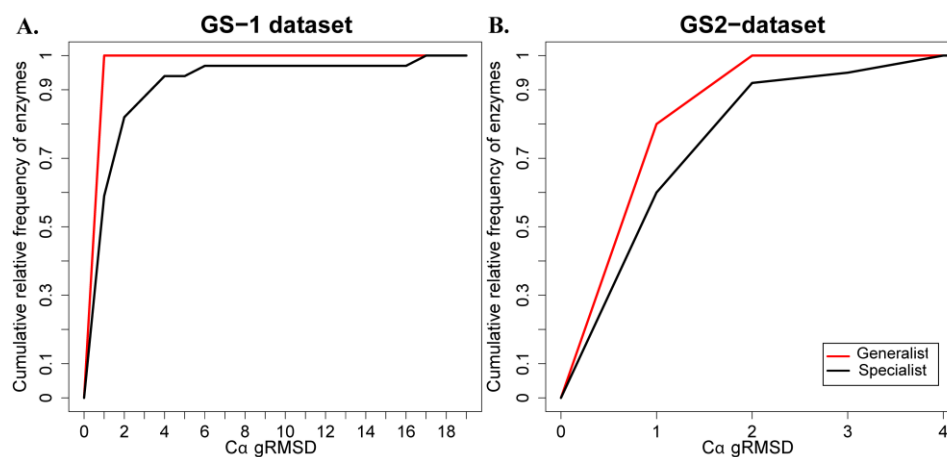


Figure 2.19 Cumulative relative frequency of enzymes with varying Ca-gRMSD in A) GS-1 dataset B) GS-2 dataset.

The comparison of local structural changes of catalytic and binding site residues as assessed by IRMSD is shown in Figure 2.20. As is illustrated in Figure, IRMSD of catalytic/binding site of specialist is slightly higher in comparison to generalist suggesting a greater local conformation change in specialist.

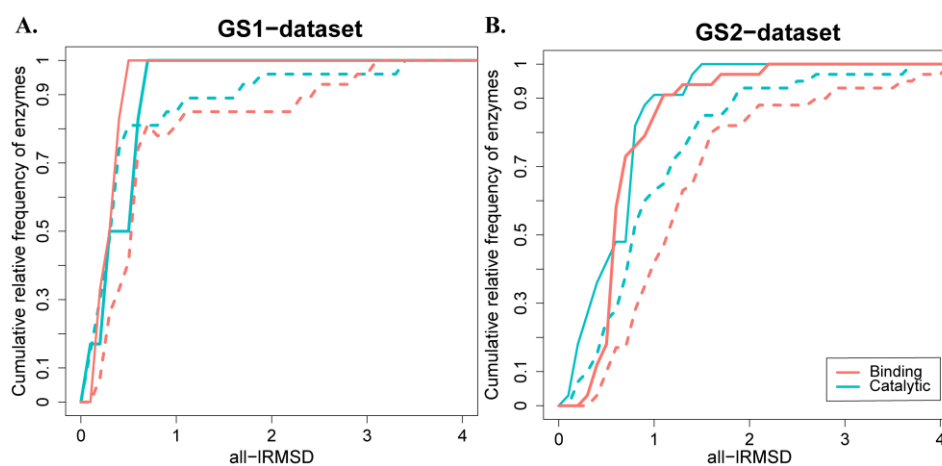


Figure 2.20 Cumulative relative frequency of enzymes with varying all-I-RMSD for binding/catalytic residues in A) GS-1 dataset B) GS-2 dataset. Here the solid lines represent generalists while dashed lines represent specialist enzymes.

The comparison of $\Delta C-C\alpha-Fg$ angle of catalytic/binding site residues between specialist and generalist is shown in Figure 2.21. Almost ~80-90% of binding/catalytic residues show a change of up to only 10° in $\Delta C-C\alpha-Fg$ angle. At any given cut-off, higher number of generalists show the given change in $\Delta C-C\alpha-Fg$ angle in their catalytic/binding residue comparison to specialist and thus reaches saturation before. Thus, generalist tends to have less the less change in catalytic/binding residues in comparison to specialists. For instance, 98% of binding residue in generalist shows a change of 40° in $\Delta C-C\alpha-Fg$ while in specialist the percentage of such binding residues is 94%. Thus, remaining 6% show large change in specialist in comparison to only 2% of residues in case of generalist.

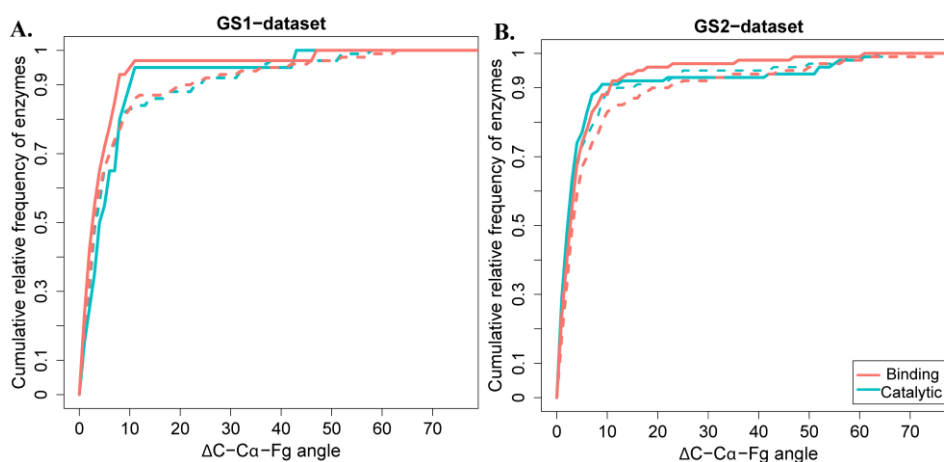


Figure 2.21 Cumulative relative frequency of binding/catalytic residues with varying $\Delta C-C\alpha-Fg$ angle for A) GS-1 dataset B) GS-2 dataset. Here the solid lines represent generalists while dashed lines represent specialist enzymes.

The comparison of $\Delta\chi_1$ of catalytic/binding site residues between specialist and generalist is shown in Figure 2.22. As shown in Figure 2.22, 90% of catalytic residues have $\Delta\chi_1 > 44^\circ$ in case of generalist whereas 96° in case of specialists. Thus, specialist show higher conformational variability compared to generalist. Further, in case of generalists, there is similar variation seen in $\Delta\chi_1$ angles for both binding and catalytic residues. In case of specialist, mostly variation is more in binding site residues compared to the catalytic site. At any given cut-off of $\Delta\chi_1$, the fraction of binding residues with $|\Delta\chi_1| > \text{given cutoff}$ are usually higher than catalytic residue. In generalist, only 6% of catalytic residues show $|\Delta\chi_1| > 120^\circ$ which in turn is contributed by catalytic residues from a small fraction of generalists (6%). In generalist, the catalytic residues of most of the generalists do not show much of a change in their $|\Delta\chi_1|$ as seen in Figure 2.21B. We analyzed some of the examples of specialist and generalist, which show significant

structural changes in their catalytic residues to further understand the conformational variability of catalytic residues in two classes of enzymes.

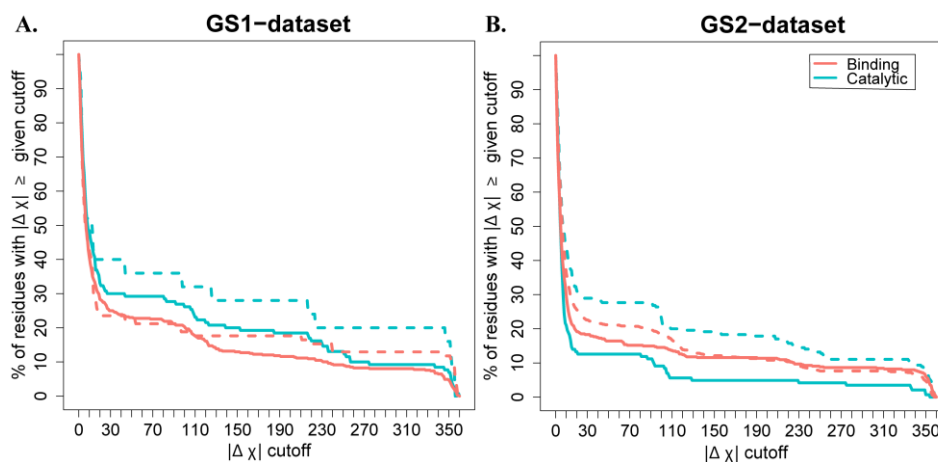


Figure 2.22 Cumulative frequency of fraction of catalytic/binding residues with varying cutoff from 0° to 360° of $|\Delta\chi| > \text{given cutoff}$ for A) GS-1 and B) GS-2 datasets. The solid lines represent generalists while dashed lines represent specialist enzymes.

We have discussed the case of specialist like ecDHFR earlier, where the M20 loop harboring catalytic residues oscillates in two conformations during catalysis. Here, we discuss a case of generalist, GDP-fucose synthetase (GFS) involved in biosynthesis of GDP-L-fucose. GFS of *E. coli* GFS is known to show catalytic promiscuity and catalyzes both epimerization and reduction of the substrate fucosyl transferase. As shown in Figure 2.23, it has small structural change (gRMSD of 0.267 and IRMSD of 0.182) upon binding to its substrate NADPH. The modelling and crystal structure studies have identified single binding site for GDP-sugar substrate, where both the reactions of epimerization and NADPH-dependent reduction occur. GFS has SER-TYR-LYS catalytic triad, which has been proposed to function in a mechanistically equivalent manner in both the reduction and epimerization reactions. Superimposition of substrate (NADPH) bound and product (NADP⁺) bound structures showed that while there is no change in the tertiary structure of protein, the conformation of substrate changes as shown in Figure 2.23. During catalysis, NADPH binds in syn-conformation forming a hydrogen bond with phosphoryl oxygen. This in turn breaks the hydrogen bond with the catalytic residue S107 and S108, with two water molecules entering into the site. This allows transfer of the pro-S hydride. However, the product (NADP⁺) binds in anti-conformation. Such a conformational difference in binding mode may lead to different

binding affinity of substrate vs. Product, thereby facilitating product release (Somers, Stahl, and Sullivan 1998).

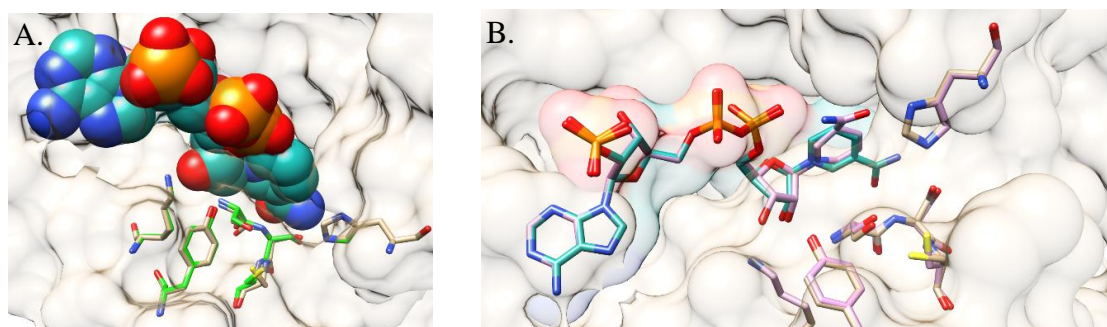


Figure 2.23 Comparison of bound and unbound states in EcGFS. A) Structural superimposition of EcGFS in bound (pdbid: 1fxsA) and unbound (pdbid: 1gfsA) states colored in pale brown and green color respective. The catalytic residue residues are shown in sticks representation, which doesn't show any drastic change upon substrate binding. B) Two different conformations of substrate (1bsv) and product (1fxs) bound states of EcGFS. Here the substrate NADPH bound in syn-conformation is shown in pink colored sticks representation, with doesn't form H-bond with catalytic S107 and S108. The product NADP+ bound in anti-conformation is shown in seagreen colored sticks representation.

Despite the limitations of numbers of known specialist/generalist and lack of statistically significant structural differences, the above analysis marginally provide hint towards the proposition that catalytic site residues might be primed to facilitate reaction in generalist enzymes. However, due to severe limitation on of the dataset, this cannot be generalized for all promiscuous enzymes as yet.

Conformation variability in the structure upon reactant between product bound structures

One of the limitations of investigating structural changes in enzyme due to ligand (substrate/product) binding is that it is not trivial to experimentally determine structures with all relevant bound substrates. Hence, tertiary structures in PDB dataset are usually bound to substrate/product/cofactor analogues or are bound to only one of the reactants.

In our analysis, we ignored a. whether structure is bound to substrate or product b. whether is single substrate or multi-substrate enzyme. These will probably influence local conformations in following manner. For cases belonging to type a., product bound structure may not show structural change as the reaction has already finished compared

to substrate bound structure as the initial stage of catalysis would need orientation of residues. For cases belonging to type b., enzymes may need all or some other substrate bound before catalytic residues will undergo structural change to catalyze reaction. In order to study both a. and b. cases, we analyzed each possible enzyme and categorized ligands (all-wt-enz dataset) into substrates and products using the EC-PDB database. Further, we identified set of enzyme structures for enzymes in all-wt-enz dataset, which are bound to one or more substrates. Next, we computed gRMSD for two datasets. Figure 2.24, shows difference in gRMSD of same enzyme bound to one or more substrates.

In general, structural variation in substrate bound structures is more with respect to product bound structures. As seen in Figure 2.24A, 90% of enzymes bound to product have gRMSD within 1.2 Å while in specialist this value is 78%. This indicates towards two key things. One, if the structural change doesn't happen in case of one substrate binding, this doesn't mean there won't be change in the structure upon substrate binding. There is always a possibility that the second substrate might bring structural change. As the reaction proceeds, n number of substrate might bind sequentially or together to an enzyme, and any one of them can bring upon the structural change depending upon the need of the reaction. However, once the reaction is completed, no more structural reorganization is needed to orient the substrate and catalyze the reaction.

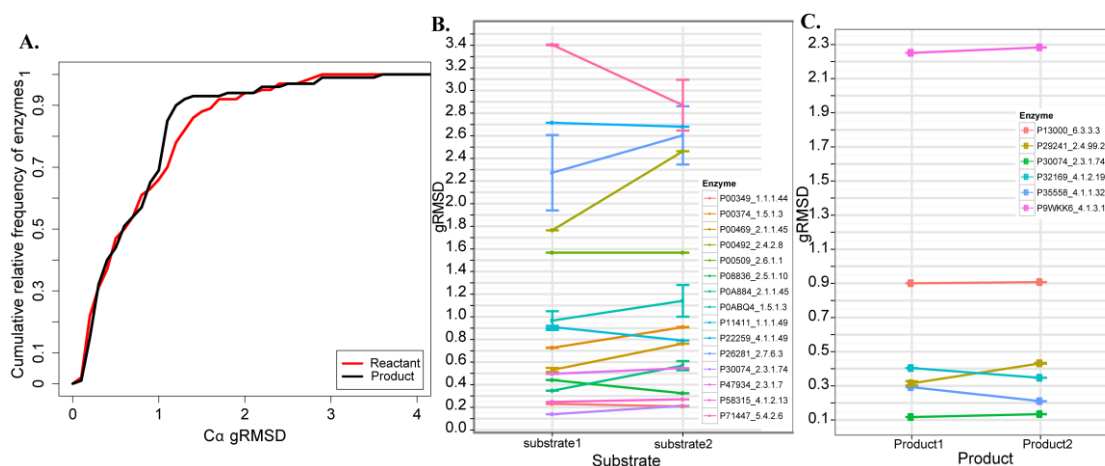


Figure 2.24 A) gRMSD for substrates and products in all-wt-enz dataset. Difference in gRMSD of the same protein with structure bound to two different C) substrates and C) products colored based on proteins.

Thus, when structures with two different products are compared to the respective unbound state, the structural variation is negligible compared to substrates which show more variation as seen in Figure 2.24B and 2.24C. Further, if one has different structures

each bound to the substrate and products, comparing their gRMSD, one can decipher the sequence of binding of the substrate and understand the reaction mechanism more clearly and to greater extent.

2.3.3.4 Allosteric changes in enzymes upon substrate binding

The binding of ligand can propagate structural changes in spatially distantly located residues such as in case of allosteric enzymes. For instance as seen in allosteric enzymes such as aspartate transcarbamoylase (ATCase) (Weber 1968), glutamine synthetase (Jiang, Peliska, and Ninfa 1998) and glycogen phosphorylase (Johnson et al. 1989).

In this study, we investigated whether residues spatially distant from binding/catalytic residues undergo any change upon substrate binding in enzymes. Here, we used all enzymes from all-wt-enz dataset irrespective of their known allosteric behavior. For this, we use side-chain network approach to identify region in protein structures undergoing structural changes. In this, first we constructed protein side-chain network for the bound and the unbound for using PSN-Ensemble program (Bhattacharyya, Bhat, and Vishveshwara 2013). Then, used change in degree of equivalent residues $|\Delta \text{ degree}|$ between holo (bound) and apo (unbound) form as a metric to quantitate the structural change occurring upon ligand binding. A degree of a node (protein residue) in a protein network is defined as the number of connections (total number of contacts) it has with other nodes (other protein residue) at the non-covalent interaction strength used for PScN construction (see methodology section 2.3.2). As a way of measuring spatial distance from the binding site, we constructed concentric spheres of increasing 5 Å radius by taking the centroid of binding site as center of sphere. For instance, first sphere consists of residues within 5 Å from the centroid of the binding site, second sphere will have residues lying between 5-10 Å of centroid and so on. The distribution of difference in change in degree from holo to apo form is shown in Figure 2.25A. Here, the median change in degree is 1 and many residues as far as 35 Å from the binding site shows change in their side-chain network property.

Importantly, change in degree only gives information about connectivity of a residue but does not change in their nature of contacts. For instance, despite having same degree of a residue it may interact with different set of residues in holo and apo form of enzymes.

To assess, change in nature of contacting residues we calculated the fraction overlapping contact of each equivalent residue in apo and holo form of structures. As shown in Figure 2.25B, of all residues having zero degree only ~48% of residues have just 5% overlap in contacts.

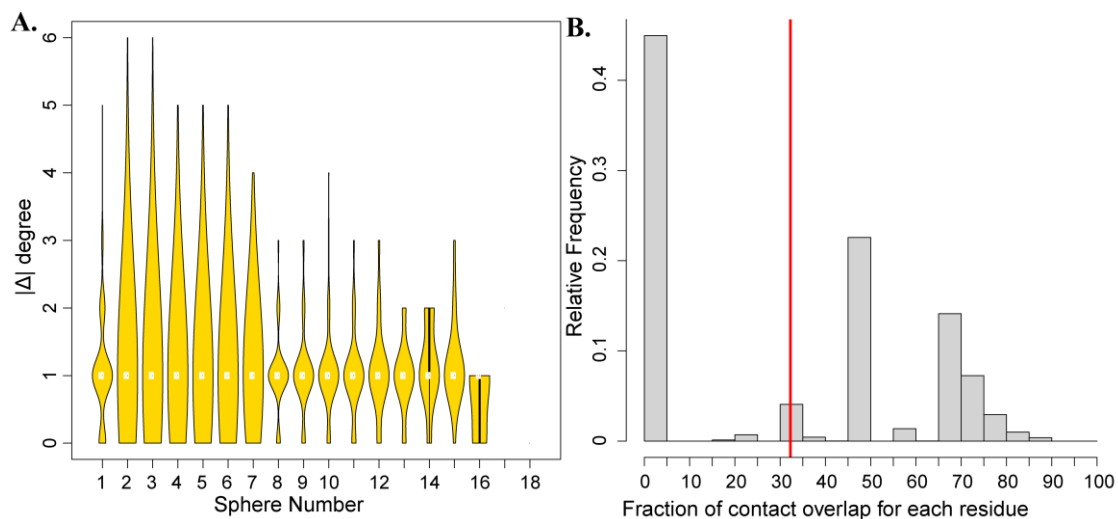


Figure 2.25 Allosteric change in enzymes upon substrate binding. A) Violin plot showing the distribution of $|\Delta \text{ degree}|$ of equivalent residues in apo and holo form of the enzyme with increase consecutive sphere constructed from the centroid of the binding site. B) Distribution of fraction of contact overlap of each equivalent residue in apo and holo form of the enzyme

To further explore regions of enzyme undergoing maximum structural change upon ligand binding, we analyzed cases where change of degree is zero. Of total equivalent residues (15028), residues having no contact overlap are 6755 (~45%). The distribution of degree change for these residues is shown in Figure 2.26A The distribution is similar to that of Figure 2.25 and average (SD) degree change is 1.1 (0.66). Next, we analyzed cases, which undergo large change both in terms of contacts and their degree as well. Here we looked at the cases, where $|\Delta \text{ degree}| > 3$ and residues have zero overlap in the contacts. Only, 50 residues (from 29 BUB complexes) out of 6755 have $|\Delta \text{ degree}| > 3$. On an average, these residues showing large change in both the degree and contact overlaps belongs to sphere 4 as shown in Figure 2.26A. This means, in these few cases, structural changes is observed at least 20 Å away from the binding site. Mostly, these include enzymes with two catalytic centers, generalists and enzymes with allosteric regulation.

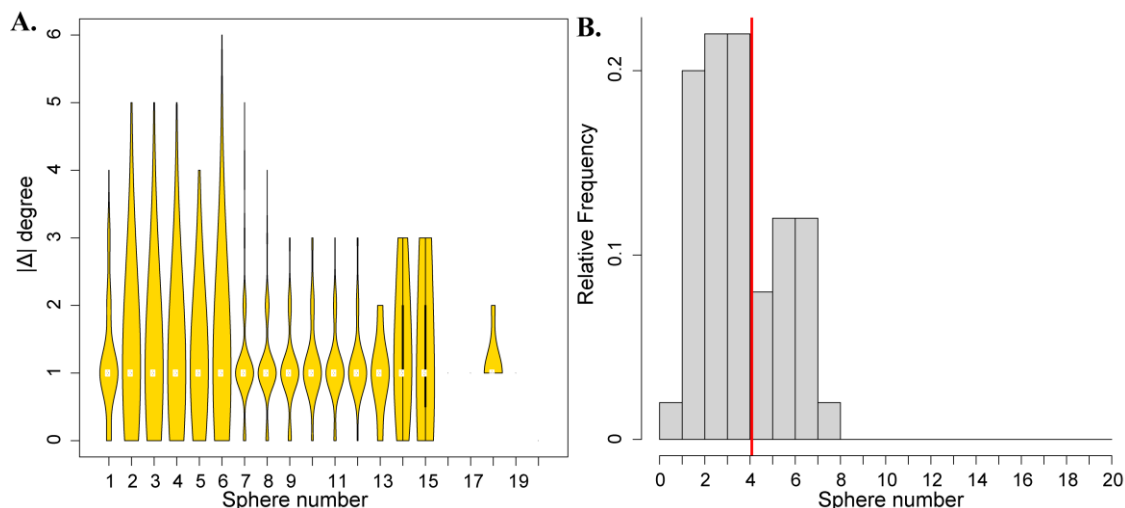


Figure 2.26 Maximum allosteric change in enzymes upon substrate binding. *A) Violin plot showing distribution of $|\Delta \text{ degree}|$ of equivalent residues which have zero overlap in contacts in apo and holo form of the enzyme with increase consecutive sphere constructed from the centroid of the binding site. B) Distribution of sphere to which residues with zero contact overlap and $|\Delta \text{ degree}| > 3$ of equivalent residue in apo and holo form of the enzyme belong.*

Next, we studied dynamics of contacts by analyzing residues having zero degree but has fraction overlapping residue in contact < 1 . We observed these in 2394 (~16%) residue have $|\Delta \text{ degree}| = 0$, but contact overlap < 1 . These residues were spread across almost all proteins (185 out of 187 proteins). Among these, 790 residues have no contact overlap and these were present in most proteins (166/187). The detailed analysis showed that these 790 residue cases mostly constituted surface residues and usually make one or two contacts. Among these cases, we found two enzymes, which does not undergo large conformational globally (gRMSD up to 1.5 Å) but residue in their remote site undergoes large change contacts with degree remaining the same. These include protein Deacetoxycephalosporin C synthase (P18548) whose residue 77 present in 4th sphere have 55, 63, 67 and 99 residues as contacts in holo (1uofA) form and 54, 66, 80 and 101 residues as contacts in apo (1w28A) form. The other protein was Pyruvate dehydrogenase kinase isozyme 2 (Q64536) whose residue 374 present in 6th sphere have 27, 28, 30 and 371 residues as contacts in holo (1jm6A) form and 66, 67, 70 and 379 residues as contacts in apo (3crkA) form.

In this present analysis of side-chain contact dynamics due to ligand binding showed that side-chain contacts do not change much (degree average of 1) between

equivalents residues from holo and apo structures. This indicates slight structural rearrangements of residues not only lying in the vicinity of substrate binding site but also of residues located distantly from the binding site.

2.3.4 Conclusions

In this study, we analyzed ligand induced conformational changes in enzymes and extended it to find any characteristic feature in this aspect among promiscuous enzymes. For this analysis, we constructed all-wt-enz dataset of wild-type enzyme pairs having ligand unbound (apo) and bound (holo) structures with only one bound ligand, which is similar (score ≥ 0.8) to cognate substrate/product/cofactor of the enzyme using EC-PDB database. Subsequently, a subset of this dataset is generated with enzymes classified as specialist and generalist. On an average, global C α Root Mean Square Deviation (RMSD) and local C α RMSD (binding/catalytic residues) does not show large change (≤ 1 Å) upon ligand binding as has been reported previously. Further, local conformational changes in the binding and catalytic residues are analyzed using measures such as side-chain torsional angle ($\Delta\chi_1$) and change in functional group angle ($\Delta C-C\alpha-Fg$). In general, only small set of binding/catalytic residues (~20%) show $\Delta\chi_1 > 20^\circ$ between apo and holo structures. Of these, a greater fraction binding residues has large ($>120^\circ$) $\Delta\chi_1$. Similarly, $\Delta C-C\alpha-Fg$ is slightly more for binding site residues. This shows substrate binding is mostly facilitated by conformational changes involving small number of residues and catalytic residues are relatively conformationally restrained. Further, comparison of the same measures between generalist and specialist showed similar trend. Interestingly, in some specialist enzymes catalytic residues undergo greater structural change whereas little/no structural change is observed in generalist. This indicates a possibility that conformationally restrained catalytic residues in generalist may facilitate catalysis/binding of alternate reaction/substrate.

Chapter 3

CSmetaPred: a meta-approach for prediction of catalytic residues

3.1 Introduction

Enzymes catalyze diverse biochemical reactions involved in almost all cellular processes of an organism and constitute a large proportion of genes encoded in genomes. For instance, in prokaryotes and eukaryotes ~30-40% and ~18-29% fraction of proteins are enzymes respectively (Freilich et al. 2005). Enzymes facilitate biochemical reactions by lowering the activation energy, which is essentially to enable formation of reaction transition state. Much of this comes from bringing substrates together in their favorable orientation and suitable local environment to promote formation of transition state. Usually, the active site of an enzyme is the region involved in binding of substrates and harbor catalytic residues, which directly participate in catalyzing enzymatic reaction. Catalytic residues can act as electrophile, nucleophile or general acid-base, can exert effect on another residue or water which is directly involved in catalytic mechanism, can stabilize proposed transition state intermediate and can influence another co-substrate or cofactor which aids in catalysis (Bartlett et al. 2002). Identification of catalytic residues and deciphering catalytic mechanism has been a daunting task that involves tedious experimental studies to elucidate active site residues and further studies to establish roles of residues in catalysis. Having experimentally determined three-dimensional structure of enzyme can expedite to identify active sites. Then, biochemical information of enzymes can be used to propose catalytic mechanism, which can be confirmed experimentally by studying the effect of catalytic residue mutants on biochemical reaction or kinetics. Therefore, identification of catalytic residues is the first essential

step to characterize reaction mechanism of enzyme. Moreover, knowledge of catalytic residues and understanding of reaction mechanism is not only important to gain insights into enzymatic processes but also crucial for designing enzyme inhibitors, protein engineering and predicting protein functions.

In past decades, many experimental and computational studies have been dedicated to decipher catalytic reaction mechanism and identification of catalytic residues. These have been documented carefully in databases such as Mechanism, Annotation and Classification in Enzymes (MACiE) (Holliday, Almonacid, Bartlett, et al. 2007; Holliday et al. 2012), Catalytic Site Atlas (CSA) (Furnham et al. 2014; Porter, Bartlett, and Thornton 2004), Structure, Function and Linkage Database (SFLD) (Holliday et al. 2017) and EzCATDB (Nagano 2005). Of these, primary databases of catalytic residues are MACiE and CSA. MACiE documents manually curated list of catalytic residues with their putative roles in mechanistic steps of enzymatic reaction and has 335 non-redundant entries (Holliday, Almonacid, Bartlett, et al. 2007). Since manual curation is not feasible for every enzyme, CSA resource attempts to increase the coverage of catalytic residues by documenting catalytic residues in enzymes with known tertiary structures in PDB (Furnham et al. 2014). CSA entries are classified into two types: a) enzymes with hand-annotated entries derived from literature, and b) catalytic residues extended from original entries to homologous sequences. In CSA version v2, catalytic residues have been documented as 928 literature annotated entries and ~24,000 for homologous enzymes. In a recent work, MACiE and CSA databases have been merged to create unified resource Mechanism and Catalytic Site Atlas (M-CSA) (Ribeiro et al. 2018) having 964 enzymes. This database has removed redundancy between two primary databases and included the option to search for active sites. Brief summary of evolution of catalytic site databases is already shown in Figure 1.3 of chapter 1.

In previous studies it has been observed that the annotations in the CSA often omit catalytic residues in one family that have been implicated in another family (even when they are present and co-located in the structures) (Furnham et al. 2016). This reflects the challenge of identifying “catalytic residues”, with different authors in the literature using different criteria in describing residues as “catalytic”.

3.1.1 Present status of catalytic residue prediction tools

Since experimental characterization of catalytic residues is still a challenging task, the computational approaches to predict catalytic residues from protein sequence/structure can greatly aid in these efforts of enzyme enhanced function annotation. Moreover, in the post genomic era with deluge of protein sequences computational prediction methods will play a crucial role in the process of function association. In the past decade, several catalytic residues prediction methods have been developed based on sequence or/and structural features of enzymes. Many methods utilize properties derived only from protein sequences or/and structures. Among sequence-based methods, initial approaches relied on establishing homologous relationships among sequences and identifying motifs to annotate catalytic residues (Hofmann et al. 1999), (Mistry, Bateman, and Finn 2007). Other sequence-based catalytic site prediction tools used sensitive sequence-based scoring function (Dou et al. 2012; J. D. Fischer, Mayer, and Söding 2008) or conservation scores like Jensen-Shannon divergence, Von Neumann entropy, relative entropy to predict catalytic residues (Capra and Singh 2007). Many approaches employed difference in amino acid propensities between catalytic and non-catalytic residues and conservation of residues in multiple sequence alignment (Petrova and Wu 2006; T. Y. Chien et al. 2008). One of the best known sequence-based methods is CRpred, which used several types of sequence-based features including position-specific scoring matrix and entropy of weighted observed percentages extracted from multiple sequence alignment using PSI-BLAST (T. Zhang et al. 2008b). As three-dimensional structure is more conserved than sequence, prediction methods were further improved by including structural information (Kato and Nagano 2011; Fajardo and Fiser 2013a). These were combined with phylogenetic motifs (La and Livesay 2005; Tobi and Bahar 2005; Dukka Bahadur and Livesay 2008), phylogenetic trees (Sankararaman and Sjölander 2008a), amino-acid stereochemical properties (Dou, Zheng, and Wang 2009; Dou et al. 2010; X.-S. Liu and Guo 2008), protein hydrophobicity distribution (Bryliński et al. 2007a). Many methods have been developed are based on searching the active-site template in the pre-computed library of known active-sites structural motifs (Nilmeier et al. 2013b; Kirshner, Nilmeier, and Lightstone 2013; Izidoro, de Melo-Minardi, and Pappa 2015). Network properties were also used to predict catalytic residues and it has been shown that catalytic residues are usually the central hubs or their neighbors (Gil Amitai et al. 2004). Some catalytic site prediction

methods are based on geometric properties such as catalytic residues are usually moderately exposed and located to the protein centroid (Ben-Shimon and Eisenstein 2005). THEMATICIS (Theoretical microscopic titration curves) calculates theoretical residue electrostatic properties from protein structure in order to detect catalytic residues (Ondrechen, Clifton, and Ringe 2001; Shehadi and Uzun 2004) that is based on electrostatic properties of active site residues. The performance of THEMATICIS was improved by including structure geometric features in another method called POOL (Partial Order Optimum Likelihood) which used monotonicity-constrained maximum likelihood approach in order to detect catalytic residues (Tong et al. 2009). EXIA uses the side-chain orientations of protein residues in order to predict catalytic site (Y. T. Chien and Huang 2012; C. Lu et al. 2014). Many recent methods combine features derived from both sequence and structure to further improve the accuracy of prediction of catalytic residues. (Petrova and Wu 2006) used a SVM (Support Vector Machine) approach to integrate both sequence and structural features. Youn et al., reviewed many frequently used features and ranked their performance based on their ability to distinguish catalytic residues from non-catalytic ones; the top-ranked features found in his study are sequence conservation, structural conservation, residue hydrophobicity, solvent accessibility and uniqueness of a residue's structural environment (Youn et al. 2007a). Profunc (R. a Laskowski, Watson, and Thornton 2005), SitesIdentify (Bray et al. 2009) and ResBoost (Alterovitz et al. 2009). A comprehensive list of catalytic site prediction methods is tabulated in Table 3.1.

Table 3.1 List of catalytic residues prediction methods.

Catalytic site prediction method	Property based on which prediction is done	References
Sequence based Methods		
MINER	Phylogenetic motif, conserved sequence fragments assessed by Partition Around Medoids Clustering (PAMC)	(La and Livesay 2005)
SMO(Sequential Minimal Optimization) algorithm	Sequence conservation, catalytic propensities of amino acids, relative position of the residue on protein surface, the number of hydrogen bonds between the residue main chain atoms and other atoms in the protein	(Petrova and Wu 2006)
	conservation scores such as Von Neumann entropy, relative entropy, Jensen-Shannon divergence and sum-of-pairs measure	(Capra and Singh 2007)

Active site residue annotation in Pfam database	Search Active site patterns in sequences in alignment with unannotated catalytic site for a given homologous family in Pfam	(Mistry, Bateman, and Finn 2007)
E1DS	Sequential blocks(conserved segments)	(T.-Y. Chien et al. 2008)
FRpred	Use conditional probability density estimation to calculate the probability of each site to be functional given its conservation, the observed amino acid distribution, and the predicted ss and rsa states.	(J. D. Fischer, Mayer, and Söding 2008)
INTREPID	An information-theoretic approach uses a traversal of the phylogeny in combination with a positional conservation score, based on Jensen–Shannon divergence, to rank positions in an MSA.	(Sankararaman and Sjölander 2008a)
CRpred	Uses various sequence features such as residue type, hydrophobicity, and PSI-BLAST profiles in a Support Vector Machine (SVM) based binary classification of residues into catalytic and non-catalytic residues.	(T. Zhang et al. 2008a)
L1Pred	L1-logreg classifier to integrate eight sequence-based scoring functions which include residue type, overlapping properties, averaged cumulative hydrophobicity, predicted protein secondary structure, predicted accessible surface area, Jensen-Shannon divergence (JSD) conservation score, the combination of relative entropy of Venn diagram and JSD conservation score (VJSD), and Consurf score.	(Dou et al. 2012)
Structure based methods		
FOD (Fuzzy oil drop)method	Hydrophobicity distribution in protein	(Bryliński et al. 2007b)
	Structural neighborhood	(Cilia and Passerini 2010)
Matching in pre-calculated active site structural motif/template library		
Weighted Mean Deviation (WMD) and DALI Score- based Discriminative Similarity (DSDS)	Machine-learning-based similarity or deviation measurements for comparison of template structures with local site structures in proteins	(Kato and Nagano 2011)
CATSID	Identify structural matches to a library of catalytic sites	(Nilmeier et al. 2013a; Kirshner, Nilmeier, and Lightstone 2013)
GASS (Genetic active site search)	Genetic algorithm for active site matching	(Izidoro, de Melo-Minardi, and Pappa 2015)
Network based methods		
GANN(Genetic algorithm integrated	Network closeness centrality	(Tang et al. 2008)

neural network) predictor		
	Weighted contact networks	(S.-W. Huang et al. 2011)
	Network centrality measures	(Fajardo and Fiser 2013b; Chea and Livesay 2007; Mitternacht and Berezovsky 2011)
Geometry based methods		
EnSite	Catalytic residues close to molecular centroid	(Ben-Shimon and Eisenstein 2005)
SurpResi	Probabilistic analysis of global radial distributions of atoms	(Kochańczyk 2011)
Electrostatics based methods		
THEMATICS (theoretical microscopic titration curves)	Residues exhibiting perturbed titration function are putative catalytic residues, applicable to only ionized residues	(Shehadi and Uzun 2004; Ondrechen, Clifton, and Ringe 2001)
POOL (Partial Order Optimum Likelihood)	All residue extension of THEMATICS, combined with cleftsize, sequence conservation	(Tong et al. 2009)
Distal catalytic residue prediction	Based on the Partial Order Optimum Likelihood (POOL) machine learning method, using computed electrostatic properties, surface geometric features, and information obtained from the phylogenetic tree as input features	(Brodkin et al. 2015)
Sequence and structure base methods		
ProFunc	Both sequence and structure based features to identify structural motifs	(R. a Laskowski, Watson, and Thornton 2005)
	Measure of sequence conservation, a measure of structural conservation, a degree of uniqueness of a residue's structural environment, solvent accessibility, and residue hydrophobicity	(Youn et al. 2007a)
MFS (Meta-Functional Signatures)	Combines Sequence, Structure, Evolution, and Amino Acid Property Information	(K. Wang et al. 2008)
SitesIdentify	Combining sequence conservation information with geometry-based cleft identification	(Bray et al. 2009)
ResBoost	Residue evolutionary conservation, 3D clustering, solvent accessibility, and hydrophilicity	(Alterovitz et al. 2009)
DISCERN	Statistical models based on phylogenomic conservation score of sequence and several structural features	(Sankararaman et al. 2010)
EXIA,EXIA2	Side-chain orientation of catalytic residues	(C. Lu et al. 2014; Y.-T. Chien and Huang 2012)

3.1.2 Overview of the study

In the present work, we have analyzed structural properties of catalytic residues and their geometrical relationship to substrate binding sites/active site residues to gain insights into three-dimensional features of active sites.

Despite significant development in catalytic residues prediction methods, one of the issues in current prediction results is the ranked position of predicted catalytic residues, which is usually ranked high (poor). Moreover, ability to predict catalytic non-polar residues is rather limited. To address these issues, we have developed meta-approach based method for catalytic residues prediction. Having improved ranked positions of catalytic residues usually also results in higher prediction accuracy. The main motivation to use consensus (meta) approach was to harness good prediction results from already existing methods and further improve predictions by combining them. Such meta-predictors have been used to improve performance of prediction for protein structure prediction- 3D-Jury system (Ginalski et al. 2003) and for binding site prediction metaPocket (B. Huang 2009).

Here, we have reported development of meta-predictors (CSmetaPred and CSmetaPred_poc), which have improved prediction accuracy as well as ranks of putative catalytic site residues. CSmetaPred ranks protein residues based on meta-score calculated as an average of scores obtained from 4 different methods *viz.* EXIA2, CATSID, DISCERN and CRpred. CSmetaPred_poc incorporates predicted pocket information in this meta-score, resulting in improved catalytic residue ranking. Both meta-predictors have been benchmarked on two comprehensive benchmark datasets and three legacy datasets. Both meta-predictors are freely available for public use as webservers at <http://14.139.227.206/csmetapred/>.

3.2 Methods

3.2.1 Dataset to study anatomy of active site

Usually, the enzyme active site harbors substrate/cofactor binding residues as well as catalytic residues, which directly participate in biochemical reaction. This suggests substrate binding sites and catalytic residues need to be proximal to each other.

Hence, the knowledge of substrate binding sites can be exploited to predict catalytic residues. To investigate three-dimensional geometric relationship between substrate/cofactor and catalytic residues, we generated non-redundant dataset (at 60% sequence identity) of enzymes with known tertiary structure bound to substrate/cofactor or their analogs, which are from derived EC-PDB database (<https://www.ebi.ac.uk/thornton-srv/databases/enzymes/>). This database documents experimentally known tertiary structures of enzymes and their associated EC numbers. For ligand bound to enzyme structure, importantly, EC-PDB provides similarity score between the enzyme cognate substrate/cofactor to the chemical compound bound to enzyme structure (Roman A Laskowski, Chistyakov, and Thornton 2005). This similarity score varies from 0 to 100% where the similarity score of 100% indicates that the ligand bound to the enzyme structure is indeed its cognate ligand. The steps used for construction of datasets are summarized in Figure 3.1.

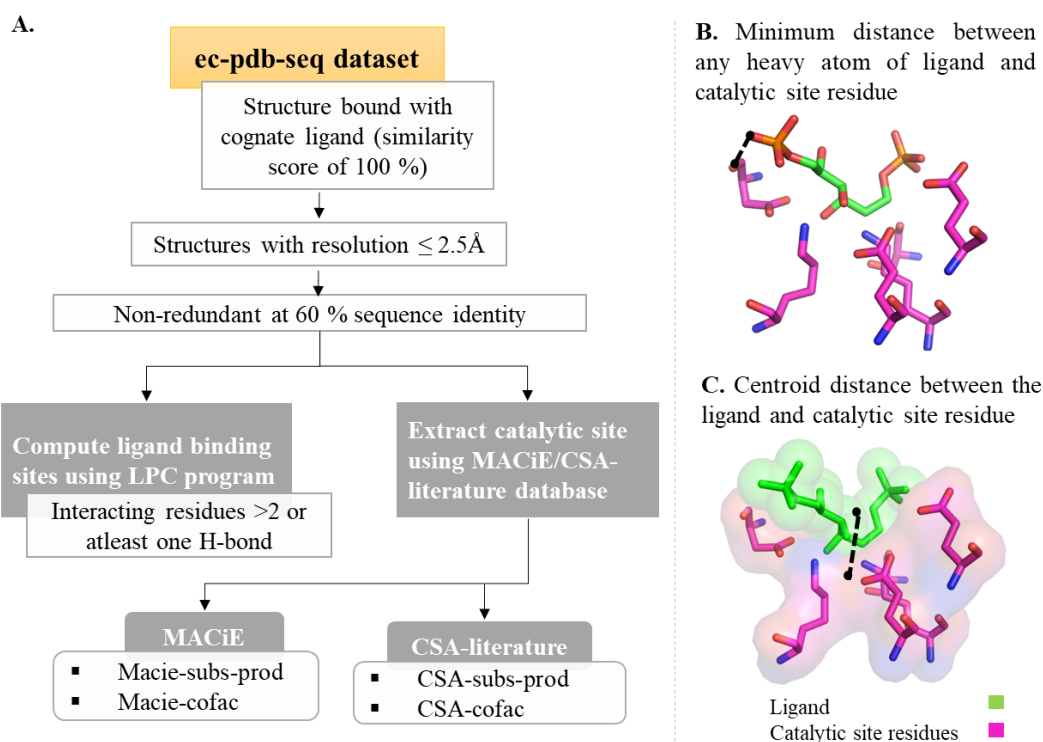


Figure 3.1 Dataset construction and quantitative metrics used for analysis. A) Schematic for construction of Macie-subst-prod, Macie-cofac, CSA-subst-prod, CSA-cofac datasets. Schematic diagram showing two metrics B) minimum and C) centroid distances used for quantifying distance between active site and their cognate ligand.

Briefly, we took structures bound with ligand having similarity score of 100%, which corresponds to the cognate substrate/cofactor of the enzymes. Subsequently,

interaction between ligand and enzyme structure was obtained using LPC (Sobolev et al. 1999). To consider biologically relevant interacting ligands, those having heavy atoms > 6 and interacting with at least 3 residues or is involved in at least one hydrogen bond are considered for further analysis (Brylinski and Skolnick 2008). The catalytic residues were obtained from MACiE (Holliday, Almonacid, Bartlett, et al. 2007) and CSA-literature (Catalytic Site Atlas) (Furnham et al. 2014) databases. This resulted in ligand-enzyme pairs, which depending on closeness of ligand to substrate/product or cofactors are categorized in two groups. In order to make our dataset non-redundant, we took each ligand type and clustered the enzyme sequences to which it is bound at 60% sequence identity using cd-hit (W. Li and Godzik 2006) program. Finally, this gave rise to following datasets: Macie-subst-prod, Macie-cofac, CSA-subst-prod, CSA-cofac having 71, 30, 169 and 39 enzyme-ligand pairs. This dataset is provided in supplementary material (*c3.1_ligand_distance.xlsx*) for chapter 3 provided in a CD along with this thesis.

To characterize geometric relationship between ligands and catalytic residues, we have used simple distance measures: a) Minimum distance between heavy atom of ligand and catalytic residues, b) Centroid distance between ligand (only heavy atoms) and main chain atoms of catalytic residues, and c) Centroid distance between ligands (non-hydrogen atoms) and side chains atoms of catalytic residues. These metrics are shown in Figure 3.1B and 3.1C.

3.2.2 Overview of CSmetaPred/CSmetaPred_poc methodology

CSmetaPred generates ranked list of residues based on their scores, which is average of per-residue scores obtained from following well-known catalytic site predictors:

1. **CATSID**: It is a structure based method, which predicts catalytic residues by matching the query structure to template/s with known catalytic residues. Based on the match score catalytic residues are transferred from template/s to query enzyme (Kirshner, Nilmeier, and Lightstone 2013; Nilmeier et al. 2013a).
2. **CRpred**: It is purely sequence based method, which uses SVM classifier on sequence-based features such as residue type, average cumulative hydrophobicity, custom-designed sequence motifs and sequence-derived PSI-BLAST profiles to predict catalytic residues (T. Zhang et al. 2008a).

3. **EXIA:** Uses both sequence and structure based properties such as amino-acid combination and theoretical structural flexibility with SVM classifier to predict catalytic residue. An important feature of this method is that it uses residue side-chain orientation angle property in prediction (C. Lu et al. 2014; Y.-T. Chien and Huang 2012).
4. **DISCERN:** It uses logistic regression model on both features derived from sequence and structure such as evolutionary measures of positional conservation, relative and absolute solvent accessibility, presence in a cleft or pocket, secondary structure, polarity and charge of a residue in prediction. It also employs INTREPID's phylogenomic method which in turn use of Jensen-Shannon(JS) divergence and phylogenetic tree traversal to estimate evolutionary conservation of each residue of the protein (Sankararaman et al. 2010; Sankararaman and Sjölander 2008b).

The above methods were chosen for meta-approach based on their performance, variability in property used for catalytic site prediction, difference in the input type of query and their availability as a source-code or webserver. Figure 3.2 outlines overview of CSmetaPred/CSmetaPred_poc methodology and is discussed in details below.

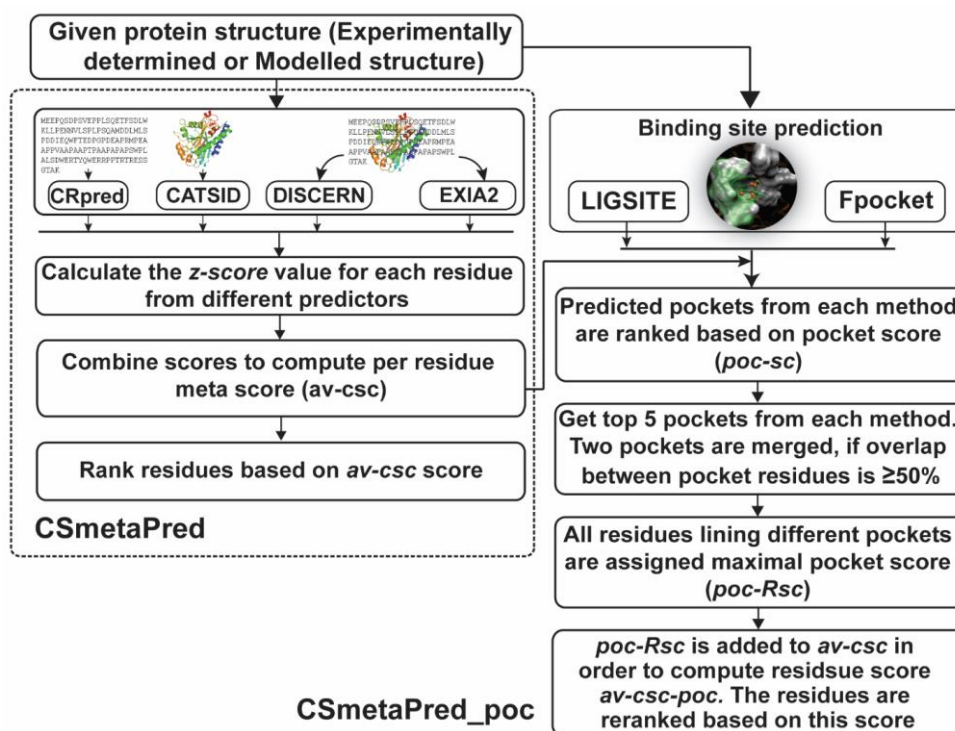


Figure 3.2 Schematic diagram showing the overview of CSmetaPred and CSmetaPred_poc.

3.2.3 Processing of prediction server outputs and meta-score computation

As we need to combine varied outputs from various predictors, the raw outputs from other methods were processed to assign residue score. We followed following procedure to process raw outputs:

1. CATSID outputs a score for each hit template, which is measure of the likelihood of structural match of the query protein with library of catalytic site templates. To obtain CATSID score (S_{ca}) of a residue in the query protein, we simply assign the score of the template hit to which the residue belongs. In case the residue belongs to more than one template hit, we sum score for all hits and assign this summed score to the residue. We used all the templates hits in order to obtain S_{ca} for a residue. It should be noted that, CATSID results in residue score only for a subset of query residues because the method matches template.
2. EXIA2 uses structural property of side chain orientation, weighted contact network (WCN), PSI-BLAST derived PSSM and amino acid type to predict catalytic residues. These properties are combined to calculate the rank score, which in turns reflects the probability of residue to be catalytic. Higher the rank score, higher the chance of it being a catalytic residue. EXIA server gives output only for its phase 1, which includes only the residues with side chain functional atoms in its calculations. These are ranked based on rank score and include polar/charged amino acid residues (R, N, D, C, Q, E, H, K, S, T, Y) and tryptophan. In order to rank the remaining non-functional side chain (non-polar) amino-acids including glycine, we followed the second phase approach of EXIA2 (Y.-T. Chien and Huang 2012). For every polar residue, in the ranked order given from EXIA webserver, we selected its neighboring non-polar residues with WCN score >0.9 and further ranked these neighbors based on their WCN scores. While ranking neighbors, we follow the ranked order list of polar residues of phase 1. Thus, non-polar neighbors of ranked 1 polar residue will be ranked before the neighbors of ranked 2 polar residues. Every non-polar residue will be ranked only once in its first occurrence as a neighbor to a polar residue. This results in a ranked list of non-polar residues, which we refer to as NP-1. The remaining non-polar residues which are not the neighbors of any ranked polar residues or $WCN < 0.9$, are ranked after NP-1 based on their WCN score and referred as NP-3. Thus, first we take EXIA ranked polar residues, followed by NP-1 and finally, NP-

2 to generate the final list of ranked residues based on EXIA rank score (S_{rs}). Further, WCN scores measures structural flexibility of residues (Lin et al. 2008). As catalytic residues tend to be rigid, they usually have higher WCN scores. Thus, WCN score (S_{wcn}) was used as an independent score to compute meta-score. WCN score was either obtained from EXIA server or calculated locally using previously described algorithm (Lin et al., 2008).

3. CRpred is one of the best sequence based catalytic site prediction method. It uses SVM classifier trained on sequence features such as residue type, hydrophobicity and PSI-BLAST profiles. We used residue SVM score (S_{cr}) directly to rank the residues from CRpred.
4. DISCERN uses features derived from both sequence and structure such relative and absolute solvent accessibility, cleft/pocket, secondary structure, polarity, charge and evolutionary measures of positional conservation to predict catalytic residues. The final residue level scores (S_{di}) are used directly to rank the residues from DISCERN.

To compute a single score to rank residues in CSmetaPred, first the residue score from each method is normalized with respect to its respective mean and standard deviation. The normalized residue score is defined as:

$$zSc(ij) = (S(ij) - \mu(j)) / \sigma(j)$$

where, $zSc(ij)$ and $S(ij)$ are normalized and raw scores of residue i for method j respectively, $\mu(j)$ and $\sigma(j)$ are mean and standard deviation for method j scores respectively. Then, we calculate mean of normalized residue scores for each residue referred to as meta-score or *av-csc* score, which is defined as:

$$av - csc(i) = \frac{\sum_{j=1}^5 zSc(ij) * p(j)}{\sum_{j=1}^5 p(j)}$$

where, $zSc(ij)$ is z-score of residue i for method j and $p(j)$ is binary function with $p(j)=1$ for residue having an assigned score, or 0 otherwise. The *av-csc* score is used in CSmetaPred to rank residues for every protein, wherein high score represents a greater chance for it to be a catalytic residue.

As has been shown from previous studies (Cilia and Passerini 2010) as well as in our study that catalytic residues are spatially proximal to substrate/cofactor binding sites, we

have combined meta-score obtained from CSmetaPred with predicted pocket residues to further improve prediction (CSmetaPred_poc) accuracy. In CSmetaPred_poc, residues lining the predicted pockets are assigned pocket score, which is summed with meta-score for ranking residues. Here, first we have used two binding site prediction methods Fpocket (Le Guilloux, Schmidtke, and Tuffery 2009) and LIGSITE (B. Huang and Schroeder 2006) and then merged these pocket to generate a combined list of pockets. In order to merge the pockets predicted from these two different prediction methods, we first ranked these pocket based on pocket score (poc_sc). For each pocket i , $poc_sc(i)$ is defined as:

$$poc_sc(i) = \left(\sum_{j=1}^{Nres(i)} av - csc(j) \right) / Nres(i)$$

where, $av-csc(j)$ is meta-score of pocket residue j , $Nres(i)$ is number of residues in a given pocket i . From ranked list of pockets, we select top five pockets from both Fpocket and LIGSITE, and merge two pocket if they have an overlap of more than 50%. The parameters for pocket ranking and merging were optimized after looking at cumulative distribution of catalytic residues present in predicted pockets for proteins in macie-254 dataset. As seen in Figure 3.3, at pocket rank 5 both LIGSITE and Fpocket have achieved close to the maximum catalytic residues identified within predicted pockets. Infact, there is a drastic increase in catalytic residues fraction after re-ranking in LIGSITE could also be due to merging of pockets within LIGSITE. Thus, we selected top five pockets for merging pockets from both the methods.

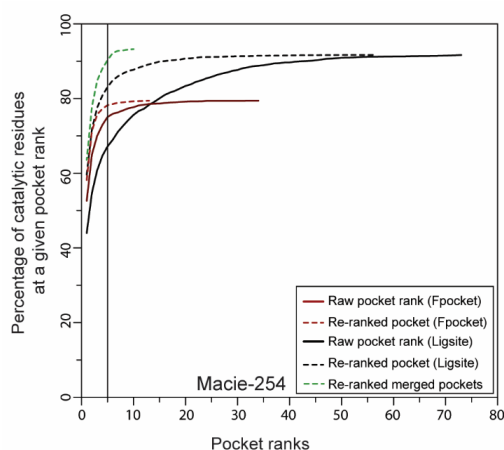


Figure 3.3 Cumulative distribution of catalytic residues within a given pocket rank. Data shown for macie-254 dataset for: Pockets output from LIGSITE/Fpocket, re-ranked pockets using poc_sc score and merged top 5 re-ranked pockets.

Next, we assign the residue pocket score (poc_Rsc) to each of the residue present in these merged top 5 re-ranked pockets. Poc_Rsc is essentially the poc_sc score of the pocket to which the residue belongs. In case a residue belongs to more than one pocket, the maximum poc_sc is assigned as poc_Rsc . If a residue doesn't belong to any pocket, then poc_Rsc of 0 is assigned to that residue. Finally, this poc_Rsc score is linearly combined with $av-csc$ to calculate $av-csc-poc$ score, which is defined as:

$$av - csc - poc(i) = av - csc(i) + poc_Rsc(i)$$

Residues are ranked based on $av-csc-poc$ score, in CSmetaPred_poc.

3.2.4 Benchmark datasets for meta-predictor

In this study, we have used five datasets to benchmark meta-predictor and for performance comparison with other constituent methods of meta-predictor. Of five datasets, three are legacy datasets, which were derived from previous works (Tong et al. 2009; Youn et al. 2007b; Petrova and Wu 2006) while two benchmarking datasets were collated in this study. The details of the construction of these datasets are discussed below:

Three legacy datasets have been mainly used to compare CSmetaPred with previously developed methods. From previous work, we took entries from POOL-160, EF-Fold and PW-79 datasets (Tong et al. 2009; Youn et al. 2007b; Petrova and Wu 2006) and removed enzymes having catalytic site present in more than one Protein Data Bank (pdb) chain. Any obsolete pdb entry was either replaced with updated pdb entry or removed from the dataset. Subsequently, these datasets were referred to as POOL-148, EF-Fold-164 and PW-79 depending on number of pdb entries, which are 148, 164 and 79 respectively. Next, entries from these 3 datasets were merged to form EF_POOL_PW, which was further made non-redundant at 60% sequence identity using CD-HIT.

Two datasets macie-254 and csalit-688 were compiled in this study using catalytic site definition from MACiE and CSA databases respectively. Briefly, 335 MACiE entries having catalytic site defined in single pdb chain were used to prepare a non-redundant set of 254 proteins at 60% sequence identity using CD-HIT. Similarly, from CSA-literature annotated pdb structures after removing MACiE entries, a non-redundant csalit-688 dataset (60% sequence identity) was constructed using CD-HIT. Additionally, we

merged two or more catalytic sites in a single pdb chain that have at least one common residue between them. Further, a combined non-redundant dataset (60% sequence identity) CSAMAC was generated by taking all structures from macie-254 and csalit-688. Since CSmetaPred_poc uses predicted pocket residues, a structure bound with ligands (possibly substrate or analogue) can bias pocket prediction. To remove this bias, we constructed dataset of pdb structures without ligands (apo form) from entries in CSAMAC. This resulted in 137 unbound structures and was called UB-137 dataset. This dataset is provided in supplementary material (*c3.2_datasets.pdf*) for chapter 3 provided in a CD along with this thesis.

Generation of homology models

To estimate performance of meta-predictor in the cases, where no tertiary structure is available for the protein, we generated a set of template based modelled structures and used these as input for CSmetaPred_poc. We used MODELLER to build models of sequence from CSAMAC and sequence identity between templates and query ranged from 40-90%. The templates were searched in Template library (LIB_TEM), which was built using PISCES server (G. Wang and Dunbrack 2003) with following criteria: Structures having resolution ≤ 2 Å, sequence length of 40-1000 residues and were non-redundant at 60% sequence identity.

We searched full length protein sequence of each pdb structure in CSAMAC dataset against LIB_TEM library and identified all the templates with sequence identity ranging from 40 to 90%, sequence coverage of 70% between query and template using `profile_build()` module of MODELLER (Webb and Sali 2016). The templates with the above mentioned criteria were grouped into sequence identity bin of 40–50%, 50–60%, 60–70%, 70–80% and 80–90%. For each query sequence present in each sequence identity bin, we selected the best template for it based on maximum sequence identity. This results in total 468 query-template pairs, where 235, 135, 53, 22 and 23 query-template pairs belonged to 40–50%, 50–60%, 60–70%, 70–80% and 80–90% sequence identity bins respectively. Further, query sequence was aligned with template structure using `align2d()` module of MODELLER. Further, we build 10 models for each query sequence using these query-template alignments following *automodel* class of MODELLER. These models were ranked based on DOPE score and the model with best (energetically the lowest) DOPE score was taken as a representative for catalytic site

prediction. Thus, we generated a total of 468 models for 335 query protein sequences, which constituted model dataset used for analysis. We used a more conventional approach for modelling, which can be improved by including multiple templates or using better structure prediction methods.

3.2.5 Metrics used in evaluation meta-predictor

To evaluate performance of predictors we have used measures employed to assess typical binary classifiers such as Receiver Operating Characteristic (ROC) and Precision Recall (PR) curves. We have defined true positive and true negative by creating a binary classification by selecting top n ranked list as predicted catalytic residues and rest as non-catalytic residues, where n is rank cut-off whose value varies from 1 to the length of the query protein. Thus, true positives (TP) are correctly predicted catalytic residues; false negatives (FN) are catalytic residues predicted as non-catalytic; false positives (FP) are non-catalytic residues predicted as catalytic; true negatives (TN) are correctly predicted non-catalytic residues. The other quantitative measures of binary classification are defined as given below:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{TPR (recall)} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{FPR (1-specificity)} = \text{FP} / (\text{FP} + \text{TN})$$

We have used ROC curves to visually represent and compare the performance of catalytic residue predictors. ROC curve is essentially a plot between TPR and FPR and depicts relative trade-off between the numbers of correctly classified positive examples with the number of incorrectly classified negative examples. In order to generate a single vertical average ROC curve for all the proteins in the dataset, we averaged recall at all the FPR values (0-1). In case a recall is not computed at a given FPR value, it is linearly interpolated. In order to quantitate the performance evaluation done by ROC curves, we also computed Area Under Curve of ROC curve (AUCROC) and Mean Average Specificity (MAS) (Tong et al. 2009), which is mean of Average Specificity (AveS):

$$\text{AveS} = \frac{\sum_{r=1}^N S(r) * \text{pos}(r)}{N_{\text{pos}}}$$

where, r is rank, N is number of residues in a protein, $\text{pos}(r)$ is binary function with $\text{pos}(r)=1$ for known catalytic residue or 0 otherwise and $S(r)$ is the specificity at a given cutoff rank r , N_{pos} is the total number of positive examples (catalytic residues in this case).

PR Curve is an alternative to ROC curve for method evaluation when datasets have large skew in total numbers of negative with positive counts (Saito and Rehmsmeier 2015). As the number of positives (catalytic residues) is much lower than negatives (non-catalytic) residues, we have also used PR curve to evaluate the performance of the meta-predictor. We used AUCCalculator to generate PR curves. It generates average PR curve for all the proteins in the dataset by averaging all the precision values a given recall values. If a protein doesn't have precision at a given recall value, we interpolate its values using local skew (J. Davis and Goadrich 2006). Further, to quantify the performance assessment by PR curves, we have calculated AUCPR (Area Under PR curve). Another metric to quantitate the performance of meta-predictors is Mean Average Precision (MAP), which is frequently used in the information retrieval and has been suggested to have good discrimination and stability (Manning, Raghavan, and Schütze 2008). MAP is mean of average precision (AP), which is defined as the arithmetic mean of precisions for a set of top n residues after each true positive (catalytic residue) is retrieved. MAP is calculated as:

$$MAP = \frac{1}{N} \sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} \text{Precision}(R_{ij})$$

where, N is total number of proteins in the dataset; for protein i , n_i is the number of true positives and $\text{Precision}(R_{ij})$ is precision calculated at the rank R_{ij} at which true positive j for protein i is retrieved in the ranked list.

The above metrics were used to compare the performance of CSmetaPred, CSmetaPred_poc with CRpred, EXIA, DISCERN and WCN. In our assessment, we have not included CATSID because it ranks only subset of residues, whereas rest all methods assign ranks to all residues.

3.3 Results

3.3.1 Analysis of distance geometries between catalytic residues and bound substrates/cofactors

In enzyme active sites, substrates/cofactors involved in enzymatic reaction needs to be spatially proximal for catalytic residues to facilitate biochemical reactions. Based on this assumption, we analyzed the overlap between substrate/cofactor binding and catalytic residues. The number of overlapping residues is defined as number of overlapping residues/number of catalytic residues. As is evident in Figure 3.4, catalytic residues overlap with binding residues in most enzymes (>50% enzymes has greater than >55% overlapping residues). On an average ~59% of residues are common between catalytic and binding sites. There were ~11(19)% of enzyme-ligand pairs which have less than <10% overlap in their catalytic and binding site in MACiE (CSA literature) dataset. Detailed analysis of these cases showed that most of these are the enzymes (14 out of 32) where there is only one catalytic residue in CSA literature dataset. Some of these enzymes are bound to the product. For example, anthranilate synthase of *Serratia marcescens* (pdb 1i7qA) is bound to its product (PYR). This enzyme has two distinct catalytic centers due to which the distance between PYR and catalytic residue varies from 5.38 Å (HIS_A_398) to 32.45 Å (CYS_B_85).

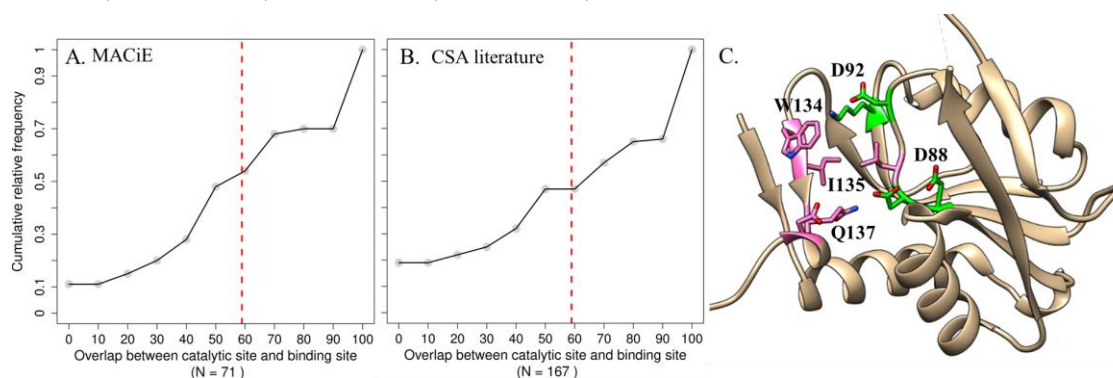


Figure 3.4 Cumulative distribution plot of overlapping residues between binding and catalytic residues for A) MACiE and B) CSA literature. The red line shows the mean of the distribution. C) 75% overlap in Binding site (pink color) and catalytic site (green color) of *E. coli* xanthine-guanine phosphoribosyltransferase with D92, K115 and D89 common among them.

Having observed that there is usually a significant overlap between binding and catalytic residues, we analyzed the euclidean distance between substrate/cofactors (ligands)

occupying the binding site and catalytic residues. The intention of this analysis is to identify search space for docking of ligand (substrate/cofactor) when we know enzyme catalytic residues or elucidate possible catalytic residues from ligand bound enzyme structure. Moreover, this analysis will provide general understanding of distance geometry requirement between substrate/cofactor and catalytic residues. As mentioned in methods section 3.2.1, we calculated minimum distance between heavy atoms ligand and catalytic residues as well as distances between centroids of ligand and catalytic residues. The minimum distances distribution for substrates and cofactors are shown in Figures 3.5 A and B. As shown in Figure, substrates/products are mostly proximal ($>82\%$ of ligands are within 5 \AA distance) to catalytic residues in comparison to cofactors ($>80\%$ of ligands within 5 \AA distance). The median distance observed for substrates and cofactors are 2.85 and 2.91 respectively in Macie-sub-prod and Macie-cofac datasets and 2.96 and 3.17 in CSA-sub-prod and CSA-cofac dataset. The cofactors are mostly located farther from catalytic residues could be because many of these do not participate in reaction but facilitate reaction through substrate. It is important to note that definition of substrate or products depends on the way reaction equation is written. Hence, in tertiary structure it is difficult to find whether a substrate is yet to be positioned in binding site or a product is leaving binding site after reaction when we observe large distances between catalytic residues and ligands.

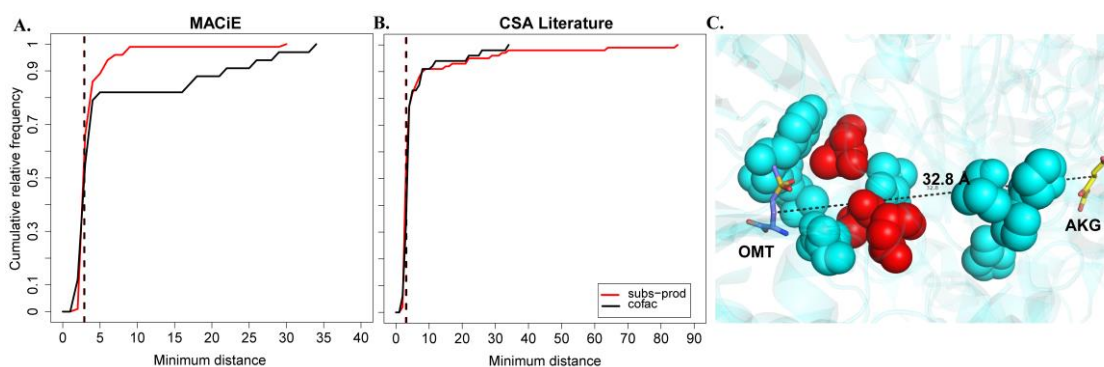


Figure 3.5 Cumulative distribution of minimum distance between catalytic residues and any heavy atom of ligands/cofactors in A) MACiE and B) CSA literature for-sub-prod cofac datasets. The red dotted line indicates the median of the distribution. C) Active site of enzyme Glutamate synthase (PDB 1EA0) bound to oxoglutaric acid (AKG) and L-Gln analogue (OMT), ammonia tunnel is shown in sphere representation. Cross talk and ammonia-channeling occur between the active sites present at the two ends of the tunnel

We performed detailed analysis of cases when we observed large distances either for substrate or cofactors and are discussed below.

1. Intermediate step of enzymatic reaction; In reaction with more than one reactant, the bound ligand might interact with other reactant, which in turn interact with the catalytic residues. For example, 2,4-dienoyl-CoA reductase (DCR) in *E. coli* is an iron-sulfur flavoenzyme, which contains FMN, FAD, and a 4Fe-4S cluster. It is also a monomer, unlike that of its eukaryotic counterparts, which form homotetramers and lack the flavin and iron-sulfur cofactors. DCR utilizes NADPH to remove the C4-C5 double bond of unsaturated fatty acids. This reaction is initiated by hydride transfer from NADPH to FAD, which in turn transfers electrons, one at a time, to FMN via the 4Fe-4S cluster. The fully reduced FMN provides a hydride ion to the C5 atom of substrate, and Tyr and His are proposed to form a catalytic dyad that protonates the C4 atom of the substrate and completes the reaction. Thus, in the crystal structure 1ps9, we observe the first reactant NADP(+) involved in catalysis far from catalytic site (with minimum distance of 16.56 Å from catalytic residue HIS_A_252), while the second reactant MDE (81.25% similar to enzyme reactant Trans-2,3-didehydroacyl-CoA) is bound close to catalytic site with minimum distance of 2.9 Å from catalytic residue HIS_A_252 (P. A. Hubbard et al. 2003).
2. Ligand bound in non-productive subsite in catalytically incompetent enzymes **Glutamate-tRNA ligase**(EC 6.1.1.17). To avoid aminoacyl-AMP formation in absence of tRNA(GLU), ATP is accommodated in a non-productive subsite within the ATP-binding site (Sekine et al. 2003). The α -phosphate of ATP bound in this mode is too far from the α -carboxyl group of glutamate to react with it (pdb 1J09). tRNA binding causes conformational changes and ATP is bound to productive subsite (pdb 1N77).
3. Involvement of distinct active centers in catalysis: Certain enzymes have many distinct active centers for e.g. Glutamate synthase (1.4.1.13): The enzyme functions through three distinct active centers carrying out L-glutamine hydrolysis, ammonia dependent synthesis of 2-oxoglutarate to L-glutamate and oxidation of NADPH (Binda et al. 2000). The crystal structure 1EA0 shows that 2-oxoglutaric acid is bound at 27 Å from amidotransferase active center and amidotransferase and 2-oxoglutarate sites are 31 Å apart. These two catalytic centers constituting site of ammonia production and L-Glu synthesis site where ammonia is being used are connected via a tunnel (Figure 3.5C).
4. Ligand bound can be product: In pdbid 1DBT the ligand bound is U5P which is at a minimum distance of 11.96 Å is actually the product (Appleby et al. 2000) leaving

the enzyme after reaction. This enzyme is **Orotidine 5'-phosphate decarboxylase** (EC 4.1.1.23):

5. Involvement of some mediators for travelling of substrates during catalysis: Certain reactions involve some mediators like cofactors, water molecules which help in transfer of substrates during catalysis. For e.g. **BOVINE F1-ATPASE** (3.6.3.14) has three active sites. In crystal structure 1E79 ATP is bound at a minimum distance of 20.34 Å. Such large distance can be explained by water acting as a mediator. The γ -phosphate of ATP deprotonates the first water, which deprotonates a second water, which attacks the gamma phosphate in a nucleophilic addition resulting in a pentavalent phosphate intermediate (Gibbons et al. 2000).
6. Ligand bound is a cofactor: When cofactors are required in enzyme catalysis, they usually bind away from catalytic site as they are not directly involved in catalysis. *E. coli* pyridoxine 5'-phosphate oxidase is the enzyme catalyzing the final step in the synthesis of pyridoxal 5'-phosphate, a vital cofactor in many metabolic processes including amino acid metabolism. The reaction involves the oxidation of PNP to PLP using the cofactor FMN. It proceeds through hydride transfer from the 4'Carbon to the N7 of FMN, with steric strain from Arg 197 acting to place the substrate and cofactor in correct orientation for this to occur. This forms an electron deficient transition state; the oxygen lone pair then forms a bond to the 4'Carbon to result in the product. Thus, in the crystal structure 1g79, the cofactor FMN is bound far (minimum distance 25.34 Å) from catalytic residue R197, while the enzyme product PLP is bound close to the active. Interestingly, in addition to the active site, pyridoxine 5'-phosphate oxidase contains a non-catalytic site, which is 11 Å away from the active site and it binds to another molecule of pyridoxal 5'-phosphate tightly. It has been suggested that a possible tunnel exists between the two sites so that pyridoxal 5'-phosphate formed at the active site may transfer to the non-catalytic site without passing through the solvent. This second binding site of PLP protects the product of the reaction from release into the cell so it can be transferred directly onto the enzymes that require it (Safo et al. 2001).

In predicting position of ligand or catalytic residues from such distance information, it will be challenging to identify residue/ligand to compute appropriate search space as it only gives the lower limit of the distance. Moreover, having incorrectly annotated catalytic residue or wrong substrate can lead to greater chance of incorrect predictions.

To circumvent this and have an idea about average distance space between catalytic and binding site, we computed distances between centroid of ligand and catalytic centroid calculated considering only main or side chain atoms of residues. Figure 3.6 summarizes the centroid distances for main and side chain respectively. The centroid distance between main chain atoms of catalytic residues and any heavy atom of ligand/cofactor varies from ~ 8 - 10 Å. Compared to main chain atoms, the side chain atoms of catalytic residues are closer to the ligands/cofactors are show slightly less centroid distance of ~ 6 - 8 Å. This can be due to direct interaction of side chain atoms of catalytic residues during catalytic reaction. The above observation concurs for ligands/cofactors in both MACiE and CSA Literature datasets.

We have exploited this spatial proximity of ligand binding site and catalytic site in improving the performance of our meta-predictor CSmetaPred, and developing another version of it, called CSmetaPred_poc, which shows improved performance. The details of the same are discussed in later section.

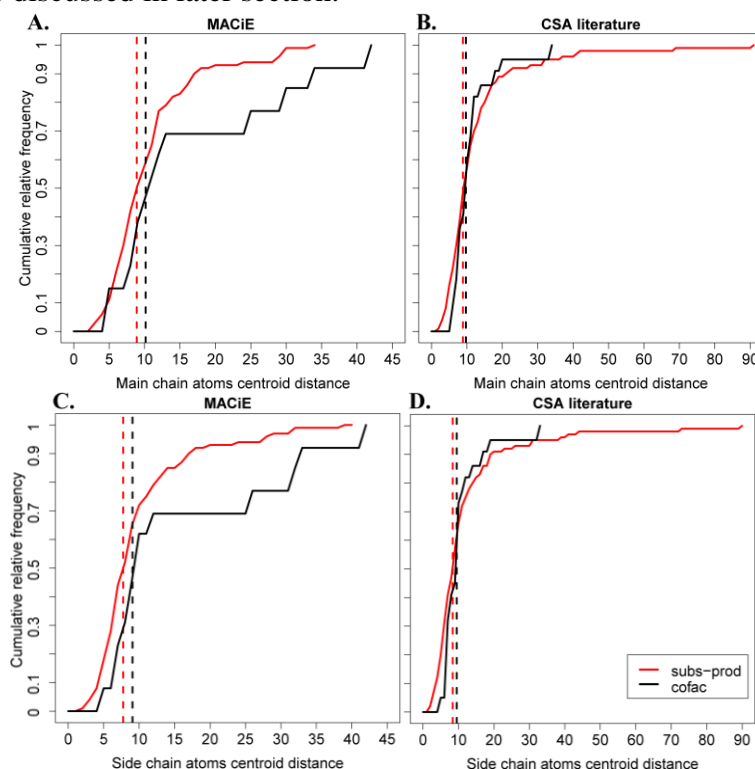


Figure 3.6 Cumulative relative frequency of distance distribution of centroid distance between A) and B) main chain atoms and C) and D) side chain of catalytic residue and any heavy atom of ligands/cofactors in Macie-subprod/cofac and CSA-subprod/cofac datasets respectively. The red dotted line indicates the median of the given distribution.

3.3.2 Evaluation of meta-predictor prediction

As mentioned before, we have assessed the performance of our meta-predictors using three legacy datasets (compiled from previous studies) and two benchmark datasets (compiled in this study) using average ROC and PR curves. In this assessment, we have excluded CATSID method (see methods).

3.3.2.1 ROC curve comparison of CSmetaPred and its constituent methods

First we used average ROC curves to evaluate and compare the performance of CSmetaPred and its constituent methods. ROC curve shows relative trade-off between the numbers of correctly classified positive examples with the number of incorrectly classified negative examples. It is curve between True Positive Rate (recall) and False Positive Rate as shown in Figure 3.7. Any point on the diagonal of this curve represents a random prediction and diagonal divides the ROC space. Points above the diagonal represent good classification results (better than random), points below the line represent poor results (worse than random). As seen in Figure 3.7, for any given FPR, CSmetaPred always have a higher recall value. Thus, it outperforms all its constituent methods. Apart from visual inspection, we also computed quantitative metrics AUROC and MAS (discussed in methods) to represent performance of ROC curves. CSmetaPred performs best among its constituent methods in terms these quantitative metrics with AUCROC and MAS (Table 3.2) having the highest value of 0.960 and 0.961 respectively in CSAMAC dataset. Similar trend was seen in other datasets as well (see Table 3.2).

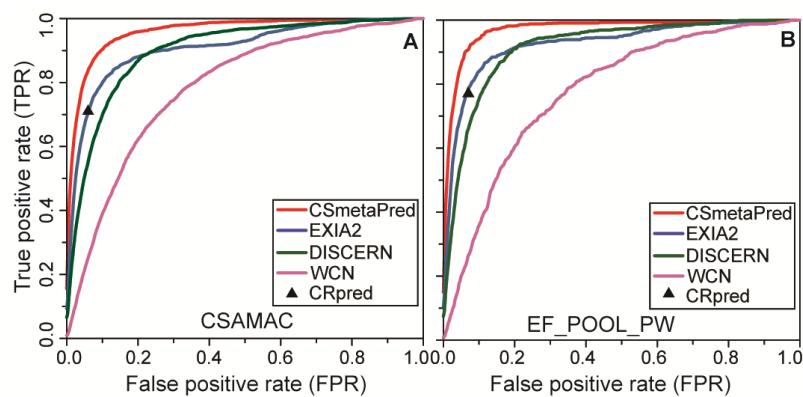


Figure 3.7 Average ROC plots to show comparison among various predictors (EXIA2, DISCERN and WCN) on A) CSAMAC and B) EF_POOL_PW. CRpred SVM performance is shown as filled triangle.

Since MAS provides comparison of average performance among methods, we estimated statistical significance of performance difference between CSmetaPred and other methods considering all pairwise comparison *i.e.* on the per protein basis. Using Wilcoxon signed-rank test performance difference between CSmetaPred and other methods is statistically significant (p -value < 0.0001). We observed similar CSmetaPred performance on EF_POOL_PW and individual datasets (Figure 3.7B). Other datasets also showed similar performance (Table 3.2 and 3.11).

Table 3.2 summarizes quantitative measures for ROC and PR curves.

Method	AUCROC	AUCPR	MAS	MAP	Median Rank	Average rank
CSAMAC dataset (884 protein)						
CSmetaPred_poc	0.967	0.347	0.968	0.514	6.0	12.0
CSmetaPred	0.960	0.324	0.961	0.489	7.0	14.4
EXIA2	0.908	0.167	0.910	0.317	14.5	33.0
CRpred	--	--	--	--	14.0	21.2
DISCERN	0.900	0.103	0.901	0.226	23.0	36.3
WCN	0.785	0.034	0.786	0.081	53.4	68.6
EF_POOL_PW dataset (286 protein)						
CSmetaPred_poc	0.974	0.366	0.975	0.531	5.5	9.9
CSmetaPred	0.970	0.338	0.972	0.502	6.3	11.1
EXIA2	0.926	0.172	0.927	0.333	12.8	25.9
CRpred	--	--	--	--	11.6	17.9
DISCERN	0.916	0.110	0.918	0.241	21.0	30.6
WCN	0.777	0.031	0.779	0.073	55.1	71.9
POOL-148 dataset (148 protein)						
CSmetaPred_poc	0.975	0.426	0.976	0.571	5.5	9.2
CSmetaPred	0.971	0.403	0.972	0.547	6.0	10.2
EXIA2	0.919	0.192	0.920	0.342	14.0	26.0
CRpred	--	--	--	--	13.2	17.9
DISCERN	0.911	0.117	0.913	0.241	22.5	30.7
WCN	0.793	0.034	0.795	0.078	52.9	64.4
PW-79 dataset (79 protein)						
CSmetaPred_poc	0.972	0.457	0.973	0.599	5.0	9.6
CSmetaPred	0.969	0.445	0.970	0.584	5.0	10.0
EXIA2	0.918	0.219	0.920	0.378	12.2	24.8
CRpred	--	--	--	--	12.5	16.7
DISCERN	0.914	0.129	0.916	0.261	20.5	27.5
WCN	0.772	0.035	0.775	0.084	46.5	62.2
EF-Fold-164 dataset (164 protein)						
CSmetaPred_poc	0.969	0.329	0.970	0.506	5.5	11.6
CSmetaPred	0.966	0.300	0.967	0.476	6.5	13.0
EXIA2	0.930	0.161	0.931	0.322	12.3	25.8
CRpred	--	--	--	--	11.0	17.2

DISCERN	0.913	0.104	0.914	0.237	18.0	32.6
WCN	0.759	0.028	0.761	0.069	59.9	80.5
macie-254 dataset (254 protein)						
CSmetaPred_poc	0.961	0.335	0.962	0.486	8.0	15.4
CSmetaPred	0.947	0.308	0.949	0.458	9.8	20.2
EXIA2	0.899	0.172	0.900	0.304	19.6	38.2
CRpred	--	--	--	--	19.3	26.6
DISCERN	0.886	0.097	0.887	0.201	28.3	44.3
WCN	0.791	0.039	0.793	0.082	57.7	68.7
csalit-688 dataset (688 protein)						
CSmetaPred_poc	0.971	0.366	0.972	0.534	5.3	10.4
CSmetaPred	0.966	0.343	0.967	0.509	6.0	12.0
EXIA2	0.911	0.169	0.913	0.325	12.7	31.0
CRpred	--	--	--	--	12.3	19.1
DISCERN	0.905	0.107	0.906	0.237	21.0	33.8
WCN	0.785	0.032	0.787	0.081	51.8	67.6
UB-137 dataset (137 protein)						
CSmetaPred_poc	0.974	0.468	0.976	0.620	5.0	7.9
CSmetaPred	0.970	0.433	0.971	0.582	5.6	9.0
EXIA2	0.908	0.220	0.910	0.378	14.7	25.0
CRpred	--	--	--	--	9.8	15.3
DISCERN	0.905	0.148	0.907	0.288	19.5	30.0
WCN	0.786	0.042	0.788	0.093	46.0	58.9

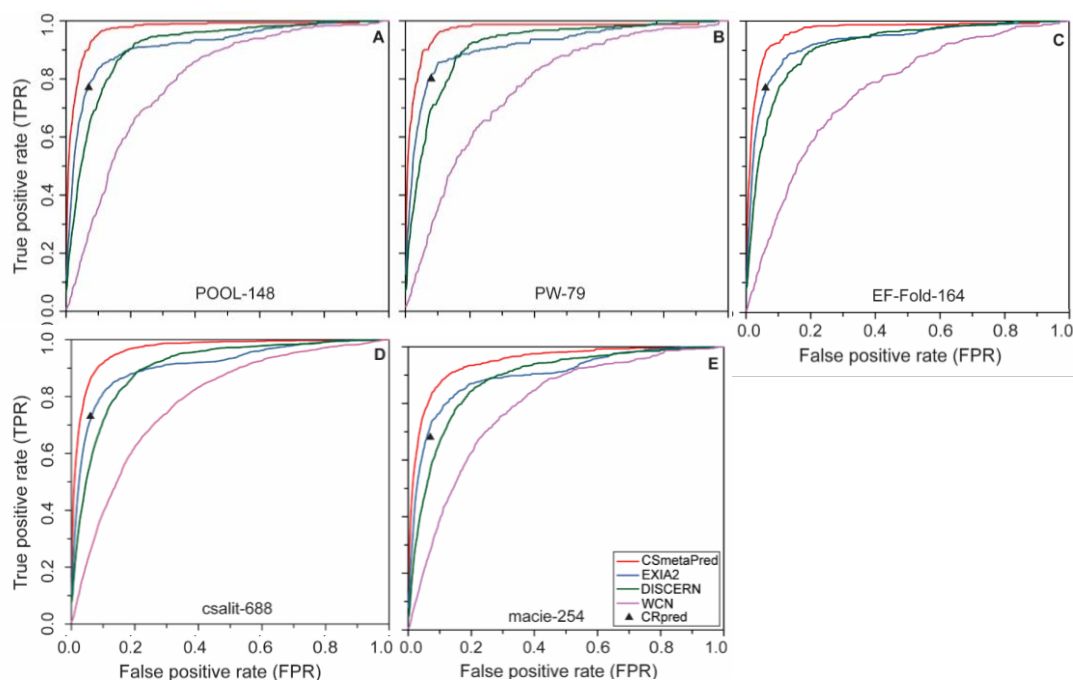


Figure 3.8 Average ROC curves to show comparison among various predictors (EXIA2, DISCERN and WCN) on POOL-148, PW-79, EF-Fold-164, csalit-688 and macie-254 datasets. CRpred SVM performance is shown as filled triangle.

As the results are consistent among various datasets, here onwards results will only be shown for CSAMAC and EFPOOLPW only (Figure 3.8). The results for other datasets are provided in chapter 3 supplementary material (*c3.1_figures.pdf*) provided in a CD with this thesis.

3.3.2.2 PR curve comparisons of CSmetaPred and its constituent methods

PR curve is an alternative to ROC curve for method evaluation when datasets have large skew in total numbers of negative with positive counts. It is a curve between recall and precision on x-axis and y-axis respectively. We compared the PR curves of CSmetaPred and its constituent methods, with an intention to evaluate how well the given predictor classified the positives unlike ROC curves, which also considers mis-classification of the negatives. Figure 3.9 shows that CSmetaPred has higher precision values at any recall values compared to its constituent methods. Further it leads its constituent methods in terms of other quantitative metrics of AUCPR and MAP, having the highest values of 0.324 and 0.489 respectively for CSAMAC dataset (Table 3.2). The performance is consistently observed in other datasets (Figure 3.9).

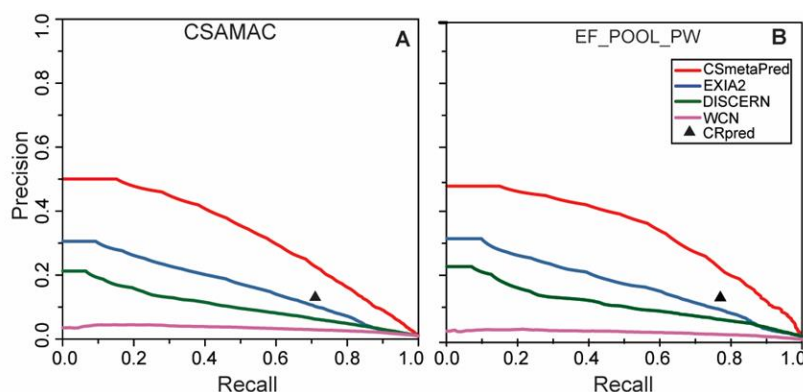


Figure 3.9 Average PR curves to show comparison among various predictors (EXIA2, DISCERN and WCN) on A) CSAMAC and B) EF_POOL_PW. *CRpred SVM* performance is shown as filled triangle.

In order to find if the difference in the performance of CSmetaPred observed is statistically significant, we compared the distribution of pairwise comparison of AveS (MAS) and AP (MAP), calculated for each pdb entry in a given dataset, from two methods using Wilcoxon signed-rank test. The p-value were < 0.0001 for all the datasets

indicating that the differences observed in performance of predictors are statistically significant.

As seen in Table 3.2, CSmetaPred also has lowest median and mean rank for catalytic residues compared to its constituent methods. Thus, CSmetaPred not only shows best performance in terms of ROC and PR curves, but also improves the rank of the catalytic residues in comparison to its constituent methods.

Further, we compared catalytic residue predicted ranks from CSmetaPred to their best possible ranks derived from all methods. Here, the best possible rank is the minimum of ranks assigned to residues in five scores. Such comparison of the best possible and predicted ranks will provide an upper bound of meta-approach performance. For this analysis, we used CSAMAC dataset with 2912 catalytic residues. With respect to the best possible rank, CSmetaPred showed improved or no change in ranked positions for 42.9% of catalytic residues. Of these, most (79.42%) residues have the best possible rank ≤ 10 and the mean and median decrease in CSmetaPred predicted ranks with respect to the best possible rank are 3.3 and 1.0 respectively. CSmetaPred shows a marginal decrease (better) in ranks from the best possible scenario for residues ranked ≤ 10 . The best possible rank of most (80.73%) catalytic residue is ≤ 10 . Of these, CSmetaPred predicted ranks of 57.80% of catalytic residues are higher (poorer) than the best possible rank. However, this does not drastically increase rank (decrease performance with respect to best) as the mean and median increase in rank is 7.8 and 3 respectively. The detailed analysis of cases with large increase in CSmetaPred ranks showed that in most instances only one or two methods have a high residue scores, whereas other methods scores are relatively low, which lead to a decrease in the meta-score with subsequent increase in their ranked position.

3.3.2.3 Comparison of catalytic site prediction performance for polar and non-polar residues

Since 90% of the catalytic residues are polar/charged (Bartlett et al. 2002) with only small fraction of non-polar residues classified as catalytic. Hence, prediction of non-polar residues is rather difficult. We have assessed the performance of meta-predictor for polar and non-polar catalytic residues separately. In this analysis, any amino-acid having functional side-chain atom as polar/charged set, which consists of 12 amino acids namely

(R, N, D, C, Q, E, H, K, S, T, Y) and tryptophan as defined previously by EXIA. The remaining 8 amino-acids (P, F, A, V, I, L, M and G) constituted non-polar set. Table 3.3A shows that CSmetaPred outperforms its constituent methods with highest values of MAS, MAP, AUCPR and AUCROC and lowest values for average and median rank of polar set of catalytic residues. Further, the consistent best performance of CSmetaPred is apparent from Figure 3.10. Similarly, CSmetaPred is the best performing method for non-polar amino acids. (Figure 3.11 and Table 3.3B. More importantly, the performance differences between CSmetaPred and its constituent methods by pairwise comparison of AveS/AP on per protein basis is found to be statistically significant (p -value < 0.0001 using Wilcoxon signed-rank sum test) for both the polar and non-polar set of residues. These analyses suggest that improve the performance of catalytic site prediction in case of both polar as well as non-polar set of residues.

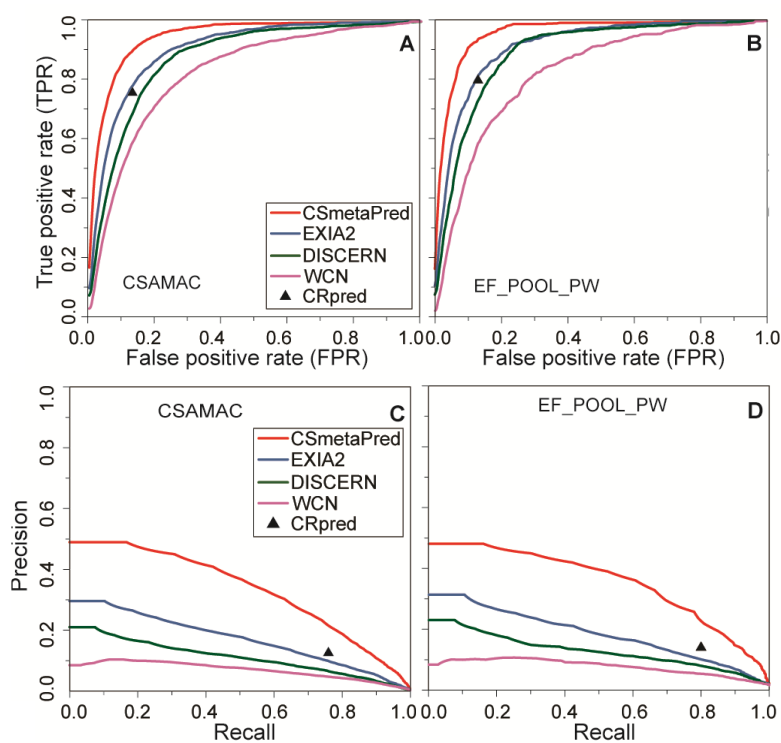


Figure 3.10 Average ROC (A-B) and PR(C-D) curves to show comparison among various predictors (EXIA2, DISCERN and WCN) on CSAMAC and EF_POOL_PW datasets for polar residues. CRpred SVM performance is shown as filled triangle.

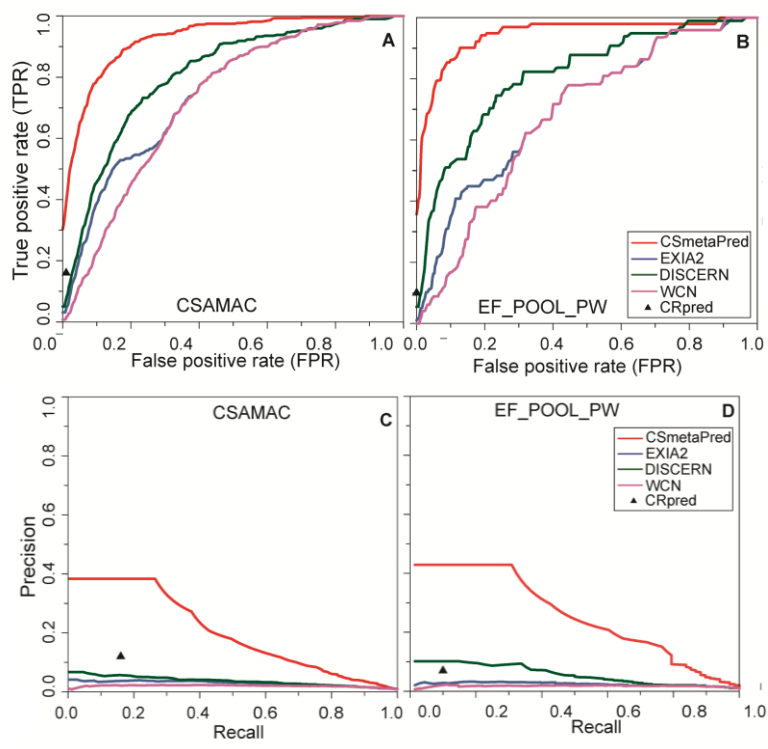


Figure 3.11 Average ROC (A-B) and PR(C-D) curves to show comparison among various predictors (EXIA2, DISCERN and WCN) on CSAMAC and EF_POOL_PW datasets for non-polar residues. CRpred SVM performance is shown as filled triangle.

Table 3.3 Comparison of ROC/PR curves quantitative measures when only (A) polar/charged amino acids and (B) non-polar amino acids are ranked. Quantitative measure of ROC is AUCROC and MAS, whereas PR curves are compared using AUCPR and MAP. Median and average ranks of catalytic residues are also summarized.

A) Polar/charged amino acids

Method	AUCROC	AUCPR	MAS	MAP	Median rank	Average rank
CSAMAC Polar dataset (873 protein)						
CSmetaPred_poc	0.958	0.371	0.961	0.545	5.0	8.0
CSmetaPred	0.95	0.347	0.953	0.519	5.5	9.4
EXIA2	0.908	0.185	0.911	0.343	10.5	16.9
CRpred	--	--	--	--	11.0	15.7
DISCERN	0.879	0.126	0.883	0.265	15.0	22.7
WCN	0.829	0.084	0.832	0.186	20.3	28.1
EF_POOL_PW Polar dataset (286 protein)						
CSmetaPred_poc	0.964	0.382	0.967	0.555	4.9	7.3
CSmetaPred	0.959	0.354	0.962	0.528	5.1	8.3
EXIA2	0.916	0.184	0.919	0.349	9.8	16.4
CRpred	--	--	--	--	10.0	14.3

DISCERN	0.894	0.129	0.898	0.268	15.0	20.0
WCN	0.825	0.078	0.828	0.170	21.4	29.9
POOL-148 Polar dataset (148 protein)						
CSmetaPred_poc	0.966	0.446	0.968	0.601	4.5	6.8
CSmetaPred	0.961	0.420	0.963	0.577	4.9	7.6
EXIA2	0.914	0.207	0.917	0.364	9.8	15.6
CRpred	--	--	--	--	10.0	14.0
DISCERN	0.888	0.133	0.892	0.266	16.5	20.0
WCN	0.840	0.085	0.844	0.175	20.5	26.4
PW-79 Polar dataset (79 protein)						
CSmetaPred_poc	0.962	0.479	0.964	0.630	3.7	6.8
CSmetaPred	0.958	0.465	0.960	0.615	4.0	7.2
EXIA2	0.909	0.235	0.913	0.400	7.7	14.6
CRpred	--	--	--	--	10.0	13.2
DISCERN	0.889	0.148	0.893	0.292	12.5	17.5
WCN	0.824	0.087	0.828	0.185	18.5	25.0
EF-Fold-164 Polar dataset (164 protein)						
CSmetaPred_poc	0.960	0.342	0.962	0.526	5.0	8.3
CSmetaPred	0.955	0.315	0.957	0.498	5.5	9.5
EXIA2	0.915	0.172	0.919	0.338	9.6	16.6
CRpred	--	--	--	--	9.5	13.9
DISCERN	0.891	0.124	0.894	0.266	13.8	21.3
WCN	0.807	0.073	0.811	0.162	22.8	33.9
macie-254 Polar dataset (251 protein)						
CSmetaPred_poc	0.954	0.365	0.957	0.522	6.2	9.5
CSmetaPred	0.942	0.338	0.945	0.494	7.0	11.7
EXIA2	0.904	0.193	0.908	0.334	13.0	18.5
CRpred	--	--	--	--	13.8	18.8
DISCERN	0.872	0.122	0.876	0.240	18.5	25.9
WCN	0.834	0.095	0.837	0.188	22.5	28.3
csalit-688 Polar dataset (679 protein)						
CSmetaPred_poc	0.962	0.389	0.965	0.565	4.5	7.1
CSmetaPred	0.955	0.365	0.958	0.538	5.0	8.3
EXIA2	0.910	0.187	0.913	0.350	9.5	16.2
CRpred	--	--	--	--	10.0	14.6
DISCERN	0.883	0.129	0.886	0.275	14.0	21.5
WCN	0.830	0.082	0.834	0.190	19.5	27.5
UB-137 Polar dataset (136 protein)						
CSmetaPred_poc	0.968	0.498	0.971	0.650	4.0	5.6
CSmetaPred	0.962	0.462	0.966	0.611	4.5	6.2
EXIA2	0.906	0.241	0.910	0.406	9.0	13.4
CRpred	--	--	--	--	7.3	11.5
DISCERN	0.887	0.177	0.891	0.334	12.2	19.4
WCN	0.830	0.108	0.834	0.220	17.9	24.2

B) Non-polar amino acids

Method	AUCROC	AUCPR	MAS	MAP	Median rank	Average rank
CSAMAC Non-polar dataset (193 protein)						
CSmetaPred_poc	0.951	0.265	0.953	0.537	3.3	8.5
CSmetaPred	0.929	0.204	0.931	0.465	4.5	11.8
EXIA2	0.753	0.029	0.757	0.111	25.0	37.3
CRpred	--	--	--	--	13.0	18.9
DISCERN	0.808	0.037	0.812	0.150	21.0	30.7
WCN	0.725	0.020	0.729	0.068	32.3	42.0
EF_POOL_PW Non-polar dataset (49 protein)						
CSmetaPred_poc	0.953	0.301	0.955	0.588	2.0	6.3
CSmetaPred	0.945	0.260	0.948	0.537	3.0	7.7
EXIA2	0.706	0.024	0.710	0.090	30.0	38.1
CRpred	--	--	--	--	11.0	15.2
DISCERN	0.812	0.057	0.816	0.195	15.5	30.9
WCN	0.678	0.018	0.682	0.055	37.0	43.6
POOL-148 Non-polar dataset (29 protein)						
CSmetaPred_poc	0.974	0.353	0.976	0.641	1.5	4.6
CSmetaPred	0.971	0.328	0.973	0.611	2.0	4.8
EXIA2	0.725	0.029	0.729	0.111	26.0	34.9
CRpred	--	--	--	--	8.0	12.4
DISCERN	0.826	0.065	0.830	0.214	10.0	28.5
WCN	0.694	0.020	0.699	0.058	32.0	40.6
PW-79 Non-polar dataset (18 protein)						
CSmetaPred_poc	0.986	0.401	0.987	0.656	1.3	3.9
CSmetaPred	0.982	0.390	0.984	0.645	1.3	3.6
EXIA2	0.689	0.020	0.694	0.079	27.5	39.7
CRpred	--	--	--	--	10.5	13.6
DISCERN	0.868	0.060	0.873	0.189	8.5	29.7
WCN	0.673	0.016	0.678	0.061	28.5	43.1
EF-Fold-164 Non-polar dataset (27 protein)						
CSmetaPred_poc	0.938	0.291	0.941	0.572	3.0	7.4
CSmetaPred	0.929	0.235	0.932	0.503	3.0	9.6
EXIA2	0.676	0.020	0.680	0.085	37.5	45.0
CRpred	--	--	--	--	15.5	16.7
DISCERN	0.792	0.044	0.795	0.160	21.0	35.5
WCN	0.645	0.014	0.649	0.043	43.0	50.4
macie-254 Non-polar dataset (68 protein)						
CSmetaPred_poc	0.945	0.179	0.948	0.417	6.0	10.6
CSmetaPred	0.902	0.140	0.905	0.354	8.5	18.4
EXIA2	0.783	0.031	0.786	0.113	22.0	36.9
CRpred	--	--	--	--	15.8	24.3
DISCERN	0.771	0.031	0.775	0.118	28.0	39.0
WCN	0.752	0.022	0.756	0.066	33.3	40.6
csalit-688 Non-polar dataset (133 protein)						

CSmetaPred_poc	0.951	0.333	0.954	0.610	2.0	7.6
CSmetaPred	0.941	0.267	0.944	0.543	3.0	8.6
EXIA2	0.738	0.028	0.742	0.108	25.0	36.4
CRpred	--	--	--	--	9.0	15.5
DISCERN	0.824	0.042	0.828	0.173	16.0	26.7
WCN	0.709	0.019	0.714	0.063	31.0	42.0
UB-137 Non-polar dataset (33 protein)						
CSmetaPred_poc	0.938	0.214	0.941	0.452	4.3	7.1
CSmetaPred	0.922	0.162	0.926	0.366	5	8.9
EXIA2	0.716	0.028	0.721	0.088	26.0	31.0
CRpred	--	--	--	--	11.0	13.7
DISCERN	0.778	0.045	0.783	0.168	17.0	28.5
WCN	0.691	0.022	0.696	0.064	31.0	35.7

3.3.2.4 Comparison of CSmetaPred and CSmetaPred_poc

As discussed earlier in section 3.3.1, catalytic residues are known to be spatially proximal to substrate/cofactor binding sites, we evaluated whether including predicted pocket residue information could increase prediction accuracy. For this, we have developed meta-approach CSmetaPred_poc, which combines meta-score with pocket score (poc-Rsc) harboring information of combined predicted binding pockets from Fpocket (Le Guilloux, Schmidtke, and Tuffery 2009) and LIGSITE (B. Huang and Schroeder 2006) (see methods). We compared the performance of CSmetaPred and CSmetaPred_poc using both ROC and PR curves. Visual comparison of both ROC (Figure 3.12A-) and PR (Figure 3.12B) curves showed that CSmetaPred_poc performs better compared to the CSmetaPred. Further based on quantitative metrics – AUCROC, MAS, AUCPR and MAP are the highest for CSmetaPred_poc. In fact, it achieves the lowest (best) 12 and 6 as mean and median rank of catalytic residue (Table 3.2).

The performance of CSmetaPred_poc is found to be statistically significant (p -values < 0.0001 using paired Wilcoxon signed-rank test) better than CSmetaPred, when we considered the statistical significance of pairwise differences in AveS/AP calculated for each protein between these two meta-predictors. Thus, CSmetaPred_poc is able to take advantage of spatial proximity of binding and catalytic sites to improve prediction performance. For example, the catalytic residues viz. H334, Y95, S550, and P108 of rat

choline acetyltransferase (pdbid: 1q6x chain B) are ranked at position 1, 6, 8, and 27 respectively by CSmetaPred_poc (Figure 3.13).

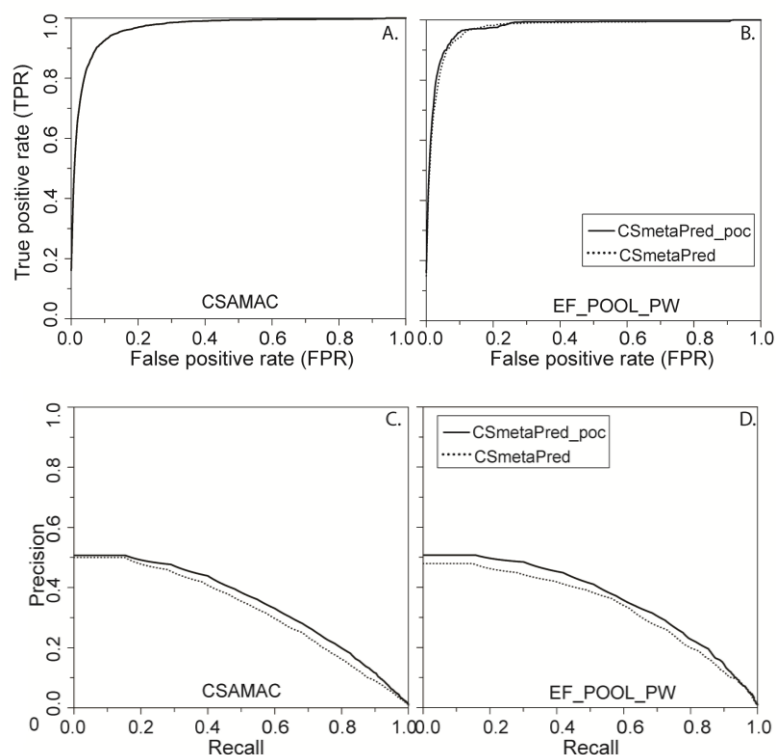


Figure 3.12 Average ROC (A-B) and PR(C-D) curves to show comparison among CSmetaPred_poc and CSmetaPred on CSAMAC and EF_POOL_PW.

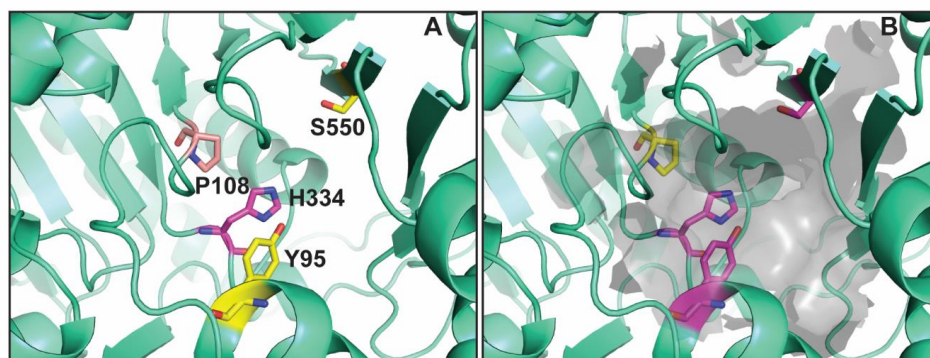


Figure 3.13 Comparison of prediction results for enzyme rat choline acetyltransferase (PDB: 1q6xB) from CSmetaPred (A) and CSmetaPred_poc (B) after including pocket information. Tertiary structure and known catalytic residues are shown in cartoon and licorice representations respectively. Catalytic residues are colored based on their meta-predictor predicted ranks: magenta for residues with rank ≤ 5 , yellow for rank > 5 and ≤ 10 and pink for rank > 20 . Top pocket ranked by pocket score is shown in gray transparent surface representation.

As these residues are present in the top re-ranked merged pockets (see methods), when the pocket information was included in the meta-predictor CSmetaPred_poc, their ranked were improved to positions 1, 2, 3, and 11 respectively.

3.3.2.5 Prediction performance for enzymes with single or multiple catalytic residues

Mostly, there are more than one catalytic residue identified for an enzyme. However, in some cases, only one catalytic residue is defined. Since predicting only one catalytic residue is challenging task, we assessed performance of CSmetaPred_poc on enzymes have single (SS dataset) and multiple catalytic residues (MS dataset). As has been observed before, CSmetaPred_poc achieves high TPR at any given FPR for both SS and MS enzyme datasets shown in Figures 3.14 and 3.15 respectively.

Similarly, it is evident in PR curves that meta-predictor outperforms its constituent methods for both multiple residue catalytic site and single residue catalytic site. This implies that meta-predictors, in general, can predict catalytic site irrespective of the number of catalytic residues present in them.

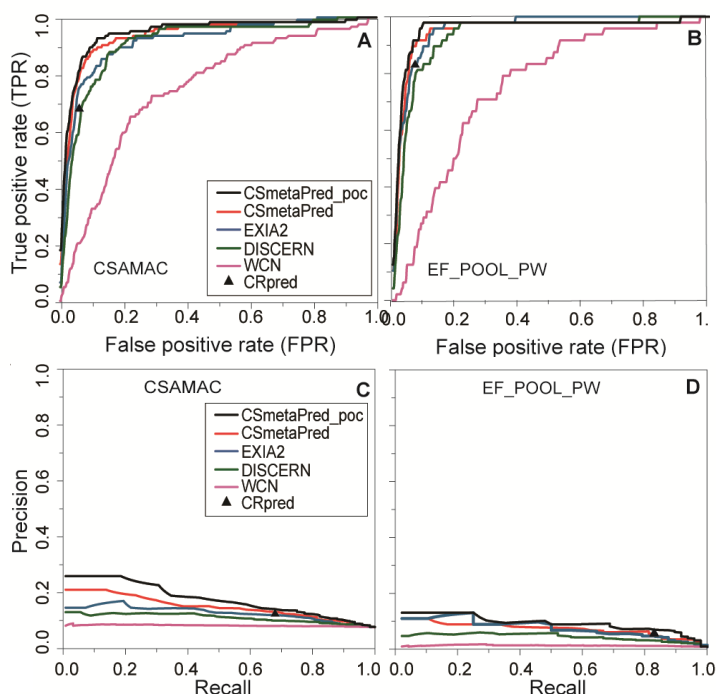


Figure 3.14 Average ROC (A-B) and PR(C-D) curves to show comparison among various predictors on CSAMAC and EF_POOL_PW datasets for single residue catalytic site (SS) dataset.

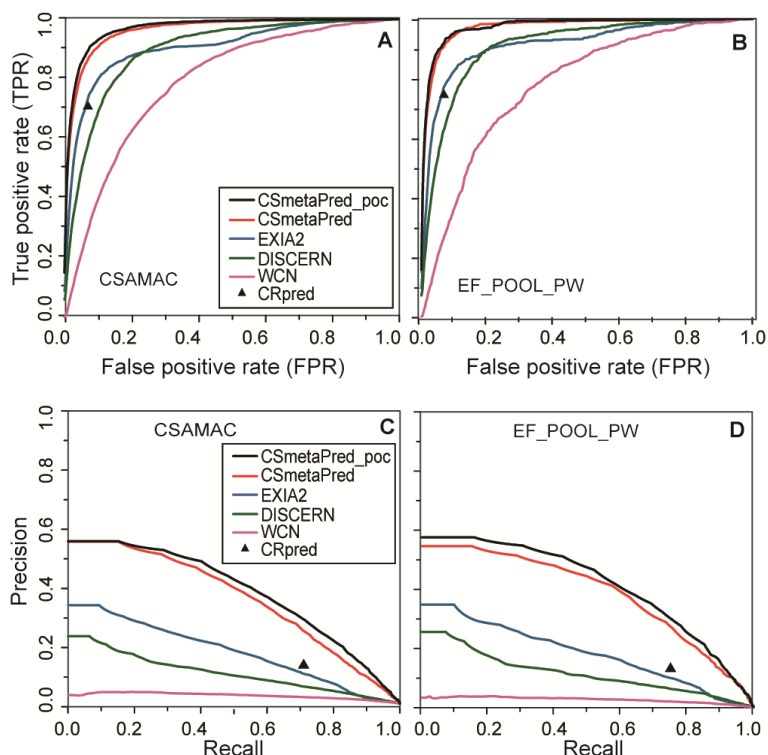


Figure 3.15 Average ROC (A-B) and PR(C-D) curves to show comparison among various predictors on CSAMAC and EF_POOL_PW datasets for multiple residue catalytic site (MS) dataset.

3.3.2.6 Effect of ligand binding on the prediction by CSmetaPred_poc

As CSmetaPred_poc exploits the pocket information in giving additional preference to the catalytic residues present in the binding pocket, and uses the information of predicted binding pocket, it is worth to check if there is any bias due to ligand bound structures in the dataset in the prediction performance of CSmetaPred_poc. For this analysis, we constructed UB-137 dataset (as discussed in the methods section), in which no ligand is bound to any structure and all the structures present in their apo form. The detailed analysis of average ROC and PR curves show that CSmetaPred_poc is still the best performing method (Figure 3.16). CSmetaPred_poc achieves MAS and MAP values on UB-137 dataset of 0.976 and 0.620 respectively (Table 3.2). Thus, CSmetaPred_poc performs similarly for UB-137 dataset compared to other benchmarking dataset. Hence, most likely there is no bias in the prediction due to ligand bound to the structure of query enzyme.

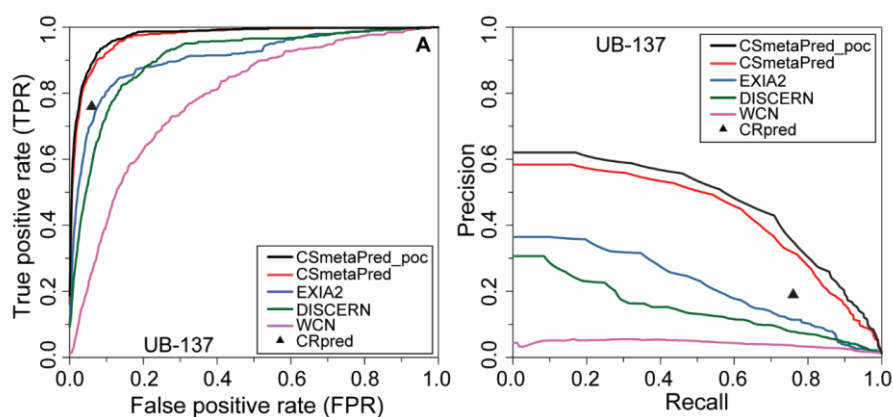


Figure 3.16 Average ROC (A) and PR (B) curves to show comparison among various predictors (EXIA2, DISCERN and WCN) on UB-137 dataset. *CRpred SVM performance is shown with filled triangle.*

3.3.2.7 Comparison of CSmetaPred_poc with previously developed other classifiers

We compared the performance of CSmetaPred_poc with previously developed classifiers, which have computed their results on exactly the same dataset as that used in our study. However, as described in the methods, we excluded some of the obsolete entries while compiling legacy datasets. Among all the three legacy datasets, only PW-79 dataset remained unchanged and hence such a comparison can be done only on this dataset. On PW-79 dataset, Cilia and Passerini method achieves average recall and precision of 0.46 and 0.28 respectively (Cilia and Passerini 2010). With the same dataset, at a recall of 0.46 CSmetaPred_poc has precision of 0.54 and at a precision 0.28 it has recall of 0.87. CRpred achieves average recall of 0.54 and precision of 0.175 on PW-79 dataset (T. Zhang et al. 2008a). CSmetaPred_poc achieves a precision of 0.50 at same recall of 0.54 and a recall of 0.94 at same precision of 0.175.

3.3.2.8 Assessment of catalytic residue rank by meta-predictors

In order to compare the catalytic residue, rank of various predictors used in this study, we used mean and median catalytic residue rank as a metric. Both meta-predictors achieve lower (better) median/mean rank in comparison to other methods across all datasets (Table 3.2 and Table 3.3). In fact, CSmetaPred_poc achieves the lowest catalytic residue median rank of 6. The same is observed when either polar/charged or non-polar residues are ranked separately.

We are not committed to any specific cut-off to select active site residues. However, to prioritize residues for experimental studies, we analyzed two different criteria a) select top n percent of residues from ranked list; and b) select top m ranked residues. In the first criteria, we select top n percent of residues as true positives referred as filtration ratio, which is calculated at varying rank cut-off. We calculated mean recall value for every filtration ratio value and plotted them as Recall Filtration Ratio (RFR) curve. As seen from Figure 3.17, at any given filtration ratio, CSmetaPred_poc achieves higher recall compared to CSmetaPred. For instance, taking 5% of residues from the ranked list give an average recall of 0.83 and 0.80 for CSmetaPred_poc and CSmetaPred respectively. The same is also observed for other datasets.

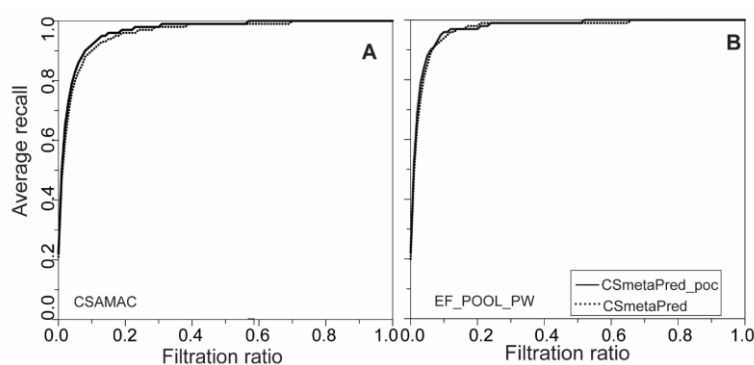


Figure 3.17 Recall filtration curve for a) CSAMAC and b) EF_POOL_PW datasets.

In the second criteria, we calculated the fraction of proteins having at least 0.5, 0.8 and 1.0 catalytic site coverage (fraction of known catalytic residue predicted by CSmetaPred_poc at a given rank cut-off) at varying rank cut-offs. It is clearly evident from Figure 3.18 that there is rapid increase in number of enzymes with increasing rank, which reaches at plateau around rank 30.

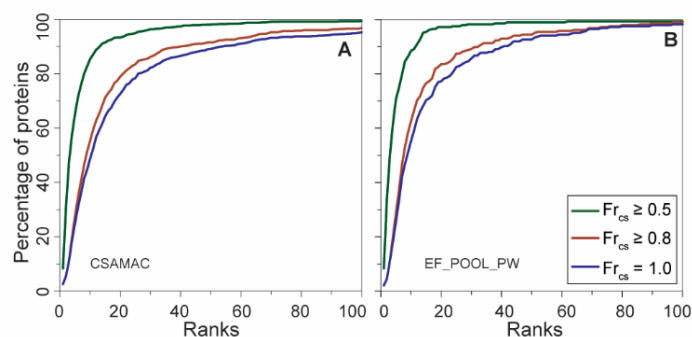


Figure 3.18 Cumulative fraction of proteins (shown in percentage) having catalytic residue coverage of at least 0.5, 0.8 and 1.0 calculated at ranks ≤ 100 for A) CSAMAC and B) EF_POOL_PW.

Interestingly, at rank ~30 all catalytic residues are identified in ~82% of enzymes. Moreover, at lower ranks, such as within rank 20 CSmetaPred_poc correctly predicts $\geq 50\%$ of catalytic residues for ~93% of proteins and all catalytic residues for ~73% of proteins. Moreover, in most enzymes more than 50% catalytic residues are within top 20 ranks in CSmetaPred_poc. This is consistently observed in individual datasets as well (Figure 3.S18 provided in supplementary material for chapter 3).

Based on the above analyses and average precision, average recall, and average accuracy from all datasets, we suggest residues with ranks ≤ 20 or ranks $\leq 4\%$ filtration ratio cut-off as catalytic residues. On CSAMAC dataset, with 4% filtration ratio cut-off CSmetaPred_poc achieves the average precision, recall, and accuracy of 0.2, 0.79, and 0.96 respectively. Using same dataset and method the average precision, recall and accuracy with rank 20 are 0.14, 0.87, and 0.94 respectively. Similar average values are observed in other datasets. These cut-off values are to be used as an indicator rather than a rule to predict catalytic residues. Using these criteria, CSmetaPred_poc is able to rank ~87% and ~76% of known catalytic residues within top 20 ranks and 4% filtration ratio respectively. This selection criterion can be used to prioritize the residues for experimental studies.

3.3.2.9 Prediction performance of meta-predictor on modeled structures

As the experimental structure for many enzymes is still not yet known, next, we evaluated if the CSmetaPred_poc prediction could be reliably used for prediction of catalytic site in homology-based models. Previous comparison showed that usefulness of modeled structures in catalytic site prediction and has also suggested that low quality predicted structures could be used for catalytic residues prediction (Carbajo and Tramontano 2012). Here, we compared prediction performance based on sequence to structure/sequence based prediction, where structure can either be a native/modeled structure.

In order to evaluate the quality of the model, we computed the Root Mean Square Deviation (RMSD) between the native and its corresponding model structure. The average RMSD between the native and its corresponding model was found to be 3.0 Å.

Most of these high RMSD cases were from low sequence identity bin of 40-50%. The analysis of model cases with large RMSD usually involved cases with a long N/C terminal region or part of query sequences without any template aligned regions. In some extreme cases, there was a large conformational change observed between template and native structure. Both visual and quantitative comparison of ROC/PR curves showed that the prediction using model structure as an input is comparable to their corresponding native structure (Figure 3.19). Moreover, as seen in Table 3.4, the median rank of model dataset is slightly higher (8.4) in comparison to their corresponding native structure (6.3).

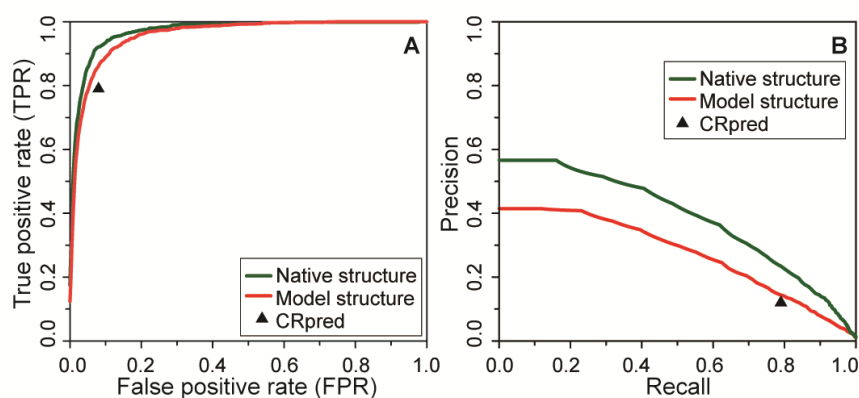


Figure 3.19 Average ROC (A) and PR (B) curves to show comparison of CSmetaPred_poc performance for model and their corresponding native structure dataset. CRpred SVM performance is shown with filled triangle.

Table 3.4 Summary of quantitative analysis of ROC and PR curves using AUCROC/MAS and AUCPR/MAP respectively, for CSmetaPred_poc prediction on model and native structures.

Method	AUCROC	AUCPR	MAS	MAP	Median rank	Average rank
Modeled structures (468 structures)						
CSmetaPred_poc	0.959	0.274	0.961	0.441	8.4	15.2
Native structures (468 structures)						
CSmetaPred_poc	0.971	0.386	0.973	0.542	6.3	10.6

Detailed analysis showed that poor performing models usually belonged to low sequence identity bin of 40-50%. As seen in Figure 3.19, in comparison to CRpred, which is only sequence-based method, the CSmetaPred_poc performs better than CRpred. In fact, the mean and median rank of catalytic residue of CRpred is higher which is 11 and

23.2 respectively. This implies that, the CSmetaPred_poc is indeed gives a better prediction performance compared to the predictor like CRpred, where only sequence based information is utilized for catalytic site prediction. Further, the modeled structure could be used for reliable prediction from CSmetaPred_poc in case of unavailability of native/experimentally solved structure for query enzyme.

3.3.2.10 CSmetaPred: evaluation as a meta-predictor

Given that meta-predictor combines results from several methods, it can be expected to perform the best among its underlying methods for all proteins. However, for cases when only one method outperforms rest all meta-predictor will not be able to achieve the best performance. This is essentially because consensus of outputs from methods with poor prediction except one is going to be lower than the best output. For instance, a residue may not show improved ranked positions when it gets a high score only from one method whereas rest other methods assign low residue scores. Importantly, meta-predictor may not be able to achieve theoretically the best possible ranked positions for all catalytic residues, it results in most catalytic residues within top 20 ranks. Considering percentage of catalytic residues within top 20 ranks, CSmetaPred is able to rank ~83% of residues in comparison to ~87% of residues taking the best rank (not including CATSID ranks). Thus, the meta-predictor does show improvement in prediction considering consensus among methods over its constituent methods.

In order to understand the underlying reason of improved performance of meta-approach, we compared CSmetaPred predicted rank for every catalytic residue to the best possible rank from any of its constituting methods. The 'best possible rank' is a theoretical best scenario for selecting ranks for catalytic residue and this provides an upper bound of meta-approach performance. In this analysis, the best possible rank for a given catalytic residue is the minimum ranks assigned to it from any one of methods- be it CRpred, DISCERN and EXIA2. We have not included CATSID in this analysis as we could only rank subset of residues (see methods) for which their corresponding templates match were found as hit. Hence, residues might be ranked lower because of inability to rank all residues, which may not imply better performance necessarily. This analysis was done using CSAMAC dataset, which possess 2912 catalytic residues.

The performance of CSmetaPred will be majorly affected by low ranked catalytic residues, hence we analyzed catalytic residues having the best possible rank less than 20. Infact, most (86.9%) of catalytic residues have the best possible rank ≤ 20 . Of these, for ~51% of catalytic residues CSmetaPred predicted ranks are either unchanged or improved marginally having median and mean decrease in rank of 2 and 3.4 respectively. Further, CSmetaPred predicted ranks are higher (poorer) for ~49% of catalytic residues compared to the best possible rank. Importantly, the increase in CSmetaPred predicted rank is not large as evident from median and mean rank increases of 3 and 7.6 respectively. The detailed analysis of catalytic residues with increase in CSmetaPred ranks showed that in most instances these residues are predicted only by one or two methods, which is exhibited in their higher normalized scores, whereas other methods assign lower residue scores as predictions from other methods are not good (Table 3.5). This indicates that even though meta-predictor is not able to achieve the best possible scenario in meta-approach, it does not decrease ranks of catalytic residues drastically from the best possible scenario.

Table 3.5 Illustrates examples where only one method predicts catalytic residue within top 25 ranks (highlighted in bold), whereas other methods comparatively rank these residues higher (worst). CSmetaPred still ranks these residues lower (better).

S.No.	PDB	Catalytic residue (CS)	EXIA2 rank of CS	CRpred Rank of CS	Discern rank of CS	WCN rank of CS	CSmetaPred rank of CS
Polar residues							
1.	1reqA	K604	67	52	22	120	8
2.	1rblA	D203	1	101	94	110	9
3.	1ofdA	K972	2	92	385	53	4
4.	2pdaA	N996	25	103	360	91	10
5.	1ohhC	K209	14	51	56	287	9
6.	2c3mA	N996	15	103	227	90	8
Non-polar residues							
1.	1pjhB	A70	138(--)	27	88	55	9
2.	2tplA	F123	300(--)	23	70	116	4
3.	1dd8A	F392	290(--)	52	32	175	10
4.	1ecfA	G102	266(--)	93	168	3	10
5.	1cgkA	F215	225(--)	87	59	40	7

Further, we also did critical evaluation of CSmetaPred ranking performance with respect to the best possible rank has been performed by deriving best rank from all the methods i.e. EXIA2, DISCERN, CRpred and CATSID. Similar criteria were followed as stated above earlier for this analysis. Importantly, most (94.8%) of catalytic residues have the best possible rank ≤ 20 . Of these residues, ~26% and ~74% of catalytic residues showed no change/decrease (better performance) and increase (poor performance) in ranks with respect to the best rank respectively. For ~74% of residues with higher (poorer) ranks than the best possible rank cases, these do not have large increase in ranks as exhibited by mean and median increase in rank of 10.1 and 4 respectively. The detailed analysis of cases with large increase in CSmetaPred ranks showed that in most instances only one or two methods have a high residue scores, whereas other methods scores are relatively low, which lead to a decrease in the meta-score with subsequent increase in their ranked position.

As discussed before, the best possible rank analysis provides the upper bound of meta-approach implemented in CSmetaPred. We analyzed CSmetaPred predicted ranks of some catalytic residues, which show large increase (poor predicted CSmetaPred ranks) or decrease (better ranks from CSmetaPred) in ranked positions with respect to the best possible rank. One of the catalytic residues (LYS-150) of enzyme phosphatidylinositol phosphate kinase (1b01B) is ranked at 28, 17, 25 and 17 by EXIA2, CRpred, WCN and DISCERN respectively (CATSID did not provide rank for this residue) that improves to rank 8 by CSmetaPred. Even though LYS-150 is not top residue in all methods, it is among top ranked residues and has consistent normalized residue scores of 1.1230, 2.4744, 1.4825 and 1.8339 from EXIA2, CRpred, WCN and DISCERN respectively. An example of increase in CSmetaPred predicted ranks is residue LYS-591 from enzyme isoleucyl-trna synthetase (1ileA) that is ranked at 120, 96, 569 and 6 by EXIA2, CRpred, WCN and DISCERN respectively. This residue (LYS-591) is ranked at 110 in CSmetaPred predicted ranks, mostly because normalized scores are not consistent and only one method (DISCERN) assigns relatively high scores as shown by normalized scores of 0.3521, 1.0424, -0.5579 and 2.647 from EXIA2, CRpred, WCN and DISCERN respectively. We have provided predicted ranks and scores of all benchmark proteins in our webserver available at <http://14.139.227.206/csmetapred>

In order to explore the extent of contribution of each of its constituent methods in the performance of CSmetaPred, we computed a modified meta-score after excluding one

score at a time. Figure 3.20 shows that all methods contribute to a different extent towards the performance of the CSmetaPred. Maximum decline in performance of modified CSmetaPred is seen upon removing either CRpred or CATSID. Thus, these two methods majorly contribute in enhancing the accuracy of the CSmetaPred prediction.

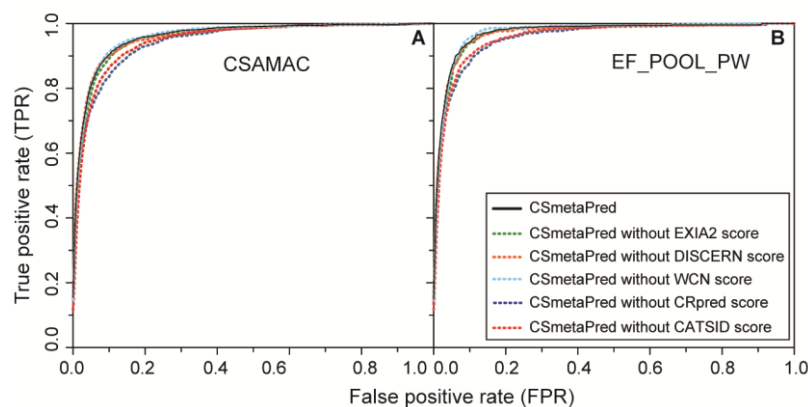


Figure 3.20 Average ROC plots for modified CSmetaPred, wherein one score is excluded from meta-score computation for a) CSAMAC and b) EF_POOL_PW datasets.

3.3.2.11 Catalytic residue prediction for protein structures deposited in PDB after development of CSmetaPred

We analyzed CSmetaPred_poc prediction results for structures, with known catalytic residues, deposited in RCSB PDB (Berman et al. 2000) database subsequent to development of our method. For this analysis, we manually searched PDB database for recently determined tertiary structures of enzymes having curated list of catalytic residues. As seen in Table 3.6 that CSmetaPred_poc. Further, we discuss few interesting examples having best CSmetaPred_poc performance below. β -keto-acid cleavage enzyme family KCE (DUF849) is involved in the anaerobic fermentation of lysine. Recently, H46, H48, E143, R226, D231 are identified as crucial catalytic residues and S82, T106, and E230 as important functional residues (Bastard et al. 2013). Interestingly, CSmetaPred_poc ranks H46, H48, E143, R226, D231, S82, T106, and E230 residues (pdbid: 2y7f), at 2, 4, 6, 1, 5, 14, 17, and 3 ranked positions respectively (Figure 3.21). The experimental site-directed mutagenesis in thioesterase enzyme YbdB from *E.coli* has identified H89, E63, S67, H54, and Q48 as putative catalytic residues (Wu et al. 2014). Using YbdB structure (pdbid: 4k4c), CSmetaPred_poc is able to predict residues H89, E63, S67, H54, and Q48 residues at ranks 1, 2, 4, 5, and 19 respectively.

Table 3.6 Meta-predictor prediction performance on pdb entries, with experimentally known catalytic residues, submitted in RCSB PDB database subsequent to development of meta-approach method. Catalytic residue ranks from CSmetaPred_poc are summarized in this table.

S.no.	PDB	UNIPROT	Catalytic residue	Rank
1	2n6jA	Q183R7	E185	1
			E143	2
			H146	3
			H142	4
			Y178	8
2	2nbqA	Q9UH17	H253	2
			E255	3
			C284	4
			C289	7
3	4ufoA	Q41415	H300	1
			D105	2
			L106	3
			Y235	4
			Y154	9
			F33	10
4	2ruqA	Q13526	H13	1
			H111	2
			A67	3
			K17	8
			Q85	11
			T106	17
			S69	19
5	4zamA	P0AD64	S70	11
			K234	7
6	5b6aA	Q9HT57	D225	1
7	5b6sA	A8NI40	D109	2
			D224	5
			E276	9
8	5cqmX	P16442	C303	1
9	5dn5A	P15931	E184	5
			E223	81
10	5e2jA	Q9AJS0	D143	3
			E515	6
			D146	15
11	5e9eA	P42527	D766	3
12	5c17A	Q7DJN2	C165	1
			S105	2
			C102	3
13	4ywiA	Q07412	E165	4
14	5c0uA	P77072	C159	1
			D99	2
			C96	3
15	5ej3A	P26515	E177	2
			E87	3
16	5g2gA	P07445	D40	1
			Y16	5
17	5gmtA	E7FLQ2	Y142	3

			Y190	9
			K99	13
18	5j9qE	Q12692	Q338	13
19	5jadA	Q13093	H351	6
			S273	8
			D296	13
20	5jmdA	Q89YS4	Y301	1
			H431	3
			N247	6
21	5ccdA	Q9ZMY2	E175	1
			E13	6
22	4wyiA	O81770	H155	2
23	4x22A	Q8F5I5	H97	1
			E169	4
24	4pixA	P21816	H155	1
			Y157	2
			S153	9
25	5idiA	B9K7M5	E164	2
			G349	10
26	5j7xA	B8N653	R337	2
			D63	5
27	5j8cA	Q9H7Z6	S316	1
			Q350	16
28	5lb1A	I6Y9J2	C354	1
			H336	2
29	5kf6A	F7X6I3	C844	5
30	5h38A	F5HBQ9	C219	3
31	5haiA	P05364	G64	18
32	5grrA	A0A0R6L508	T285	21
33	5ezqA	P27282	C477	21
34	5jciA	Q652L6	R320	22
			Y349	49
35	5d0nA	Q195N6	G298	42
36	4z85A	A4UVY1	Y193	74
			C194	13
				7

Glutathione transferase from *Nilaparvata lugens* delta-class is responsible for the intracellular detoxification of diverse xenobiotic and endogenous substances. We used pdbid: 3WYW for prediction and found the ranks of 2, 7 and 52 for its known catalytic residue S11, E66 and H52(Yamamoto et al. 2015).

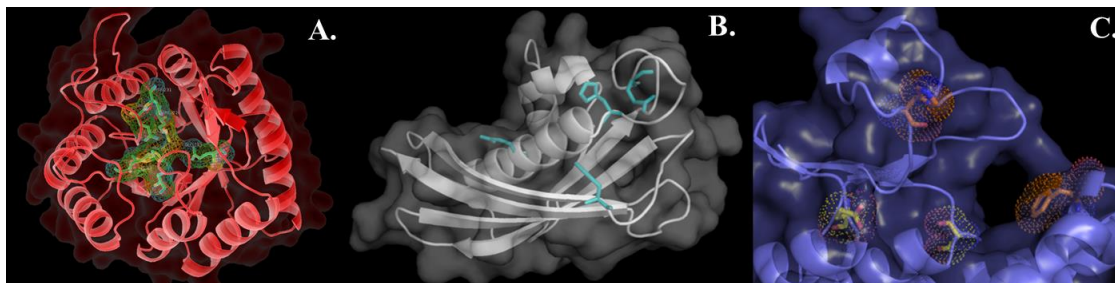


Figure 3.21 Examples of CSmetaPred_poc predictions. Catalytic residues of A) KCE B) YdbB and C) Glutathione transferase from *Nilaparvata lugens* delta-class.

3.3.2.12 Experimental validation of catalytic site prediction for *E. coli* γ -glutamylcysteine ligase (GCL)

The catalytic residues of *E. coli* γ -glutamylcysteine ligase (GCL) are still not known. We used pdbid 1v4g as *E. coli* tertiary structure of GCL as an input to predict catalytic residue by CSmetaPred_poc. The top 20 predicted catalytic residues of *E. coli* GCL by CSmetaPred_poc are shown in Table 3.7.

Among the top 20 predicted residues (Table 3.7), we randomly selected and mutated R330 (rank 1), R235 (rank 11), Y131, (rank 16) and R132 (rank 20) to investigate their role in catalysis using previously described in vivo and in vitro assays (Shailesh Kumar et al. 2013). Preliminary studies show no enzymatic activity for R330A mutant and reduced activity for mutants of R235, R132, and Y131 (Figure 3.22 and Table 3.8) suggesting these could play a role in catalysis. Interestingly, R330 structural equivalent in GCL homologue from *S. cerevisiae* (Sc- γ -GCL) (pdbid: 3ig5) is R472, which has also been suggested to be a catalytic residue (Biterova and Barycki 2009). Further detailed study is required to investigate specific role of R330 during catalysis.

Table 3.7 List of top 20 predicted catalytic residues of γ -glutamylcysteine ligase from *E. coli* (pdbid: 1v4gA) by CSmetaPred_poc.

Rank	Residue name	Residue number
1	ARG	330
2	HIS	150
3	GLU	328
4	ARG	304
5	GLU	29
6	GLU	27
7	LYS	306
8	ASP	60
9	GLU	296
10	ASP	333
11	ARG	235
12	GLU	67
13	ARG	32
14	ASN	297
15	LYS	128
16	TYR	131
17	HIS	44
18	ASN	152
19	TRP	100
20	ARG	132

Table 3.8 Enzyme activity (in vitro) of EcGCL mutants calculated with respect to wild type activity of enzyme.

Mutant	Relative enzyme activity (in% with respect to wild type)
Wild type	100
R330K	Not determined
R330A	0
Y131S	1.3
R132A	0.4

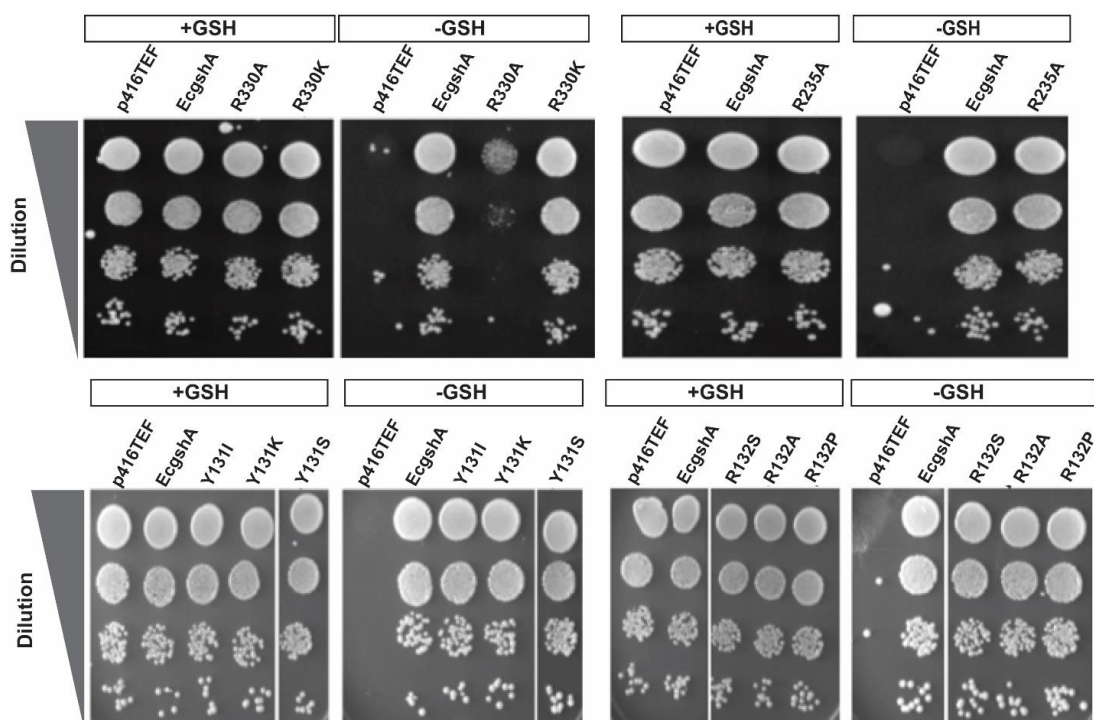


Figure 3.22 In vivo complementation assay of predicted catalytic residues mutants of GCL enzyme. *Saccharomyces cerevisiae* strain ABC1195 plasmids bearing WT GCL or the different cysteine binding residues GCL mutant gene cloned under TEF promoter. The transformants were grown overnight in SD+GSH medium and used to re-inoculate secondary culture. Cells were harvested at OD600 = 0.6 and serially diluted (0.2 to 0.0002 OD600). 10 μ l was spotted on SD medium with or without GSH as sole source of organic sulphur. The vector pTEF416 and EcGCL were used as negative and positive control respectively. (Adapted from Kumar, Shailesh 2014, “Structure-function studies on the γ -glutamylcysteine synthetase enzyme of bacteria”, PhD thesis, submitted to Jawaharlal Nehru University, New Delhi (as a PhD scholar of Institute of Microbial Technology, Sector 39-A, Chandigarh).

3.3.2.13 Availability of meta-predictors are webserver

CSmetaPred and CSmetaPred_poc are provided as a webserver, which is freely accessible at <http://14.139.227.206/csmetapred/> for public use. Initially we relied on EXIA2 server, but due to technical issues in this server we have recoded EXIA2 and optimized parameters to our best ability. We use this in-house recode version of EXIA2 in our webserver. The residues rank comparison between EXIA2 server and in-house recoded EXIA2 can be seen in Figure 3.23. The Pearson correlation coefficient between ranks for all residues obtained from EXIA2 and in-house program is 0.86, and the same for catalytic residues is 0.55. In due course of time, we will recode CATSID and

implement in our server. Hence, in long run we will have all four methods executed locally. We will maintain our server and update as and when required.

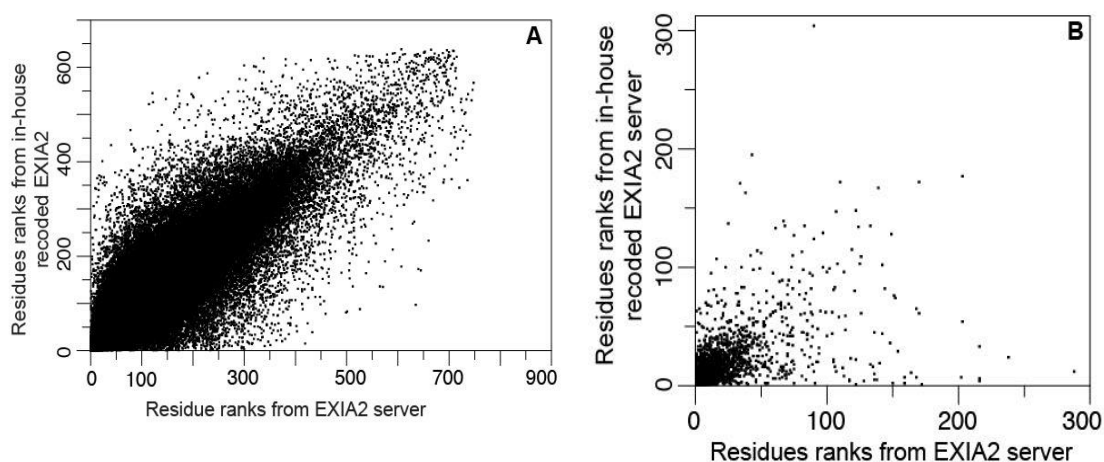


Figure 3.23 Comparison of residue rank from EXIA2 server output and in-house recorded EXIA2 for (A) all residues, and (B) catalytic residues.

3.3.2.14 Comparison of Catalytic site annotation in CSAMAC and M-CSA datasets

As mentioned earlier, M-CSA (Mechanism and Catalytic Site Atlas) is a major update in catalytic site annotation databases, which combines both MACiE and the CSA, providing a unified resource to facilitate the searching of catalytic sites. We compared our CSAMAC dataset with M-CSA dataset, and found total 647 common enzyme structures. M-CSA dataset is more refined and annotates additional catalytic residues for these enzymes. For these 647 enzymes, M-CSA and CSAMAC define total 2869 and 2205 catalytic residues respectively. All the catalytic residues defined in CSAMAC dataset are present in M-CSA dataset. Interestingly, 664 additional catalytic residues of M-CSA are usually present in top 20 ranks of CSmetaPred_poc. As seen in Figure 3.24, the fraction of catalytic residues presents at rank 20 is much higher in case of M-CSA (~83%) dataset compared to that of CSAMAC (~66%) dataset.

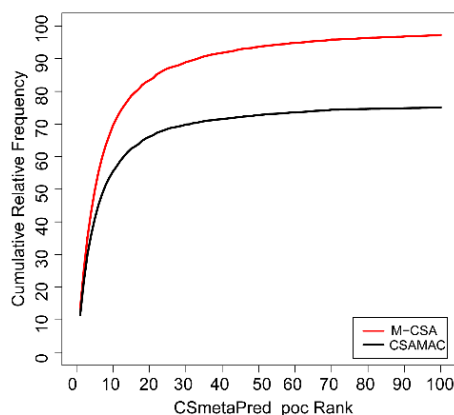


Figure 3.24 Cumulative fraction of catalytic residues present at varying CSmetaPred_poc rank cutoffs for common set of enzymes in CSAMAC and M-CSA datasets.

This analysis shows that, among top 20 predicted ranked residues by CSmetaPred_poc, the residues which were earlier not defined as catalytic residue in CSAMAC dataset, were later found out to be catalytic, when a much more pruned dataset (M-CSA) was made available.

3.3.2.15 Catalytic site prediction for CSA-Homology dataset

Further, we used CSmetaPred_poc to predict the catalytic site for the enzymes in CSA-homology dataset (Furnham et al. 2014) with an intention of enrichment of catalytic site residues in this previously annotated catalytic site dataset. To remove obvious redundancy in the dataset, we culled the PDB chains present in CSA-homology dataset at 95% sequence identity. We predicted the catalytic residues of every pdbchain present in this dataset using CSA-homology. Further, we analyzed the fraction of catalytic residues present in top 20 ranks for each protein present in the dataset. Figure 3.25 shows that on an average 70% of catalytic residues are present in top 20 predicted ranks assigned by CSmetaPred_poc. ~41% of the proteins have 90-100% of catalytic residues within top 20 predicted ranks.

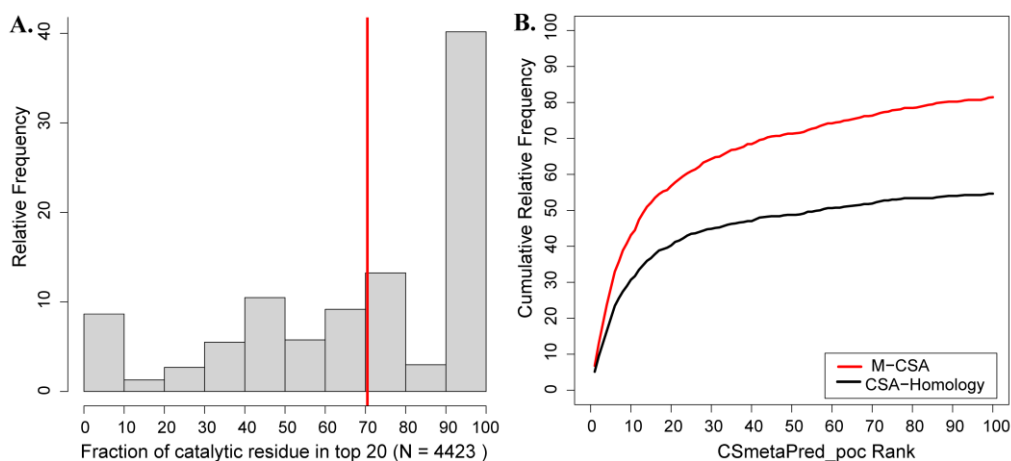


Figure 3.25 A) Relative frequency of proteins with a given fraction of catalytic residues present in top 20 predicted ranked residues by CSmetaPred_poc in CSA-Homology B) Cumulative fraction of catalytic residues present at varying CSmetaPred_poc rank cutoffs for common set of enzymes in CSA-Homology and M-CSA datasets.

Having analyzed that most of the previously defined catalytic residues (~70%) are within top 20 predicted ranks of the CSmetaPred_poc, we analyzed the fate of the remaining residues which are present in our top 20 predicted ranks but not listed in the CSA-homology dataset. We suggest that these residues are either catalytic but are missed by CSA-homology or are functionally important and have role in binding of substrates as well. To analyses if this is correct, we again looked at the common set of enzymes present in CSA-homology and M-CSA datasets. We found total 125 enzymes common between CSA-homology and M-CSA datasets, which define 553 and 798 catalytic residues for these enzymes respectively. Analysis showed that CSA-homology is a subset of M-CSA, which in addition has 245 more catalytic residues. Most (~54%) of these additional catalytic residues have catalytic residue within top 20 predicted ranks and median catalytic rank for these additional catalytic residue is 15. Further we analyzed the fraction of catalytic residues at varying predicted ranks and is shown in Figure 3.25B. As seen in Figure 3.25, the fraction of catalytic residues is more in M-CSA dataset compared to CSA-homology at any given CSmetaPred_poc predicted rank. For instance, at rank 20, 57% and 40% of catalytic residues are present in M-CSA and CSA-homology dataset respectively. These analyses suggest that the more than half of the additional catalytic residues present in M-CSA dataset are present in top 20 CSmetaPred_poc predicted ranks. Thus, reannotation of catalytic residues for enzymes in CSA-homology dataset using our meta-predictor CSmetaPred_poc indeed has enriched the catalytic residues

compared to previous annotation. The chances of finding correct catalytic residues are higher in this re-annotated CSA-homology dataset.

3.4 Conclusions

Analysis of enzyme catalytic site distance geometry with their cognate reactants/cofactors showed that usually ligand bind in the close vicinity of catalytic site. The minimum distance between the reactant and the catalytic site usually varies from 2-6 Å. In comparison to substrates/products, cofactors are found to be located farther from active sites. This can be explained as cofactors are not usually involved directly in the reaction. The distance between centroids of active sites residues and their reactants is usually within 20 Å.

Further, we have developed catalytic residue prediction method called CSmetaPred with higher prediction accuracy as well as known catalytic residues ranked at better ranked positions, when all residues are ranked by a method. In this work, we have developed a meta-approach to combine varied outputs from various prediction methods and developed a scoring scheme to calculate residue meta-score to improve ranking of catalytic residues. The ranking of residues based on this meta-score (av-csc score) resulted in most catalytic residues ranked within top ranked positions (within top 20 ranks).

Relying on the fact that catalytic residues are spatially proximal to substrate/cofactor binding sites (as seen in the previous analysis of enzyme catalytic site distance geometry with their cognate reactants/cofactors), we have developed another meta-approach called CSmetaPred_poc, which incorporates predicted pocket information with meta-score to further improve catalytic residue prediction accuracy. In this approach, first we use average of residue meta-score of pocket residues (pocket score *i.e.* poc-sc score) to re-rank predicted pockets from Fpocket and LIGSITE and merge top re-ranked pockets to generate a list of merged pockets. Further, pocket score is linearly combined with meta-score (av-csc) of residues lying in predicted pockets. This final residue score (av-csc-poc) is used for ranking residues. This method resulted in improved average/median ranks of catalytic residues compared to CSmetaPred. Thus, both the meta-predictors showed improved performance over its constituting methods. Comparison of ranks of

CSmetaPred_poc with previous studies showed that it has much lower (improved) rank. Moreover, ranks of catalytic residues assigned by CSMetaPred were usually within top 20, for protein structures deposited in PDB after development of CSMetaPred. Further, the assessment of performance done separately for polar and non-polar set of residues, showed that meta-predictor has ability to improve the ranks for both polar and non-polar residues. Homology based model structure showed comparable performance to their corresponding native structures. More importantly, the CSMetaPred_poc assigned catalytic rank is lower (better) for model structure compared to the prediction done using CRpred (which only uses sequence information). Thus, CSMetaPred/CSmetaPred_poc can be reliably used when there is no structure available. Our webserver, will generate the homology based model for the input sequence using MODELLER and further use it for prediction.

We used CSMetaPred_poc to annotate the catalytic residues in CSA-Homology dataset. We found that for in general, 70% of catalytic residues are present in top 20 CSMetaPred_poc predicted ranks for CSA-Homology dataset. We also compared the CSMetaPred_poc performance for same set of proteins but an improved catalytic site definition from M-CSA, and used catalytic residues definition compared to the other (CSAMAC dataset). Here, we found that among top 20 predicted ranked residues by CSMetaPred_poc, the residues which were earlier not defined as catalytic residue in CSAMAC dataset, were later found out to be catalytic, when a much more pruned dataset (M-CSA) was made available. This further supports the observation that usually catalytic and functionally critical residues are ranked within top 20 ranks by our meta-predictor.

Both the meta-predictors are free available for public use as webserver at <http://14.139.227.206/csmetapred/>

Chapter 4

Understanding substrate promiscuity in *E. coli* γ -glutamyl cysteine ligase (EcGCL)

4.1 Introduction

Glutathione (GSH) is a tripeptide (Glu-Cys-Gly), having unusual γ -peptide linkage between L-Glu and L-CYS, is the most important low molecular weight antioxidants in eukaryotes and has suggested to have similar role in some prokaryotes (Galant et al. 2011; Fahey 2013; Meister and Anderson 1983; Masip, Veeravalli, and Georgiou 2006; Pompella and Corti 2015). GSH protects cells mostly against free radicals, oxidative damage, primarily caused by reactive oxygen species such as free radicals, peroxides and reactive oxygen species (Pompella et al. 2003). Moreover, GSH can also provide protection against xenobiotic compounds/electrophiles or metal toxicity (Pompella et al. 2003; Jozefczak et al. 2012). Usually, GSH achieves this by their sulfhydryl group of cysteine that is involved in reduction and conjugation reaction. The glutathione peroxidase family of enzymes catalyze reduction of hydrogen peroxides by GSH into GSSG and water and phospholipid hydroperoxide glutathione peroxidase reduces lipid peroxides to lipid alcohols (Brigelius-Flohe and Maiorino 2013). Moreover, radical species are removed by non-enzymatic reduction by GSH. Many xenobiotic compounds conjugates with GSH to form its adduct, which is mostly catalyzed by glutathione-S-transferases and subsequently this adduct form is eliminated from the cell. GSSG can be reverted back to its reduced form (*i.e.* glutathione) by the enzyme glutathione reductase in the presence of NADPH (which acts as an electron donor) (Couto et al. 2013). The

redox state of cell is often measured by the ratio of GSH/GSSG (Pastore et al. 2001; S. C. Lu 2013). In recent studies it has been shown that glutathione provides protective ability towards oxidative stress in cancerous cells and confers resistance to tumors against chemotherapeutic drugs (Backos, Franklin, and Reigan 2012; Traverso et al. 2013). The peptide linkage in GSH and γ -GC is resistance to cleavage by cellular proteases and require cellular peptidases and metabolized into its constituent amino-acids only by enzyme γ -glutamyl transpeptidase (γ GT) (S. C. Lu 2009).

The *de novo* glutathione biosynthesis is a two-step pathway, which involves ATP dependent enzymatic reactions. The first step is catalyzed by γ -glutamyl cysteine ligase (GCL) also known as γ -glutamyl cysteine synthetases, which forms peptide bond between γ -carboxylate group of L-glutamate (L-GLU) and α -amino group of L-Cysteine (L-CYS) with ATP and Mg^{2+} as cofactors to produce γ -glutamyl cysteine (γ -GC). The next and last step in GSH biosynthesis is formation of peptide bond between γ -GC and L-Glycine catalyzed by glutathione synthetase (GS), which requires ATP hydrolysis (Meister and Anderson 1983). The first step is essential and a rate-limiting step in GSH biosynthesis. In eukaryotes (except Plants), GCL enzyme is a heterodimer of catalytic (GCLC) and modulatory (GCLM) subunits, where the GCLC is responsible for enzymatic activity and GCLM regulates its catalytic activity (Franklin et al. 2009). However, there is only one single catalytic unit in prokaryotes and plants that is usually referred to as GshA and GCL respectively. Interestingly, in some pathogenic bacteria, such as *Streptococcus agalactiae*, *Listeria monocytogenes*, and *Pasteurella multocida* the GCL and GS enzymes are fused together as one bifunctional enzyme, which is referred to as GshF or GshAB (Vergauwen, De Vos, and Van Beeumen 2006; Biterova et al. 2013; Stout et al. 2012; Janowiak and Griffith 2005). GCL activity regulates cellular GSH levels and dictates GSH biosynthetic capacity. Imbalance in GCL functional activity is known to be involved in many human diseases such as cancer, Parkinson's disease, Alzheimer's disease and diabetes (S. C. Lu 2009; Franklin et al. 2009). In these diseases, usually GCL's impaired function leads to decreased levels of glutathione, thereby reducing cellular anti-oxidant capacity and inducing oxidative stress. On contrary, in cancer, GCL's levels are elevated and its levels are enhanced, supporting high cellular proliferation and also imparts resistance towards many chemotherapeutic agents (Backos, Franklin, and Reigan 2012).

GCL enzymes have been identified and experimentally characterized from both prokaryotes and eukaryotes organisms. It was observed that GCL enzymes from eukaryotes (except plants) do not share significant sequence similarity to prokaryotic sequences. In 2002, seminal work of Copley and Dhillon, classified GCL enzymes into three groups: Group1 (consists mostly sequences from γ -proteobacteria), Group 2 (sequences are from non-plant eukaryotes) and Group 3 (sequences are from mostly plants and α -proteobacteria) (Copley and Dhillon 2002a). In their work, they could relate these groups based on three common conserved regions among them. Subsequent experimental tertiary structure determination of at least one member from each group clearly showed that these groups are evolutionary related as assessed by TM-score, which for EcGCL and BjGCL is 0.68, EcGCL and ScGCL is 0.61 and BjGCL and ScGCL is 0.79. Ec, Sc and Bj represent *E. coli*, *S. cerevisiae* and *B. juncea* respectively. Ec, Sc and Bj belong to group 1, group 2 and group 3 respectively.

The initial insights into the enzymatic mechanism of GCL was elucidated with experimental structure determination of *E. coli* GCL (EcGCL) in two forms: ligand unbound (apo), pdbid: 1v4g, and sulfoximine based transition state analogue (2s)-2-amino-4-[[[(2r)-2-carboxybutyl](phosphono)sulfonimidoyl]butanoic acid (P2S) bound, pdbid: 1va6 (Hibi et al. 2004). EcGCL structure has two domains: a catalytic domain (residues 18–387 and 442–518) and a small domain (residues 1–16 and 388–441). The catalytic domain consists of curved partial β sheet arranged in barrel forming a funnel-shape cavity. γ -phosphate of ATP moiety of P2S is already phosphorylated and its carbonyl group interacts with R330 suggested to orient its si-face towards L-CYS binding site. The putative nucleotide interacting residues also involve three Mg^{2+} (termed n1, n2, n3) found around phosphate-binding pocket that interacts with E29, D60, E67; E27, H150, E328 and E27, E67 respectively. The glutamate moiety of P2S showed interaction with I1476, R235, H150, Y241 while the cysteine moiety of P2S interacts with F61, Y131, R132, L135, Y300, Y241, Q144. The comparison of apo and P2S bound structures showed that there is ligand induced conformational change involving switch loop (residue 240-249), which leads to orienting key residues in substrate binding sites. Additionally, there is change in side chain torsional angle χ_1 of two residues Y241 and Y300 by 108° and 137° respectively. This facilitates hydrogen bond formation of Y300 with Y241 and Q144 as well as with carboxyl group of cysteine equivalent region of P2S. Based on this, it has been suggested that putative cysteine binding site comprise of

residues Y131 and Y300 involved in hydrogen bond interaction with carboxyl group of cysteine and residues F61, Y131 and L135 accommodates side chain of cysteine. Subsequent to structure determination of eukaryotic GCL (group 2) from *Saccharomyces cerevisiae* (ScGCL) (Biterova and Barycki 2009) and plant GCL (group 3) *Brassica juncea* (BjGCL) (Hothorn et al. 2006), suggested that binding site motifs are structurally conserved, however, a similar structural transition has not been observed in ScGCL and BjGCL.

The experimental studies have shown that GCL from various organisms exhibit substrate promiscuity. Previous studies on mammalian GCL (rat and bovine) have shown that GCL can catalyze wide variety of substrates (L-CYS) including natural and unnatural amino acids (Griffith and Mulcahy 1999). The cyanobacterium *Synechocystis sp.* has been shown to produce norophthalmate and ophthalmate in GshA and GshB dependent manner. This suggests that *Synechocystis sp.* GshA could accept Alanine and 2-aminobutyrate instead of L-CYS to produce γ -glutamylalaninyl and γ -glutamyl-2aminobutyryl, which are precursor of norophthalmate and ophthalmate (Narainsamy et al. 2016). The work of Kelly *et al.* on EcGCL showed that it could catalyze variety of natural/unnatural amino acids instead of L-CYS. On the contrary, a limited compounds could be catalyzed instead of L-GLU by EcGCL (Kelly, Antholine, and Griffith 2002). EcGCL has been shown to accept amines as substrates to form γ -glutamylamides including theanine (γ -glutamylethylamide), which is component of green tea (MIYAKE and KAKITA 2009). GCL from different lineages have different affinity towards its inhibitors. For instance, BSO is slow binder in case of EcGCL and is unable to form a strongly inhibiting phosphorylated derivative (Tokutake et al. 1998; Kelly, Antholine, and Griffith 2002). Mammalian GCL on the other hand are inhibited potently by BSO (Griffith 1982).

Although, GCL enzymes from various lineages have been shown to exhibit substrate promiscuity, the underlying molecular mechanism have not yet been investigated. The main objective in this chapter is to investigate structural basis of substrate promiscuity shown by GCL enzymes. We have used EcGCL (*E. coli* GCL) for this study, as it a well characterized enzyme, especially in terms of availability of experimental data such as enzyme kinetics of various alternate substrates/analogues apart from its natural substrate cysteine. Due to unavailability of substrate (L-CYS) bound structure, our initial work objective was to a) characterize L-CYS pockets from various enzymes by evaluating

pocket similarity and ligand similarity, and b) use docking approach to identify L-CYS binding in EcGCL enzyme. Moreover, this will also provide putative conformation of L-CYS for docking studies. For this study, all the L-CYS bound known tertiary structures were collated, followed by comparison of binding pockets of these structures among themselves to identify any common characteristics of cysteine binding sites such as residues involved in interaction with cysteine. Further, these sites were compared with EcGCL to identify putative substrate binding site as well as substrate conformations using approach of ligand docking. Next, in order to understand the substrate promiscuity of EcGCL, we have used the docking to find the best docked poses of alternate substrates and analyzed conservation of interactions between enzyme and substrates. Further, Molecular Dynamics (MD) simulation approach was used to analyze the dynamics of substrates in their bound states to the enzyme.

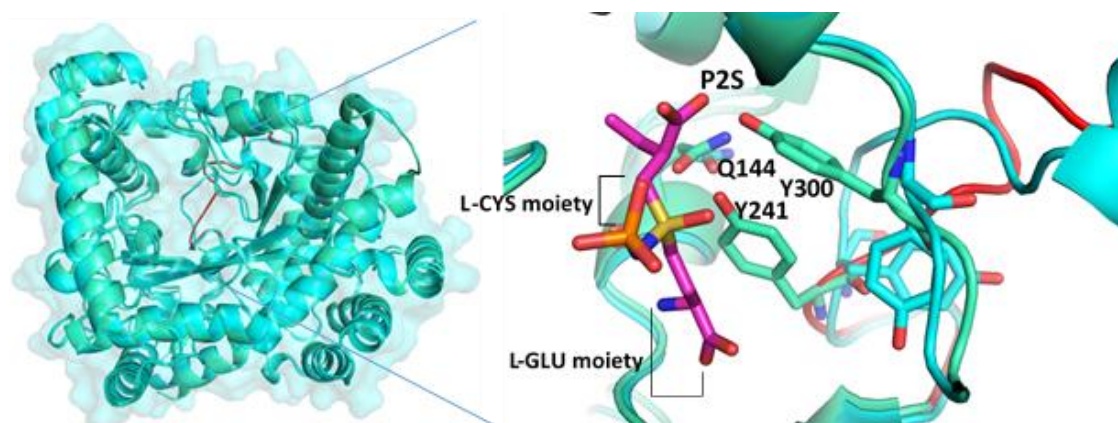


Figure 4.1 Conformational change in EcGCL between transition state analog (P2S) bound structure (1va6) and unbound structures (1v4g). The unbound structure 1v4g and P2S structure are shown in cyan and sea green color. P2S is shown in ball and stick representation with magenta color. Switch loop is shown in red. The residues with change in side-chain torsional angle and involve in hydrogen bond network are shown in licorice representation.

Glutathione plays a role in the acquisition of resistance to anti-cancer drugs, such as a cisplatin, adriamycin, melphalan, or taxol (Kondo and Iida 1997; Backos, Franklin, and Reigan 2012). Often GCL is suppressed for the depletion of GSH and are new molecular targets in cancer treatment (Kondo and Iida 1997; Griffith and Mulcahy 1999). Understanding structural basis of GCL substrate promiscuity will help in selecting/designing inhibitor candidates that have a greater likelihood of being less promiscuous binders alleviating the adverse off-target effect of inhibitor promiscuity.

4.2 Methods

4.2.1 Analysis of L-CYS binding enzymes

4.2.1.1 Construction of L-CYS bound enzymes/transporters dataset

In order to analyze tertiary structures of enzymes, which catalyze L-CYS, we extracted all structures (enzymes/transporters) with L-CYS bound as cognate ligand and followed steps given below to construct cys-bound dataset:

1. A list of enzymes bound to L-CYS was extracted from PDBeChem database (Dimitropoulos, Ionides, and Henrick 2006) and we removed structures with resolution $> 2.5 \text{ \AA}$. This resulted in 62 structures.
2. The L-CYS binding site residues was computed using LPC (Sobolev et al. 1999) program and structures having at least four residues interacting with L-CYS were retained. We also excluded enzymes with ligand bound at surface having most interaction with solvents. This resulted in 43 structures.
3. These structures (43) were mapped to Uniprot ID and we selected only one representative structure, which has maximum residue involved in ligand interaction, for each enzyme. In case an enzyme structure has L-CYS bound to different pdb chains, both chains are retained if the L-CYS binding sites are distinct from on two chains. This step resulted in 34 structures. These were made non-redundant at 60% sequence identity using CD-HIT program that resulted in 22 structures, which constitute cys-bound dataset. This dataset is provided in supplementary material (*c4.1_cys_dataset.xlsx*) for chapter 4 provided in a CD along with this thesis.

4.2.1.2 Structural alignment of L-CYS and L-CYS binding pockets and

We extracted the ligand L-CYS (having seven atoms) from the ligand bound enzymes from cys-bound dataset. Further, these L-CYS conformations were aligned using OBAAlign program from Open Babel toolkit (O'Boyle et al. 2011). RMSD was used as

metric to identify the extent of structural difference/deviation among these conformations.

Since cys-bound dataset has structures without any structural relatedness to each other, we have used APoc to structurally align L-CYS binding pockets. However, APoc needs a minimum of 10 residues for aligning two given pockets. In many instances, there were less binding site residues than required number for structural alignment. Hence, we extended the pocket definition by including residues, which are exposed (relative surface area $> 5 \text{ \AA}$) and involved in binding ligand (L-CYS) till number of residues is 10. Thus, obtained pocket residues were aligned using APoc (Gao and Skolnick 2013b) program. The pocket similarity was assessed using PS-score (Pocket Similarity score), which varies from 0 to 1. The identical pockets have PS-score as 1 (upper bound).

Given two pockets with query pocket of length L_Q and template pocket of length L_T , APoc aligns the pockets iteratively superimposing the residues using Kabsch algorithm (Kabsch 1976) minimizing the RMSD of aligned residues until optimal alignment is found. PS-score is calculated using following equation:

$$PS - score = \frac{S + s_0}{1 + s_0} \quad (1)$$

where scaling factor s_0 scales the score of the random pockets of similar length such that its mean score is independent of its length and is calculated as follows:

$$s_0 = 0.23 - \frac{12}{L_Q^{1.88}}$$

where L_Q is the length of the query pocket. S in equation (1) is calculated as follows:

$$S = \frac{1}{L_Q} \max_{sup} \left[\frac{\sum_{i=1}^{N_a} p_i r_i}{1 + \frac{d_i^2}{d_0^2}} \right] \quad (2)$$

where N_a is the number of aligned residue pairs, d_i is the distance in \AA between the $C\alpha$ atoms of the i^{th} aligned residue pair and d_0 is empirical scaling factor standardized by fitting the distribution of $C\alpha$ distances in random alignment of pockets and p_i is a measure of the directional similarity between two $C\alpha$ to $C\beta$ vectors. r_i measures the chemical similarity of two aligned amino-acids. The notation \max here denotes PS-score that gives maximum of all scores.

4.2.1.3 Hierarchical clustering of similar L-CYS bound pockets

We used complete agglomeration method and maximum distance measure as a metric for hierarchical clustering of similar L-CYS binding pockets using following scores to derive distance matrix for clustering:

- Similarity matrix based on PS-Score cut at 0.36 score is converted to distance matrix by subtracting each value from 1.
- Dissimilarity matrix based on RMSD cut at 0.5 Å

4.2.2 Overview of Docking procedure

4.2.2.1 Preparation of initial structures (receptor (EcGCL) and ligands)

The coordinates of EcGCL (1v4g and 1va6) were obtained from RCSB. Both structures have missing coordinates of residues: 1v4g has missing residues from 164-167 and 210-214, while 1va6 has missing residues from 210-213 and 459-462. The missing residues were modeled using MODELLER (Webb and Sali 2016) using standard *loopmodel* class function followed by refinement using energy minimization done using GROMACS (Lindahl and Hess 2001).

For docking, the polar hydrogens were added to the above modelled protein structure and then gasteiger charges were computed using GUI of AutoDock. Further, non-polar hydrogens in the molecule were merged and total charge on the system was kept as integral. The ligand was treated in same manner. All dockings were performed with AutoDock Tools (G. Morris and Huey 2009) (version 1.5) using the Lamarckian and Genetic Algorithms and results were visualized in visualization program of AutoDock/UCSF Chimera (Pettersen et al. 2004).

4.2.2.2 Parameters used for docking

The AutoDock parameter set and distance-dependent dielectric functions were used for calculating the van der Waals and the electrostatic terms, respectively. The initial position, orientation and torsions, in flexible docking of the ligand molecules, were set

randomly. Each docked ligand was derived from 100 independent docking runs that were set to terminate after a maximum of 2.5×10^6 energy evaluations with mutation rate of 0.02 and crossover rate of 0.8. The population size was set to use 250 randomly placed individual. The low-energy binding orientations were search by Lamarckian Genetic Algorithm using standard parameters of AutoDock.

In order to select the appropriate GCL structure among 1v4g and 1va6, we used blind docking approach. For blind docking a grid box size of 126 x126 x 126 Å points with a grid spacing of 0.569 Å was generated using AutoGrid. L-CYS was kept flexible allowing all its rotatable bonds to move during docking.

In our limited docking approach, we kept the ligand flexible allowing their respective rotatable bonds to move during docking. Here we used grid box size of 60 x 54 x 68 Å points with a grid spacing of 0.375 Å. After selecting the appropriate binding mode based on lowest binding energy and its ability to facilitate the enzymatic reaction, we performed another set of docking called restricted limited docking. Here, we kept ligand as rigid and allowed zero degree of freedom to its rotatable bond and performed the docking in the selected binding site only with appropriate ligand binding mode. In this docking, we used a grid box size of 12 x12 x 12 Å points with a grid spacing of 0.375 Å.

4.2.2.3 Selection of best docked conformation

Subsequent to docking, Autodock ranks various (usually, 100 in our docking studies) independent docked poses based on their binding energies. Usually, the lowest energy docked conformation is considered the most favorable binding pose. Alternatively, docked conformations can be clustered into groups having similar conformations and the most populated cluster with lower mean energy can be analyzed. We clustered docked conformation based on RMSD and used the command *rmsnosym* of AutoDock. The RMSD cut-off to cluster docked poses is varied using a parameter (*rmstol*). We have typically used a 2 Å cut-off to cluster docked ligand conformations. Further, these clusters are ranked based on the lowest mean binding energy of each cluster members. Often, the lowest energy member of the most populated cluster is also considered the most favorable binding pose. It should be noted that by default, AutoDock tries to compute the minimum RMSD by taking into consideration the symmetry in the molecule, and works well, if the two conformations are very similar. However, to ensure 1-to-1

correspondences of atoms, we have used *rmsnosym* command. Secondly, if the difference in the binding energies between the mean binding energies of the two clusters is less than about 2.5 kcal/mol, this is within the standard deviation of the AutoDock force field, and it is difficult to say which one is the "correct" one.

However, in our study, we could not use either of these criteria for selecting the best docked conformation for L-GLU/L-CYS, essentially, because the best ligand conformation was not in appropriate orientation to facilitate catalysis based on the P2S structure. The experimental structure suggested that L-GLU binds in deep cavity of EcGCL funnel with its γ -carboxylate group facing outwards in the open funnel and L-CYS is lying close to L-GLU binding site, slightly on funnel rim (Figure 4.1). For catalysis to progress, the γ -carboxylate group of L-GLU and α -amino group of L-CYS should be in close contact. In L-GLU docking, we often observed the lowest energy conformation had its γ -carboxylate group facing away from funnel cavity. Since such L-GLU conformation can participate in reaction, we selected the lowest energy structure of L-GLU having appropriately oriented functional group. Similarly, while docking L-CYS, often the lowest energy structure had catalytic α -amino group of L-CYS oriented away from the γ -carboxylate group of L-GLU that would not favor catalytic reaction. In order to assess docked poses of L-CYS favorable for catalysis, we measured the distance between the γ -carboxylate group of L-GLU and α -amino group of docked L-CYS that we referred to as NoE distance. The L-CYS docked poses having NoE within 4.5 Å was selected for finding the lowest binding energy structure. Hence, keeping physiologically relevant information we have filtered out reaction incompetent conformations and selected the lowest energy docked conformation in top ranked cluster among the remaining docked conformations. We compute NoE distances for both OE1 and OE2 of L-GLU and use the lowest as NoE distance.

4.2.2.4 Docking of alternate substrate for L-CYS

In order to understand substrate promiscuity of EcGCL, we have docked two categories of alternate substrates: a) amino acids instead of L-CYS and b) polyamines shown Figures 4.2A and 4.2B respectively. These were chosen because on the basis of available experimental data (Kelly, Antholine, and Griffith 2002; MIYAKE and KAKITA 2009). The docking approach was used to elucidate the binding mode of these alternate substrates, which can be accommodated in the cysteine pocket. Each of alternate

substrates was docked in EcGCL-glu complex (discussed in section 4.2.2.4) using limited docking approach using the parameters mentioned before. We allowed rotations across all rotatable bonds of alternate substrates, which we refer to as flexible limited docking.

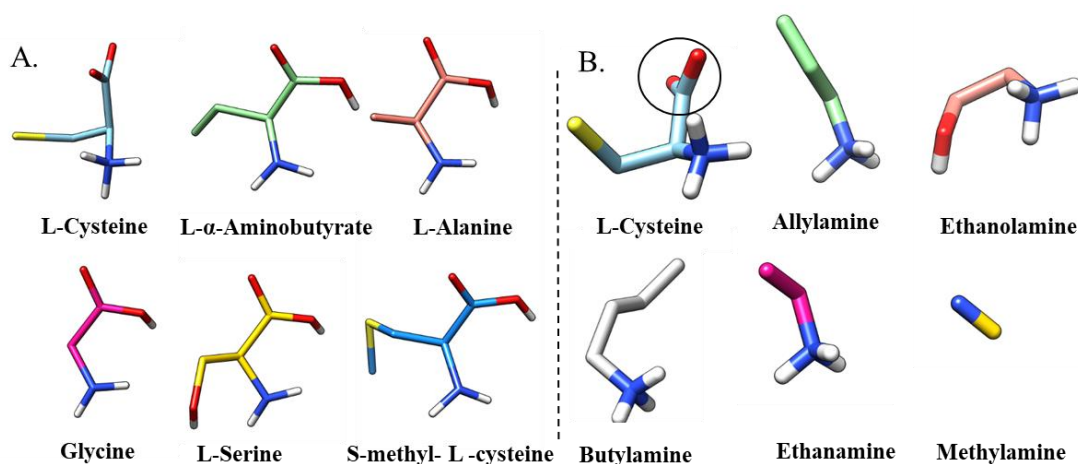


Figure 4.2 Comparison of native L-cys substrate with A) set-A- various natural and unnatural amino-acids B) Set-B- polyamines used for docking to EcGCL.

EcGCL in catalysis using alternate substrates will require functional group L-GLU substrate lying close to α -amino group of these alternate substrates, because they also involve γ -peptide bond formation. Thus, these two catalytically active chemical groups of substrate (α -amino group of alternate substrate and γ -carboxylate group of L-GLU) should be in close vicinity to facilitate catalysis. Similar to criteria used to select L-CYS conformation, we selected the best docked structure of alternate substrate by finding the lowest energy conformation having NoE distance within 4.5 Å (Figure 4.3A). We measured the overlap between the binding site of L-CYS (wild type) and alternate substrates in order to quantitate the difference in orientation of alternate substrates that is referred to as “OV” (Figure 4.3B). Between two categories of docked alternate substrates, polyamines lack carboxyl group and sulfhydryl group. For polyamines, we additionally calculated the distance between the α -amino group of docked polyamine and L-CYS (bound in wild type EcGCL) and referred this distance as N-N*. Further, we also calculated overlapping residues interacting with the α -amino group in docked polyamine and L-CYS and referred this overlapping measure as OV1. We compared of docking binding energies of the selected docked pose of alternate substrate and their corresponding experimentally observed relative activities (from earlier studies) and further assessed the role of various interacting residues (of both ligand and protein) which

might play a crucial role in substrate recognition and elucidated the structural components of EcGCL facilitating it to be a substrate promiscuous enzyme.

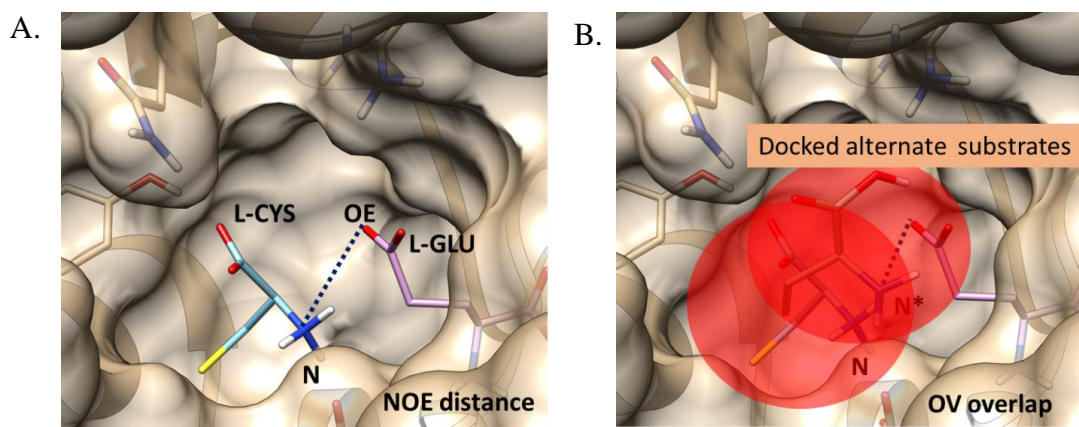


Figure 4.3 Representation of measures (NoE and OV) used in selection of the best docked poses of alternate substrate favorable for catalysis, A) NoE distance is between α -amino group of substrate and γ -carboxylate group of L-GLU. B) OV overlap between the binding site of L-CYS (wild type) and alternate substrates.

4.2.2.5 Overview of docking protocol to generate wild type EcGCL bound to L-GLU and L-CYS

There are two structures available for EcGCL currently in PDB: 1va6 (bound to transition state analog P2S) and 1v4g (unbound form). Although many studies refer to the cysteine binding residues proposed based on the cysteine moiety of transition state analog-P2S bound to EcGCL (pdbid: 1va6), the exact residues involved in binding SH and other chemical groups of L-CYS is not yet known. In the present study, we have docked L-CYS to EcGCL (1va6 and 1v4g) first using blind docking approach *i.e.* without using any prior knowledge of the binding site, followed by limited docking approach *i.e.* using predicted binding residues (Hibi et al. 2004) to elucidate the mode of L-CYS binding. Following were the main steps involved in docking of L-CYS to EcGCL:

- 1) **Selection of EcGCL starting structure for docking:** To select structure (1va6 or 1v4g) for docking analysis, we used blind docking approach and docked L-CYS to both 1va6 and 1v4g using parameters discussed before. From both structures, we selected the structure having the maximum percentage of docked conformations in the cysteine binding cavity. In PDB, a structure (2d32) of EcGCL is deposited bound to L-CYS and L-GLU. However, this structure has no structural change as observed

in 1va6. Moreover, there is no publication discussing this structure. In absence of any direct way to know more detail about 2d32, we have not considered this structure for further analysis.

- 2) **Docking of L-GLU to EcGCL:** In our previous blind docking experiments, we observed that many L-CYS conformations docked in the site suggested for L-GLU. This is because both amino acids have same backbone constituted by C α , C β , CH $_2$ and NH $_3$ groups. However variable group R of L-CYS is shorter compared to L-GLU, thereby can be easily accommodated in L-GLU site. In order to prevent L-CYS occupying L-GLU site, we first dock L-GLU to EcGCL using the limited docking parameters as given before. This docked EcGCL-glu complex was further used to dock L-CYS or alternate substrates. More importantly, such a complex would mimic reaction mechanistic state as well.
- 3) **Docking of L-CYS to EcGCL-glu docked complex:** L-CYS was docked in EcGCL-glu complex using the limited docking parameters as given before. This EcGCL-glu-cys docked complex is regarded as wild type state.

4.2.3 MD simulations of bound and unbound states of EcGCL

For Molecular Dynamics simulation of EcGCL, we have used both unbound (1v4g) and transition state bound structures (1va6). We performed two set simulations for EcGCL: a) Unbound (1v4g and unbound 1va6) and b) Substrate bound (ATP, L-GLU and L-CYS docked complex). In first (unbound) simulation set, the unbound structure (1v4g) was used a starting structure for simulation. In the second simulation, a substrate bound structure consists of L-GLU, L-CYS, ATP and Mg $^{2+}$. In order to obtain substrate, bound structure, first we removed P2S and ADP from 1va6 and then use AutoDock to dock the ATP in the same spatially location as of ADP (in 1va6) by limiting the docking space (see parameters for docking). The ATP docking resulted in the extension of γ - phosphate with rest of conformation similar to that of bound ADP. The other 2 substrates L-GLU and L-CYS were docked as described previously in previous sections. We refer this substrate bound 1va6 structure as 1va6-sb. Thus, we have used 1v4g and 1va6-sb structures for MD simulations. The simulations were performed using Nanoscale Molecular Dynamics (NAMD) (Phillips et al. 2005) package. The input structures was processed and all accessory files were generated using plugins available in VMD (Humphrey, Dalke, and Schulten 1996). First, we used AutoPSF plugin to generate PSF

files and the structure was solvated with TIP3P water model (Jorgensen et al. 1983) using solvate plugin. Further, the solvated system was neutralized by Na⁺ ions using autoionize plugin. All MD simulations were performed using NAMD (CVS-2016-04-11 for Linux-x86_64-multicore-CUDA version) with CHARMM22 All-Hydrogen topology (Soteras Gutiérrez et al. 2016) and force field parameters for proteins and nucleic acid having CMAP corrections (Mackerell, Feig, and Brooks 2004). In all simulations, the initial system was subjected to 5000 steps of energy minimization followed by a 1 ns equilibration MD run, under constant temperature and pressure at 300 K and 1.01 bar (1 atm), respectively. Once the system was equilibrated, the simulation was extended and production run for 50 ns was performed. All systems were simulated in periodic boundary conditions with electrostatics interactions computed using Particle Mesh Ewald (PME) method (Essmann et al. 1995) by specifying grid sizes. A switching function with cut-off distance of 12 Å was used in computing the van der Waals interactions. The constant pressure simulations at 1 atm were conducted using Noé-Hoover Langevin piston method (Feller et al. 1995) with piston period of 200 fs, a damping timescale of 100 fs and piston temperature of 300 K. A constant temperature of 300 K was maintained using the Langevin dynamics, with the damping coefficient set to 5 ps⁻¹ for all the heavy atoms. In all cases, time-step was set to 2 fs in both equilibration and production runs. The trajectory was visualized and analyzed mostly using VMD or UCSF Chimera program.

4.3 Results

4.3.1 Analysis of L-CYS binding sites in enzymes

4.3.1.1 Conservation of L-CYS binding pockets in enzymes/transporters

Since the structure of EcGCL bound to L-CYS is not yet known and knowledge of cysteine binding pocket is derived from transition state analogue structure, we explored the possibility of deciphering the cysteine binding mode (conformation) and its interacting residues from enzymes, which use L-CYS as their cognate substrate. For this analysis, we considered only enzymes/transporters bound to L-CYS, which use it as their cognate substrates/products and analyzed cysteine binding similarity. Further, we also

compare conformations of L-CYS bound to these enzymes. To assess the cysteine binding pocket similarity, we have used two metrics: RMSD and PS-score. Here, RMSD is dissimilarity metric whereas PS-score is similarity metric. For appropriate alignment of L-CYS binding pockets, we had to extend the binding site to include spatially proximal residues (see methods section 4.2.1.2). For EcGCL, we relied on the binding site residues of L-CYS as proposed by (Hibi et al. 2004) and shown in Figure 4.1.

As mentioned in methods, APoc was used to align cysteine binding pockets. For each pocket, the best aligned pocket is considered for analysis. The distribution of C α RMSD of the best matched (least RMSD) cysteine binding pocket residues of 22 enzymes after performing all-against-all comparison is shown Figure 4.4A. As can be seen in the Figure, most pockets are similar with mean (SD) RMSD of 0.93 Å (0.59 Å). This suggests that pockets are not very variable at least among the proteins analyzed in cys-bound dataset. The RMSD was obtained from APoc alignment, which structurally aligns pocket residues. Further, we used pocket score (PS-score) as a measure to find similar pockets among 22 enzymes and also pocket similar to EcGCL. The distribution of the best PS-scores of cysteine pockets of each enzyme is shown in Figure 4.4B. The mean (SD) and median PS-score of the cysteine binding pocket is ~ 0.5 (0.19) and 0.43 respectively. The statistical significance is associated with each PS-score generated using comparison with millions of randomly generated pockets suggests that PS-score of 0.4 is statistically significant at $p\text{-value} < 1 \times 10^{-3}$ representing similar binding pockets (binds to same/similar ligands). Based on this, ~ 36 (08/22) and ~ 23 (5/22) % of proteins have p -values associated with PS-score ≥ 0.4 having p -values $< 10^{-2}$ and $< 10^{-3}$ respectively. Using PS-score, EcGCL was found to be similar to L-CYS binding pocket of Metallothiol transferase FosB (pdbid: 4jh8) with PS-score (P-value) and RMSD of 0.4 (0.01) and 1.05 Å respectively. However, the PS-score is not statistically significant. FosB enzymes are M(2+)-dependent thiol transferases that catalyze nucleophilic addition of either L-CYS or bacillithiol to the antibiotic, resulting in a modified compound with no bactericidal properties (Thompson et al. 2013). Even though many cysteine binding pockets showed similarity, we could not find L-CYS binding site of EcGCL.

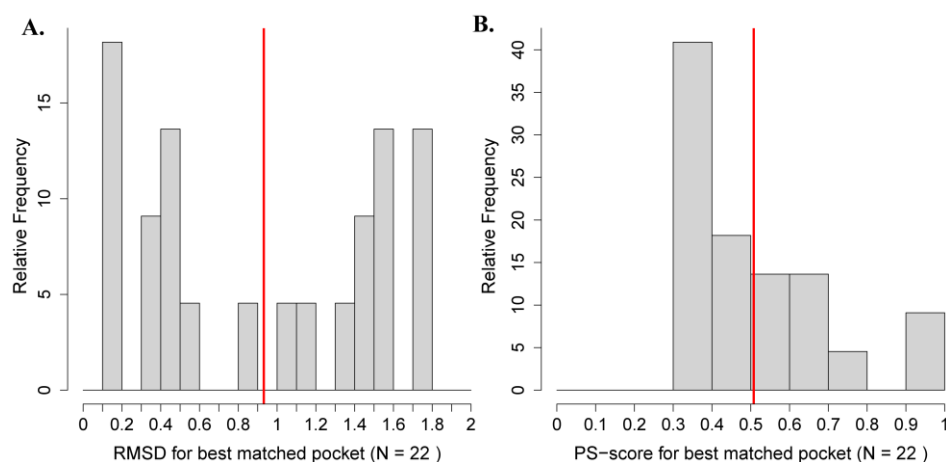


Figure 4.4 Comparison of cysteine binding pockets of enzymes using A) RMSD and B) PS-Score.

4.3.1.2 Conformational space explored by the substrate L-CYS

In previous analysis, we observed that most cysteine binding pockets are able to find at least related pocket. Next, we aligned L-CYS conformations to assess conformational variability of cysteine in bound state. For this, we aligned L-CYS bound conformations from various enzyme structures using OAlign and considered L-CYS conformation from 2ibn as a reference structure. Such analysis will assist in deciding the mode of L-CYS for docking. The RMSD distribution is shown in Figure 4.5A.

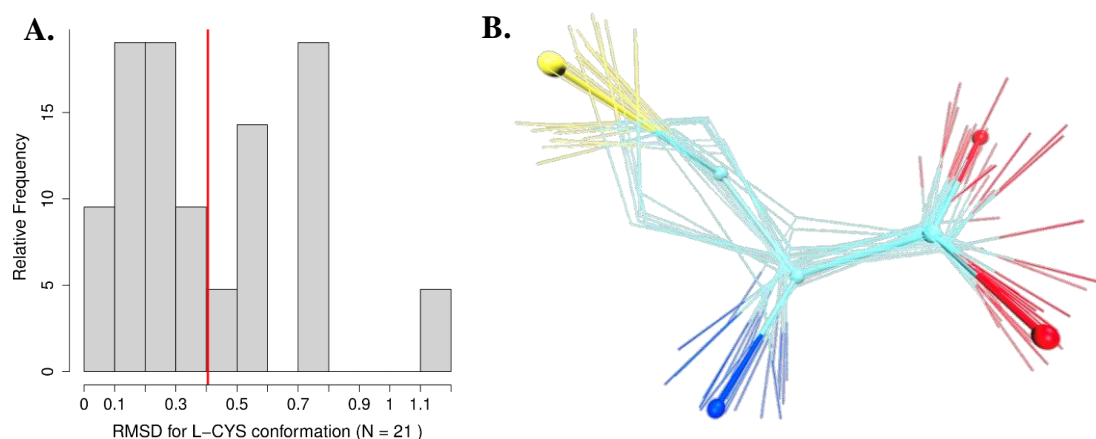


Figure 4.5 Conformational variability of L-CYS bound in 22 enzymes. A) RMSD distribution of L-CYS bound to the enzymes in NR dataset. B) All L-CYS conformations aligned to cysteine bound to 2ibn. The structures are shown in stick representation.

As is evident from the Figure 4.5A, the mean (standard deviation) and median RMSD for a pair of L-CYS conformations bound to any enzyme is ~ 0.4 (0.28) and 0.35 Å. Thus,

in general L-CYS does not show structural variation in the bound state (Figure 4.5B) suggesting that while docking one can technically start with any one of the cysteine conformations with a believe that bound state is not going to be very different.

4.3.1.3 Conservation of residue contacts in L-CYS binding pockets of 22 enzymes

Given that cysteine binding pockets, bound cysteine does not show extensive variability, we analyzed the conservation of interactions between cysteine binding residues and L-CYS. Here, we essentially asked whether interacting residues for cysteine chemical moiety is conserved. For this, we used LPC program to extract interacting residues involved in atomic contacts with individual ligand atoms of cysteine. Here, we did not consider the nature of interaction between residues and ligand. The distribution of counts of occurrence of atomic contacts between ligand atoms and residue side chain or main chain atoms is shown in Figure 4.6.

As can be seen in Figure 4.6, that side-chain of ARG is the most frequently interacting residue of C, O and OXT atoms of the L-CYS ligand. The C α and C β atom of substrate L-CYS mostly interacts with CYS residues, while SG atom frequently interacts with HIS residue. Since P2S (in 1va6) does not have equivalent atom for SG of cysteine, it is difficult to find residues involved in interaction with sulfhydryl group of cysteine. We wanted to generate list of putative residues surround the SG atom of cysteine in the docking of L-CYS to EcGCL such that we can exploit this information to suggest likely binding residues of this group.

4.3.1.4 Hierarchical clustering of the L-CYS binding pockets based on their pocket similarity

Further, we clustered the similar L-CYS binding pockets based on their pocket similarity score. The basic idea behind this clustering was to find similar binding pocket of EcGCL, and subsequently used its binding site information to transfer information to EcGCL.

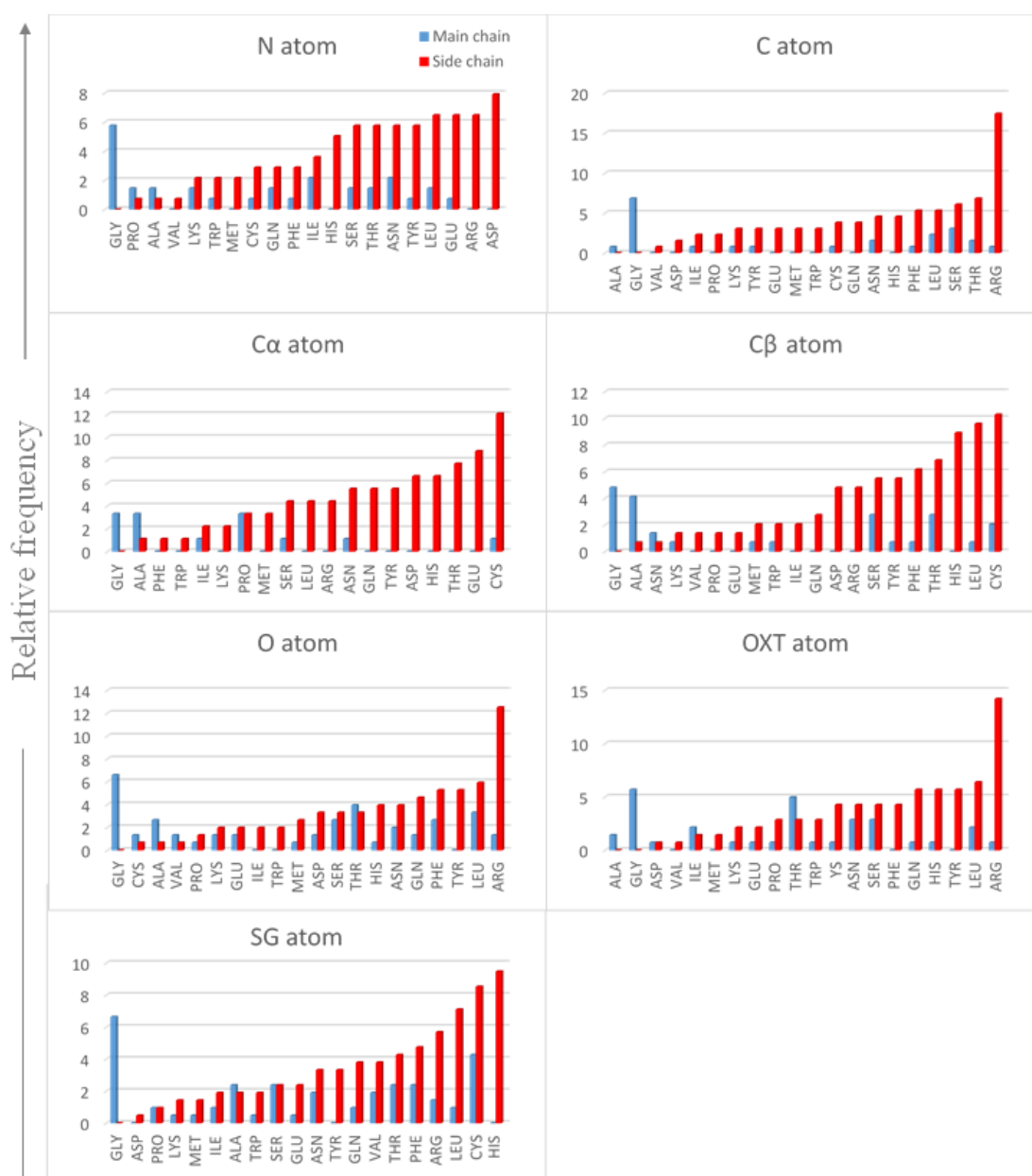


Figure 4.6 distribution of frequency of various residue type making atomic contact with various atom of substrate L-CYS.

We used hierarchical clustering approach using complete agglomeration method and maximum distance measure. The PS-score was converted to distance by subtracting it from 1. From the hierarchal clustered tree, we used empirical tree height cut at 0.36 to obtained clusters of enzymes (Figure 4.7). As can be seen from Figure 4.7, the enzymes with same E.C number are clustered together. This indicates the enzymes with same function essentially have similar binding sites. The Figure 4.8. shows the structurally

aligned binding pocket of cysteine dioxygenase enzymes with EC number 1.13.11.20, from *Rattus norvegicus* (PDBID: 4xf0) and *Bacillus subtilis* (PDBID: 4qm9) by APoc program. The PS-score for these aligned pockets is 0.675 with a p-value of 0.465×10^{-7} . Since EcGCL is not grouped with any functionally related enzymes, we could not exploit this information for EcGCL. However, such clustering shows that functionally related enzymes could be grouped based on pocket similarity of their substrates.

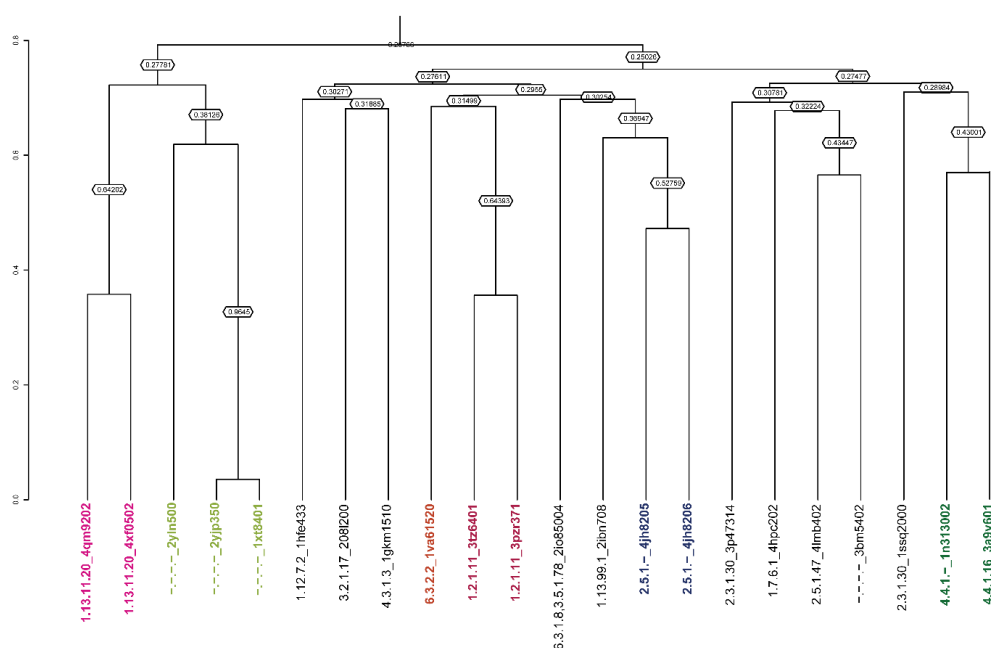


Figure 4.7 Hierarchical clustering of the L-CYS binding pockets using complete agglomeration method. Clusters are obtained based on height of tree cut at 0.36.

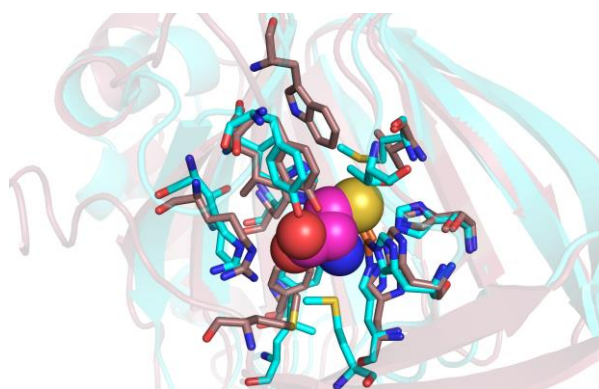


Figure 4.8 Example of same EC number enzymes having similar pockets. Pocket alignment of the enzyme cysteine dioxygenase from *Rattus norvegicus* (pdbid 4xf0) and *Bacillus subtilis* (pdbid 4qm9) colored in raspberry and cyan respectively. The binding residues are shown in CPK representation and the substrate L-CYS bound to 4xf0 is shown in sphere representation

4.3.2 Docking studies to understand substrate promiscuity of EcGCL

4.3.2.1 Selection of EcGCL structure for docking

As mentioned in methods section 4.2.2.1, there are two EcGCL experimentally determined structures: 1v4g (unbound) and 1va6 (transition state P2S bound structure). Their structural comparison showed that the binding site of L-CYS is a large open cavity in unbound form (1v4g), however, in 1va6 the conformational transition of switch loop towards this open cavity, nicely partitions the large open site into two sub sites, such that these two sites could comfortably accommodate L-GLU and L-CYS (Figure 4.9). Since ligand induced conformational change is observed in EcGCL, selection of structures for docking studies is critical. For this, we performed blind docking of L-CYS with both 1v4g and 1va6 by considering solvent exposed surface of protein as the search space for docking of rigid L-CYS conformation obtained from structure PDBID: 2d32. It should be noted that choosing L-CYS conformation should not have effect in docking as L-CYS conformation is conserved (Figure 4.5). This was performed 10 times with generating 10 docked conformations in each experiment, thereby, generating 100 independent docked poses of rigid L-CYS. The analysis of these docked conformations showed that on an average L-CYS docked more frequently in funnel open cavity in 1va6 (62%) compared to 1v4g (55%). Based on this and observed switch loop transition we have used 1va6 for rest docking studies.

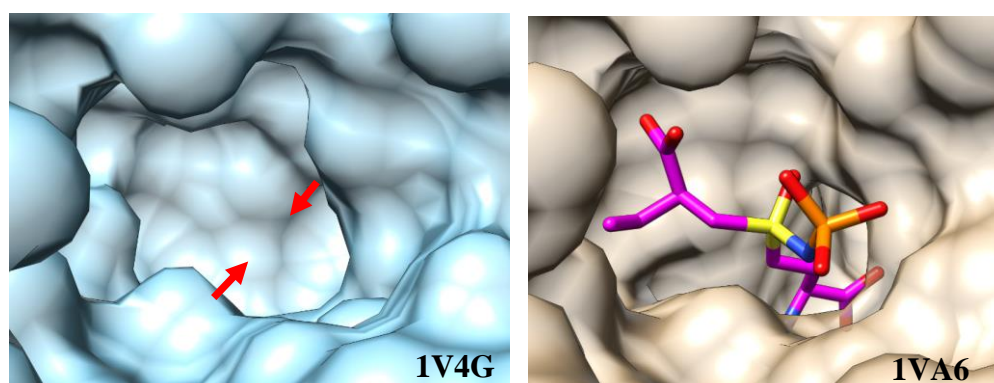


Figure 4.9 Comparison of substrate binding cavity of EcGCL in unbound (1v4g) and P2S ligand bound (1va6). The red arrows show region of structure that moves to partition into subsites.

4.3.2.2 Docking EcGCL with L-GLU followed by L-CYS generating wild type bound form of EcGCL

As described in methods section, we first docked L-GLU in the substrate binding funnel cavity of EcGCL. Among the docked conformations, the lowest energy conformation of L-GLU has flipped orientation with equivalent L-GLU moiety of P2S bound to 1va6. Moreover, to facilitate reaction the phosphorylated γ -carboxylate of L-GLU needs to face outside the cavity and should be spatial proximal to α -amino group of L-CYS. Hence, we have selected the L-GLU conformation (binding energy: -6.8 kcal/mol) with appropriate orientation to facilitate catalysis and have the lowest energy as the L-GLU conformer for subsequent analysis. This EcGCL structure with WT-L-GLU is referred to as EcGCL-glu complex. Further, we docked the L-CYS in the active site funnel-shaped cavity of EcGCL-glu complex, keeps its torsional degree of freedom free. This resulted in 100 docked conformations and 7 clusters upon RMSD based clustering of L-CYS. The careful visual inspection of these clusters showed that predominantly there are 3 sites, which broadly occupied by L-CYS. Of these one site is quite open and does not provide sufficient interacting residues; second site has α -amino group of docked cysteine located ~ 6.5 Å away from γ -carboxylate group of docked conformation of L-GLU. Thus, these two set of conformations may not be able facilitate reaction. The third site has L-CYS conformation with α -amino group of docked cysteine is ~ 3.9 Å away from γ -carboxylate group of docked L-GLU. This docked L-CYS carboxylate group is within hydrogen bonding distance with Tyr-131, Asn-297 and Tyr-300 and α -amino group of L-CYS is hydrogen bonded to Asp-60. The hydrophobic part of L-CYS probably can form hydrophobic interaction with Phe-61, Tyr-131, Arg-132 and Leu-135. We performed limited docking with this site and obtained lowest energy structure of (-4.8 kcal/mol). This shares same overlapping site as cysteine equivalent region of P2S. This docked conformation is EcGCL-glu-cys. We further docked ATP for MD simulations. This structure with all substrates is referred to as 1va6-sb (Figure 4.10A). The putative interacting residues with docked L-CYS, L-GLU and ATP are shown in Figure 4.10B, C and D.

To understand substrate promiscuity of EcGCL, we used simple approach of docking approach and docked 2 set of substrates a) L-CYS and its analogs and b) polyamines of varying methylene chain length to the EcGCL-glu complex.

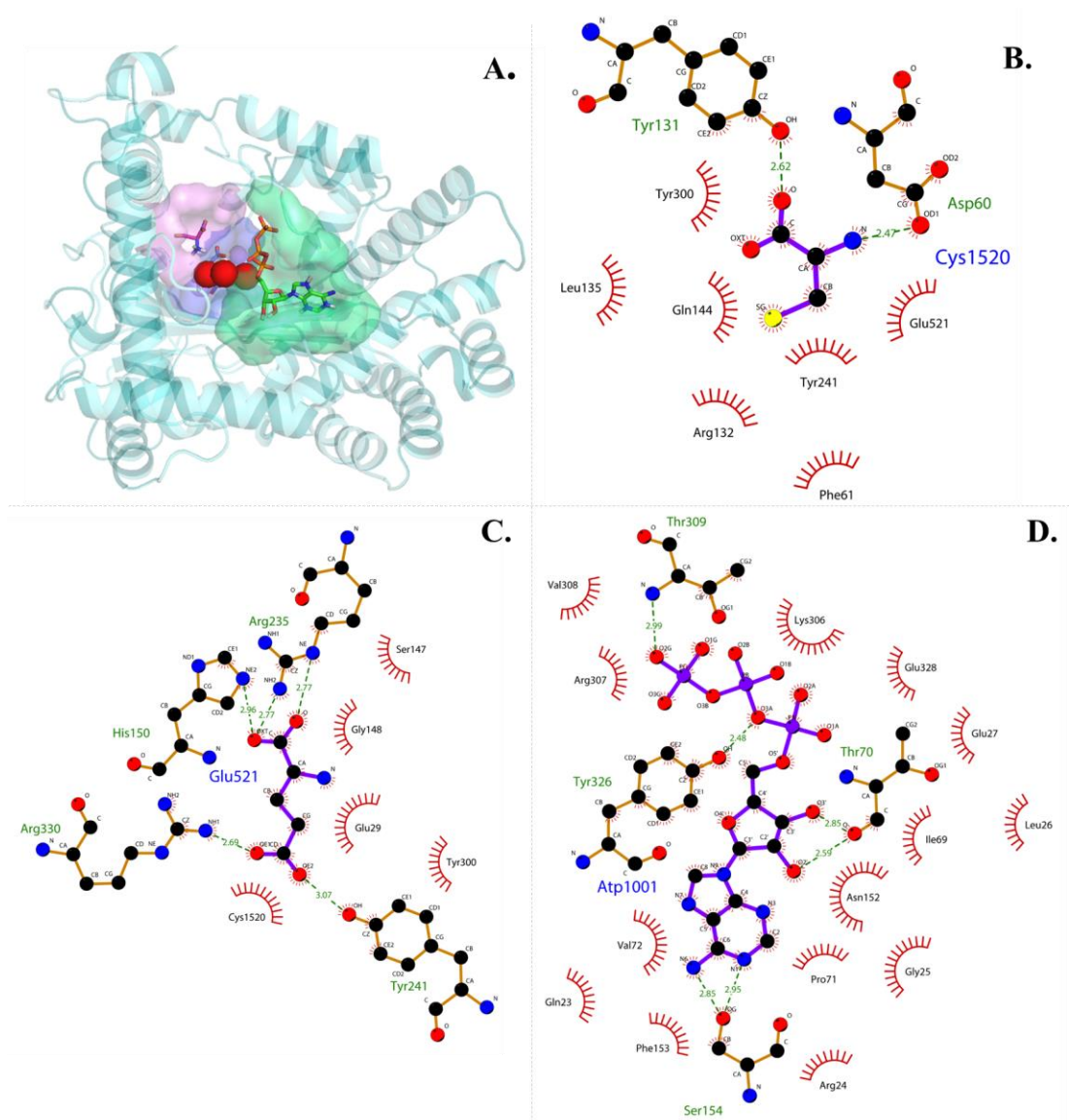


Figure 4.10 EcGCL (1va6-sb) structure showing docked substrates Mg²⁺, ATP, L-GLU and L-CYS. A) Binding cavity shown in transparent surface representation. B, C and D) shows putative interacting residues with docked L-CYS, L-GLU and ATP respectively. The ligands Mg²⁺, ATP, L-GLU and L-CYS are shown by red spheres, green, blue and magenta ball CPK representations respectively.

4.3.2.3 Docking of alternate substrates into the EcGCL-glu complex

To elucidate the putative mechanism of substrate promiscuity exhibited by L-CYS binding pocket, we have docked two categories of alternate substrates a) amino-acids and

unnatural amino acids (set-A) b) polyamines (set-B). The alternate natural and unnatural amino-acids docked include: β -Chloro-L-alanine, L- α -Aminobutyrate, S-methyl- L –cysteine, Glycine, L –Alanine, L –Alanine and L –Valine. From polyamines, we docked Methylamine, Ethanamine, Ethanolamine, Butyl amine and Prop-2-en-1-amine/ Allylamine. These were chosen based on the availability of experimental kinetic data (K_m and V_{max}) or relative enzyme activities with these compounds as substrates (Kelly, Antholine, and Griffith 2002; MIYAKE and KAKITA 2009). We performed docking of these natural/unnatural amino acids and polyamines with AutoDock by keeping these ligands as flexible and grid space for docking is of limited docking conditions. The docking resulted in 100 docked poses and we selected the best conformer primarily using NoE (distance between the α -amino group of L-CYS alternate substrate and γ -carboxylate group of L-GLU) distances and OV (overlap between the binding site of alternate substrate and wild-type docked L-CYS).

Docking of natural/unnatural amino acids and with EcGCL-glu complex

The results of docking of L-CYS analogues (set-A) are summarized in Table 4.1. Based on the overlapping parameter (OV), their α -amino group are close to (NoE distances within 3 Å except in case of L-Valine) the γ -carboxylate group of L-GLU, followed by the best lowest energy docked structure of all alternate substrates binds close to WT-L-CYS binding space, we could find catalytically favorable docked pose for all docked ligands in set-A except L-Valine. Previous studies have shown the role of methylene chain length of the amino-acid substrate in controlling the binding affinity owed to its hydrophobic interaction with the enzyme (Hiratake et al. 2002). The experimentally observed V_{max} for various L-CYS analogs in decreasing order (relative to WT-L-CYS) is as follows S-methyl- L –cysteine > β -Chloro-L-alanine > L- α -Aminobutyrate > L-Serine > L –Alanine > Glycine > Valine (Kelly, Antholine, and Griffith 2002). Among good substrates with comparable relative activities usually have the length of carbon chain up to 3 carbons. For instance, S-methyl- L –cysteine, β -Chloro-L-alanine and L- α -aminobutyrate fit properly in the binding site and have comparable relative activity and binding affinities. Moreover, the binding site of these docked L-CYS analogs and L-CYS (wild-type) share more than 60% of interacting residues as depicted by OV values (Table 4.1).

Interestingly, in these analogs, α -amino group is slanted upwards and closer to γ -carboxylate of L-GLU compared to WT-L-CYS. This is due to additional hydrogen bond between R304 and Q297 with OXT and O atom of the analog respectively. Since this set-A substrates shares the common moiety, *viz.* α -amino group and γ -carboxylate group, residues interacting with these common moieties are found to be conserved for all analogs. However, the side chain (R-) group seems to be responsible for observed reduced binding affinity of the substrate. In general, the amino acids with higher carbon chain length are poor instance. For instance, L-Valine having carbon chain length of 4 has relative V_{max} of just 1.6% in comparison to L-CYS (wild-type). In docking, the lowest energy docked conformation binds away from the WT-L-CYS binding site. This can be explained by the shallow binding site accommodating side-chain (R-group) of the substrate as shown in Figure 4.11. Having a substrate with large R-group may result in steric clashes. On the contrary, substrates with smaller R-group such as L-Ala and Glycine will bind loosely bound to the cys-cavity. These can act as substrate as NoE distance is within 3 Å (Table 4.1), which would facilitate the interaction of α -amino group of L-CYS and γ -carboxylate group of phosphorylated L-GLU, however, may have poor affinity. The analysis of interacting residues revealed that docked L-Ala has lost hydrophobic contacts with residues such as F61 and L135; hydrogen bond interactions with Q144, Y241, and Y300. In case of docked Gly conformer, in addition to lost interaction as with L-Ala, substrate Gly does not have interaction with Y131 and R132. This is also reflected in the overlap between the binding site of L-Alanine and glycine with binding site of WT-L-CYS is 54.55% and 36.36% respectively. This observation indicates that these hydrophobic residues play a major role in the binding of the ligand.

Table 4.1 Summary of natural/unnatural amino acids with EcGCL-glu complex

Ligand name	Km	Vmax	B.E	N-OE	OV
L-Cysteine	0.10 ± 0.02	3590 ± 200 (100%)	- 4.38	3.6(OE2)	--
β -Chloro-L-alanine	0.17 ± 0.04	3120 ± 210 (87%)	- 4.84	2.71(OE1)	63.64
L- α -Aminobutyrate	3.9 ± 0.4	2970 ± 120 (83%)	- 4.79	2.74(OE1)	63.64
S-methyl- L -cysteine	8.1 ± 0.5	3200 ± 55 (89%)	- 5.09	2.75(OE1)	100
Glycine	17.6 ± 0.2	251 ± 25 (7%)	- 3.91	2.67(OE1)	36.36

L -Alanine	21.7 ± 3.6	433 ± 52 (12%)	- 4.34	2.7(OE1)	54.55
L -Serine	24.6 ± 2.8	625 ± 44 (17%)	- 4.24	2.64(OE1)	45.45
L -Valine	27.1 ± 5.5	59 ± 9 (1.6%)	- 3.84	5.53(OE2)	27.27

Interestingly, serine has the same size as cysteine with only difference of hydroxyl group instead of sulfhydryl group, yet it is poor substrate. Despite similarity between Serine and Cysteine, docked L-SER has poor overlap with the L-CYS binding site (~45%). This is mostly because OH group of serine is able to make hydrogen bond with D60. Moreover, there are additional hydrogen bond between R304 and Q297 with OXT and O atom of the L-SER respectively. It is quite likely that the reaction proceeds poorly because of unavailability of free binding site due to its inability to release product because serine is involved in making additional hydrogen bonds.

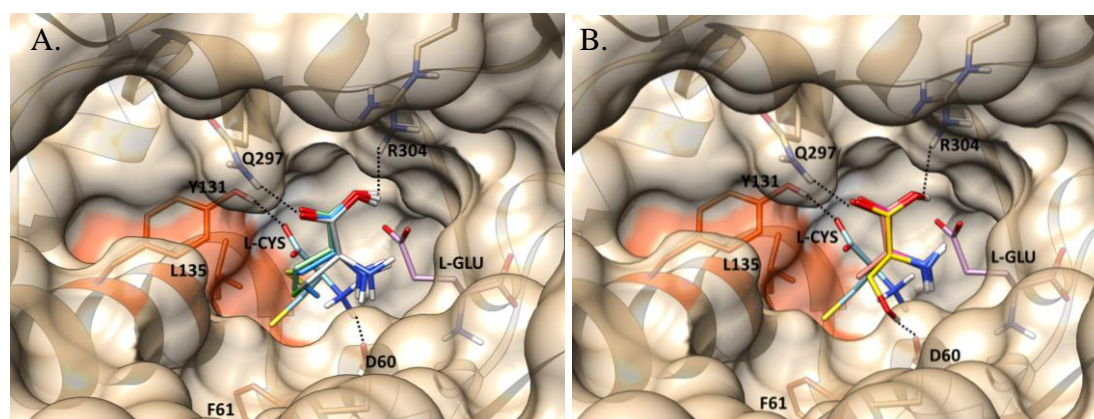


Figure 4.11 Docked conformations of set A substrate in the cysteine binding pocket of EcGCL. Docked conformation of A) good substrate B) poor substrates The hydrophobic pocket involved in interaction with L-CYS highlight with orange color.

The docking studies show that compounds having functional amino group and carboxyl group as in amino acids can act as a EcGCL substrate as far as the side-chain of these could be accommodated in the shallow binding cavity of cysteine. Thus, suggesting EcGCL exhibits substrate promiscuity due to common interactions available for functional groups of ligands and non-polar nature of the interaction site.

Docking of polyamines into the EcGCL-glu complex

EcGCL can catalyze polyamine substrates (set-B) to generate γ -glutamylamides. The relation between carbon chain length and enzymatic relative activity has been observed before. The decreasing order of experimentally observed relative activity of various polyamines (carbon chain length) with respect to WT-L-CYS is as follows: Allylamine (C3) > Butylamine (C4) > Ethylamine (C2) > Methylamine (C1). The results of docking studies are summarized in Table 4.2. The binding energies of all good substrates (Allylamine and Butylamine) are comparable to cysteine binding energy. The amino group of docked polyamines, is within 3 Å of γ -carboxylate group of L-GLU suggesting reaction can proceed with amino group (Figure 4.12). In fact, the interacting residues with N atom between docked polyamines and WT-L-CYS are same (100% OV1) except methylamine, which is a poor substrate. However, in case of ethylamine and butylamine, does not interact with the hydrophobic site of cys-cavity and hence are loosely bound and show relatively poor activity (Figure 4.12B).

Table 4.2: Docking results of polyamines with EcGCL-glu complex).

Ligand name	R.A	B.E	N OE	N-N*	OV	OV1
L-Cysteine	100	-4.38	3.6	--	--	--
Methylamine	19	-1.25	6.55	3.77	54.55	66.67
Ethanamine	48	-4.34	3.87	2.39	81.82	100
Ethanolamine	59	-3.79	3.64	0.29	81.82	100
Butylamine	50	-4.84	3.71	2.12	81.82	100
Allylamine	78	-4.76	3.84	2.43	81.82	100

*R.A- Relative activity with respect to L-CYS (100 %) as observed experimentally

*B.E: Binding energy of lowest docket conformation of docked ligand

*N-OE: Distance between N atom of the ligand and OE1/OE2 atom of L-Glu

N-N: Distance between N atom of the ligand and N atom of docked L-Cys

*OV: Fraction of overlap of ligand binding site with docked cysteine binding site

*OV1: Fraction of overlap of interacting residues with the N atom of ligand with docked cysteine N atom interacting residue

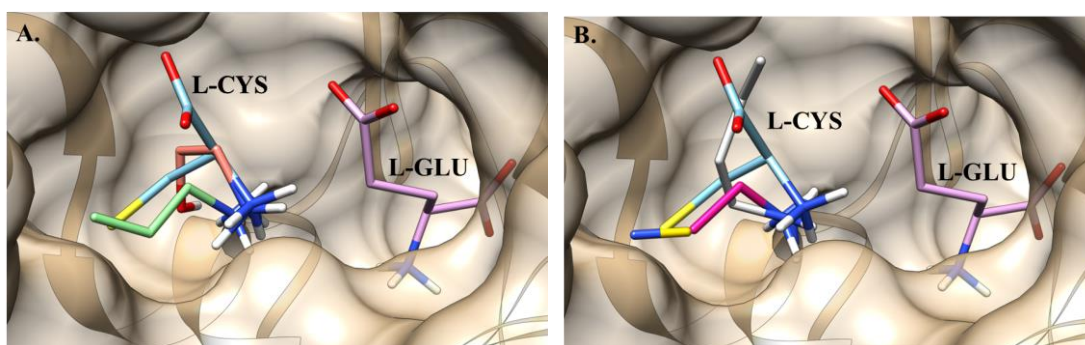


Figure 4.12 Docked conformations of various polyamines in the cysteine binding pocket of EcGCL. A) Good substrate, which have relative activity (>50%) with respect to L-CYS. The α -amino group of these analogs is present at equivalent position of α -amino group of WT-L-CYS. B) Docked conformations of all poor substrates (butylamine, ethylamine and methylamine).

The docking studies of set-A substrates suggested that substrates having both α -amino and carboxy moieties and sufficient long R-group, which can be accommodate in shallow binding cavity of cysteine can be accepted as substrate of EcGCL. However, experimental and docking of set-B substrates showed that interaction with γ -carboxy group is not essential and substrates, which can be accommodated in binding cavity of cysteine having functional α -amino can be catalyzed by EcGCL. This suggests major determinant of substrate promiscuity in EcGCL is the hydrophobic substrate pocket of EcGCL. Additionally, there are many potential hydrogen bond partners in the vicinity of carboxylate group of cysteine, which can stabilize alternate substrates.

4.3.3 MD simulations of substrate bound and unbound structures of EcGCL

In order to investigate the dynamics of various substrates of EcGCL, we performed MD simulations for 50 ns for EcGCL in completely bound state (1va6-sb) *i.e.* when substrates are bound to EcGCL (see section 4.2.3). The simulations with 1va6-sb was performed for 50 ns (production run) for which we analyzed various dynamical properties. First, we analyzed the stability of the three bound substrates (ATP, L-CYS and L-GLU) during the simulation time. We quantitated this using RMSD of ligand throughout the simulation (Figure 4.13).

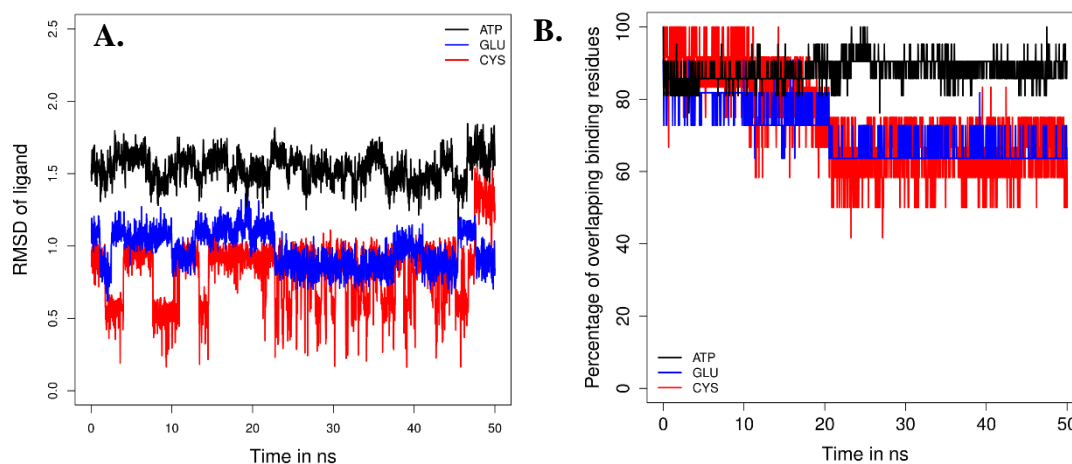


Figure 4.13A) RMSD of all three ligands in GCL group with respect to time in MD simulations. B) Fraction of overlapping residues with respect to the binding residues at starting/reference state.

As seen in Figure 4.13A, L-CYS shows highest fluctuations compared to ATP and L-GLU. Further, we analyzed the overlapping interacting residues of ligands as a function of simulation time to understand their interaction dynamics using starting state as a reference structure. Essentially, this provides the changing interaction pattern between ligand and protein. As can be seen from Figure 4.13B, ATP binding site seems to be similar to the initial state with 80% or more interacting residues conserved during simulations. The change in L-CYS binding site is more compared to L-GLU. Thus, L-CYS is more loosely bound to EcGCL, where binding site overlap drops to 40% in MD simulations. Since L-GLU is also involved in interaction with L-CYS, it also follows a similar trend as L-CYS. However, with a lower fluctuation. The list of interacting residues in both initial state and at least observed at least 50% of time in MD simulation is summarized in Table 4.3.

Further, for the loosely bound L-CYS substrate we computed the percentage of persistent interacting residues with each functional atom (N, SH, OE1, and OE2) during simulation. The mean (SD) persistent interacting residues for atoms N, SH, OE1 and OE2 are 73.92 (6.10), 70.09 (10.92), 40.01 (25.48), and 49.81 (8.58) respectively. Of these, the most dynamical interactions are of the functional atom OE1 and OE2 or carboxylate group. Table 4.4 lists interacting residues for each of these atoms. Despite so fluctuating

carboxylate group in terms of interacting residues, the interaction between α -amino group of L-CYS and γ -carboxylate group of L-GLU is maintained throughout the simulation.

Table 4.3 List of interacting residues of ATP, L-GLU and L-CYS

Ligand	Initial interacting residues	Interacting residues conserved for 50% of time during simulations
ATP	Q23, R24, G25, L26, E27, I69, T70, P71, V72, N152, F153, S154, R304, K306, R307, V308, T309, P315, E325, Y326, E328	Q23 (100), R24 (100), G25 (100), L26 (100), E27 (100), K55 (100), E67 (100), I69 (100), T70 (100), P71 (100), V72 (100), K128 (75), N152 (100), F153 (100), S154 (100), N297 (54), R304 (96), K306 (100), V308 (65), T309 (80), E325 (100), Y326 (100), E328 (100)
L-GLU	E29, D60, L65, I146, S147, G148, H150, R235, Y241, Y300, R304, R330 (putative catalytic residue identified in chapter 3)	E29 (100), D60 (100), L65 (96), E67 (100), Q144 (96), I146 (100), S147 (100), G148 (100), H150 (100), R235 (100), R330 (100)
L-CYS	E29, D60, F61, Y131, R132, L135, Q144, Y241, N297, E298, Y300, R330	E29 (88), D60 (100), F61 (100), Y131 (100), R132 (100), L135 (100), Q144 (83), N297 (88)

Table 4.4 interacting residues with the α -amino group, γ -carboxylate group and Sulfhydryl (SH) group atoms of L-CYS substrate during simulations.

L-CYS atom	Initial interacting residues	Interacting residues conserved for 50% of time during simulations
N	D60, F61, Y241	D60(100), F61(95)
SH	D60, F61, Y131, R132, L135, Q144	F61(92), Y131(83), R132(100), L135(88)
OE1	Y131, L135, Y241, N297, E298, Y300	Y131(92), R132(52), N297(72)
OE2	D60, Y131, R132, L135, N297	D60(59), R132(50), N297(50)

This large interaction dynamics of OE1/OE2 of L-CYS lead us to perform detailed investigation into this because previous work has suggested that γ -carboxylate group of L-CYS equivalent would be stabilized by hydrogen bonds with Y131 and Y300, which in turn makes hydrogen bond with Y241 and Q144 (Hibi et al. 2004). These residues constitute hydrogen bond network. The analysis showed that this hydrogen bond network is not preserved during the simulations. Moreover, while Y131 interacts with the γ -carboxylate group of L-CYS in 92% of the time, Y300 interacts only 35% of the time in MD simulations. Further, L-CYS remains bound to EcGCL in favorable orientation for catalysis because of the hydrophobic interaction of its SH group with L135, F61 and Y131 and polar interaction with R132 (Figure 4.14 and Table 4.3).

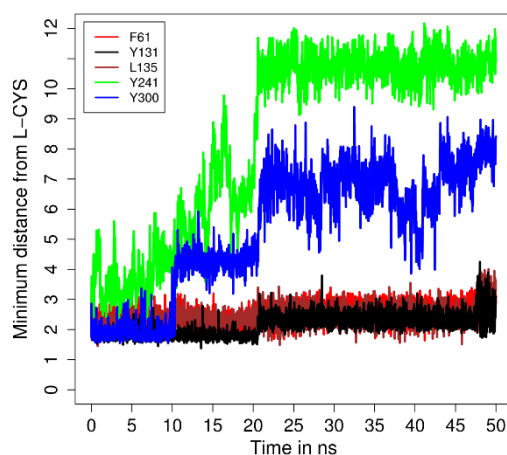


Figure 4.14 Minimum all-atom distance variation between L-CYS and interacting residues (F61, Y131, L135, Y241 and Y300) during simulations.

Moreover, α -amino group is orientated because of D60, which in turn also coordinated with Mg ions. In the initial state, Y300 and Y241 are involved in interaction with γ -carboxylate group of L-CYS. However, due to the motion involving the switch loop of which Y241 is part of the interacting residues are fluctuating. The minimum distance of residues interacting with L-CYS are shown in Figure 4.14. Of these, Y241 and Y300 shows maximum variations.

To assess whether such fluctuations of residues are restricted to ligand bound structures, we performed MD simulations of unbound structures for 50 ns. As discussed in methods section, we have 1v4g as unbound form. Overall, RMSD of bound and unbound structures are similar over simulation time (varying within ~ 2 Å (Figure 4.15A).

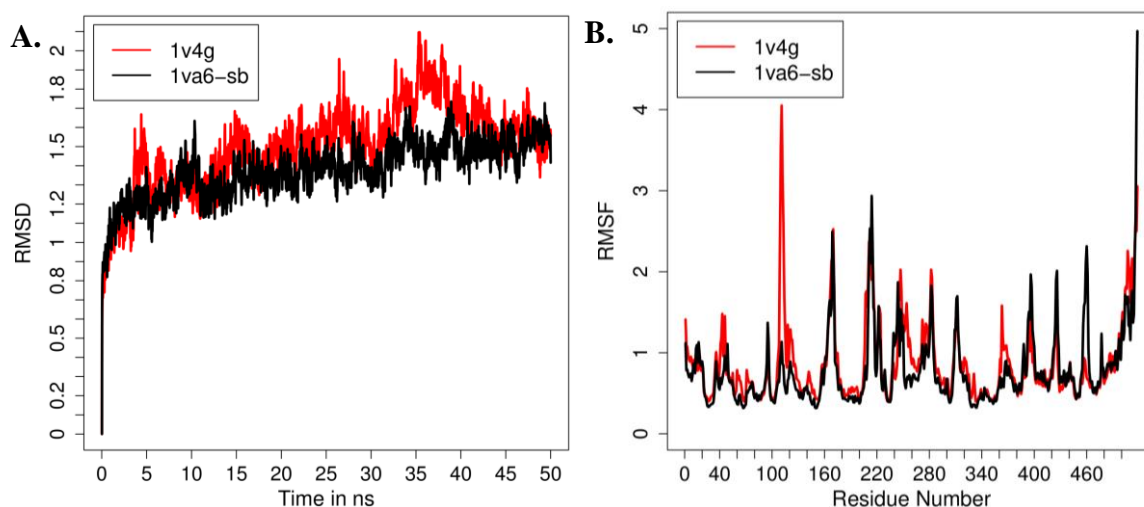


Figure 4.15 A) RMSD values for the residues in unbound and bound structure of EcGCL: 1v4g and 1va6 docked complex with all the substrates bound: Mg²⁺, ATP, L-GLU and L-CYS (1va6 (bound)) and 2) unbound state of 1va6 (unbound) without any ligand bound. B) RMSF values for the protein structures in these three set of simulations.

Next, we computed RMSF to analyze residue variation over a period of simulation time. As shown in Figure 4.15B, RMSF values of residues are comparable in both bound and unbound structures. Especially, RMSF of residue Y300 are comparable and for residue, Y241 it is relatively more for unbound 1v4g compared to 1va6-sb, however, difference is minimal (for Y300 the RMSF values are 0.57/0.59 and 0.68 in MD simulations for 1v4g (unbound state) and 1va6-sb (completely bound state), these values are 0.67/0.72 and 1.14 for Y241). In comparison to other regions of the protein, the dynamic switch loop (residue 240-249) which undergoes conformational change in transition state analog bound state (pdbid 1va6) compared to its unbound state (pdbid 1v4g) does not show significantly high RMSF values (within 2 Å). The other large residue fluctuations are not involved in interactions with ligand. Given that we observed fluctuations in residue interactions involving Y300 and Y241, these residues may play specific functional role in EcGCL such as involved during catalysis. This is plausible as EcGCL is determined for transition state analogue.

Although, RMSF values of switch loop are not significantly high in comparison to rest of the protein residues, the conformation of switch loop deviates significantly from its initial conformation (as evident from fluctuations in residue interactions involving Y300

and Y241, Figure 4.15B and visual analysis of MD simulations). Further, in order to quantitate these deviations in switch loop conformation from its initial state, we computed the local RMSD of the switch loop during the simulations of bound (1va6-sb) and unbound structures (1v4g) and is shown in Figure 4.16. Despite being bound to the ligand, the bound state (1va6-sb) exhibit maximum conformational variability (upto ~ 5 Å RMSD) in comparison to unbound state (1v4g). Further, as seen in crystal structure of 1va6, upon binding to transition state analog (P2S), Y241 and Y300 by undergo significant change in side chain torsion angle χ_1 by 108° and 137° respectively and switch loop undergo transitional state change, next we looked at the conformational change in switch loop. Thus, we computed the all-atom local RMSD of the switch loop (Figure 4.16). The bound state -1va6-sb, despite being bound to ligands undergoes large deviations (upto ~ 5 Å) compared to its unbound state (within ~ 3 Å) (Figure 4.16).

Interestingly, while visual analysis of MD simulation of completely bound state 1va6-sb, we observed certain time steps where the conformation of Y241 is very similar to the conformation observed in crystal structure of unbound state (1v4g). Similarly, while analyzing unbound state simulation (1v4g), Y300 does attain a conformation very close to that observed in 1va6. Thus, during simulations, the bound and unbound state can attain each other's initial conformation of Y300/Y241. For instance, during unbound state-1v4g simulations, Y300 spans the conformational space of bound state- 1va6 as early as around ~ 1.7 ns (frame 84), with $\Delta\chi_1$ of -19.7° . Later again around 25 to 35 ns in 1v4g simulation, Y300 attains conformations very similar to that observed in 1va6 crystal structure. Figure 4.17 shows one such instance at ~ 31 ns (1547 frame), the conformation of Y300 observed in 1v4g (unbound state) is very close to that observed in bound state (transition state analog bound structure) with centroid distance of 0.61 Å and $\Delta\chi_1$ of 1.65° . This suggests that the conformational change specially observed in crystal structure of EcGCL in transition state analog (P2S) bound state (1va6) is not restricted to ligand binding and can also be observed in simulation of unbound state structure. These observations suggest that EcGCL exists in various conformations, which is maximally populated by 1va6 conformation in bound state and 1v4g's conformation in unbound state, with these states interchangeable without the dependence on the presence/absence of ligand. Thus, ligand recognition and binding occurs through conformational ensemble approach rather than induced fit theory as previously suggested (Hibi et al. 2004).

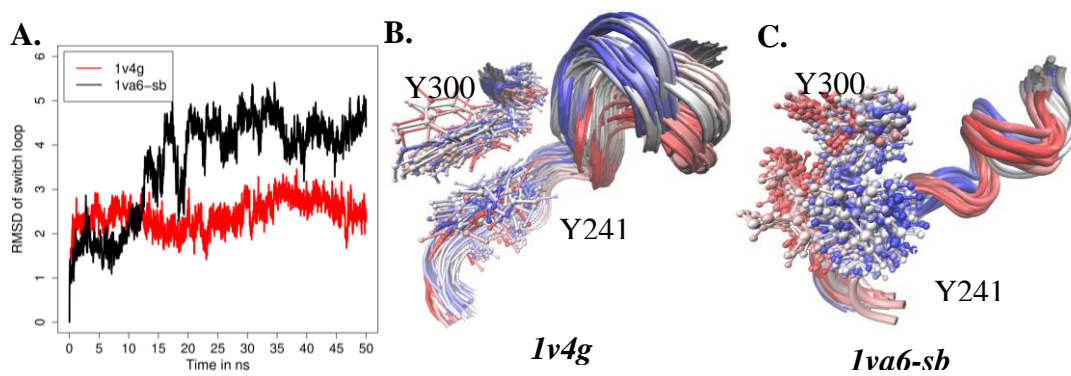


Figure 4.16 A) Local RMSD of switch loop (residues 240-249) in 1v4g, 1va6 (bound) and 1va6 (unbound) state simulations. B) Switch loop conformations in EcGCL bound state -1v4g and bound state -1va6-sb during MD simulations were aligned to illustrate the relative switch loop motions. Using VMD, all the molecular dynamics trajectory structures are superimposed on each other and the structures after every 1 ns are shown using smoothing step size of 20. Color indicates time, with red being the early stages of the simulations and blue indicating the later stages of simulations.

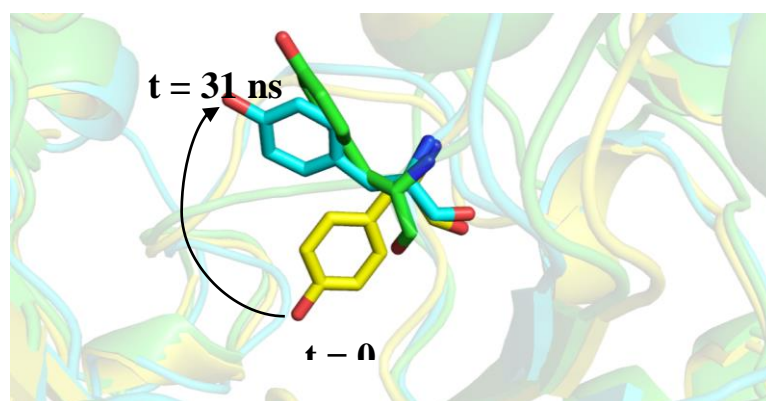


Figure 4.17 The conformation of Y300 observed in 1v4g (unbound state) is very close to that observed in transition state analog bound structure (bound state) with centroid distance of 0.61 \AA and $\Delta\chi_1$ of 1.65° at 31 ns. Similar conformation is attained just at the beginning of the simulations and then at 25-35 ns. Here the green color Y300 residue is of 1va6 crystal structure, while yellow and blue represents 1v4g's Y300 conformation observed at $t = 0$ and $t = 31$ ns during MD simulation.

4.4 Conclusions

The analysis of cysteine binding sites from various enzymes, which use cysteine as substrates showed that cysteine binding pockets are similar as assessed by the pocket score (PS-score) with mean (standard deviation) of ~ 0.5 (0.19). Furthermore, hierarchical clustering of these binding pockets based on PS-score clearly clustered enzymes with

same EC class suggesting there exists a high pocket similarity when enzymes are involved in similar catalytic reactions. Interestingly, the conformational diversity of L-CYS is limited indicating that one should be able to use any of these conformations for docking studies and flexible ligand docking may not be essential. The comparison of suggested EcGCL cysteine binding pocket did not show similarity to any known binding site. Hence, we used docking approach to identify and characterize EcGCL cysteine binding site that showed that suggested EcGCL site (Hibi et al. 2004) is indeed the cysteine binding site and residues Phe-61, Tyr-131, Arg-132 and Leu-135 are involved in hydrophobic interactions and Tyr-131, Asn-297 and Tyr-300 are involved in hydrogen bond interactions with the ligand.

We investigated the substrate promiscuity of EcGCL using docking and molecular dynamics approach. EcGCL is known to bind alternate substrates instead of L-CYS with varying binding affinity. The two sets of substrates a) natural and unnatural amino acids and b) polyamines were docked on EcGCL-glu complex structure. The comparison of overlapping binding sites of these alternate substrates with docked L-CYS conformation and distance between functional groups of L-CYS and L-GLU elucidated that hydrophobic interactions in the shallow cavity of cysteine binding site is primarily responsible for substrate promiscuity of EcGCL. Additionally, there are many potential hydrogen bond partners in the vicinity of carboxylate group of cysteine/alternate substrates, which can stabilize cysteine/alternate substrates. Another important feature of substrates is that it should have hydrophobic chain of sufficient length to be accommodated in shallow hydrophobic cavity of funnel in EcGCL for efficient catalysis.

The molecular dynamics studies of EcGCL docked with all cognate substrates (L-CYS, L-GLU and ATP) showed that among all substrates L-CYS is most loosely bound during the simulation. The detailed analysis MD simulation trajectories showed that residues Y241 and Y300, which undergo ligand induced structural change show conformational variability. This hints to a possibility that these may be required in specific catalytic step of enzymatic reaction. Moreover, these conformations observed in transition state analogue structure have also been observed in trajectory of unbound structures that clearly suggests that loop conformation in EcGCL can exist in both productive (reaction catalyzing state) and non-productive states and upon substrate binding it favors a productive state to facilitate enzymatic reaction.

Thus, this study provides clues towards elucidating the substrate recognition in EcGCL in case of alternate substrates and highlights crucial role of both the hydrophobic binding pocket and the size of binding substrate regulating the binding of substrate in EcGCL. As mentioned earlier, given the central role of GCL in glutathione homeostasis, GCL is an attractive target for drug design. The knowledge gained in this study can be used to design better inhibitors against EcGCL or engineer the active site to accommodate desired alternate substrates. For instance, given the non-specific hydrophobic nature for L-CYS binding pocket in EcGCL, one may have to mutate more than one binding residues in order to prevent EcGCL's binding to L-CYS.

Chapter 5

Enhanced function annotation and phylogenetic analysis of γ -glutamyl cysteine ligase (GCL) superfamily

5.1 Introduction

γ -glutamyl cysteine ligase (GCL) catalyzes the first and rate-limiting enzymatic step in the biosynthesis of anti-oxidant glutathione (GSH) (Seelig and Meister 1984; Meister and Anderson 1983). GSH biosynthesis is two step pathway catalyzed by GCL and GS (glutathione synthetase). In some pathogenic organisms such as *Listeria monocytogenes*, *Streptococcus agalactiae*, and *Pasteurella multocida*, GCL (prokaryotes gene: gshA) and (prokaryotes gene: gshB) are fused into a single enzyme “glutathione bifunctional” enzyme GshF/GshAB/ γ -GCL-GS which catalyzes both of these steps (Janowiak and Griffith 2005; Stout et al. 2012; Vergauwen, De Vos, and Van Beeumen 2006). As mentioned earlier in chapter 4, GCL, in an ATP-dependent manner and presence of cofactor (Mg^{2+}/Mn^{2+}), forms a peptide bond between γ -carboxylate group of L-Glutamate and α -amino group of L-Cysteine to form γ -glutamyl cysteine (γ -GC), which is subsequently conjugated to glycine by GS enzyme to form GSH (Meister and Anderson 1983). The peptide bond of γ -GC is resistant to cleavage by most cellular peptidase (McIntyre and Curthoys 1980, Giannis. and Kolter 1993). Both γ -GC and GSH requires γ -glutamyl transpeptidase (γ -GT) for their cleavage into their respective constituent amino-acids (S. C. Lu 2009). GCL enzymes have been identified in almost all eukaryotes and some prokaryotic phylum. In eukaryotes, GCL has been shown as an essential enzyme required for their growth (Dalton et al. 2004). However, bacteria do not show

any significant growth defect in absence of GCL (Veeravalli et al. 2011a). Given its essentiality in the survival of the cell, GCL is subjected to multi-level regulation of its expression, activity and function viz. transcriptional, post-transcriptional and translational level regulation (S. C. Lu 2009; Franklin et al. 2009; Backos et al. 2013, 2011). Apart from the regulation of its constitutive expression essential for cell survival, GCL levels are also induced by various transcription factors such as Nrf2, AP-1, and NF- κ B in response to GSH depletion caused by oxidative stress and exposure to toxic compounds (S. C. Lu 2009; Franklin et al. 2009). GCL activity regulates cellular GSH levels and dictates GSH biosynthetic capacity. Owing to its rate-limiting capability of GSH biosynthesis, alteration in GCL activity directly equate with alteration in cellular GSH biosynthetic capacity (Franklin et al. 2009). Many therapeutic strategies alter GSH production exploit GCL. Imbalance in GCL functional activity is known to be involved in many human diseases such as cancer, Parkinson's disease, Alzheimer's disease and diabetes (S. C. Lu 2009; Franklin et al. 2009). In these diseases, usually GCL's impaired function leads to decreased levels of glutathione, thereby reducing cellular anti-oxidant capacity and inducing oxidative stress. On contrary, in cancer, GCL's levels are elevated and its levels are enhanced, supporting high cellular proliferation and also imparts resistance towards many chemotherapeutic agents (Backos, Franklin, and Reigan 2012).

Even though GCL sequences from metazoans, plants, and bacteria have same function and share similar catalytic features, relationship among these could not be reliably established purely based on sequence information. This is mostly because of large sequence divergence among these as evident from their low sequence identity. The relationship among members of GCL sequences had been be shown in a previous work using PSI-BLAST searches, albeit with low e-values (0.05) (Abbott et al. 2001). Subsequently, a study on establishing evolutionary relatedness among GCL members was performed by Copley and Dhillon using three conserved regions among GCL members and generated Maximum Parsimony phylogenetic tree to trace its evolutionary history (Copley and Dhillon 2002a). Further, depending on sequence relationship they had classified GCL into three groups viz. Group1 (consists mostly sequences from γ -proteobacteria), Group 2 (sequences are from non-plant eukaryotes) and Group 3 (sequences are from mostly plants and α -proteobacteria). These studies were solely based on sequence information. It was only after the experimental structure determination of at least one member from these three groups that a clear relatedness between them could be

established. Currently there are 13 structures available in PDB for GCL along with their source and their respective group are listed in Table 5.1. The structural similarity computed using TM-align (Yang Zhang and Skolnick 2005) and assessed by all-against-all TM-score pair-wise comparison among representatives from each of the three families clearly shows that these are related despite having insignificant sequence identity (all-against-all TM-score ≥ 0.60). The TMscore between EcGCL and BjGCL is 0.68; EcGCL and ScGCL is 0.61 and BjGCL and ScGCL is 0.79. Ec, Sc and Bj represent *E. coli*, *S. cerevisiae* and *B. juncea* respectively. Ec, Sc and Bj belong to group 1, group 2 and group 3 respectively (Figure 5.1).

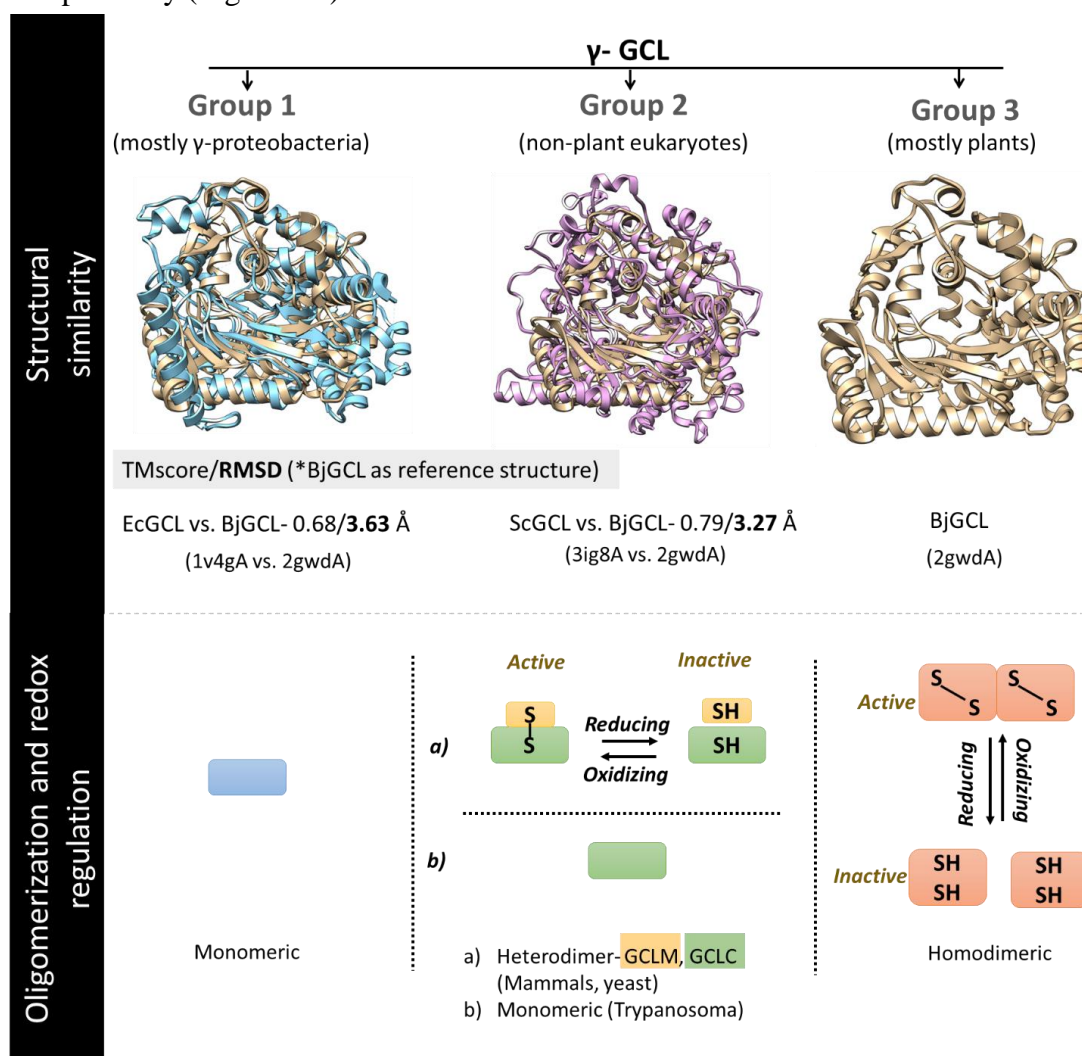


Figure 5.1 Description of various general features of three GCL families

Moreover, members of each group exhibit different oligomeric organization and redox regulation as shown in Figure 5.1. The group 1 representative *E. coli* functions as a monomeric protein (Hibi et al. 2004).

Table 5.1 List of experimentally determined structures of GCL superfamily

pdbid	Source	Family	Ligand bound
1V4G	<i>E. coli</i>	Group 1	--
1VA6	<i>E.coli</i>	Group 1	ADP, MG, P2S, P6G
2D32	<i>E. coli</i>	Group 1	ANP, CYS, GLU, MG
2D33	<i>E. coli</i>	Group 1	ADP, AF3, CYS, GLU, MG
3NZZ	<i>F. tulaensis</i>	Group 1	AMP, SO4
3LN6	<i>S. agalactiae</i>	Group 2 (GshAB)	SO4
3LN7	<i>P. multocida</i>	Group 2 (GshAB)	-
3IG5	<i>S.cerevisiae</i>	Group 2	CSO, GLU, MG, PGE
3IG8	<i>S.cerevisiae</i>	Group 2	ADP, GLU, MG, PGE
3LVV	<i>S.cerevisiae</i>	Group 2	ADP, LBP, MG, PGE
3LVW	<i>S.cerevisiae</i>	Group 2	GSH, PGE
2GWC	<i>B.juncea</i>	Group 3	BSC, MG
2GWD	<i>B.juncea</i>	Group 3	ACT, GLU, MG
1R8G	<i>E. coli</i>	Group 3 (ybdk, weak GCL activity)	--

The group 2 representatives (mammals and yeast) exists as heterodimer composed of a heavy catalytic subunit GCLC (Mr ~70 kDa) and a light regulatory modifier subunit GCLM (Mr ~ 30kDa) (Seelig and Meister 1984; Fraser, Saunders, and McLellan 2002) where while GCLC catalyzes the formation of γ -GC, the GCLM increases the affinity towards L-GLU and decreases the inhibition by glutathione (Seelig and Meister 1984; C. S. Huang et al. 1993; Fraser, Saunders, and McLellan 2002). The association of these two subunits is via intermolecular disulphide bond (Seelig and Meister 1984; Fraser et al. 2003; Fraser, Saunders, and McLellan 2002) and further respond to changes in cellular redox environment to control γ -GC production. Although GCLC protein sequence of *Trypanosoma brucei* exhibit ~45% sequence similarity with mammalian GCLC, *T. brucei* functions as monomer. Group 3 representative plants exists as homodimer and harbor a unique regulatory mechanism based on two intramolecular redox-sensitive disulphide bonds (CC1 and CC2) (Hothorn et al. 2006). The first disulphide bond CC1 is limited to plants from Rosids clade and its reduction allows a β -hairpin motif to shields active site preventing the entry of substrates. The reduction of second disulphide bond

CC2 reversibly controls dimer to monomer transition of GCL converting it from active to inactive state. CC2 is conserved across various representative members of group 3 (Gromes et al. 2008).

GCL is widely studied enzyme and myriad of literature is available on its characterization and properties. Since initial classification of GCL into groups (Copley and Dhillon 2002), numerous GCL sequences from each group are characterized and their crystal structures are available (Table 5.1). However, these studies are limited to a specific group of organism or any one GCL family. Most of the evolutionary sequence analysis also suffers from similar limitations. Moreover, previous studies have not exploited the structural similarity among various GCL groups to understand evolutionary relationships among them.

With the increasing number of completely sequenced genomes, number of GCL sequences have also increased and many experimental studies have characterized functions not involved in glutathione biosynthesis such as EgtA enzyme is involved in ergothioneine biosynthesis (Seebeck 2010). This necessitated detailed classification of GCL superfamily. In the present work, we have systematically performed detailed large-scale sequence analysis and sub-family classification of GCL members for enhanced function annotation. The motivation of classifying sub-families of GCL is to prepare cohesive group of sequences, such that once the function of any one member is elucidated, the annotation for other members is given by inferred from electronic/automated annotation. Previous work has relied on small stretch of conserved sequence motifs or limited set of sequences to derive phylogenetic relationship among three groups. In this work, we have used structure guided sequence alignment with manual curated alignment of relevant motifs to investigate phylogenetic relationship among members of each group and subsequently extended the same to derive evolutionary relatedness between groups to explore evolution of GCL family. Due to sequence divergence among 3 groups, here after we refer compilation of GCL enzymes as GCL superfamily and each group is referred to as GCL family. Hence, we have 3 GCL families, and sequence divergence of each family studied individually and further classified into sub-families.

Thus, the main objectives of this study is to perform a systematic study to: a) analyze sequence divergence of GCL families b) enhance function annotation of GCL

superfamily by classifying family into subfamilies, c) GCL family and subfamily annotation in completely sequenced genomes and d) derive possible ancestral relationship among 3 families using phylogenetic analysis.

5.2 Methods

5.2.1 Construction of GCL sequence dataset

To construct representative set of sequences for each GCL family, the experimentally characterized sequences were taken as a query to search against Uniprot database using PSI-BLAST (Altschul et al. 1997) with an e-value and inclusion threshold for profile generation value of 10^{-3} . The iterations in PSI-BLAST were performed until convergence or up to 20 rounds. The query sequence used for searches was of *E. coli* (Uniprot ID: P0A6W9), *S. cerevisiae* (Uniprot ID: P32477) and *B. juncea* (Uniprot ID: O23736) to retrieve homologs for GCL families 1, 2, and 3 respectively. From the PSI-BLAST output, sequences having an e-value $\leq 10^{-5}$, query coverage $\geq 60\%$ and having a minimum length of 200 residues were considered as homologues of query sequence. The length cut-off was imposed to reduce ambiguous annotations as well as reduce false positives. This cut-off was chosen critically after detailed preliminary analysis such that it allowed inclusion of certain small GCL, which were commonly observed in plant or plant-like GCL, *Phragmites australis* (Uniprot ID: B7U9Z4, length 374 residues).

Modelling of these sequences using I-TASSER (Yang Zhang 2008) revealed these usually have missing first β -strand of catalytic funnel which harbors two glutamate residues critical for Mg^{2+} binding (Hothorn et al. 2006) but rest of the catalytic fold is conserved indicating that these are legitimate GCL sequences. Further from the PSI-BLAST output, we also removed sequence having 'sequence status' in Uniprot as Whole Genome Shotgun (WGS), which are annotated sequences of draft genome. The issue with these is that they can be obsolete in the next update of Uniprot. Moreover, the objective here was to collect as many as true members of GCL families and drastically minimize number of false positives or incorrectly annotated sequences. The domain boundary of GCL was taken as that found in PSI-BLAST results. In case of GCL sequences from plants, we removed transit peptide sequence if present. Pfam (Finn et al. 2016) has three sequence domain families Glu_cys_ligase (PF04262), GCS (PF03074) and GCS2

(PF04107), which corresponds to families 1, 2 and 3 respectively. However, many members of PF04107 profiles does not include first crucial Mg binding site in the alignment, Thus, we have not relied completely on Pfam profiles to extract homologues of GCL rather we performed additional check by running each sequence against these Pfam profiles and found that all sequences were also be identified by them. We did initiate multiple PSI-BLAST runs by taking other member of GCL family as query sequence. However, it did not affect list of homologues. Finally, GCL representative dataset has 1083, 288 and 2325 sequences in each of group 1, 2 and 3 respectively.

5.2.2 HMM profiles of GCL sub-family

5.2.2.1 Generation of HMM profile of GCL group 1 sub-family GshAB

GCL family 1 (group 1) consists of two sub-families, classical GshA and GshAB, which has GshA and GshB fused into single polypeptide chain (as mentioned in section 5.1). Based on domain assignment in the previous step, we identified a total of 195 GshAB sub-family members or bifunctional Gsh, which have N-terminal GshA gene and domain boundaries were derived from PSI-BLAST output. We considered only N-terminal encoding GCL for generating HMM profiles and completely ignored the C-terminal harboring GS. These sequences were clustered using CD-HIT (W. Li and Godzik 2006) at 50% sequence identity that resulted in 8 representative sequences. These were aligned using MUSCLE (Edgar 2004). The sequence of *Simkania negevensis* (Uniprot id F8L825) has an insertion of 30 residues (residue 490-519) toward C-terminal domain region of GshA, which was edited manually. Further, we used ‘hmmbuild’ program from HMMER package (<http://hmmer.org>) to generate HMM profile (Eddy 2011). Using similar approach, we generated HMM profile for GshA sub-family. Each of these HMM profile was compressed and indexed using ‘hmmcompress’. We searched 1083 group 1 members using hmmsearch against these sub-family profiles and found that these could correctly identify and distinguish between GshA and GshAB members.

5.2.3 Subfamily classification of GCL group 3

The GCL family 3 (group 3) is the largest among GCL families and it has 2325 sequences. From the previous works, it is recognized that this family based on function

or sequence relationship can be clearly classified into at least 3 subfamilies, viz. Plant-like, EgtA and YbdK. First, we identified and classified putative members of these subfamilies using following approach. To identify Plant-like subfamily members, we simply searched for all sequences belonging to Viridiplantae taxonomic lineage. This is resulted in 103 sequences. The members of EgtA subfamily members are involved in biosynthesis of ergothioneine. Hence, the members of this subfamily are identified using following criteria: a. sequences having comments ‘used as substrate for the biosynthesis of the low-molecular thiol compound ergothioneine’ included in the CC section of Uniprot summary file, b. sequences having InterPro IPR017809 domain (Finn et al. 2017), which is annotated as putative glutamate-cysteine ligase EgtA, Actinobacteria. Thus, it resulted in sequences classified as EgtA. The YbdK subfamily is based on *E. coli* ybdK sequence, which has weak-glutamate cysteine-ligase activity (Lehmann et al. 2004). However, the native substrates apart from cysteine could not be identified experimentally. To identify YbdK members, we searched for term “weak glutamate--cysteine ligase activity” in the CC section of Uniprot summary file for a given sequence. Further, these can have InterPro IPR011793 domain, which has been annotated as putative glutamate--cysteine ligase YbdK family. This resulted in 1225 YbdK sequences.

Subsequent to identifying and classifying sequences into 3 subfamilies, we generated HMM profiles of representative sequences using same approach as has been described before. The sequences were clustered at 50% sequence identity using CD-HIT (W. Li and Godzik 2006). This resulted in 9, 34 and 240 representative sequences for Plant-like, EgtA and YbdK subfamily respectively. These representative sequences were aligned using muscle and the alignment, which was used in generating HMM profile using ‘hmmbuild’ program of HMMER package. Following ‘hmmcompress’ of 3 HMM profiles, these were searched using ‘hmmsearch’ to find whether these profiles can reliably classify these 3 subfamilies. For this, we searched 2325 sequences against these profiles. The known subfamily members had the best hit to their respective subfamily profiles. Apart from these, we also identified other sequences having significant e-values to one the subfamily profile suggesting they can be classified as one the members of subfamily. For this, we obtained e-values (‘hmmsearch’ of profiles against sequences) for a profile match of a subfamily match to its known members and found the lowest (worst) e-value. We used this e-value as a cut-off for assigning a sequence to any one of 3 subfamilies.

After identifying members belonging to 3 subfamilies using criteria mentioned above, we were left with 459 sequences without any subfamily classification. In order to classify these left-out sequences, we used sequence-based clustering and grouped closely related sequences into subfamilies with an objective that function characterization of any one member would facilitate annotation of rest other family members as well. For clustering sequences, we relied on sequence similarity network (SSN) generated by Enzyme Similarity Network tool (EFI-EST) (Gerlt et al. 2015). The 459 sequences were submitted to EFI-EST server to generate SSN, which generated it by searching sequences among submitted sequences only (“option C: User supplied FASTA file” in EFI-EST server “<https://efi.igb.illinois.edu/efi-est/>”). Thus, obtained SSN of 459 sequences was used to generate initial set of clusters using an alignment score 40 as a cut-off that corresponds to (10^{-40}) e-value. Essentially, this score is used as a threshold to cluster sequences in the network. This step resulted in 8 clusters having 4 or more members and rest 6 sequences could not be clustered within these were not considered in subsequent steps. Since these clusters are related, we performed another grouping of these clusters by using profile search method (Eddy 2011). For this step, we aligned members of each cluster using MUSCLE and generated HMM profile for each cluster. The members of each cluster were used to search against these cluster profiles. We merged those clusters, where most members of a given cluster could be recognized by profile of another cluster with significant e-values. This merging step was performed manually that resulted in a total of 4 clusters. These represents another 4 subfamilies of different GCL group 3, hereafter referred to as: subgroup1 (sb1), subgroup2 (sb2), subgroup3 (sb3), and subgroup4 (sb4). Thus, we classified GCL group 3 in seven putative subfamilies- YbdK, EgtA, Plant-like, sb1, sb2, sb3 and sb4 (Figure 5.2)

5.2.4 GCL family/subfamily annotation in completely sequenced genomes

For protein function annotation in completely sequenced genomes, we have used a total of 5635 completely sequenced genomes (5609 prokaryotic, 24 non-plant eukaryotic and 2 plant (green-algae) completely sequenced genomes) available on NCBI ftp on 09 January 2017. The protein sequences encoded in completely sequenced genomes were obtained from NCBI database (Geer et al. 2009) ftp site (<ftp://ftp.ncbi.nlm.nih.gov/genomes/all/>). Since completely sequenced genomes of many

plants were unavailable, to identify GCL group 3 plant subfamily members, we included genome assemblies available at various draft levels: chromosome (82), contigs (38) and scaffold (175). This resulted in inclusion of 295 plant genomes in our dataset. Thus, finally we have 5930 completely sequenced genomes and draft genome sequences of plants in final dataset for genome annotation. These protein sequences were searched against all GCL families/subfamilies HMM profiles using ‘hmmsearch’ of HMMER package. Subsequently, we also included all Pfam profiles in our profile database to enhance domain annotation of GCL sequences.

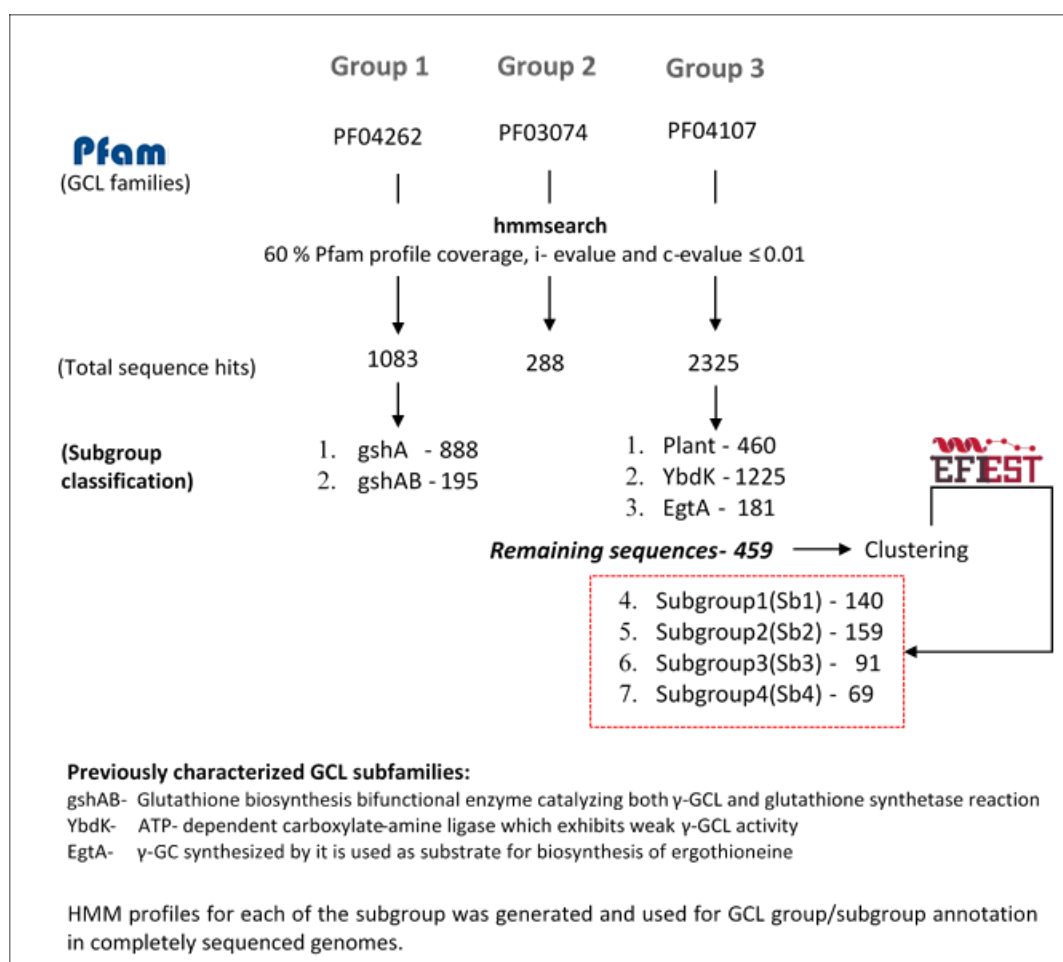


Figure 5.2 represents the schematic for the construction of GCL family/subfamily HMM profiles.

Following approach was used for domain annotation and classification of protein sequences into GCL family/subfamily. In our analysis, a sequence is defined to be GCL member if the sequence has profile coverage of $\geq 60\%$, and an e-value ≤ 0.01 as well as additional domain e-values of c-evalue and i-evalue also less than ≤ 0.01 . The profile coverage is defined as ratio of domain length defined in the sequence divided by profile

length. Since there can be more than one domain in a sequence or a domain can be split into more than one segments, we devised following criteria for domain annotation: a. for non-overlapping domains, the domain boundary is defined as given from ‘hmmScan’, b. For overlapping domains, first we calculate percentage overlap regions, defined as number of overlapping residues between two domains divided by longer domain length. If this overlapping region is $\geq 30\%$, then only the domain with the most significant e-value (numerically lowest e-value) was assigned to the sequence. Otherwise, both domains are assigned with overlapping region assigned to the domain with the most significant e-value. Subsequent to domain assignment, we relied on the ranking of domains based on their e-values to classify sequence to a GCL family/subfamily. A sequence is classified to a family/subfamily of the best-ranked (numerically lowest e-value) domain profile.

This resulted in domain assignment and GCL family/subfamily classification of 5790 sequences encoded in 3596 genomes. These sequences constitute complete genome GCL sequences (CGS) dataset. In comparative genomic analysis, we analyzed enrichment of family/subfamily in any specific taxonomic lineage at some level of taxonomic classification such as class, phylum. Further, we also investigated conserved neighborhood among members of GCL subfamily for function annotation. To facilitate this analysis, we constructed taxonomically non-redundant CGS dataset.

5.2.4.1 Construction of NR CGS dataset

In completely sequenced genome dataset, there are genome sequences available for different strains of same bacteria such as various stains of *E. coli* or *Mycobacterium tuberculosis*. Using such overrepresentation of a certain genus will likely bias the gene neighborhood or functional family/subfamily enrichment analysis. In order to address this, we reduced over counting organisms classified at same taxonomic level by simply representing many genomes by a representative genome such that it encapsulated all distribution features of organisms at that level. For example, different species of *Escherichia* has varied numbers of GCL family/subfamily. Consider strains of *E. coli* genomes GCA_001617645.1_ASM161764v1 has only GCL group 1 and GCF_001623605.1_ASM162360v1 has both GCL group 1 and group 3 YbdK subfamily. Here, the taxonomically redundant data (CGS) for *Escherichia* genus will be defined to consist of both GCL family and GCL subfamily 3. At the taxonomic level of genus, 759

CGS non-redundant genomes were constituted of 1115 GCL sequences. This contains 197 group1-gshA, 41 group1-gshAB, 9 group 2, 254 group 2–YbdK, 207 group 2–plant-like, 142 group 2–EgtA, 128 group 2–sb1, 81 group 2–sb2, 39 group 2–sb3 and 17 group 2–sb4 sequences.

5.2.4.2 Neighborhood analysis of GCL family-3 subfamilies

For function annotation of GCL family-3, we performed gene neighborhood analysis to find any gene conservation pattern around GCL sequences. Here, we define neighboring genes as 10 genes upstream and 10 genes downstream of GCL genes in a given genome. To find conservation pattern or overrepresentation of functional family in genes lying close to GCL genes, we performed Pfam domain annotation following the same criteria as mentioned before in beginning of section to all neighboring GCL genes.

5.2.5 Multiple sequence alignment of representative sequences of each GCL group

As has been mentioned before, the representative sequences for profile generation or phylogenetic studies were obtained by clustering sequences using CD-HIT at 60% sequence identity for each GCL family/subfamily. Table 5.2 summarized the statistics of the redundancy levels. The representative sequences were aligned using MUSCLE. The alignment was minimally manually adjusted to align substrate or metal binding residues.

Table 5.2 Summary of total and representative members of GCL family/subfamily

S.No.	GCL family/subfamily	Representative dataset	Working dataset (NR at 60% sequence identity)
1.	group1-GshA	888	123
2.	group1-GshAB	195	42
3.	group2	288	40
4.	group3-YbdK	1225	235
5.	group3-EgtA	181	37
6.	group3-Plant	460	46
7.	group3-Sb1	140	27
8.	group3-Sb2	159	34

9.	group3-Sb3	91	38
10.	group3-Sb4	69	13

5.2.5.1 Multiple sequence alignment of representative sequences of 3 GCL families

Due to large sequence diversity among GCL family members, automated multiple sequence alignment (MSA) of these does not always result in good alignment as assessed by alignment of conserved structural features, substrates or metal binding sites. In many instances, metal (Mg^{2+}) binding residues or highly conserved glutamate binding residues are not aligned in MSA among three families. We used MUSCLE, ClustalX (Larkin et al. 2007) and T-COFFEE (Di Tommaso et al. 2011) and all these resulted in one or the other issues in MSA. However, the conservation of binding sites is apparent in structure alignment as well as has been reported previously. In order to generate a reliable MSA of all three GCL families, we selected 63 representative sequences from GCL family and subfamilies. Moreover, we also included sequences of known tertiary structure in the list of representative sequences. These sequences were aligned using MUSCLE that resulted in misaligned regions. This starting MSA was modified using Jalview program (Waterhouse et al. 2009) to accommodate features from structural alignments, secondary structure alignment in absence of tertiary structure, and conserved substrate/metal binding residues or motifs mentioned in the work of Copley and Dhillon. To align secondary structures, we used PSIPRED (McGuffin, Bryson, and Jones 2000) to predict secondary structures of all sequences.

The structure alignment of known tertiary structures was performed all possible pairwise structures representing each family. Using the structure-based alignment as a guide, we obtained structurally equivalent residues and used these to initiate the manual curation of the MUSCLE derived MSA. In the process of alignment, it was ensured that predicted secondary structures are aligned in MSA. In case, it requires manual adjustments we followed topology of structures to align secondary structures. While manually editing the alignment, we maintained that automated aligned regions having conserved features are minimally modified. In editing special precaution was taken such that make alignment scores are improved or remain same as given by Jalview. Thus, we

obtained an improved MSA having conserved structural and functional features aligned among all members of GCL. This manually adjusted alignment was further extended to nr60 dataset constituting 632 sequences. The manual alignment of sequences is provided as Jalview project. This alignment is provided in supplementary material (*gsh_grp1_grp3_ybdk_plant.jvp*) for chapter 5 provided in a CD along with this thesis.

5.2.6 Phylogenetic reconstruction of GCL family/subfamily

To understand evolutionary and functional relationship among subfamilies of GCL family and between families we reconstructed phylogenetic tree for each GCL family as well as a combined GCL evolutionary tree of GCL families.

In order to reconstruct phylogenetic tree for each family, first we manually selected representative sequences from diverse set of organisms that maximizes representation of taxonomic diversity. This resulted in 49 representative sequences of group 1 GCL family, of which 19 sequences belonged to gshAB subfamily. A total of 34 sequences were selected as representative for GCL group 2. For group 3, we selected 186 sequences. Of these 30, 39, 22, 29, 38, 19 and 9 sequences belonged to YbdK, Plant-like, EgtA, sb1, sb2, sb3 and sb4 respectively. The set of sequences of each family were aligned using MUSCLE and aligned manually adjusted in case motifs are not aligned for any subfamily members. For a combined tree of all GCL families, we used manually curated aligned generated previously in our work to reconstruct phylogeny.

The phylogenetic trees were reconstructed using Neighbor-Joining (NJ) and Maximum Likelihood (ML) method as implemented in MEGA (Sudhir Kumar, Stecher, and Tamura 2016) suite of programs. For NJ tree, the Jones-Taylor-Thornton (JTT) matrix was used model amino acid substitutions with rate variation among sites modeled using a gamma distribution of alpha value of 1. The partial deletion method was used with site coverage of 80% to handle gaps. For ML tree, the amino acids substitution was modeled using Wheelan and Goldman substitutions matrix (WAG model). The rate variation among sites was modelled with a Gamma distributed with Invariant sites (G+I) (Number of discrete Gamma categories= 5). All the remaining parameters were kept same as used for generating NJ tree. The nodes of tree were assessed using 500 numbers of Bootstrap replications.

The YbdK group 3 subfamily phylogenetic tree was constructed using db60 YbdK GCL group 3 subfamily dataset constructed as follows. Since there are many genomes having multiple members of YbdK subfamily, selecting representative sequences could be a potential issue in phylogenetic tree because we will not be able to distinguish whether selected sequence is an orthologous or paralogous gene. Hence, we constructed a subset of YbdK members, which are present in genome as a single copy gene and clustered them at 60% sequence identity using CD-HIT. These non-redundant sequences (Ybdk-SC) were aligned using MUSCLE and HMM profile was constructed using HMMER package. This Ybdk-SC profile was searched against other members to this subfamily to identify only one gene from each taxonomic lineage.

5.3 Results

5.3.1 Classification of GCL into families and subfamilies

A representative number of 3696 γ -glutamyl cysteine ligase (GCL) sequences were identified using PSI-BLAST searches in Uniprot and combined with Pfam domain family annotation. Of these, 1083, 288 and 2325 sequences could be classified into GCL family-1 (group1), family-2 (group 2) and family-3 (group 3) respectively as shown in Figure 5.2. We quantitated sequence similarity using needleall program of EMBOSS package (Rice, Longden, and Bleasby 2000) with default parameters (gap extension penalty 0.5, gap opening penalty 10 and EBLOSUM62 substitution matrix) within GCL families to understand sequence divergence in a given GCL family. The mean pairwise sequence similarity among sequences of family-1, family-2 and family-3 are 51%, 58% and 36% respectively. The distribution of sequence similarities is shown in Figure 5.3A. As can be seen from Figure 5.3A and mean sequence similarity, both families 1 and 2 are cohesive groups with most sequences (>~75%) having high sequence similarity (>40%). The family-3 shows extensive sequence divergence with ~40% of sequences have sequence similarity <30%.

Since GCL family-3 has such diverged sequences, we classified family-3 into subfamilies to better explore functional divergence in this group. Moreover, such subfamily classification would facilitate better function annotation of sequences. As has been described in methodology section 5.2.3, we have classified family-3 into seven

subfamilies using previous work on characterization of GCL members and sequence similarity network (SSN) clustering approach (Gerlt et al. 2015). Subsequent to classification into subfamilies, we calculated distribution of sequence similarity (Figure 5.3B) for each subfamily. As can be seen from the Figure, most subfamilies are homogenous with respect to sequence similarity. Of seven subfamilies, Plant-like and EgtA are experimentally known to show γ -glutamyl cysteine ligase activity (Hothorn et al. 2006; Seebeck 2010; Musgrave et al. 2013). However, only members of Plant-like subfamily have been suggested to be involved in biosynthesis of glutathione. We identified 460 sequences belonged to Plant-like subfamily, which included most sequences from plants and some prokaryotic sequences of α -proteobacterial origin. Apart from glutathione biosynthesis, GCL homologues may play role in disease resistance (Kular et al. 2004) and participates in detoxification process (Schäfer et al. 1997) in plants.

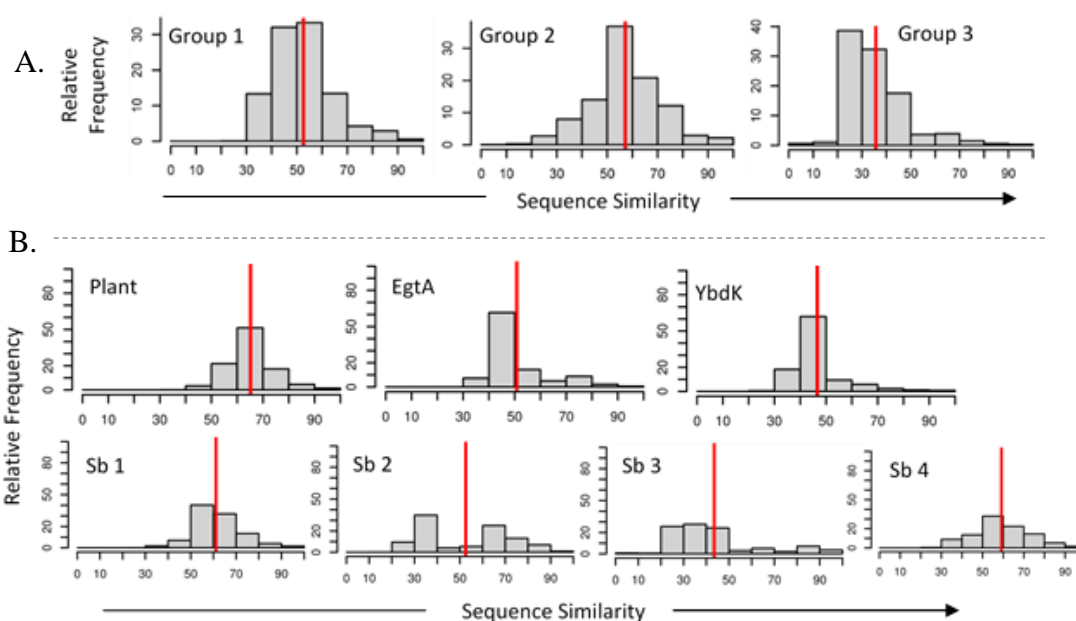


Figure 5.3 Pairwise sequence similarity distribution of A) GCL family and B) subfamilies of GCL family-3.

Even though members of EgtA subfamily show same enzymatic activity to produce γ -glutamyl-cysteine, these are recruited in ergothioneine biosynthesis (Harth et al. 2005). We classified 181 sequences in this subfamily. It has been shown that EgtA lies within EGT operon (EgtABCD). All members classified in this subfamily are similarly present as part of operon. The YbdK subfamily is characterized and named after *E. coli* gene ybdK. Its structure determination in 2004 showed that YbdK shares structure

similarity to GCL enzyme (Lehmann et al. 2004). However, experimental studies suggested it is ATP-dependent carboxy-amine ligase that exhibits weak γ -GCL activity (Lehmann et al. 2004). A recent study on finding alternate pathways of glutathione biosynthesis in *E. coli* suggested that YbdK may not be involved in glutathione biosynthesis (Veeravalli et al. 2011b). This leaves possibility of YbdK enzyme involved in ATP dependent conjugation of glutamate with other substrates such as other amino acids. However, we do not know the precise function of YbdK. The idea of classifying sequences in this subfamily is that experimental function characterization of any one YbdK would facilitate annotation of rest subfamily members as well. As described in methods section, we classified remaining 459 sequences of family-2 into 4 subfamilies based on Sequence Similarity Network approach. These subfamilies are referred to as subgroup1 (sb1), subgroup2 (sb2), subgroup3 (sb3), and subgroup4 (sb4) with each having 140, 159, 91 and 69 sequences respectively. Of these, sb1 sequences are mostly belong to Actinobacteria, sb2 have sequences from cynaobacteria (few Euryarchaeota also present), sb3 sequences are mostly from γ -proteobacteria and Euryarchaeota sequences are prevalent in sb4 subfamily.

5.3.2 Analysis of conservation of substrate/metal binding residues

Having classified GCL superfamily into 3 families, subsequently into subfamilies, we analyzed the degree of binding site residue conservation of substrates (L-Glu, L-Cys, and ATP) and Mg^{2+} , even though this has always been assumed to be conserved. For this, we performed MSA of non-redundant representative sequences from each family (Table 5.2). The conservation of motif is assessed using Information Content (IC) of motif, which is simply summation of IC motif position. Essentially, this will quantitate how different a given motif is from its uniform distribution. Since IC is opposite of entropy, we first calculated the Shannon entropy (Strait and Dewey 1996) of each position in the given motif. Shannon entropy (H_n) of each binding site position is given by the equation:

$$H_n(p_1, p_2, \dots, p_n) = - \sum_{i=1}^n P_i \log_2(P_i)$$

where P_i is the frequency of amino acid i in that motif and n is the number of characters, which is 20 amino acids in case of proteins. The Shannon entropy ranges from 0 to 4.32 for a complete conserved position (Only one amino acids occur in a position) to

occurrence of 20 amino acids with equal frequency. Information content of a motif position is defined as maximum entropy minus the entropy given at that position, given by equation:

$$IC_i = 4.32 - H_i$$

and IC of motif is given by:

$$IC_{motif} = \sum_i^w IC_i$$

where H_i is entropy of a given i^{th} position and w is number of positions in a motif. In order to compare sequence motifs from two families of different lengths we normalize IC_{motif} with respect to length and calculated IC/residue (IC_r) of a motif, given by more generalized equation:

$$IC = \frac{4.32 \times L - \sum_{n=1}^L (-H_n)}{L}$$

where L is the length of the motif, H_n is the Shannon entropy of the amino acid residue i in the given motif. Thus, IC_r varies from 0 to 4.32, which shows no conservation to completely conserved motif residues.

To define residues involved in binding substrates/metals and subsequently suggest these as binding site motifs, we used known tertiary structures to find substrate/metal binding site residues, which were derived from LPC (Sobolev et al. 1999) or literature. In absence of structure not bound with substrates, we performed limited docking to find interaction residues (see chapter 4).

Mg²⁺ binding motif: There are two/three Mg²⁺ bound to GCL. Of these, one Mg²⁺ lies close to both L-GLU and L-Cys and other two are involved in coordination of phosphates of ATP. There are six residues involved in coordination of 3 Mg²⁺ and a residue can be involved in interaction with more than one metal ion. It is evident from Figure 5.4, that metal binding site is highly conserved as $IC_r > 4.1$. Figure 5.4 MSA of each GCL family representative sequences showing conserved substrate/metal binding residues as well as their neighboring residues. Group 1-gshA and gshAB sequences are shown in black and blue color respectively. In group 3 subfamily, YbdK is colored black, Plant-like is green, EgtA is salmon, sb1 is magenta, sb2 is cyan, sb3 is pink and sb4 is ice blue in color. The

motifs are represented using sequence logos. The residues are not necessarily consecutive in the MSA and residue numbers are given below the alignment.

However, sb2 subfamily consisting mostly cyanobacteria it is substituted by Histidine. The only variation in motif is observed across family-2 and 3 is that Histidine is substituted by Glutamine. The structural equivalent residue of this residue in *E. coli* is H150. In family-3, residue Glutamine is predominantly present in Plant-like, EgtA, sb1, sb3 subfamilies, and some Euryarchaeota.

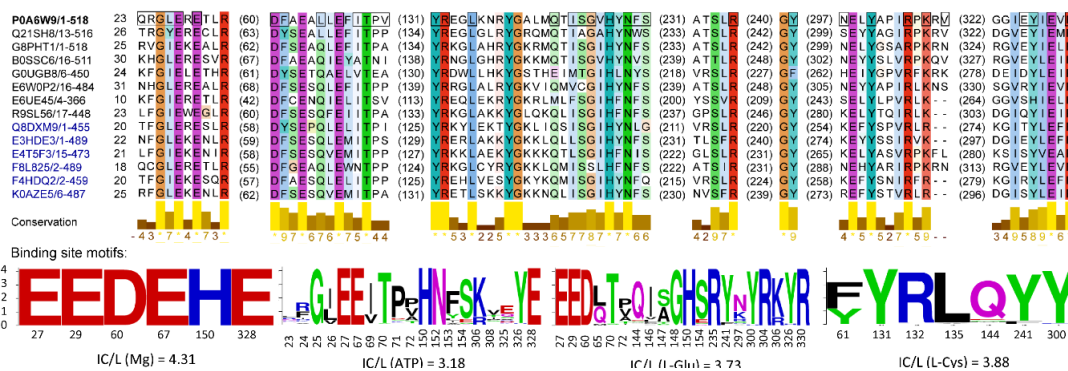
L-Glu binding motif: The residues involved in binding L-Glu are mostly conserved. As can be seen in Figure 5.4, it is highly conserved in family-2 and poorly in family-3. It is mostly attributed to many subfamilies in family-3, which would have diverged function and similarly binding site residues may not be conserved.

ATP binding site: The conservation of binding residues is relatively conserved in family-1 and family-2 than family-3 (Figure 5.4).

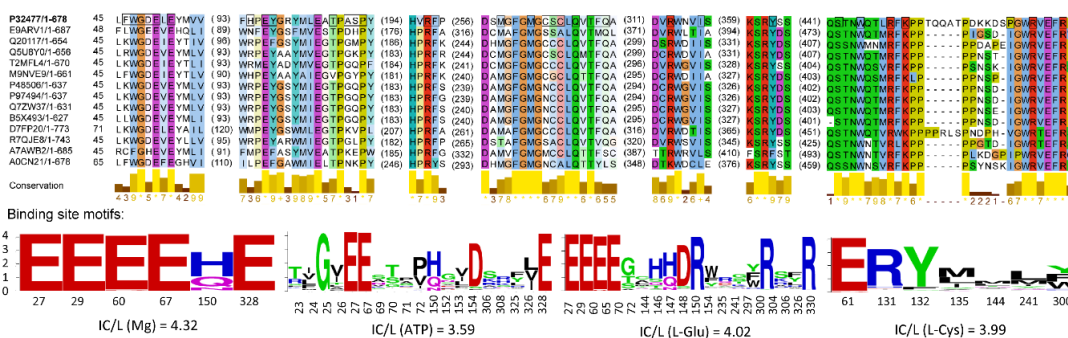
L-CYS binding motif: the conservation among families does not vary much, except family-3, which has slight poor conservation. This indicates the selection pressure on residues involved is not there because the substrate may have diverged or binding affinities among subfamilies is variable.

Apart from this, the motif EXR consisting the putative catalytic residue (R330/R472/R387 in *E. coli*/*S. cerevisiae*/*B. juncea*) is highly conserved as well indicating the even though GCL families bind to different substrate with different affinities, the catalytic reaction mechanism remains the same. Among GCL families, family-3 seems to be highly divergent followed by family-1 and family-2 being the most conserved one. The detailed analysis of Group 3 subfamilies, showed that Plant-like and EgtA subfamilies have similar binding motifs as shown in Figure 5.5. The putative catalytic residues are conserved in all subfamilies (Table 5.3).

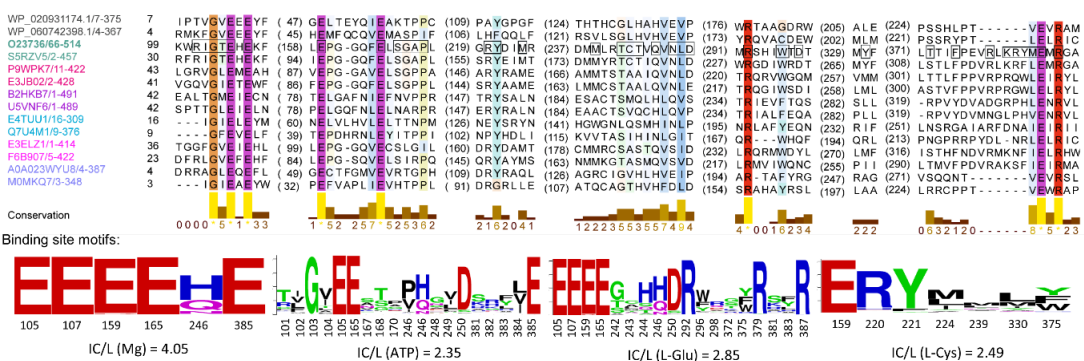
Multiple sequence alignment for representative sequences of group 1:



Multiple sequence alignment for representative sequences of group 2:



Multiple sequence alignment for representative sequences of group 3:



*IC/L= Information content of the motif (IC)/ Length of the motif (L)

Figure 5.4 MSA (multiple sequence alignment) for representative sequences of each GCL group, encompassing binding/catalytic and their neighboring residues. Here group 1-*gshA* and *gshAB* sequences are shown in black and blue color respectively. In group 3 subfamily, *YbdK* is colored black, *Plant* is green, *EgtA* is salmon, *Sb1* is magenta, *sb2* is cyan, *sb3* is pink and *sb4* is ice blue in color.

Table 5.3 List of conserved residues across GCL group 3 subfamilies and their putative role in substrate/metal binding.

S.No.	Residue	Specific role of the residue
1.	E105, E107, E159, E165, E385	Mg binding
2.	P170 (conserved except in YbdK)	ATP binding
3.	E385	ATP binding
4.	R292	L-GLU binding
5.	R379(Except in sb4)	L-GLU binding
6.	R387	L-GLU (catalytic role)
7.	R220 and Y221(except in Sb2)	L-CYS binding

We analyzed the conservation of binding site among members of subfamilies of family-3. The metal (Mg) binding site in GCL family-3/group3 subfamilies is mostly conserved, except residue Q246 (in *B. juncea*). In the sequence alignment, this position is mostly Gln, but in some sequences it is substituted with His. Among subfamilies, Plant-like, EgtA and sb1 predominantly have Gln, whereas in YbdK and sb4 it is mostly Histidine. The same is observed in structurally equivalent position in family-1.

The ATP binding site in GCL group 3 subfamilies is variable. The crucial ATP binding residues are discussed further. The residue G103 involves in hydrophobic interaction with the sugar moiety of the ATP is highly conserved in all the group3 subfamilies except in sb4, which have Ser substituted at this position in few members. The residue S167 is highly conserved only in EgtA and Plant-like subfamilies. sb3 has Ala substituted in few cases where as sb1 and sb2 have Asn and Lys substituted at this position. This position is not conserved in YbdK subfamily. G168 position is conserved only in Plant-like. EgtA has S/G and YbdK has T/S at this position. Other binding residues, which differ in GCL group 3 subfamilies are highlighted in the *Table_5.S11_grp3_subgrp_motifs.xlsx* provided in supplementary material for chapter 4 provided in a CD along with this thesis.

5.3.2.1 Conservation of residues involved in regulatory mechanism in Plant subfamily in other Plant-like GCL group3 subfamilies- sb3 and EgtA subfamilies

As shown in Figure 5.5 that EgtA and sb3 substrate binding motifs are similar to Plant-like subfamily. The Plant-like subfamily is known to have unique β hairpin motif (326-346 residues in *B. juncea*). This motif undergoes conformational change depending on the disulfide bond CC1 formed between C341 and C356 (residue number is from *B. juncea*). In their reduced state, the motif lies above the catalytic funnel thereby shielding entry of substrates resulting in slowing down binding of new substrates and release of products (Hothorn et al. 2006). However, CC1 is conserved only in plant species from Rosids clades. Another disulfide bond CC2 between residues C178 and C398 (residue number is from *B. juncea*) has been suggested to stabilize homodimeric configuration by bringing two helices together in CC2 oxidized state as these helices lie at the dimer interfaces (Gromes et al. 2008). Interestingly, most amino acid contributing to the homodimer interface in BjGCL are highly conserved in GCL of Viridiplantae group, however, these are not conserved in related proteobacterial GCL (Gromes et al. 2008). Given these, we asked whether a) the unique β hairpin motif and b) homodimer interface residues in BjGCL are conserved in subfamily members of sb3 and EgtA. The β hairpin motif is conserved in EgtA subfamily and in few members of sb3. Among EgtA members, the equivalent residues of C341 and C356 in sequence alignment are not conserved. This suggests the similar regulation involving β hairpin motif may not be present in EgtA or sb3 subfamilies. The equivalent residues lying at the homodimer interface of BjGCL is not conserved in EgtA and sb3 suggesting that members of this subfamily may not form homodimer, however, if they do so would involve interface region different than observed interface in Plant-like subfamily.

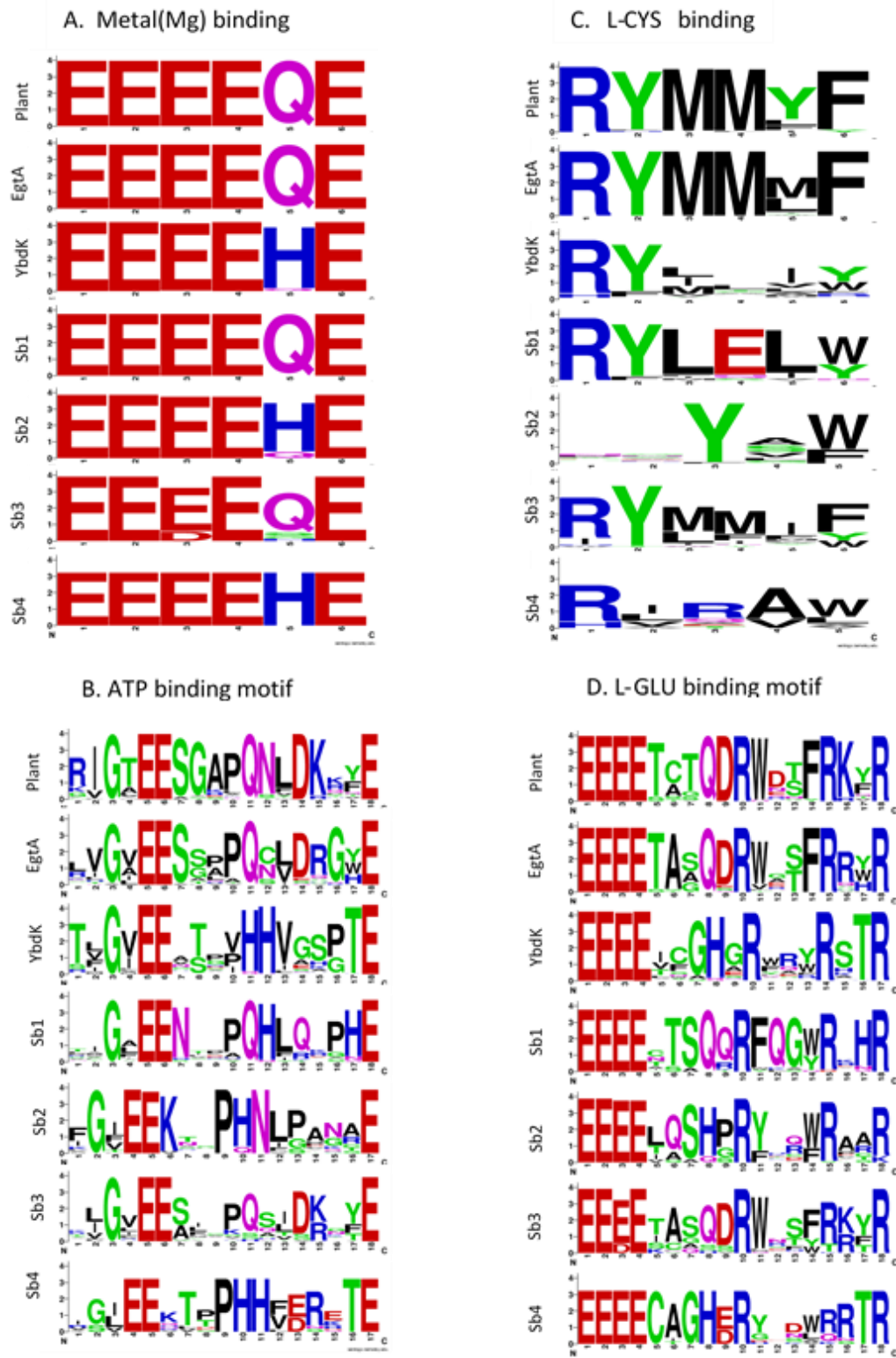


Figure 5.5 Binding site residue conservation in subfamilies of group 3. Motifs are represented using sequence logos. The residue numbers of each position in the motif is followed as in MSA of group 3 shown in Figure 5.4.

5.3.3 Function annotation of GCL homologues encoded in completely sequenced genomes

Subsequent to analysis of GCL family/subfamily, we performed enhanced function annotation of genes encoded in completely sequenced and draft plant genomes. For this, we used ‘hmmscan’ to identify and classify identified GCL homologues from the proteome of each of these 5930 genomes by searching against GCL family/subfamily HMM profiles. As described in methods section 5.2.4, a sequence is assigned as GCL homologue, if it has 60% Pfam profile coverage having both i-evalue and c-evalue ≤ 0.01 . Following these criteria, we have identified 5228 GCL homologues encoded in 3596 genomes. These are further classified into 3 families and their respective subfamilies. Table 5.4 summarizes statistics of members of GCL families/subfamilies identified in genomes. This dataset (*c5_csg_gcl_annotation.xlsx*) is provided in supplementary material along with their fasta sequences in *c5_csg_gcl_fasta* folder for chapter 5 provided in a CD along with this thesis. The gcl annotation for these genomes is also available at <http://14.139.227.206/gcl/csg/>.

Table 5.4 List of genomes sequences classified into GCL families/subfamilies.

S.No.	GCL subgroup	Total number of CSG	Total number of significant sequence hits
1.	group1-gshA	1777	1812
2.	group1-gshAB	325	327
3.	group2	18	19
4.	group3-YbdK	1831	2137
5.	group3-EgtA	264	272
6.	group3-Plant	550	741
7.	group3-Sb1	194	205
8.	group3-Sb2	168	169
9.	group3-Sb3	88	88
10.	group3-Sb4	13	21

From Table 5.4, it can be observed that there is more than one gene annotated as YbdK subfamily in many organisms suggesting extensive duplication of Ybdk subfamily

members has taken during the course of evolution. There are at least 1831 YbdK members identified in 1613 genomes (considering only one YbdK annotated genes from each genome). There are 218 genomes having multiple proteins classified as belonging to YbdK subfamily. These include Actinobacteria members like *Mycobacterium*, *Streptomyces*, *Amycolatopsis*, *Rhodococcus* and few γ -proteobacteria like *Legionellaceae* and *Pseudomonas*. We also observed YbdK subfamily in multiple copies of *Methylobacterium* (α -proteobacteria) and *Paraburkholderia* (β -proteobacteria) genomes.

To understand distribution of GCL families/subfamilies across various taxonomic lineages, we generated the dataset of GCL annotated such that counting is reduced across taxonomic lineage (see methods section 5.2.4.1). Figure 5.6 shows taxonomic distribution of GCL family/subfamilies.

As can be seen from Figure 5.6, GCL family-1 predominantly consists of sequences from γ -proteobacterial origin and this is specifically observed for classical GshA subfamily. However, most firmicutes (gram positive bacteria) contributes to fused or bi-functional Gsh subfamily (~64%) of family-1. The family-2 consists of GCL from eukaryotic origin except plants of Viridiplantae group. Interestingly, family-2 has representation from non-green algae (Rhodophyta), which should evolutionary be classified in Plant-like subfamily of GCL. Since many fungal genomes are known, dikarya group dominates in family-2 taxonomic distribution. The family-3 taxonomic profile is heterogeneous, as these constitute many prokaryotic phyla and eukaryotic plant sequences. In Plant-like subfamily of family-3, most GCL are from α -proteobacteria and Viridiplantae. Actinobacteria is predominantly present in EgtA and sb1 subfamilies. The archaea and Cyanobacteria is mostly classified within sb4 and sb2 subfamilies respectively. The sb3 subfamily has mostly firmicutes. Among all subfamilies of family-3, YbdK subfamily is taxonomically most diverged group as it consists of sequences from many taxonomic lineages.

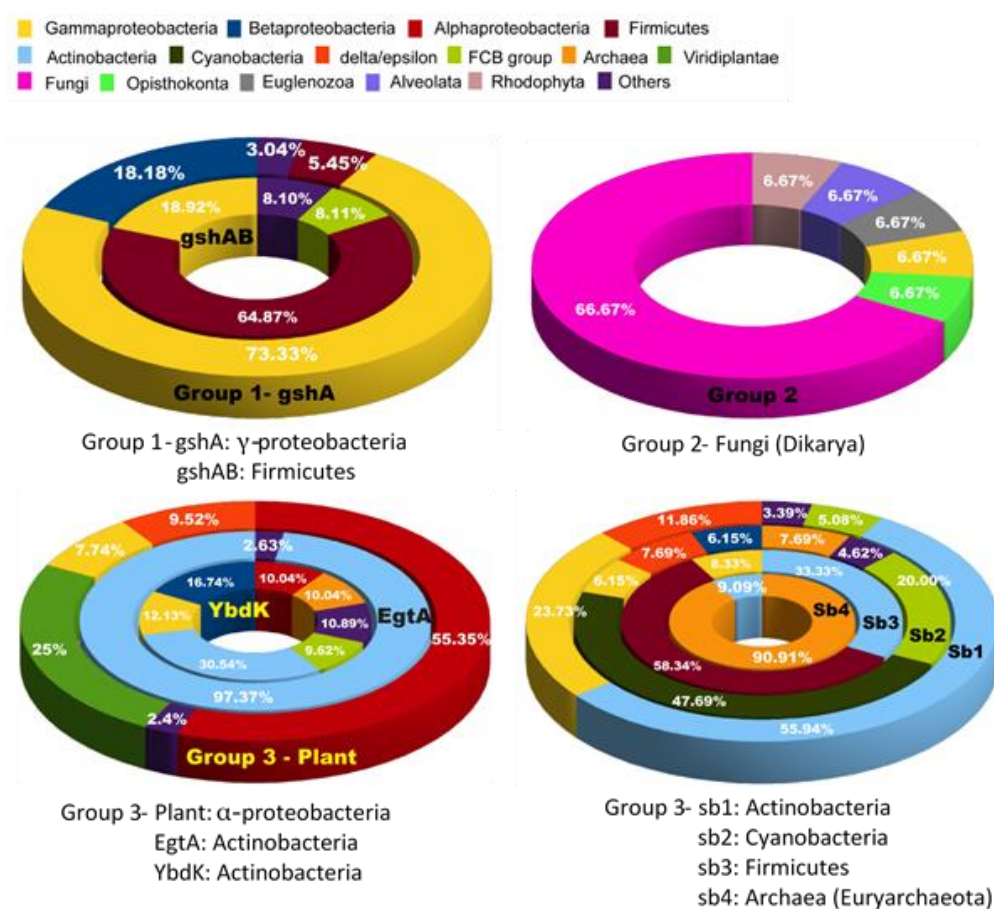


Figure 5.6 Enrichment of each GCL family/subfamily in various taxonomic lineages.

We analyzed occurrence of GCL subfamilies within an organism and it showed that like has been observed for YbdK subfamily, many organism has duplicated plant-like subfamily members. However, in these cases duplicated sequences occur as neighboring genes (with ± 10 genes upstream/downstream of GCL sequence belonging to plant-like subfamily) suggesting these are mostly likely recently duplicated genes. The comparative genomics analysis showed extensive duplicated genes classified as subfamily YbdK are present within Actinobacteria. Interestingly, YbdK subfamily members were found even in genomes not known to produce glutathione such as Mycobacteria. However, Mycobacteria and actinomycetes produce high levels of another low-molecular-weight thiol - Mycothiol for regulating their reduction or oxidation activities (Gerald L. Newton et al. 1996)(Fahey 2013). Such an expansion in bacterial lineages of YbdK subfamily, which has weak GCL enzymatic activity suggests that γ -glutamylcysteine could also serve similar role as glutathione and whole biosynthesis

pathway is not required/absent in these organism. Moreover, in the Halobacteria, an aerobic subgroup of the archaeobacterial, does not have gshB gene which codes for glutathione synthetase and catalyzes the second step of glutathione biosynthesis. In fact, Halobacteria can use γ -GC are major low-molecular-weight thiol and appears to function similarly to glutathione (G. L. Newton and Javor 1985). γ -GC presence instead of glutathione in ancestral prokaryotes suggests that glutathione biosynthesis pathway could have evolved later during evolution as organism specific requirements. This is also supported based on complex evolutionary history of GshB that suggests not a common evolutionary origin of GshB (Copley and Dhillon 2002a).

5.3.3.1 Degree of neighborhood conservation among members of GCL subfamilies

Since family-3 has many functionally not well-characterized subfamilies, we have investigated the conservation of genes or functional annotation of neighboring GCL genes to gain insight into their putative functions. We performed neighborhood analysis as described in methods section 5.2.4.2 for each of the seven subfamilies of family-3. In completely sequenced genome dataset, there are many genomes coming from same organism but of different strains, for instance, many strains of *E. coli*, *Mycobacterium tuberculosis* have been completely sequenced. To remove this bias and avoid any over count of certain organisms, we constructed the NR CGS dataset for this analysis as described in methods section 5.2.4.1.

The EgtA subfamily, a five-gene cluster (egtA, egtB, egtC, egtD and egtE) responsible for EGT production in *Mycobacterium smegmatis* has recently been identified (Seebeck 2010). This gene cluster is ubiquitously present in all Actinobacteria in our NR CGS dataset as well.

In case of plant-like subfamily, few members (11 out 207 plant GCL in NR CGS dataset) have another GCL group 3 plant-like subfamily member copy in their neighborhood. These genomes include *Citrus sinensis*, *Camelina sativa*, *Prunus mume*, *Pyrus X bretschnideri* and *Gossypium hirsutum* which have 7, 5, 3, 3 and 2 GCL plant-like subfamily member copies in its neighborhood respectively. It should be noted that plants possess multiple copies of GCL- for instance in NR CGS dataset, out of 207 genomes having GCL family-3 plant-like subfamily sequence, 140 genomes possess more than

one GCL plant-like subfamily. The neighboring genes in Plant-like subfamily show enriched Pfam domain PF01657, which is stress-antifungal domain with plays an important role in salt-stress response and has antifungal properties.

The neighborhood of YbdK subfamily annotated genes is found to be highly diverse and we could not find any specifically enriched Pfam domain family. However, Pfam domain PF00005, an ATP binding domain of ABC transporters is one of the most commonly occurring domain among neighboring genes. Interestingly, this domain is also enriched in the GCL gene neighborhood of subfamilies – sb1, sb2, and sb3 subfamilies. Other Pfam domains commonly found in the GCL neighborhood of YbdK subfamily were usually DNA-binding domains like PF02518 and PF00072. In certain group of organisms (44 genomes of 254 genomes) YbdK subfamily is found to be co-occurring with genes having MarR_2 (PF12802) domain. MarR is the repressor for MAR proteins, which are involved in multiple antibiotic resistances. These organisms include β proteobacteria from Burkholderiaceae like *Pandoraea pnomenusa* and *Burkholderia seminalis*, and actinobacteria like *Corynebacterium falsenii*, and *Streptomyces ambofaciens*. Based on this, we could not reliably associate specific function to YbdK subfamily.

Interestingly, we found ATP-grasp_3, Dala_Dala_lig_C, and/or RimK domain containing genes in the neighborhood of sb2 subfamily annotated genes (22 genomes out of 81 genomes of sb2 subfamily). These were not specific to any one lineage. These are found in delta/epsilon proteobacteria like *Haliangium ochraceum DSM 14365* and *Desulfurivibrio alkaliphilus AHT 2*; β proteobacteria like *Nitrosomonas communis* and *Azoarcus sp BH72*; γ proteobacteria like *Legionella hackeliae* and *Wenzhouxiangella marina* and Bacteroidetes like *Rufibacter tibetensis* and *Psychroflexus torquis ATCC 700755*. This raises any interesting possibility that these organisms can produce glutathione. In general, we found only 2 genomes, which did not have gshB homologues present in genomes having sb2 subfamily annotated genes. These are: *Parachlamydia acanthamoebae UV-7* (bacteria from Chlamydiae phylum) and *Methanoregula formicica SMSP* (Euryarchaeota). The sb4 neighboring genes is found to be enriched in methyltransferase and acetyltransferase domains.

5.3.4 Analysis of sequence variation and insertion/deletion regions among GCL families

Given sequence diversity of GCL families, especially divergence among family-3 sequences, automated method could not reliably align all members of GCL families. This could be notably seen in alignment of structurally conserved binding site residues. Even highly conserved Mg^{2+} not all binding residues could be aligned across three GCL families. In order to understand conservation of binding site residues, family specific insertions/deletions events and phylogenetic reconstruction, we performed structure alignment guided manually adjusted the MSA generated from aligning representative members of GCL families and is shown in Figure 5.7.

As can be seen in Figure 5.7, Mg^{2+} binding residues and EXR motif, where R is catalytic residue is conserved across all three families. Interestingly, the metal binding residue D60/E96/E159 of EcGCL /ScGCL /BjGCL and E67/E103/E165 of EcGCL /ScGCL /BjGCL have six residues in between them. However, in GCL family-3, Plant-like, EgtA and sb3 subfamily has only five residues instead of six residues as observed in rest all GCL sequences. Based on IC, all motifs are relatively conserved in MSA. Moreover, in many instances the residues are conservatively substituted.

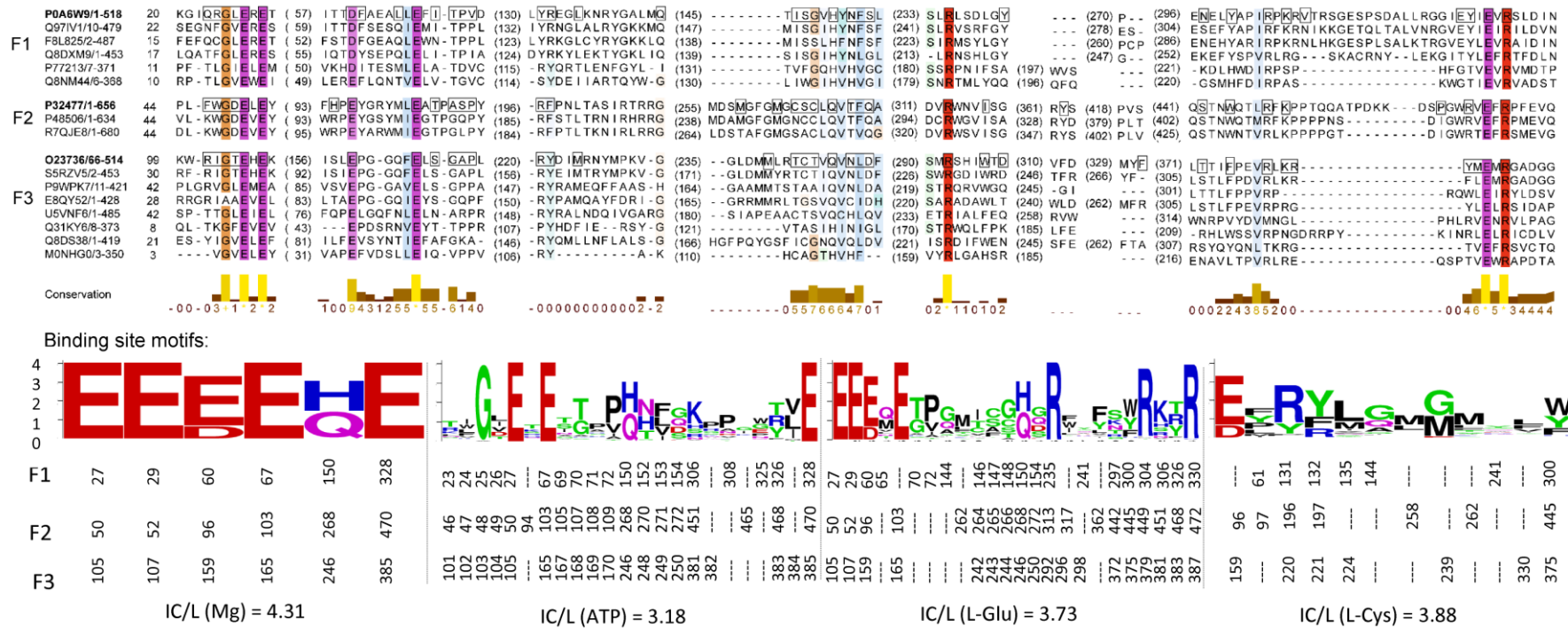


Figure 5.7 MSA of all three GCL family representative sequences showing substrate/metal binding site motifs. *EXR motif*, where *R* is probably involved in catalysis is also shown

Further, we analyzed the common insertions and deletions events specific to GCL families. Even though, most of secondary structural elements are mostly conserved across GCL families, we could find family specific insertion/deletion regions as shown in Figure 5.8. For instance, β -hairpin motif (residue 326-346) was found in *B.juncea* crystal structure. This motif is stabilized by CC1 (CYS³⁴¹-CYS³⁵⁶) disulfide bond. From various steady-state kinetics study of wild type and mutant BjGCL, it was suggested that reduction of CC1 reorients this β -hairpin motif so that it shields the entry of the substrate binding site, thereby slowing down the binding of new substrates and release of the products (Gromes et al. 2008). This β -hairpin motif is highlighted in red box in Figure 5.8. There is no such β -hairpin motif present in GCL family-2 representatives *S. cerevisiae* at structurally equivalent position. However, in GCL family-1 representative, *E. coli* a shorter β -hairpin motif (279 to 284) is observed in the structurally equivalent position. However, this β -hairpin is oriented in the opposite direction of the active site funnel, and no cysteine residues are present in the spatial proximity, which can form disulfide bond and regulate its motion. Thus, the mechanism by which this β -hairpin motif controls the binding of substrate seems to be unique only to GCL family-3. A unique insertion is observed in family-1 sequence around residue 202, 205, 318 based on *E. coli* residue numbering. Towards C-terminal region, specific insertion of anti-parallel strands is observed in family-2 and family-3.

5.3.5 Phylogenetic analysis of GCL families/subfamilies

Initially, to explore intra-family variations and to study the relationship (functional/evolutionary) within subfamilies of a given family, we performed phylogenetic analysis for each GCL family. Later to elucidate the ancestral relationship among these three diverge sequence families, we performed phylogenetic reconstruction for all three GCL families using manually curated MSA. Such analysis may shed light on the probable evolutionary origin of γ -glutamylcysteine biosynthesis



Figure 5.8 Schematic MSA of secondary structures of three tertiary structures from *E. coli*, *S. cerevisiae* and *B. juncea*. The secondary structure β strands and α helices are shown by green/green-yellow and red arrows respectively. Insertions common to groups 2 and 3 are shown by a black colored box and unique insertions in group 1 are shown by a blue colored box. The β -hairpin motif of group 3 is shown by red colored box.

5.3.5.1 Phylogenetic analysis of GCL family-1

The phylogenetic reconstruction of representative sequences of family-1 (Figure 5.9) using Maximum likelihood method shows two distinct clades at root of the tree (highlighted in blue and red boxes). These two groups of sequences correspond to two subfamilies GshA and GshAB. Previous studies have suggested that GshAB probably evolved by domain recruitment and got transferred through lateral gene transfer (Gopal et al. 2005). However, this lateral gene transfer of GshAB was limited to firmicutes and free-living pathogens. The NJ tree topology was consistent with ML tree.

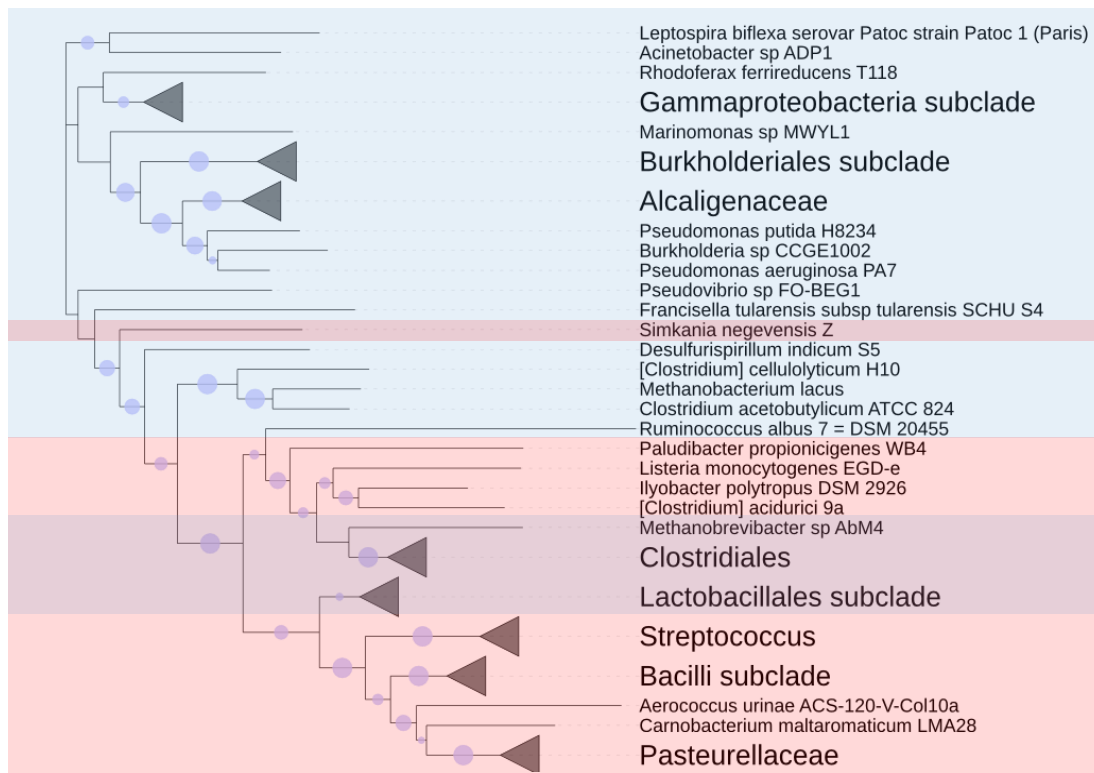


Figure 5.9 Phylogenetic tree for group 1 sequences showing evolution relationship between GshA and GshAB subfamilies. Maximum Likelihood (ML) method with WAG substitution model was used to generate tree. The rate variation among sites was modelled with a Gamma distributed with Invariant sites (G+I) (Number of discrete Gamma categories= 5). The number of bootstrap was set to 500. The width of circle represents confidence of bootstrap (1-100%), with values shown only if $\geq 50\%$. The tree is rooted using mid-point rooting method. The Figure was generated using iTOL webserver. Here GshA and GshAB are represented by red and blue boxes respectively.

There were few GshA sequences (blue) which were present in gshAB (red) clade. There are from firmicutes belonging to Clostridiales and Lactobacillales where one organism

has GshA and other has GshAB. For instance, *Clostridium botulinum* (Uniprot id B2TKY0) and *Clostridium cellulolyticum* (Uniprot id B8I2B6) harbor GshAB and GshA respectively. Interestingly, in these organisms, GshA is more similar to GshAB and usually slightly shorter in length (~470 amino-acids compared to ~518 amino-acids) compared to typical GshA.

5.3.5.2 Phylogenetic analysis of GCL family-2

1. Phylogenetic analysis of GCLC

The phylogenetic reconstruction using ML method for family-2 sequences (catalytic subunit of GCL-GCLC) is shown in Figure 5.10A. This family constitutes mostly non-plant eukaryotes. It is clearly evident from the phylogenetic tree (Figure 5.10A) that family-2 is monophyletic and it follows similar topology of universal tree of life generated from 16S rRNA sequences. Interestingly, we observed Rhodophyta, (red algae, representative member *Chondrus crispus*) and Phaeophyta (brown-algae, representative member *Ectocarpus siliculosus*) are classified within family-2 and occupies position between fungal and rest other metazoan clade. This is unusual given that GCL from algae (green algae) are present in family-3. It has been argued that Rhodophyta has undergone massive gene loss during evolution and it gained many gene during course of evolution (Qiu et al. 2015). For instance, *Chondrus crispus Stackhouse* taxa only encode ~ 5,000-10,000 genes (Collen et al. 2013). GCL must be within these genes gained in evolution by horizontal gene transfer event.

2. Phylogenetic analysis of GCLM

Most of GCL family-2 members exist as a heteromer consisting of a catalytic (GCLC) and a regulatory/modifier (GCLM) subunit that reversibly dissociate (Figure 5.1). While GCLC synthesizes γ -GC, GCLM increases the V_{\max} and K_{cat} of GCLC and decrease the K_m for ATP and L-GLU (Y. Yang et al. 2007). Next, we asked the question which one of the two subunits evolved first, or they evolved simultaneously. In order to understand the evolution of GCLC and GCLM with respect to each other, we constructed the phylogenetic tree for GCLM as well. On analysis of the phylogenetic tree for GCLM (Figure 5.10B), we found that its phylogeny is similar to that of GCLC suggesting that they both evolved together and were subjected to similar evolutionary pressures

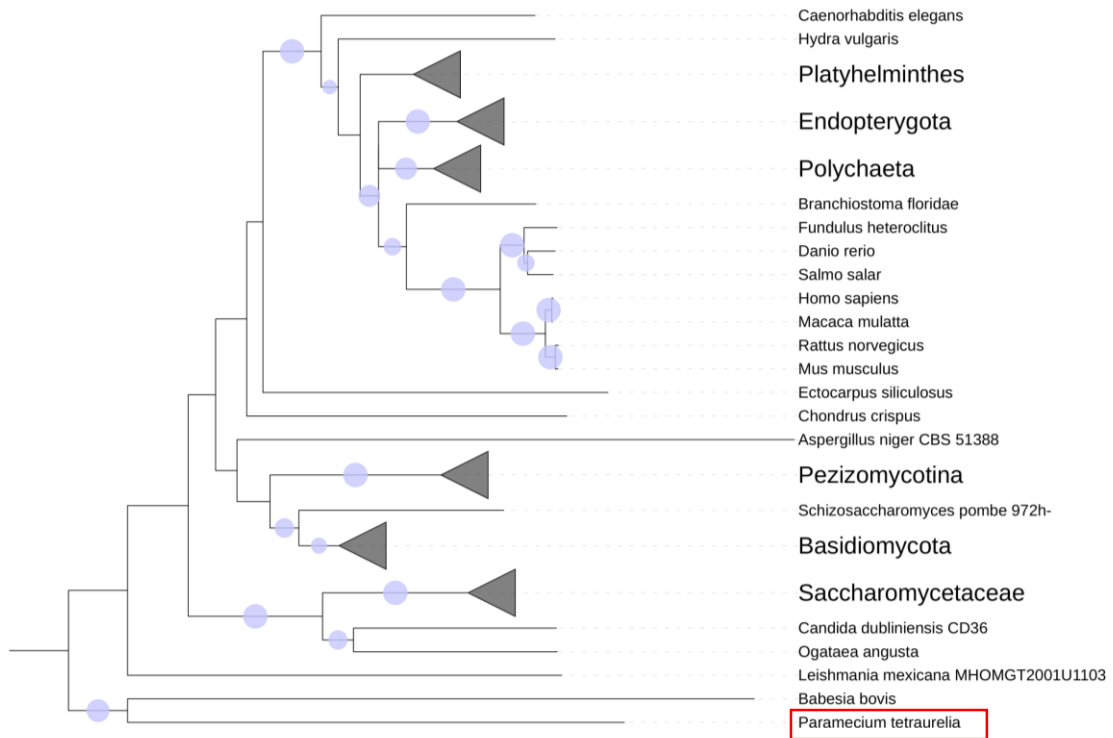


Figure 5.10A Phylogenetic tree for GCL group 2 showing the evolution of its catalytic subunit (GCLC) generated using Maximum Likelihood(ML) method. WAG was used as a substitution model and the rate variation among sites was modelled with a Gamma distributed with Invariant sites (G+I) (Number of discrete Gamma categories= 5). A bootstrap of 500 was used assess topology of tree. The width of the blue circles placed on the nodes of the phylogenetic tree represents the confidence of bootstrap (0-100%), with values shown only if $\geq 50\%$. Here the node *Paraecium tetraurelia* (highlighted in red box) was used to root the tree.

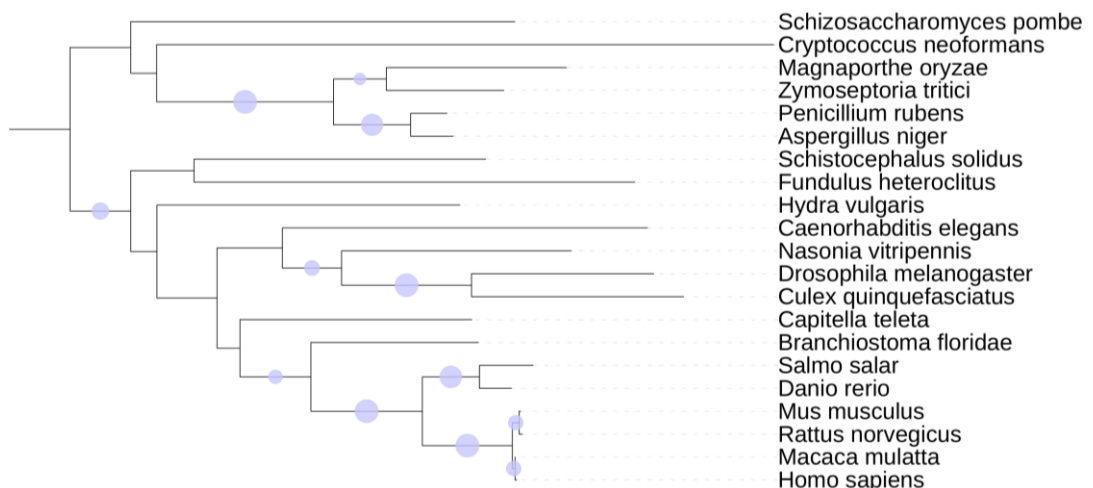


Figure 5.10B Phylogenetic tree for GCL group 2 showing the evolution of its modifier subunit (GCLM) generated using Maximum Likelihood(ML) method. WAG model was used as substitution model and the rate variation among sites was modelled with a Gamma distributed with Invariant sites (G+I) (Number of discrete Gamma categories= 5). To assess the reliability of the phylogenetic tree generated,

Bootstrap method with total 500 number of Bootstrap replications was used. The width of the blue circles placed on the nodes of the phylogenetic tree represents the confidence of bootstrap (0-100%), with values shown only if $\geq 50\%$. Here the tree was rooted using mid-point rooting method.

5.3.5.3 Phylogenetic analysis of GCL family-3

Next, we analyzed phylogeny tree of family-3 members. Among these, YbdK subfamily is the largest and consists of taxonomically diverse sequences. Hence, we reconstructed tree for YbdK subfamily. We considered this subfamily member from completely sequenced genomes and considered only sequences present in single copy in a genome (see methods section 5.2.6). The MUSCLE aligned sequences were used for tree reconstruction. Figure 5.11 shows that YbdK subfamily consists majorly of two subgroups (ybdk-sb1 and ybdk-sb2) with no taxonomically specific segregation in these subgroups. Both Actinobacteria and γ -proteobacteria were found to be present within in each of these YbdK subgroups. Among Actinobacteria, *Streptomyces*, *Pseudonocardia sp.*, *Alloactinosynnema sp.*, belongs to sb1 whereas *Corynebacterium*, belongs to other ybdk-sb2. Among γ -proteobacteria, Pseudomonadales and Xanthomonadales belongs to ybdk-sb1 whereas Legionellaes, Nitrococcus and Alteromonadales belong to ybdk-sb2. All α -proteobacteria belongs to ybdk-sb1 except YbdK sequence from endosymbiont of *Acanthamoeba sp.* UWC8 which belongs to ybdk-sb2. YbdK sequences from Bacteroidetes belong to ybdk-sb1.

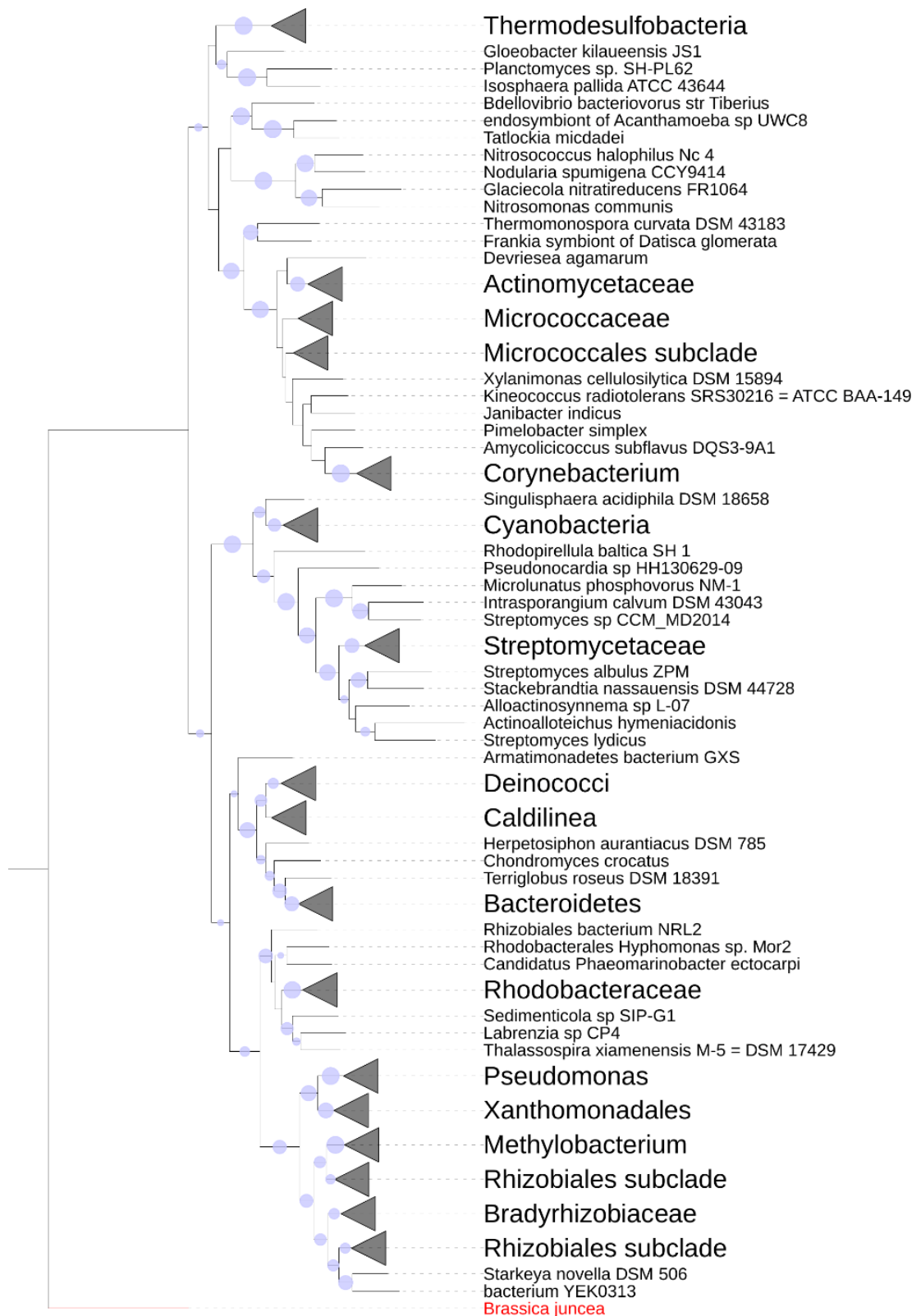


Figure 5.11 Phylogenetic tree for YbdK GCL group 3 subfamily generated using Maximum Likelihood (ML) method.

These suggest that YbdK members have probably undergone extensive duplication with horizontal gene transfer event to have such extensive distribution of its members across various phyla.

Phylogenetic tree of family-3 was reconstructed using ML method. From the phylogenetic tree of family-3 (Figure 5.12), it can be observed that in general subfamilies are mostly segregated from each other. Within each subfamily node, the evolution of GCL does not necessarily follow 16S rRNA based phylogeny. Given the complex phylogenetic relationship within these groups, we could not find whether functional divergence follows any specific trend. Rather this suggests that depending on functional requirements the GCL probably spread through horizontal gene transfer. The subfamilies Plant-like, sb3, and EgtA are more closely related to each other than other subfamily. This with other observation of conserved hairpin motif, it could be suggested that EgtA might have functionally diverged from Plant-like subfamily. Within Plant-like subfamily, plant and green algae are more evolutionary related to α -proteobacteria than to cyanobacteria. Subfamily sb2 seems to be more heterogeneous. Few sequences of sb2 are close to YbdK and other to sb1. Subfamily sb4 is mainly present in Euryarchaeota and it forms independent clade.

5.3.5.4 Phylogenetic analysis of GCL family-1, 2 and 3

To explore evolutionary relationship among three diverse GCL families, we generated phylogenetic tree for representative sequences from all three GCL families. From group 3, we used sequences from Plant-like, EgtA and YbdK subfamilies. As seen in Figure 5.13, group 1 evolved independently, whereas group 2 and group 3 showed evolutionary relatedness among themselves. Here too, EgtA and Plant-like group 3 subfamilies are more closely related to each other compared to YbdK. gshA and gshAB are well segregated in this tree as well. Further we mapped the distribution of different GCL families/subfamilies onto a universal tree of life from 16S rRNA sequences (Figure 5.14). The universal tree of life is taken from iTOL (Letunic and Bork 2016) server. Apart, from group3-Ybdk and group1-gshA co-occurring in various γ -proteobacteria (*E.coli*), Mycobacterium species have either YbdK and EgtA or YbdK and sb1 in their genome. It should be noted that not all the genomes present in iTOL's universal tree of life were present in our CSG dataset and vice-versa

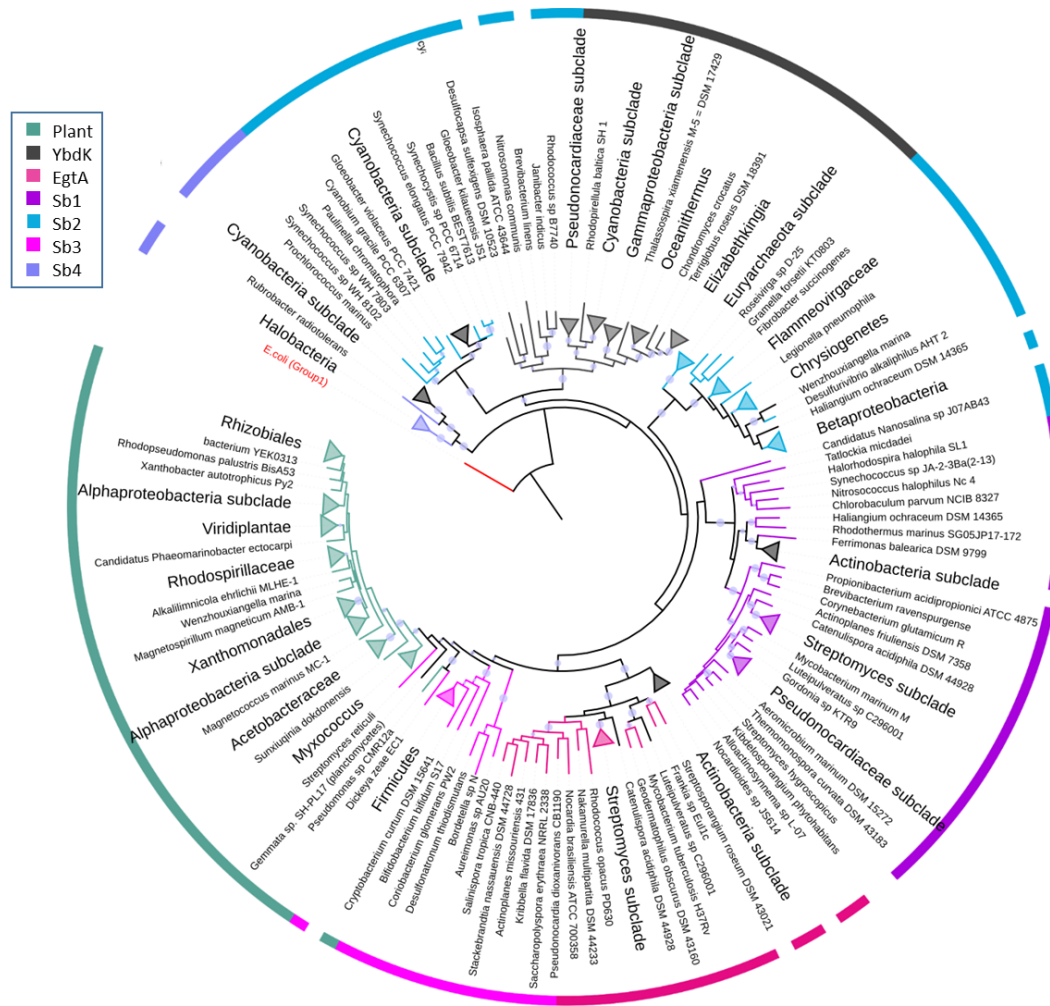


Figure 5.12 Phylogenetic tree for GCL family-3 generated using Maximum Likelihood (ML) method. *E. coli* (highlighted in red) sequences was used as an outgroup to root the tree.

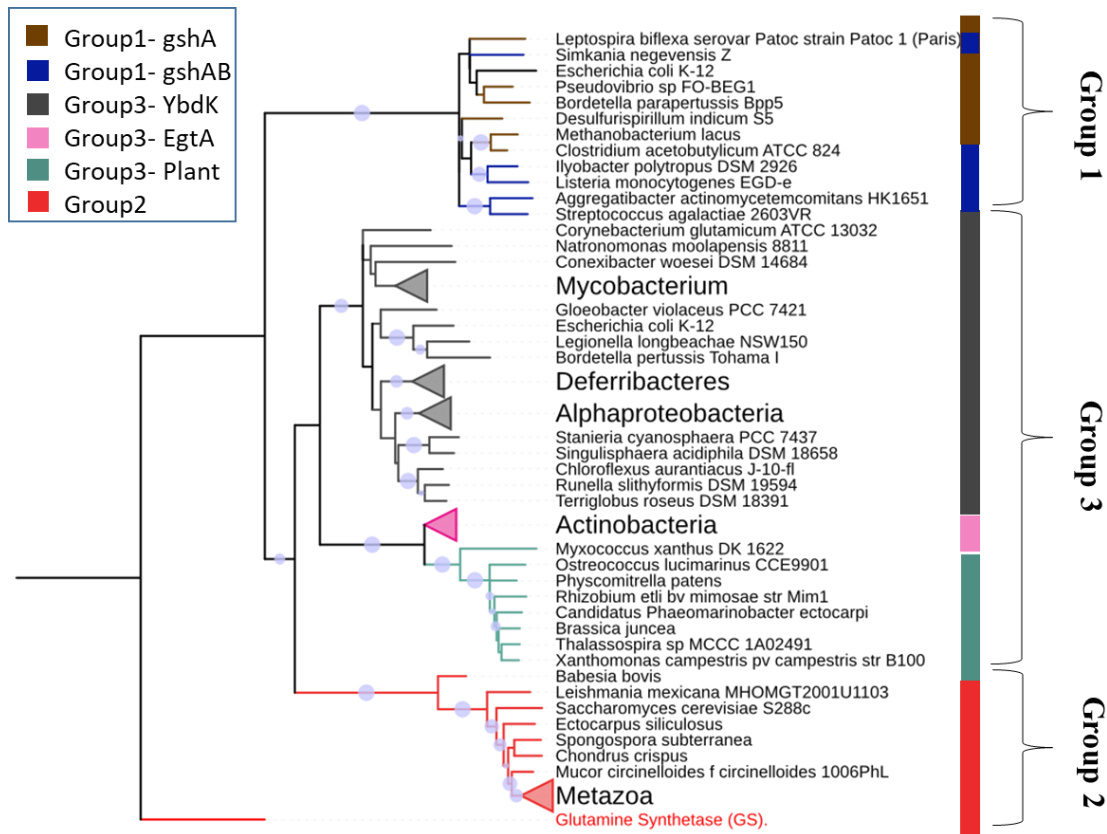


Figure 5.13 Phylogenetic tree for GCL group 1, 2 and 3 generated using Maximum Likelihood(ML) method. *JTT* model was used as substitution model and the rate variation among sites was modelled with a gamma distribution (shape parameter= 1). To assess the reliability of the phylogenetic tree generated, Bootstrap method with total 500 number of Bootstrap replications was used. The width of the blue circles placed on the nodes of the phylogenetic tree represents the confidence of bootstrap (0-100%), with values shown only if $\geq 50\%$. Here Glutamine Synthetase (PDBID:1f52A, highlighted in red) was used as an outgroup to root the tree.

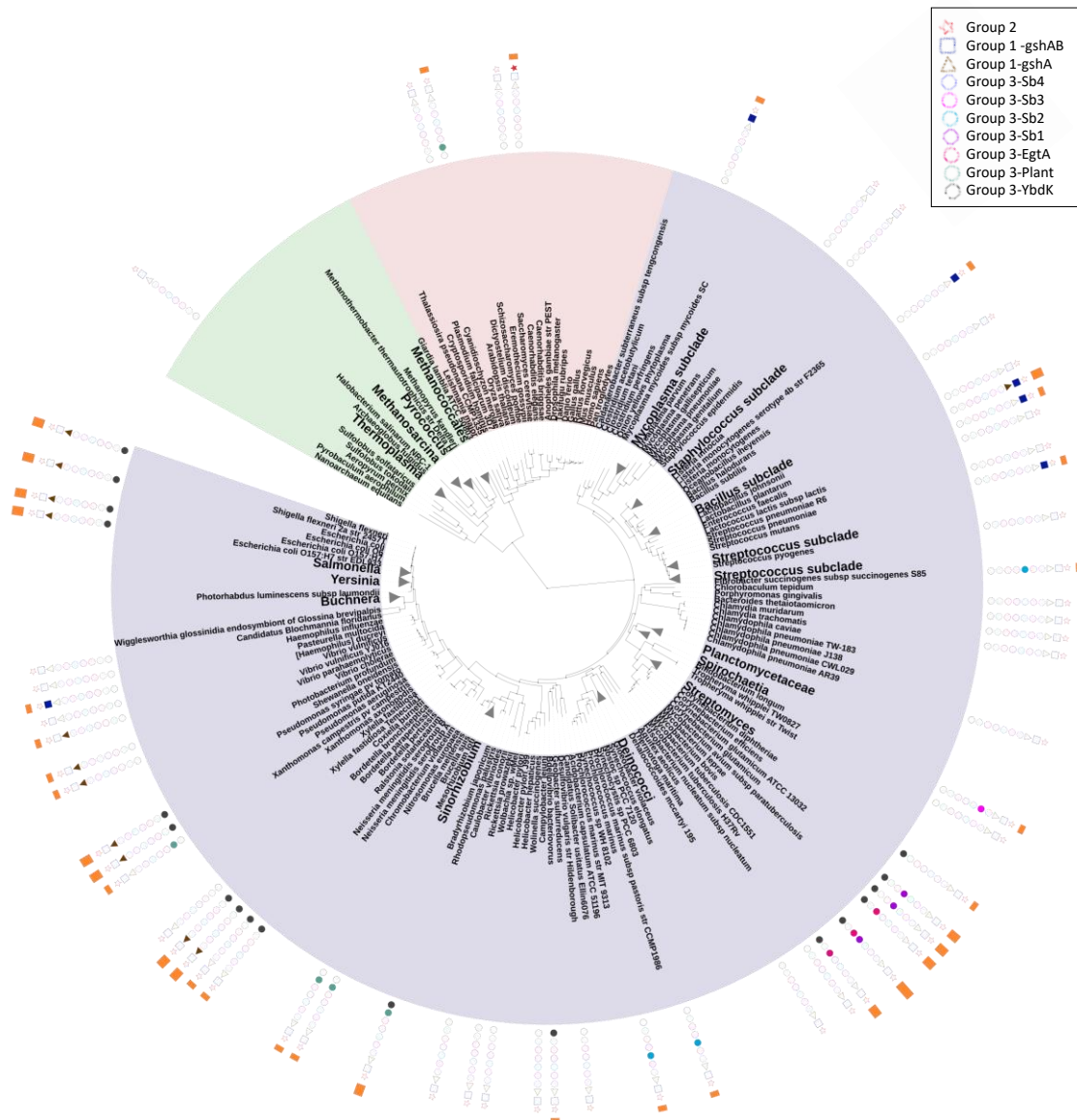


Figure 5.14 Distribution of different GCL family/subfamily in universal tree of life generated from 16S rRNA sequences. *The universal tree of life is taken from iTOL (Letunic and Bork 2016) server. The filled shape denotes the presence of the given GCL family and its absence is designated by empty shape*

5.4 Conclusions

This is a first large-scale study to classify GCL families into subfamilies sequences based on their sequences and function. In this work, we have classified GCL families into subfamilies, especially the most divergent family-3 which has been grouped into 7 subfamilies based on their sequence similarity or functions such as glutathione production or ergothioneine biosynthesis. Further, the sequence conservation analysis of substrate/metal binding motifs across three families of GCL showed that metal binding site is highly conserved, which has been acknowledged before as well, and among substrate binding sites ATP and glutamate shows relative high conservation. Importantly, these motifs show extensive conservation despite extensive sequence divergence among GCL families. We have identified catalytic residue in the motif EXR which was conserved across all GCL families. Among GCL families, family-3 showed poorly conserved substrate binding site. However, the alignment of sequences based on subfamily categories improved the conservation suggesting subfamilies may have their own characteristic binding site properties depending on their functions and also suggests these subfamilies are probably corrected classified. This classification of GCL into family and subfamily will lead to enhance function annotation. More computational and experimental studies are required to identify and characterize the function of each subfamily. The variation in the conservation of binding site of various substrates in GCL family/subfamily indicates tentatively the differences in the binding affinity for its natural/alternate substrates.

Further, we classified GCL members identified in genes encoded in completely sequenced and plant draft genomes into GCL family/subfamilies in an automated manner using curated HMM profile of each family and its respective subfamilies. This resulted in identification of 1083, 288 and 2325 sequences classified into groups 1, 2 and 3 respectively. Among group 3, YbdK has maximum members with 1225 sequences followed by Plant-like and EgtA having 460 and 181 sequences respectively. Further, the comparative genomic analysis suggested that gene duplication and possibly horizontal gene transfers events are responsible for extensive divergence of group-3 sequences, especially YbdK subfamily. Interestingly, YbdK is found in organisms which are not

known to produce glutathione, suggesting these organisms could synthesize γ -GC using weak γ -glutamylcysteine ligase function of YbdK and γ -GC could serve as anti-oxidant like glutathione in such organisms.

The phylogenetic tree reconstruction of each GCL family showed that sequences could be mostly be segregated based on their subfamily classified in our work. For instance, family-1 segregates into two distinct clades corresponding to GshA and GshAB subfamilies. The family-3 evolutionary tree also shows subfamilies as separate distinct groups. Based on the phylogenetic tree, it can be inferred that 3 subfamilies, viz. Plant-like, sb3 and EgtA are evolutionary more closely related to each other. Further, plant and green algae are evolutionary closely related to α -proteobacteria than to cyanobacteria. The phylogeny of GCL sequences of family-2 show similar topology as eukaryote evolution suggesting that GCL has been transferred in vertical manner more often than horizontal gene transfer.

The phylogenetic analyses of all GCL families together showed that group 1 might have evolved independently of group 2 and 3 and that group 2 and group 3 share more evolutionary relatedness. Moreover, structural comparison of all GCL family representatives showed that group 2 and 3 harbor many common insertions, which are absent in group 1.

Appendix I

Study of conformational variability of Conserved Recognition Elements (CoREs) in long disordered regions using molecular dynamics simulations

*Collaborative work with Dr. Kuljeet Singh Sandhu and Nitish Tayal,
IISER Mohali*

Tayal Nitish, **Choudhary Preeti**, Pandit Shashi Bhushan, and Sandhu Kuljeet Singh. 2014. “Evolutionarily Conserved and Conformationally Constrained Short Peptides Might Serve as DNA Recognition Elements in Intrinsically Disordered Regions.” *Molecular bioSystems* 10(6): 1469–80.1469–80

I.1 Background

Intrinsically disordered proteins (IDPs) are characterized by lack of stable 3D-structure (A. K. Dunker et al. 2001; A. K. Dunker et al. 2008; Dyson and Wright 2005) and exhibit spectrum of states varying from fully unstructured state such as random coils to partially folded state as in large-multi domain proteins connected *via* flexible (unstructured) linkers. IDPs may undergo conformation transition often referred as ‘couple folding and binding’ thereby attaining structure upon binding to its target (Kiefhaber, Bachmann, and Jensen 2012; Wright and Dyson 2009; Sugase, Dyson, and Wright 2007). For instance, DNA binding proteins undergo disorder to order transition when they bind to DNA. Despite more than a decade of their discovery, the mechanism via which IDPs recognize their target still remains incomprehensive. Short linear motifs have been proposed to facilitate molecular recognition in IDPs. However, any computational

method prior to this study was unable to identify these short motifs (Meszaros, Dosztanyi, and Simon 2012). In this study, work of Nitish and Kuljeet have specifically investigated the mechanism of molecular recognition by LDRs- Long Disordered Regions (> 70 amino-acids) of IDPs. The LDRs were predicted using IUPred (Dosztanyi et al. 2005) and FoldIndex (Prilusky et al. 2005) algorithms. The final dataset has 18,993 non-redundant proteins (< 70% sequence identity) harboring 27,782 LDRs. From these, hexapeptides were extracted that were enriched in different species. This resulted in final dataset referred as Conserved Recognition Elements (CoREs) comprising 877 invariant peptides of length ≥ 6 residues. Subsequently, detailed analysis of various properties of these peptides was performed and these regions were found to be evolutionary constrained and have distinct amino-acid propensity. Moreover, based on structural analysis, it was suggested that these CoREs retain their three-dimensional conformation in comparison to their adjacent regions. Moreover, significantly lower median RMSD (0.37 Å) was observed for CoREs compared to its neighboring regions (2.16 Å) in multiple structural alignments of the CoRE motifs found in non-redundant PDB entries. We have investigated the conformational variability of these short peptides and their neighboring regions of four representative proteins using explicit water MD simulations done for 50 ns (at 298 K temperature and 1 atm pressure).

I.2 Methods

MD simulations were performed using GROMACS (GRONingen MACHine for Chemical Simulations) version 4.5 with the OPLS-AA force field parameter set. Starting structure was centered in cubic box with its edges placed at a distance of 9.0 Å. This cubic box was solvated with water represented by TIP4P solvent model. Subsequently system was neutralized using with sodium and chloride ions. Particle Mesh Ewald (PME) method with a grid spacing of 1.6 Å was used to compute long-range electrostatic interactions. Coulomb and van der Waals interactions were calculated using 10 Å cut-off. These simulations were performed using periodic boundary conditions and a time step of 2 fs. The energy of system was minimized using steepest descent method followed by equilibration at a constant temperature of 298 K and then at constant pressure of 1 bar using the V-rescale and Parrinello–Rahman methods, respectively. Followed to this, a production

run of 50 ns was setup. All the analysis performed using the standard provided by GROMACS package.

I.3 Results

Comparison of RMSDs of the C α atoms of CoREs with its neighboring regions showed that CoREs are more conformationally restrained as compared to their neighboring regions. In order to define neighboring regions of CoREs, we used the six residues upstream and downstream of CoREs as pre-CoRE and post-CoRE regions. As seen in Figure A.1, CoREs tend to have lower C α RMSD in comparison to its neighboring regions. Moreover, this difference was found to be statistically significant with p -value $< 2.2 \times 10^{-16}$.

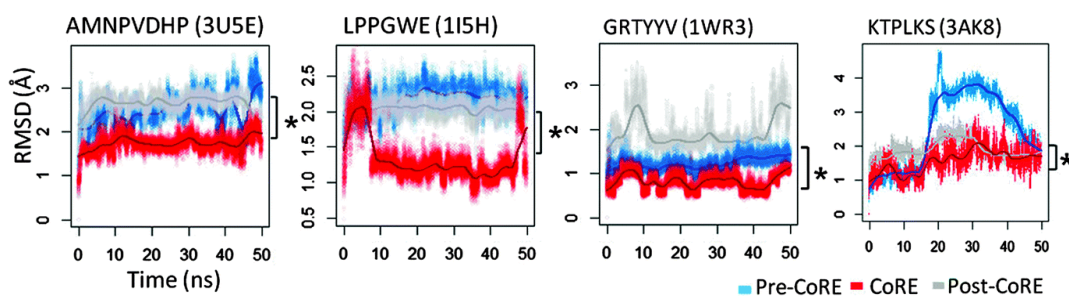


Figure A.1 C α RMSD plots of molecular dynamics trajectories for four representative examples of CoREs with pdbID 3U5E, 1I5H, 1WR3 and 3AK8. Asterisk (*) represents p -value $< 2.2 \times 10^{-16}$.

I.4 Conclusions

CoRE regions were conformationally more restrained in comparison to their neighboring regions. Other analysis suggested that these exhibits specific amino-acid properties and are enriched in DNA binding proteins. It is suggested that these might serve as “bait” for DNA recognition by long disordered regions in DNA binding proteins.

Appendix II

Understanding low pH induced structural changes of *Helicobacter pylori* TlyA using molecular dynamics simulations

Collaborative work with Dr. Kausik Chattopadhyay and Amritha Sreekumar, IISER Mohali

Sreekumar Amritha, **Choudhary Preeti**, Pandit Shashi Bhushan, and Chattopadhyay Kausik “Understanding the role of acidic pH in the structural and functional mechanism of a non-conventional hemolysin *Helicobacter pylori* TlyA.” 2018. *Manuscript to be submitted*.

II.1 Background

The protein TlyA from *Helicobacter pylori* (HpTlyA) is a membrane-damaging toxin with amyloidogenic tendencies (Lata and Chattopadhyay 2015). The mechanism of its pathogenicity is still elusive. *Helicobacter pylori* is a gastric pathogen and can withstand low pH conditions. This study focussed on detailed structural analysis in low pH conditions mimicking its physiologically relevant environment in order to understand its virulence mechanism. The experimental studies (Amritha Sreekumar) it was revealed that TlyA retains its secondary structure intact and shows no amyloid fibril formation and some structural properties are shown experimentally dependent on pH conditions. To understand low pH induced structural change, we performed explicit water MD simulations for 200 ns (300K temperature and 1atm pressure) at low and neutral pH conditions.

II.2 Methods

Due to unavailability of experimental 3D structure of HpTlyA, its starting coordinates were modeled with Robetta server (<http://www.robetta.org/submit.jsp>) using query sequence as Hemolysin (Tly) from *Helicobacter pylori* with the Uniprot id- O25718. The charged state of HpTlyA modelled structure at pH 3/pH 7 was obtained using H++ web server (<http://biophysics.cs.vt.edu/H++>), which assign H+ to titra residues based on their computed pKa at given pH (Anandakrishnan, Aguilar, and Onufriev 2012). The net charge of +8 and +23 were observed for HpTlyA at pH 7.0 and pH 3.0 respectively. We performed the MD simulations at two conditions – pH-3 and pH-7 using GROMACS (Abraham et al. 2015). The initial HpTlyA structure was centered in a cubic water-box, the edges of which were at least 16.0 Å away from the protein and this system was solvated using TIP4P water model. The energy of the system was minimized using steepest descent algorithm. System was then equilibrated first at constant temperature of 300 K, followed by equilibration at constant pressure of 1 bar using the V-rescale and Berendsen methods, respectively. This is followed by isothermal–isobaric ensemble MD simulation for 200 ns. The results were analyzed using standard programs provided with the GROMACS package. Solvent Accessible Surface Area (ASA) were calculated with NACCESS (S. J. Hubbard 1992), which implements the Lee and Richards algorithm. The Buried Surface Area (BSA) in the two domains is calculated as follows:

$$\text{BSA between two domains} = (\text{ASA}_{\text{domain1}} + \text{ASA}_{\text{domain2}}) - \text{ASA}_{\text{whole protein}}$$

II.3 Results

Based on the knowledge of its homolog structures (Witek et al. 2017), we identified two domains in HpTlyA, N-terminal domain (residue 1-60) and C-terminal domain (61-235), which corresponds to Pfam domains S4 and FtsJ respectively. A tentative linker region can be defined from residue 50 to 60 that adopts mostly coil conformation. We compared various properties of HpTlyA for its a) full-length protein, b) Domain 1 as defined from residue 1 to 49, and c) Domain 2 from residues 61 to 235 at two pH: 3.0 and 7.0 at which MD simulations were performed. First we compared of C α RMSD of HpTlyA at pH 3 and 7 conditions. We observed that it exhibits more structural variability at pH 3 as indicated by its higher median and mean (standard deviation) C α RMSD of protein of

9.54 and 8.95 (2.27) Å in comparison 6.36 and 5.92 (1.26) values observed at pH 7. Moreover, individual domains demonstrated small deviation compared to full-length protein, which shows large conformational change at low pH. Moreover, from visualization of MD trajectory we found that conformational space sampled at pH 3 is distinct from pH 7 and can be seen Figure II.1.

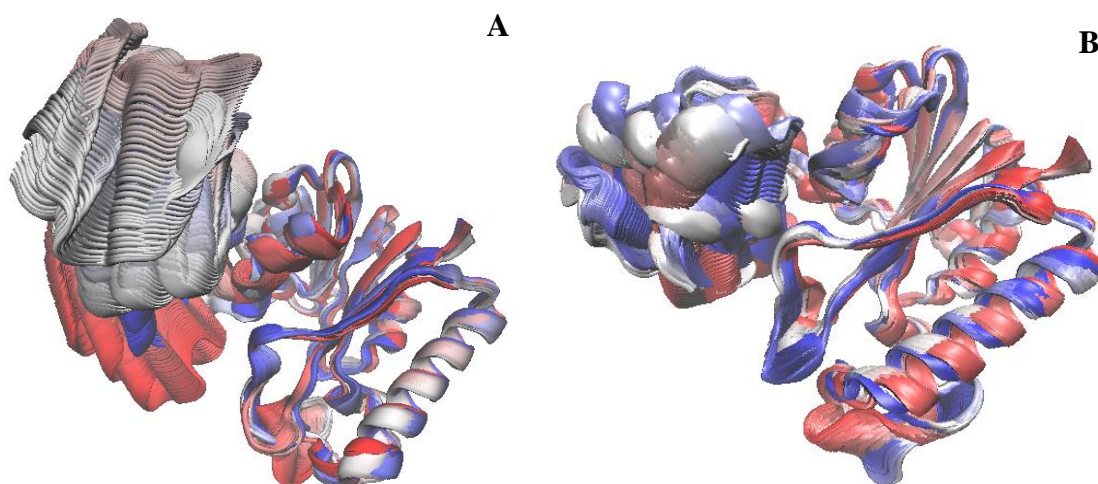


Figure II.1 Conformations of the HpTlyA at pH3 (A) and pH7 (B) during MD simulations were aligned to illustrate the relative domain motions. Using VMD, all the molecular dynamics trajectory structures are superimposed on each other by considering only domain 2 for superposition and the structures after every 1 ns are shown using smoothing step size of 20. Color indicates time, with red being the early stages of the simulations and blue indicating the later stages of simulations.

Next, we used radius of gyration (Rg) to compute compactness of protein. Figure II.2A shows that Rg of protein is high at pH 3 compared to Rg of protein at pH 7, which remains stable throughout simulations. There is drastic increase in Rg at ~50 ns time step and detailed analysis showed that at this time, two domains are disassociated from each other and their interaction interface is lost completely and the linker region from 50 to 60 residue adopts completely extended structure. This correlates with full-length RMSD variation as well.

We also analyzed the buried surface area at two pH conditions. As shown in Figure II.2B, BSA at pH 3 is reduced compared to that at pH-7. This suggests that interface is smaller and potentially hydrophobic region of a protein is exposed/accessible to solvent at pH-3. Detailed analysis showed that interaction among interface is lost in both pH conditions

leading to exposed interface to the solvent. However, at pH 3, due to extended conformation of linker region we have greater reduction in BSA especially at ~50 ns.

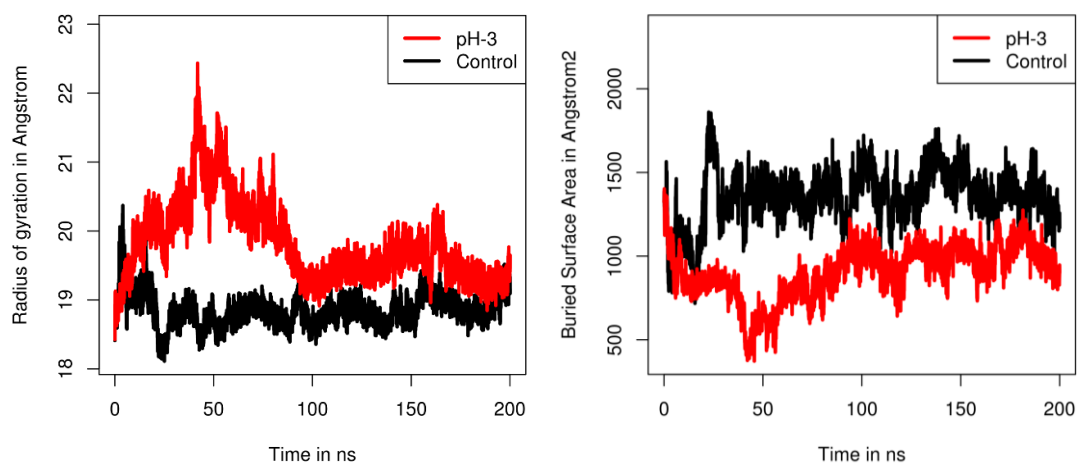


Figure II.2 Radius of gyration during MD simulations at pH3 and pH7. B) Buried Surface Area between domain-domain interfaces for MD simulations at pH3 and pH7.

II.4 Conclusions

The analysis of simulations showed that at low pH HpTlyA undergoes a large global structural change as evident by high C α RMSD compared to neutral pH. Moreover, compared to neutral pH conditions, structure at low pH has higher radius of gyration and reduced buried surface area between the domains. These suggest relative domain motion is primarily responsible for observed structural changes. This can be compared to the experimentally observed physiological properties of HpTlyA at low pH condition.

References

- Abbott, Jared J., Jimin Pei, Jennifer L. Ford, Yuan Qi, Vyacheslav N. Grishin, Lisa A. Pitcher, Margaret A. Phillips, and Nick V. Grishin. 2001. "Structure Prediction and Active Site Analysis of the Metal Binding Determinants in γ -Glutamylcysteine Synthetase." *Journal of Biological Chemistry* 276 (45): 42099–107. <https://doi.org/10.1074/jbc.M104672200>.
- Abdi, Hervé, and Lynne J. Williams. 2010. "Principal Component Analysis." *Wiley Interdisciplinary Reviews: Computational Statistics* 2 (4): 433–59. <https://doi.org/10.1002/wics.101>.
- Abraham, Mark James, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. 2015. "GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers." *SoftwareX* 1–2: 19–25. <https://doi.org/https://doi.org/10.1016/j.softx.2015.06.001>.
- Afriat, Livnat, Cintia Roodveldt, Giuseppe Manco, and Dan S Tawfik. 2006. "The Latent Promiscuity of Newly Identified Microbial Lactonases Is Linked to a Recently Diverged Phosphotriesterase." *Biochemistry* 45 (46): 13677–86. <https://doi.org/10.1021/bi061268r>.
- Aharoni, Amir, Leonid Gaidukov, Olga Khersonsky, Stephen McQ Gould, Cintia Roodveldt, and Dan S Tawfik. 2005. "The 'evolvability' of Promiscuous Protein Functions." *Nature Genetics* 37 (1): 73–76. <https://doi.org/10.1038/ng1482>.
- Ahmed, Aqeel, Richard D Smith, Jordan J Clark, James B Jr Dunbar, and Heather A Carlson. 2015. "Recent Improvements to Binding MOAD: A Resource for Protein-Ligand Binding Affinities and Structures." *Nucleic Acids Research* 43 (Database issue): D465-9. <https://doi.org/10.1093/nar/gku1088>.
- Akiva, Eyal, Shoshana Brown, Daniel E Almonacid, Alan E 2nd Barber, Ashley F Custer, Michael A Hicks, Conrad C Huang, et al. 2014. "The Structure-Function Linkage Database." *Nucleic Acids Research* 42 (Database issue): D521-30. <https://doi.org/10.1093/nar/gkt1130>.
- Aktories, K. 1997. "Identification of the Catalytic Site of Clostridial ADP-Ribosyltransferases." *Advances in Experimental Medicine and Biology* 419: 53–60.
- Alterovitz, Ron, Aaron Arvey, Sriram Sankararaman, Carolina Dallett, Yoav Freund, and Kimmen Sjölander. 2009. "ResBoost: Characterizing and Predicting Catalytic Residues in Enzymes." *BMC Bioinformatics* 10: 197. <https://doi.org/10.1186/1471-2105-10-197>.
- Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25 (17): 3389–3402. <https://doi.org/10.1093/nar/25.17.3389>.
- Amemiya, Takayuki, Ryotaro Koike, Sotaro Fuchigami, Mitsunori Ikeguchi, and Akinori Kidera. 2011. "Classification and Annotation of the Relationship between Protein Structural Change and Ligand Binding." *Journal of Molecular Biology* 408 (3): 568–84. <https://doi.org/10.1016/j.jmb.2011.02.058>.

- Amemiya, Takayuki, Ryotaro Koike, Akinori Kidera, and Motonori Ota. 2012. "PSCDB: A Database for Protein Structural Change upon Ligand Binding." *Nucleic Acids Research* 40 (D1): 554–58. <https://doi.org/10.1093/nar/gkr966>.
- Amitai, Gabriel, Leonid Gaidukov, Rellie Adani, Shelly Yishay, Guy Yacov, Moshe Kushnir, Shai Teitlboim, et al. 2006. "Enhanced Stereoselective Hydrolysis of Toxic Organophosphates by Directly Evolved Variants of Mammalian Serum Paraoxonase." *The FEBS Journal* 273 (9): 1906–19. <https://doi.org/10.1111/j.1742-4658.2006.05198.x>.
- Amitai, Gil, Arye Shemesh, Einat Sitbon, Maxim Shklar, Dvir Netanel, Ilya Venger, and Shmuel Pietrokovski. 2004. "Network Analysis of Protein Structures Identifies Functional Residues." *Journal of Molecular Biology* 344 (4): 1135–46. <https://doi.org/10.1016/j.jmb.2004.10.055>.
- Anandakrishnan, Ramu, Boris Aguilar, and Alexey V Onufriev. 2012. "H++ 3.0: Automating PK Prediction and the Preparation of Biomolecular Structures for Atomistic Molecular Modeling and Simulations." *Nucleic Acids Research* 40 (W1): W537–41. <https://doi.org/10.1093/nar/gks375>.
- Andreini, Claudia, Ivano Bertini, Gabriele Cavallaro, Gemma L Holliday, and Janet M Thornton. 2009. "Metal-MACiE: A Database of Metals Involved in Biological Catalysis." *Bioinformatics (Oxford, England)* 25 (16): 2088–89. <https://doi.org/10.1093/bioinformatics/btp256>.
- Appleby, Todd C, Cynthia Kinsland, Tadhg P Begley, and Steven E Ealick. 2000. "The Crystal Structure and Mechanism of Orotidine 5'-Monophosphate Decarboxylase." *Proceedings of the National Academy of Sciences of the United States of America* 97 (5): 2005–10. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC15744/>.
- Arora, Benu, Joyeeta Mukherjee, and Munishwar Nath Gupta. 2014. "Enzyme Promiscuity: Using the Dark Side of Enzyme Specificity in White Biotechnology." *Sustainable Chemical Processes* 2 (1): 25. <https://doi.org/10.1186/s40508-014-0025-y>.
- Atkins, William M, Weiya Doug Lu, and Daniel L Cook. 2002. "Is There a Toxicological Advantage for Non-Hyperbolic Kinetics in Cytochrome P450 Catalysis? Functional Allostery from 'Distributive Catalysis'." *The Journal of Biological Chemistry* 277 (36): 33258–66. <https://doi.org/10.1074/jbc.M204425200>.
- Babtie, Ann C., Subhajit Bandyopadhyay, Luis F. Olguin, and Florian Hollfelder. 2009. "Efficient Catalytic Promiscuity for Chemically Distinct Reactions." *Angewandte Chemie International Edition* 48 (20): 3692–94. <https://doi.org/10.1002/anie.200805843>.
- Babtie, Ann, Nobuhiko Tokuriki, and Florian Hollfelder. 2010. "What Makes an Enzyme Promiscuous?" *Current Opinion in Chemical Biology* 14 (2): 200–207. <https://doi.org/10.1016/j.cbpa.2009.11.028>.
- Backos, Donald S, Christopher C Franklin, and Philip Reigan. 2012. "The Role of Glutathione in Brain Tumor Drug Resistance." *Biochemical Pharmacology* 83 (8): 1005–12. <https://doi.org/10.1016/j.bcp.2011.11.016>.
- Backos, Donald S, Kristofer S Fritz, Debbie G McArthur, Jadwiga K Kepa, Andrew M Donson, Dennis R Petersen, Nicholas K Foreman, Christopher C Franklin, and Philip Reigan. 2013. "Glycation of Glutamate Cysteine Ligase by 2-Deoxy-d-Ribose and Its Potential Impact on Chemoresistance in Glioblastoma." *Neurochemical Research* 38 (9): 1838–49. <https://doi.org/10.1007/s11064-013-1090-4>.
- Backos, Donald S, Kristofer S Fritz, James R Roede, Dennis R Petersen, and Christopher C Franklin. 2011. "Posttranslational Modification and Regulation of Glutamate-Cysteine Ligase by the Alpha,Beta-Unsaturated Aldehyde 4-Hydroxy-2-Nonenal." *Free Radical Biology & Medicine* 50 (1): 14–26. <https://doi.org/10.1016/j.freeradbiomed.2010.10.694>.
- Baier, F, J N Copp, and N Tokuriki. 2016. "Evolution of Enzyme Superfamilies: Comprehensive Exploration of Sequence-Function Relationships." *Biochemistry* 55 (46): 6375–88. <https://doi.org/10.1021/acs.biochem.6b00723>.

- Bairoch, A. 2000. "The ENZYME Database in 2000." *Nucleic Acids Research* 28 (1): 304–5.
- Baker, Nathan A, David Sept, Simpson Joseph, Michael J Holst, and J Andrew McCammon. 2001. "Electrostatics of Nanosystems: Application to Microtubules and the Ribosome." *Proceedings of the National Academy of Sciences* 98 (18): 10037 LP-10041. <http://www.pnas.org/content/98/18/10037.abstract>.
- Bakker, Martin, Fred van Rantwijk, and Roger A Sheldon. 2002. "Metal Substitution in Thermolysin: Catalytic Properties of Tungstate Thermolysin in Sulfoxidation with H₂O₂." *Canadian Journal of Chemistry* 80 (6): 622–25. <https://doi.org/10.1139/v02-082>.
- Bar-Even, Arren, Elad Noor, Yonatan Savir, Wolfram Liebermeister, Dan Davidi, Dan S Tawfik, and Ron Milo. 2011. "The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters." *Biochemistry* 50 (21): 4402–10. <https://doi.org/10.1021/bi2002289>.
- Barker, Jonathan A, and Janet M Thornton. 2003. "An Algorithm for Constraint-Based Structural Template Matching: Application to 3D Templates with Statistical Analysis." *Bioinformatics (Oxford, England)* 19 (13): 1644–49.
- Bartlett, Gail J., Craig T. Porter, Neera Borkakoti, and Janet M. Thornton. 2002. "Analysis of Catalytic Residues in Enzyme Active Sites." *Journal of Molecular Biology* 324 (1): 105–21. [https://doi.org/10.1016/S0022-2836\(02\)01036-7](https://doi.org/10.1016/S0022-2836(02)01036-7).
- Bastard, Karine, Adam Alexander Thil Smith, Carine Vergne-Vaxelaire, Alain Perret, Anne Zaparucha, Raquel De Melo-Minardi, Aline Mariage, et al. 2013. "Revealing the Hidden Functional Diversity of an Enzyme Family." *Nature Chemical Biology* 10 (November): 42. <http://dx.doi.org/10.1038/nchembio.1387>.
- Beadle, G W, and E L Tatum. 1941. "Genetic Control of Biochemical Reactions in Neurospora." *Proceedings of the National Academy of Sciences of the United States of America* 27 (11): 499–506.
- Ben-David, Moshe, Mikael Elias, Jean-Jacques Filippi, Elisabet Dunach, Israel Silman, Joel L Sussman, and Dan S Tawfik. 2012. "Catalytic Versatility and Backups in Enzyme Active Sites: The Case of Serum Paraoxonase 1." *Journal of Molecular Biology* 418 (3–4): 181–96. <https://doi.org/10.1016/j.jmb.2012.02.042>.
- Ben-Shimon, Avraham, and Miriam Eisenstein. 2005. "Looking at Enzymes from the Inside out: The Proximity of Catalytic Residues to the Molecular Centroid Can Be Used for Detection of Active Sites and Enzyme–Ligand Interfaces." *Journal of Molecular Biology* 351 (2): 309–26. <https://doi.org/https://doi.org/10.1016/j.jmb.2005.06.047>.
- Bencharit, Sompop, Christopher L Morton, Yu Xue, Philip M Potter, and Matthew R Redinbo. 2003. "Structural Basis of Heroin and Cocaine Metabolism by a Promiscuous Human Drug-Processing Enzyme." *Nature Structural Biology* 10 (5): 349–56. <https://doi.org/10.1038/nsb919>.
- Berman, H M, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. 2000. "The Protein Data Bank." *Nucleic Acids Research* 28 (1): 235–42. <https://doi.org/10.1093/nar/28.1.235>.
- BEUBERG, C. 1921. "Über Ein Kohlenstoff Ketten Knapjendes Ferment (Carboligase)." *Biochem. Z.* 115: 282–84. <https://ci.nii.ac.jp/naid/10005796857/en/>.
- Bhattacharyya, Moitrayee, Chanda R Bhat, and Saraswathi Vishveshwara. 2013. "An Automated Approach to Network Features of Protein Structure Ensembles." *Protein Science: A Publication of the Protein Society* 22 (10): 1399–1416. <https://doi.org/10.1002/pro.2333>.
- Binda, C, R T Bossi, S Wakatsuki, S Arzt, A Coda, B Curti, M A Vanoni, and A Mattevi. 2000. "Cross-Talk and Ammonia Channeling between Active Centers in the Unexpected Domain Arrangement of Glutamate Synthase." *Structure (London, England: 1993)* 8 (12): 1299–1308.
- Biterova, Ekaterina I., and Joseph J. Barycki. 2009a. "Mechanistic Details of Glutathione Biosynthesis

- Revealed by Crystal Structures of *Saccharomyces Cerevisiae* Glutamate Cysteine Ligase.” *Journal of Biological Chemistry* 284 (47): 32700–708. <https://doi.org/10.1074/jbc.M109.025114>.
- Biterova, Ekaterina I., Joseph J. Barycki, Christopher Lehmann, Victoria Doseeva, Sadhana Pullalarevu, Wojciech Krajewski, Andrew Howard, et al. 2013. “Characterization of the Bifunctional γ -Glutamate-Cysteine Ligase/Glutathione Synthetase (GshF) of *Pasteurella Multocida*.” *Journal of Biological Chemistry* 8 (1): 4380–94. <https://doi.org/10.1074/jbc.M509517200>.
- Biterova, Ekaterina I, and Joseph J Barycki. 2009b. “Mechanistic Details of Glutathione Biosynthesis Revealed by Crystal Structures of *Saccharomyces Cerevisiae* Glutamate Cysteine Ligase.” *The Journal of Biological Chemistry* 284 (47): 32700–708. <https://doi.org/10.1074/jbc.M109.025114>.
- Blaha-Nelson, David, Dennis M Krüger, Klaudia Szeler, Moshe Ben-David, and Shina Caroline Lynn Kamerlin. 2017. “Active Site Hydrophobicity and the Convergent Evolution of Paraoxonase Activity in Structurally Divergent Enzymes: The Case of Serum Paraoxonase 1.” *Journal of the American Chemical Society*. <https://doi.org/10.1021/jacs.6b10801>.
- Blum, Marc-Michael, Frank Löhr, Andre Richardt, Heinz Rüterjans, and Julian C.-H. Chen. 2006. “Binding of a Designed Substrate Analogue to Diisopropyl Fluorophosphatase: Implications for the Phosphotriesterase Mechanism.” *Journal of the American Chemical Society* 128 (39): 12750–57. <https://doi.org/10.1021/ja061887n>.
- Bohm, Hans-joachim, Gerhard Klebe, and Hans-Joachim Böhm. 1996. “What Can We Learn from Molecular Recognition in Protein–Ligand Complexes for the Design of New Drugs?” *Angewandte Chemie International Edition In English* 35 (22): 2588–2614. <https://doi.org/10.1002/anie.199625881>.
- Bombarda, E, N Morellet, H Cherradi, B Spiess, S Bouaziz, E Grell, B P Roques, and Y Mely. 2001. “Determination of the PK(a) of the Four Zn²⁺-Coordinating Residues of the Distal Finger Motif of the HIV-1 Nucleocapsid Protein: Consequences on the Binding of Zn²⁺.” *Journal of Molecular Biology* 310 (3): 659–72. <https://doi.org/10.1006/jmbi.2001.4770>.
- Branneby, Cecilia, Peter Carlqvist, Anders Magnusson, Karl Hult, Tore Brinck, and Per Berglund. 2003. “Carbon–Carbon Bonds by Hydrolytic Enzymes.” *Journal of the American Chemical Society* 125 (4): 874–75. <https://doi.org/10.1021/ja028056b>.
- Bray, Tracey, Pedro Chan, Salim Bougouffa, Richard Greaves, Andrew J Doig, and Jim Warwicker. 2009. “SitesIdentify: A Protein Functional Site Prediction Tool.” *BMC Bioinformatics* 10: 379. <https://doi.org/10.1186/1471-2105-10-379>.
- Brigelius-Flohe, Regina, and Matilde Maiorino. 2013. “Glutathione Peroxidases.” *Biochimica et Biophysica Acta* 1830 (5): 3289–3303. <https://doi.org/10.1016/j.bbagen.2012.11.020>.
- Brinda, K. V., and Saraswathi Vishveshwara. 2005. “A Network Representation of Protein Structures: Implications for Protein Stability.” *Biophysical Journal* 89 (6): 4159–70. <https://doi.org/10.1529/biophysj.105.064485>.
- Brodkin, Heather R, Nicholas a DeLateur, Srinivas Somarowthu, Caitlyn L Mills, Walter R Novak, Penny J Beuning, Dagmar Ringe, and Mary Jo Ondrechen. 2015. “Prediction of Distal Residue Participation in Enzyme Catalysis.” *Protein Science: A Publication of the Protein Society* 24 (5): 762–78. <https://doi.org/10.1002/pro.2648>.
- Bryliński, Michał, Katarzyna Prymula, Wiktor Jurkowski, Marek Kochańczyk, Ewa Stawowczyk, Leszek Konieczny, and Irena Roterman. 2007. “Prediction of Functional Sites Based on the Fuzzy Oil Drop Model.” *PLoS Computational Biology* 3 (5): 0909–23. <https://doi.org/10.1371/journal.pcbi.0030094>.
- Brylinski, Michal, and Jeffrey Skolnick. 2008. “What Is the Relationship between the Global Structures of Apo and Holo Proteins?,” 363–77. <https://doi.org/10.1002/prot>.
- Campodonico, Miguel A, Barbara A Andrews, Juan A Asenjo, Bernhard O Palsson, and Adam M Feist. 2014. “Generation of an Atlas for Commodity Chemical Production in *Escherichia Coli* and a Novel

- Pathway Prediction Algorithm, GEM-Path.” *Metabolic Engineering* 25 (September): 140–58. <https://doi.org/10.1016/j.ymben.2014.07.009>.
- Capra, John a., and Mona Singh. 2007. “Predicting Functionally Important Residues from Sequence Conservation.” *Bioinformatics* 23 (15): 1875–82. <https://doi.org/10.1093/bioinformatics/btm270>.
- Carbajo, Daniel, and Anna Tramontano. 2012. “A Resource for Benchmarking the Usefulness of Protein Structure Models.” *BMC Bioinformatics* 13 (1): 188. <https://doi.org/10.1186/1471-2105-13-188>.
- Carbonell, Pablo, and Jean-Loup Faulon. 2010. “Molecular Signatures-Based Prediction of Enzyme Promiscuity.” *Bioinformatics (Oxford, England)* 26 (16): 2012–19. <https://doi.org/10.1093/bioinformatics/btq317>.
- Carlqvist, Peter, Maria Svedendahl, Cecilia Branneby, Karl Hult, Tore Brinck, and Per Berglund. 2005. “Exploring the Active-Site of a Rationally Redesigned Lipase for Catalysis of Michael-Type Additions.” *Chembiochem: A European Journal of Chemical Biology* 6 (2): 331–36. <https://doi.org/10.1002/cbic.200400213>.
- Chakraborty, Sandeep, Renu Minda, Lipika Salaye, Swapan K Bhattacharjee, and Basuthkar J Rao. 2011. “Active Site Detection by Spatial Conformity and Electrostatic Analysis--Unravelling a Proteolytic Function in Shrimp Alkaline Phosphatase.” *PloS One* 6 (12): e28470. <https://doi.org/10.1371/journal.pone.0028470>.
- Chakraborty, Sandeep, and Basuthkar J. Rao. 2012. “A Measure of the Promiscuity of Proteins and Characteristics of Residues in the Vicinity of the Catalytic Site That Regulate Promiscuity.” *PLoS ONE* 7 (2). <https://doi.org/10.1371/journal.pone.0032011>.
- Chang, Antje, Ida Schomburg, Sandra Placzek, Lisa Jeske, Marcus Ulbrich, Mei Xiao, Christoph W. Sensen, and Dietmar Schomburg. 2015. “BRENDA in 2015: Exciting Developments in Its 25th Year of Existence.” *Nucleic Acids Research* 43 (D1): D439–46. <https://doi.org/10.1093/nar/gku1068>.
- Chea, Eric, and Dennis R Livesay. 2007. “How Accurate and Statistically Robust Are Catalytic Site Predictions Based on Closeness Centrality?” *BMC Bioinformatics* 8 (1): 153. <https://doi.org/10.1186/1471-2105-8-153>.
- Chen, Ridao, Bingquan Gao, Xiao Liu, Feiying Ruan, Yong Zhang, Jizhong Lou, Keping Feng, et al. 2017. “Molecular Insights into the Enzyme Promiscuity of an Aromatic Prenyltransferase.” *Nature Chemical Biology* 13 (2): 226–34. <https://doi.org/10.1038/nchembio.2263>.
- Chien, Ting-Ying, Darby Tien-Hao Chang, Chien-Yu Chen, Yi-Zhong Weng, and Chen-Ming Hsu. 2008. “E1DS: Catalytic Site Prediction Based on 1D Signatures of Concurrent Conservation.” *Nucleic Acids Research* 36 (Web Server issue): W291-6. <https://doi.org/10.1093/nar/gkn324>.
- Chien, Ting Ying, Darby Tien Hao Chang, Chien Yu Chen, Yi Zhong Weng, and Chen Ming Hsu. 2008. “E1DS: Catalytic Site Prediction Based on 1D Signatures of Concurrent Conservation.” *Nucleic Acids Research* 36 (Web Server issue): 291–96. <https://doi.org/10.1093/nar/gkn324>.
- Chien, Yu-Tung, and Shao-Wei Huang. 2012. “Accurate Prediction of Protein Catalytic Residues by Side Chain Orientation and Residue Contact Density.” *PloS One* 7 (10): e47951. <https://doi.org/10.1371/journal.pone.0047951>.
- Chien, Yu Tung, and Shao Wei Huang. 2012. “Accurate Prediction of Protein Catalytic Residues by Side Chain Orientation and Residue Contact Density.” *PLoS ONE* 7 (10). <https://doi.org/10.1371/journal.pone.0047951>.
- Cilia, Elisa, and Andrea Passerini. 2010. “Automatic Prediction of Catalytic Residues by Modeling Residue Structural Neighborhood.” *BMC Bioinformatics* 11: 115. <https://doi.org/10.1186/1471-2105-11-115>.
- Coleman, Ryan G, and Kim A Sharp. 2006. “Travel Depth, a New Shape Descriptor for Macromolecules: Application to Ligand Binding.” *Journal of Molecular Biology* 362 (3): 441–58.

<https://doi.org/10.1016/j.jmb.2006.07.022>.

- Colin, Pierre-Yves, Balint Kintszes, Fabrice Gielen, Charlotte M Miton, Gerhard Fischer, Mark F Mohamed, Marko Hyvonen, Diego P Morgavi, Dick B Janssen, and Florian Hollfelder. 2015. "Ultra-high-Throughput Discovery of Promiscuous Enzymes by Picodroplet Functional Metagenomics." *Nature Communications* 6 (December): 10008. <https://doi.org/10.1038/ncomms10008>.
- Collard, François, Francesca Baldin, Isabelle Gerin, Jennifer Bolsée, Gaëtane Noël, Julie Graff, Maria Veiga-da-Cunha, et al. 2016. "A Conserved Phosphatase Destroys Toxic Glycolytic Side Products in Mammals and Yeast." *Nature Chemical Biology* 12 (June): 601. <http://dx.doi.org/10.1038/nchembio.2104>.
- Collen, Jonas, Betina Porcel, Wilfrid Carre, Steven G Ball, Cristian Chaparro, Thierry Tonon, Tristan Barbeyron, et al. 2013. "Genome Structure and Metabolic Features in the Red Seaweed *Chondrus Crispus* Shed Light on Evolution of the Archaeplastida." *Proceedings of the National Academy of Sciences of the United States of America* 110 (13): 5247–52. <https://doi.org/10.1073/pnas.1221259110>.
- Colletier, Jacques-Philippe, Alexey Aleksandrov, Nicolas Coquelle, Sonia Mraïhi, Elena Mendoza-Barbera, Martin Field, and Dominique Madern. 2012. "Sampling the Conformational Energy Landscape of a Hyperthermophilic Protein by Engineering Key Substitutions." *Molecular Biology and Evolution* 29 (6): 1683–94. <https://doi.org/10.1093/molbev/mss015>.
- Cook, Daniel L, and William M Atkins. 1997. "Enhanced Detoxication Due to Distributive Catalysis and Toxic Thresholds: A Kinetic Analysis." *Biochemistry* 36 (36): 10801–6. <https://doi.org/10.1021/bi971284b>.
- Copley, Shelley D. 2017. "Shining a Light on Enzyme Promiscuity." *Current Opinion in Structural Biology* 47: 167–75. <https://doi.org/10.1016/j.sbi.2017.11.001>.
- Copley, Shelley D. 2003. "Enzymes with Extra Talents: Moonlighting Functions and Catalytic Promiscuity." *Current Opinion in Chemical Biology* 40 (2): 265–72. <https://doi.org/10.1016/j.tibs.2014.12.004>.
- Copley, Shelley D. 2009. "Evolution of Efficient Pathways for Degradation of Anthropogenic Chemicals." *Nature Chemical Biology* 5 (8): 559–66. <https://doi.org/10.1038/nchembio.197>.
- Copley, Shelley D. 2014. "An Evolutionary Biochemist's Perspective on Promiscuity." *Trends in Biochemical Sciences* 40 (2): 72–78. <https://doi.org/10.1016/j.tibs.2014.12.004>.
- Copley, Shelley D, and Jasvinder K Dhillon. 2002. "Lateral Gene Transfer and Copley, Shelley D, and Jasvinder K Dhillon. 2002. 'Lateral Gene Transfer and Parallel Evolution in the History of Glutathione Biosynthesis Genes.' *Genome Biology* 3(5): Research0025. <https://doi.org/10.1186/gb-2002-3-5-research0025>.
- Couto, Narciso, Naglis Malys, Simon J Gaskell, and Jill Barber. 2013. "Partition and Turnover of Glutathione Reductase from *Saccharomyces Cerevisiae*: A Proteomic Approach." *Journal of Proteome Research* 12 (6): 2885–94. <https://doi.org/10.1021/pr4001948>.
- D'Ari, R, and J Casades. 1998. "Underground Metabolism." *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 20 (2): 181–86. [https://doi.org/10.1002/\(SICI\)1521-1878\(199802\)20:2<181::AID-BIES10>3.0.CO;2-0](https://doi.org/10.1002/(SICI)1521-1878(199802)20:2<181::AID-BIES10>3.0.CO;2-0).
- Dalton, Timothy P, Ying Chen, Scott N Schneider, Daniel W Nebert, and Howard G Shertzer. 2004. "Genetically Altered Mice to Evaluate Glutathione Homeostasis in Health and Disease." *Free Radical Biology & Medicine* 37 (10): 1511–26. <https://doi.org/10.1016/j.freeradbiomed.2004.06.040>.
- Davenport, R C, P A Bash, B A Seaton, M Karplus, G A Petsko, and D Ringe. 1991. "Structure of the

- Triosephosphate Isomerase-Phosphoglycolohydroxamate Complex: An Analogue of the Intermediate on the Reaction Pathway.” *Biochemistry* 30 (24): 5821–26.
- David, Aaron, Goldman Joshua, and T Beatty Laura. 2016. “The TIM Barrel Architecture Facilitated the Early Evolution of Protein-Mediated Metabolism.” *Journal of Molecular Evolution* 82 (1): 17–26. <https://doi.org/10.1007/s00239-015-9722-8>.
- Davis, Andrew M, and Simon J Teague. 1999. “Hydrogen Bonding, Hydrophobic Interactions, and Failure of the Rigid Receptor Hypothesis.” *Angewandte Chemie (International Ed. in English)* 38 (6): 736–49. [https://doi.org/10.1002/\(SICI\)1521-3773\(19990315\)38:6<736::AID-ANIE736>3.0.CO;2-R](https://doi.org/10.1002/(SICI)1521-3773(19990315)38:6<736::AID-ANIE736>3.0.CO;2-R).
- Davis, Jesse, and Mark Goadrich. 2006. “The Relationship Between Precision-Recall and ROC Curves.”
- DeFrank, J J, and T C Cheng. 1991. “Purification and Properties of an Organophosphorus Acid Anhydrase from a Halophilic Bacterial Isolate.” *Journal of Bacteriology* 173 (6): 1938–43.
- Delépine, Baudoin, Thomas Duigou, Pablo Carbonell, and Jean Loup Faulon. 2018. “RetroPath2.0: A Retrosynthesis Workflow for Metabolic Engineers.” *Metabolic Engineering* 45 (December 2017): 158–70. <https://doi.org/10.1016/j.ymben.2017.12.002>.
- Dellus-Gur, Eynat, Mikael Elias, Emilia Caselli, Fabio Prati, Merijn L M Salverda, J Arjan G M de Visser, James S Fraser, and Dan S Tawfik. 2015. “Negative Epistasis and Evolvability in TEM-1 β -Lactamase – The Thin Line between an Enzyme’s Conformational Freedom and Disorder.” *Journal of Molecular Biology* 427 (14): 2396–2409. <https://doi.org/10.1016/j.jmb.2015.05.011>.
- Dellus-Gur, Eynat, Agnes Toth-Petroczy, Mikael Elias, and Dan S Tawfik. 2013. “What Makes a Protein Fold Amenable to Functional Innovation? Fold Polarity and Stability Trade-Offs.” *Journal of Molecular Biology* 425 (14): 2609–21. <https://doi.org/10.1016/j.jmb.2013.03.033>.
- Dimitropoulos, Dimitris, John Ionides, and Kim Henrick. 2006. “Using MSDchem to Search the PDB Ligand Dictionary.” *Current Protocols in Bioinformatics* Chapter 14 (October): Unit14.3. <https://doi.org/10.1002/0471250953.bi1403s15>.
- Domingues, Francisco S, Jörg Rahnenführer, and Thomas Lengauer. 2004. “Automated Clustering of Ensembles of Alternative Models in Protein Structure Databases.” *Protein Engineering, Design and Selection* 17 (6): 537–43. <http://dx.doi.org/10.1093/protein/gzh063>.
- Dosztanyi, Zsuzsanna, Veronika Csizmok, Peter Tompa, and Istvan Simon. 2005. “IUPred: Web Server for the Prediction of Intrinsically Unstructured Regions of Proteins Based on Estimated Energy Content.” *Bioinformatics (Oxford, England)* 21 (16): 3433–34. <https://doi.org/10.1093/bioinformatics/bti541>.
- Dou, Yongchao, Jun Wang, Jialiang Yang, and Chi Zhang. 2012. “L1pred: A Sequence-Based Prediction Tool for Catalytic Residues in Enzymes with the L1-Logreg Classifier.” *PLoS One* 7 (4): e35666. <https://doi.org/10.1371/journal.pone.0035666>.
- Dou, Yongchao, Xiaoqi Zheng, and Jun Wang. 2009. “Prediction of Catalytic Residues Using the Variation of Stereochemical Properties.” *The Protein Journal* 28 (1): 29–33. <https://doi.org/10.1007/s10930-008-9161-0>.
- Dou, Yongchao, Xiaoqi Zheng, Jialiang Yang, and Jun Wang. 2010. “Prediction of Catalytic Residues Based on an Overlapping Amino Acid Classification.” *Amino Acids* 39 (5): 1353–61. <https://doi.org/10.1007/s00726-010-0587-2>.
- Duarte, Fernanda, Beat Anton Amrein, and Shina Caroline Lynn Kamerlin. 2013. “Modeling Catalytic Promiscuity in the Alkaline Phosphatase Superfamily.” *Physical Chemistry Chemical Physics : PCCP* 15 (27): 11160–77. <https://doi.org/10.1039/c3cp51179k>.
- Dudev, Todor, and Carmay Lim. 2002. “Factors Governing the Protonation State of Cysteines in Proteins: An Ab Initio/CDM Study.” *Journal of the American Chemical Society* 124 (23): 6759–66. <https://doi.org/10.1021/ja012620l>.

- Dukka Bahadur, K C, and Dennis R Livesay. 2008. "Improving Position-Specific Predictions of Protein Functional Sites Using Phylogenetic Motifs." *Bioinformatics* 24 (20): 2308–16. <http://dx.doi.org/10.1093/bioinformatics/btn454>.
- Dunker, a. Keith, J. David Lawson, Celeste J. Brown, Ryan M. Williams, Pedro Romero, Jeong S. Oh, Christopher J. Oldfield, et al. 2001. "Intrinsically Disordered Protein." *Journal of Molecular Graphics and Modelling* 19 (1): 26–59. [https://doi.org/10.1016/S1093-3263\(00\)00138-8](https://doi.org/10.1016/S1093-3263(00)00138-8).
- Dunker, A Keith, Israel Silman, Vladimir N Uversky, and Joel L Sussman. 2008. "Function and Structure of Inherently Disordered Proteins." *Current Opinion in Structural Biology* 18 (6): 756–64. <https://doi.org/10.1016/j.sbi.2008.10.002>.
- Dunn, Matthew R, Carine Otto, Kathryn E Fenton, and John C Chaput. 2016. "Improving Polymerase Activity with Unnatural Substrates by Sampling Mutations in Homologous Protein Architectures." *ACS Chemical Biology* 11 (5): 1210–19. <https://doi.org/10.1021/acscchembio.5b00949>.
- Dyson, H Jane, and Peter E Wright. 2005. "Intrinsically Unstructured Proteins and Their Functions." *Nature Reviews. Molecular Cell Biology* 6 (3): 197–208. <https://doi.org/10.1038/nrm1589>.
- Ebrecht, Ana C, Ligin Solamen, Benjamin L Hill, Alberto A Iglesias, Kenneth W Olsen, and Miguel A Ballicora. 2017. "Allosteric Control of Substrate Specificity of the Escherichia Coli ADP-Glucose Pyrophosphorylase ." *Frontiers in Chemistry* . <https://www.frontiersin.org/article/10.3389/fchem.2017.00041>.
- Eddy, Sean R. 2011. "Accelerated Profile HMM Searches." *PLOS Computational Biology* 7 (10): e1002195. <https://doi.org/10.1371/journal.pcbi.1002195>.
- Edgar, Robert C. 2004. "MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput." *Nucleic Acids Research* 32 (5): 1792–97. <https://doi.org/10.1093/nar/gkh340>.
- Essmann, Ulrich, Lalith Perera, Max L Berkowitz, Tom Darden, Hsing Lee, and Lee G Pedersen. 1995. "A Smooth Particle Mesh Ewald Method" 103 (November): 31–34.
- Fahey, Robert C. 2013. "Glutathione Analogs in Prokaryotes." *Biochimica et Biophysica Acta - General Subjects* 1830 (5): 3182–98. <https://doi.org/10.1016/j.bbagen.2012.10.006>.
- Fajardo, J Eduardo, and Andras Fiser. 2013. "Protein Structure Based Prediction of Catalytic Residues." *BMC Bioinformatics* 14 (1): 63. <https://doi.org/10.1186/1471-2105-14-63>.
- Feller, Scott E., Yuhong Zhang, Richard W. Pastor, and Bernard R. Brooks. 1995. "Constant Pressure Molecular Dynamics Simulation: The Langevin Piston Method." *The Journal of Chemical Physics* 103 (11): 4613–21. <https://doi.org/10.1063/1.470648>.
- Fernández-Gacio, Ana, Antonio Codina, Jacques Fastrez, Olivier Riant, and Patrice Soumillon. 2006. "Transforming Carbonic Anhydrase into Epoxide Synthase by Metal Exchange." *ChemBioChem* 7 (7): 1013–16. <https://doi.org/10.1002/cbic.200600127>.
- Ferraro, Daniel J, Adam Okerlund, Eric Brown, and S Ramaswamy. 2017. "One Enzyme, Many Reactions: Structural Basis for the Various Reactions Catalyzed by Naphthalene 1,2-Dioxygenase." *IUCrJ* 4 (5): 648–56. <https://doi.org/10.1107/S2052252517008223>.
- Finn, Robert D., Teresa K. Attwood, Patricia C. Babbitt, Alex Bateman, Peer Bork, Alan J. Bridge, Hsin Yu Chang, et al. 2017. "InterPro in 2017-beyond Protein Family and Domain Annotations." *Nucleic Acids Research* 45 (D1): D190–99. <https://doi.org/10.1093/nar/gkw1107>.
- Finn, Robert D., Penelope Coghill, Ruth Y. Eberhardt, Sean R. Eddy, Jaina Mistry, Alex L. Mitchell, Simon C. Potter, et al. 2016. "The Pfam Protein Families Database: Towards a More Sustainable Future." *Nucleic Acids Research* 44 (D1): D279–85. <https://doi.org/10.1093/nar/gkv1344>.
- Fischer, J D, C E Mayer, and J Söding. 2008. "Prediction of Protein Functional Residues from Sequence by Probability Density Estimation." *Bioinformatics (Oxford, England)* 24 (5): 613–20.

<https://doi.org/10.1093/bioinformatics/btm626>.

- Fischer, Marcus, Ryan G Coleman, James S Fraser, and Brian K Shoichet. 2014. "Incorporation of Protein Flexibility and Conformational Energy Penalties in Docking Screens to Improve Ligand Discovery." *Nature Chemistry* 6 (May): 575. <http://dx.doi.org/10.1038/nchem.1954>.
- Fleischmann, Astrid, Michael Darsow, Kirill Degtyarenko, Wolfgang Fleischmann, Sinéad Boyce, Kristian B Axelsen, Amos Bairoch, Dietmar Schomburg, Keith F Tipton, and Rolf Apweiler. 2004. "IntEnz, the Integrated Relational Enzyme Database." *Nucleic Acids Research* 32 (Database issue): D434–37. <https://doi.org/10.1093/nar/gkh119>.
- Franca, Tanos Celmar Costa. 2015. "Homology Modeling: An Important Tool for the Drug Discovery." *Journal of Biomolecular Structure & Dynamics* 33 (8): 1780–93. <https://doi.org/10.1080/07391102.2014.971429>.
- Franklin, Christopher C., Donald S. Backos, Isaac Mohar, Collin C. White, Henry J. Forman, and Terrance J. Kavanagh. 2009. "Structure, Function, and Post-Translational Regulation of the Catalytic and Modifier Subunits of Glutamate Cysteine Ligase." *Molecular Aspects of Medicine* 30 (1–2): 86–98. <https://doi.org/10.1016/j.mam.2008.08.009>.
- Fraser, Jennifer A, Pushpa Kansagra, Claire Kotecki, Robert D C Saunders, and Lesley I McLellan. 2003. "The Modifier Subunit of Drosophila Glutamate-Cysteine Ligase Regulates Catalytic Activity by Covalent and Noncovalent Interactions and Influences Glutathione Homeostasis in Vivo." *The Journal of Biological Chemistry* 278 (47): 46369–77. <https://doi.org/10.1074/jbc.M308035200>.
- Fraser, Jennifer A, Robert D C Saunders, and Lesley I McLellan. 2002. "Drosophila Melanogaster Glutamate-Cysteine Ligase Activity Is Regulated by a Modifier Subunit with a Mechanism of Action Similar to That of the Mammalian Form." *The Journal of Biological Chemistry* 277 (2): 1158–65. <https://doi.org/10.1074/jbc.M106683200>.
- Freilich, Shiri, Ruth V Spriggs, Richard A George, Bissan Al-lazikani, Mark Swindells, and Janet M Thornton. 2005. "The Complement of Enzymatic Sets in Different Species," 745–63. <https://doi.org/10.1016/j.jmb.2005.04.027>.
- Furnham, Nicholas, Natalie L Dawson, Syed A Rahman, Janet M Thornton, and Christine A Orengo. 2016. "Large-Scale Analysis Exploring Evolution of Catalytic Machineries and Mechanisms in Enzyme Superfamilies." *Journal of Molecular Biology* 428 (2): 253–67. <https://doi.org/10.1016/j.jmb.2015.11.010>.
- Furnham, Nicholas, Gemma L Holliday, Tjaart A P de Beer, Julius O B Jacobsen, William R Pearson, and Janet M Thornton. 2014. "The Catalytic Site Atlas 2.0: Cataloging Catalytic Sites and Residues Identified in Enzymes." *Nucleic Acids Research* 42 (Database issue): D485–9. <https://doi.org/10.1093/nar/gkt1243>.
- Galant, Ashley, Mary L. Preuss, Jeffrey C. Cameron, and Joseph M. Jez. 2011. "Plant Glutathione Biosynthesis: Diversity in Biochemical Regulation and Reaction Products." *Frontiers in Plant Science* 2 (September): 1–7. <https://doi.org/10.3389/fpls.2011.00045>.
- Galligan, James J, and Dennis R Petersen. 2012. "The Human Protein Disulfide Isomerase Gene Family." *Human Genomics* 6 (July): 6. <https://doi.org/10.1186/1479-7364-6-6>.
- Gamage, Niranjali U, Sergey Tsvetanov, Ronald G Duggleby, Michael E McManus, and Jennifer L Martin. 2005. "The Structure of Human SULT1A1 Crystallized with Estradiol. An Insight into Active Site Plasticity and Substrate Inhibition with Multi-Ring Substrates." *The Journal of Biological Chemistry* 280 (50): 41482–86. <https://doi.org/10.1074/jbc.M508289200>.
- Gao, Mu, and Jeffrey Skolnick. 2013a. "A Comprehensive Survey of Small-Molecule Binding Pockets in Proteins." *PLoS Computational Biology* 9 (10). <https://doi.org/10.1371/journal.pcbi.1003302>.
- Gao, Mu, and Jeffrey Skolnick. 2013b. "APoc: Large-Scale Identification of Similar Protein Pockets." *Bioinformatics* 29 (5): 597–604. <https://doi.org/10.1093/bioinformatics/btt024>.

- Gatti-Lafranconi, Pietro, and Florian Hollfelder. 2013. "Flexibility and Reactivity in Promiscuous Enzymes." *Chembiochem: A European Journal of Chemical Biology* 14 (3): 285–92. <https://doi.org/10.1002/cbic.201200628>.
- Geer, Lewis Y., Aron Marchler-Bauer, Renata C. Geer, Lianyi Han, Jane He, Siqian He, Chunlei Liu, Wenyao Shi, and Stephen H. Bryant. 2009. "The NCBI BioSystems Database." *Nucleic Acids Research* 38 (SUPPL.1): 492–96. <https://doi.org/10.1093/nar/gkp858>.
- Gerlt, John A., Jason T. Bouvier, Daniel B. Davidson, Heidi J. Imker, Boris Sadkhin, David R. Slater, and Katie L. Whalen. 2015. "Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A Web Tool for Generating Protein Sequence Similarity Networks." *Biochimica et Biophysica Acta - Proteins and Proteomics* 1854 (8): 1019–37. <https://doi.org/10.1016/j.bbapap.2015.04.015>.
- Gibbons, C, M G Montgomery, A G Leslie, and J E Walker. 2000. "The Structure of the Central Stalk in Bovine F(1)-ATPase at 2.4 Å Resolution." *Nature Structural Biology* 7 (11): 1055–61. <https://doi.org/10.1038/80981>.
- Gielen, Fabrice, Raphaëlle Hours, Stéphane Emond, Martin Fischlechner, Ursula Schell, and Florian Hollfelder. 2016. "Ultrahigh-Throughput-Directed Enzyme Evolution by Absorbance-Activated Droplet Sorting (AADS)." *Proceedings of the National Academy of Sciences of the United States of America* 113 (47): E7383–89. <https://doi.org/10.1073/pnas.1606927113>.
- Ginalski, Krzysztof, Arne Elofsson, Daniel Fischer, and Leszek Rychlewski. 2003. "3D-Jury: A Simple Approach to Improve Protein Structure Predictions." *Bioinformatics (Oxford, England)* 19 (8): 1015–18.
- Glaser, Fabian, Tal Pupko, Inbal Paz, Rachel E Bell, Dalit Bechor-Shental, Eric Martz, and Nir Ben-Tal. 2003. "ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information." *Bioinformatics* 19 (1): 163–64. <http://dx.doi.org/10.1093/bioinformatics/19.1.163>.
- Glasner, Margaret E, John A Gerlt, and Patricia C Babbitt. 2006. "Evolution of Enzyme Superfamilies." *Current Opinion in Chemical Biology* 10 (5): 492–97. <https://doi.org/10.1016/j.cbpa.2006.08.012>.
- Gopal, Shubha, Amos Ofer, Michaela Yanku, Gerald Cohen, Werner Goebel, and Yair Aharonowitz. 2005. "A Multidomain Fusion Protein In" 187 (11): 3839–47. <https://doi.org/10.1128/JB.187.11.3839>.
- Griffith, O. W. 1982. "Mechanism of Action, Metabolism, and Toxicity of Buthionine Sulfoximine and Its Higher Homologs, Potent Inhibitors of Glutathione Synthesis." *Journal of Biological Chemistry* 257 (22): 13704–12.
- Griffith, O W, and R T Mulcahy. 1999. "The Enzymes of Glutathione Synthesis: Gamma-Glutamylcysteine Synthetase." *Advances in Enzymology and Related Areas of Molecular Biology* 73: 209–67, xii.
- Gromes, Roland, Michael Hothorn, Esther D. Lenherr, Vladimir Rybin, Klaus Scheffzek, and Thomas Rausch. 2008. "The Redox Switch of γ -Glutamylcysteine Ligase via a Reversible Monomer-Dimer Transition Is a Mechanism Unique to Plants." *Plant Journal* 54 (6): 1063–75. <https://doi.org/10.1111/j.1365-313X.2008.03477.x>.
- Gruber, Christian W, Masa Cemazar, Begona Heras, Jennifer L Martin, and David J Craik. 2006. "Protein Disulfide Isomerase: The Structure of Oxidative Folding." *Trends in Biochemical Sciences* 31 (8): 455–64. <https://doi.org/10.1016/j.tibs.2006.06.001>.
- Guerout, Marc, Daniel Picot, Josephine Abi-Ghanem, Brigitte Hartmann, and Marc Baaden. 2010. "How Cations Can Assist DNase I in DNA Binding and Hydrolysis." *PLoS Computational Biology* 6 (11): e1001000. <https://doi.org/10.1371/journal.pcbi.1001000>.
- Guilloux, Vincent Le, Peter Schmidtke, and Pierre Tuffery. 2009. "Fpocket: An Open Source Platform for Ligand Pocket Detection." *BMC Bioinformatics* 10 (June): 168. <https://doi.org/10.1186/1471-2105-10-168>.

- Gunasekaran, Kannan, and Ruth Nussinov. 2007. "How Different Are Structurally Flexible and Rigid Binding Sites? Sequence and Structural Features Discriminating Proteins That Do and Do Not Undergo Conformational Change upon Ligand Binding." *Journal of Molecular Biology* 365 (1): 257–73. <https://doi.org/10.1016/j.jmb.2006.09.062>.
- Gutteridge, Alex, and Janet Thornton. 2004. "Conformational Change in Substrate Binding, Catalysis and Product Release: An Open and Shut Case?" *FEBS Letters* 567 (1): 67–73. <https://doi.org/10.1016/j.febslet.2004.03.067>.
- Gutteridge, Alex, and Janet Thornton. 2005. "Conformational Changes Observed in Enzyme Crystal Structures upon Substrate Binding." *Journal of Molecular Biology* 346 (1): 21–28. <https://doi.org/10.1016/j.jmb.2004.11.013>.
- Hammes, Gordon G. 2002. "Multiple Conformational Changes in Enzyme Catalysis." *Biochemistry* 41 (26): 8221–28. <https://doi.org/10.1021/bi0260839>.
- Harth, Günter, Saša Masleša-Galić, Michael V. Tullius, and Marcus A. Horwitz. 2005. "All Four Mycobacterium Tuberculosis GlnA Genes Encode Glutamine Synthetase Activities but Only GlnA1 Is Abundantly Expressed and Essential for Bacterial Homeostasis." *Molecular Microbiology* 58 (4): 1157–72. <https://doi.org/10.1111/j.1365-2958.2005.04899.x>.
- Hedstrom, Lizbeth. 2002. "Serine Protease Mechanism and Specificity." *Chemical Reviews* 102 (12): 4501–24.
- Hibi, Takao, Hiroshi Nii, Toru Nakatsu, Akira Kimura, Hiroaki Kato, and Jun Hiratake. 2004. "Crystal Structure of \square -Glutamylcysteine Synthetase : Insights into the Mechanism of Catalysis by a Key Enzyme for Glutathione Homeostasis," 1–6.
- Hiblot, Julien, Guillaume Gotthard, Mikael Elias, and Eric Chabriere. 2013. "Differential Active Site Loop Conformations Mediate Promiscuous Activities in the Lactonase SsoPox." *PloS One* 8 (9): e75272. <https://doi.org/10.1371/journal.pone.0075272>.
- Hiratake, Jun, Takayuki Irie, Nobuya Tokutake, and Jun'ichi Oda. 2002. "Recognition of a Cysteine Substrate by E. Coli Gamma-Glutamylcysteine Synthetase Probed by Sulfoximine-Based Transition-State Analogue Inhibitors." *Bioscience, Biotechnology, and Biochemistry* 66 (7): 1500–1514. <https://doi.org/10.1271/bbb.66.1500>.
- Hofmann, K, P Bucher, L Falquet, and A Bairoch. 1999. "The PROSITE Database, Its Status in 1999." *Nucleic Acids Research* 27 (1): 215–19.
- Holade, Yaovi, Mengwei Yuan, Ross D Milton, David P Hickey, Atsuya Sugawara, Clemens K Peterbauer, Dietmar Haltrich, and Shelley D Minter. 2017. "Rational Combination of Promiscuous Enzymes Yields a Versatile Enzymatic Fuel Cell with Improved Coulombic Efficiency" 164 (3). <https://doi.org/10.1149/2.0111703jes>.
- Holliday, Gemma L, Daniel E Almonacid, Gail J Bartlett, Noel M O'Boyle, James W Torrance, Peter Murray-Rust, John B O Mitchell, and Janet M Thornton. 2007. "MACiE (Mechanism, Annotation and Classification in Enzymes): Novel Tools for Searching Catalytic Mechanisms." *Nucleic Acids Research* 35 (Database issue): D515–20. <https://doi.org/10.1093/nar/gkl774>.
- Holliday, Gemma L, Daniel E Almonacid, John B O Mitchell, and Janet M Thornton. 2007. "The Chemistry of Protein Catalysis." *Journal of Molecular Biology* 372 (5): 1261–77. <https://doi.org/10.1016/j.jmb.2007.07.034>.
- Holliday, Gemma L, Claudia Andreini, Julia D Fischer, Syed Asad Rahman, Daniel E Almonacid, Sophie T Williams, and William R Pearson. 2012. "MACiE: Exploring the Diversity of Biochemical Reactions." *Nucleic Acids Research* 40 (D1): D783–89. <http://dx.doi.org/10.1093/nar/gkr799>.
- Holliday, Gemma L, Shoshana D Brown, Eyal Akiva, David Mischel, Michael A Hicks, John H Morris, Conrad C Huang, et al. 2017. "Biocuration in the Structure-Function Linkage Database: The Anatomy of a Superfamily." *Database : The Journal of Biological Databases and Curation* 2017

- (1). <https://doi.org/10.1093/database/bax006>.
- Holliday, Gemma L, Julia D Fischer, John B O Mitchell, and Janet M Thornton. 2011. "Characterizing the Complexity of Enzymes on the Basis of Their Mechanisms and Structures with a Bio-Computational Analysis" 278: 3835–45. <https://doi.org/10.1111/j.1742-4658.2011.08190.x>.
- Holliday, Gemma L, John B O Mitchell, and Janet M Thornton. 2009. "Understanding the Functional Roles of Amino Acid Residues in Enzyme Catalysis." *Journal of Molecular Biology* 390 (3): 560–77. <https://doi.org/10.1016/j.jmb.2009.05.015>.
- Honaker, Matthew T, Mauro Acchione, John P Sumida, and William M Atkins. 2011. "Ensemble Perspective for Catalytic Promiscuity: Calorimetric Analysis of the Active Site Conformational Landscape of a Detoxification Enzyme." *The Journal of Biological Chemistry* 286 (49): 42770–76. <https://doi.org/10.1074/jbc.M111.304386>.
- Hothorn, Michael, Andreas Wachter, Roland Gromes, Tobias Stuwe, Thomas Rausch, and Klaus Scheffzek. 2006. "Structural Basis for the Redox Control of Plant Glutamate Cysteine Ligase." *Journal of Biological Chemistry* 281 (37): 27557–65. <https://doi.org/10.1074/jbc.M602770200>.
- Hu, Liegi, Mark L Benson, Richard D Smith, Michael G Lerner, and Heather A Carlson. 2005. "Binding MOAD (Mother Of All Databases)." *Proteins* 60 (3): 333–40. <https://doi.org/10.1002/prot.20512>.
- Huang, Bingding. 2009. "MetaPocket: A Meta Approach to Improve Protein Ligand Binding Site Prediction." *Omics : A Journal of Integrative Biology* 13 (4): 325–30. <https://doi.org/10.1089/omi.2009.0045>.
- Huang, Bingding, and Michael Schroeder. 2006. "LIGSITE Csc : Predicting Ligand Binding Sites Using the Connolly Surface and Degree of Conservation." *BMC Structural Biology* 6 (1): 19. <https://doi.org/10.1186/1472-6807-6-19>.
- Huang, C S, L S Chang, M E Anderson, and A Meister. 1993. "Catalytic and Regulatory Properties of the Heavy Subunit of Rat Kidney Gamma-Glutamylcysteine Synthetase." *The Journal of Biological Chemistry* 268 (26): 19675–80.
- Huang, Hua, Chetanya Pandya, Chunliang Liu, Nawar F Al-Obaidi, Min Wang, Li Zheng, Sarah Toews Keating, et al. 2015. "Panoramic View of a Superfamily of Phosphatases through Substrate Profiling." *Proceedings of the National Academy of Sciences of the United States of America* 112 (16): E1974-83. <https://doi.org/10.1073/pnas.1423570112>.
- Huang, Shao-Wei, Sung-Huan Yu, Chien-Hua Shih, Huei-Wen Guan, Tsun-Tsao Huang, and Jenn-Kang Hwang. 2011. "On the Relationship between Catalytic Residues and Their Protein Contact Number." *Current Protein & Peptide Science* 12 (6): 574–79.
- Hubbard, Paul A, Xiquan Liang, Horst Schulz, and Jung-Ja P Kim. 2003. "The Crystal Structure and Reaction Mechanism of Escherichia Coli 2,4-Dienoyl-CoA Reductase." *The Journal of Biological Chemistry* 278 (39): 37553–60. <https://doi.org/10.1074/jbc.M304642200>.
- Hubbard, S J. 1992. "NACCESS: Program for Calculating Accessibilities." *Department of Biochemistry and Molecular Biology, University College of London*.
- Huddleston, Jamison P, Elizabeth A Burks, and Christian P Whitman. 2014. "Identification and Characterization of New Family Members in the Tautomerase Superfamily: Analysis and Implications." *Archives of Biochemistry and Biophysics* 564 (December): 189–96. <https://doi.org/10.1016/j.abb.2014.08.019>.
- Hult, Karl, and Per Berglund. 2007. "Enzyme Promiscuity: Mechanism and Applications." *Trends in Biotechnology* 25 (5): 231–38. <https://doi.org/10.1016/j.tibtech.2007.03.002>.
- Humphrey, William, Andrew Dalke, and Klaus Schulten. 1996. "{: {Visual} Molecular Dynamics." *Journal of Molecular Graphics* 14 (1): 33–38. [https://doi.org/10.1016/0263-7855\(96\)00018-5](https://doi.org/10.1016/0263-7855(96)00018-5).

- Izidoro, Sandro C, Raquel C de Melo-Minardi, and Gisele L Pappa. 2015. "GASS: Identifying Enzyme Active Sites with Genetic Algorithms." *Bioinformatics (Oxford, England)* 31 (6): 864–70. <https://doi.org/10.1093/bioinformatics/btu746>.
- Jacobs, Steven A., Joel M. Harp, Srikrupa Devarakonda, Youngchang Kim, Fraydoon Rastinejad, and Sepideh Khorasanizadeh. 2002. "The Active Site of the SET Domain Is Constructed on a Knot." *Nature Structural Biology* 9 (11): 833–38. <https://doi.org/10.1038/nsb861>.
- Jakoby, W B, and D M Ziegler. 1990. "The Enzymes of Detoxication." *The Journal of Biological Chemistry* 265 (34): 20715–18.
- Janowiak, Blythe E., and Owen W. Griffith. 2005. "Glutathione Synthesis in Streptococcus Agalactiae: One Protein Accounts for γ -Glutamylcysteine Synthetase and Glutathione Synthetase Activities." *Journal of Biological Chemistry* 280 (12): 11829–39. <https://doi.org/10.1074/jbc.M414326200>.
- Jensen, R A. 1976. "Enzyme Recruitment in Evolution of New Function." *Annual Review of Microbiology* 30: 409–25. <https://doi.org/10.1146/annurev.mi.30.100176.002205>.
- Jerlström Hultqvist, Jon, Omar Warsi, Annika Söderholm, Michael Knopp, Ulrich Eckhard, Egor Vorontsov, Maria Selmer, and Dan I Andersson. 2018. "A Bacteriophage Enzyme Induces Bacterial Metabolic Perturbation That Confers a Novel Promiscuous Function." *Nature Ecology & Evolution* 2 (8): 1321–30. <https://doi.org/10.1038/s41559-018-0568-5>.
- Jiang, Peng, James A Peliska, and Alexander J Ninfa. 1998. "The Regulation of Escherichia Coli Glutamine Synthetase Revisited: Role of 2-Ketoglutarate in the Regulation of Glutamine Synthetase Adenylation State." *Biochemistry* 37 (37): 12802–10. <https://doi.org/10.1021/bi980666u>.
- Jing, Qing, Krzysztof Okrasa, and Romas J. Kazlauskas. 2008. "Stereoselective Hydrogenation of Olefins Using Rhodium-Substituted Carbonic Anhydrase—A New Reductase." *Chemistry – A European Journal* 15 (6): 1370–76. <https://doi.org/10.1002/chem.200801673>.
- Jogl, Gerwald, Sharon Rozovsky, Ann E McDermott, and Liang Tong. 2003. "Optimal Alignment for Enzymatic Proton Transfer: Structure of the Michaelis Complex of Triosephosphate Isomerase at 1.2-Å Resolution." *Proceedings of the National Academy of Sciences of the United States of America* 100 (1): 50–55. <https://doi.org/10.1073/pnas.0233793100>.
- Johnson, L N, J Hajdu, K R Acharya, D I Stuart, P J McLaughlin, N G Oikonomakos, D Barford, and G Herve. 1989. "Allosteric Enzymes." *Ed. G. Herve, CRC Press Inc., Boca Raton, Florida*, 81–127.
- Jonas, S, and F Hollfelder. 2009. "Mapping Catalytic Promiscuity in the Alkaline Phosphatase Superfamily." *Pure and Applied Chemistry* 81 (4): 731–42. <https://doi.org/10.1351/PAC-CON-08-10-20>.
- Joosten, Robbie P, Tim A H te Beek, Elmar Krieger, Maarten L Hekkelman, Rob W W Hooft, Reinhard Schneider, Chris Sander, and Gert Vriend. 2011. "A Series of PDB Related Databases for Everyday Needs." *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkq1105>.
- Jorgensen, William L., Jayaraman Chandrasekhar, Jeffrey D. Madura, Roger W. Impey, and Michael L. Klein. 1983. "Comparison of Simple Potential Functions for Simulating Liquid Water." *The Journal of Chemical Physics* 79 (2): 926–35. <https://doi.org/10.1063/1.445869>.
- Jozefczak, Marijke, Tony Remans, Jaco Vangronsveld, and Ann Cuypers. 2012. "Glutathione Is a Key Player in Metal-Induced Oxidative Stress Defenses." *International Journal of Molecular Sciences* 13 (3): 3145–75. <https://doi.org/10.3390/ijms13033145>.
- Jürgens, Catharina, Alexander Strom, Dennis Wegener, Stefan Hettwer, Matthias Wilmanns, and Reinhard Sterner. 2000. "Directed Evolution of a (β)<Sub>8</Sub>-Barrel Enzyme to Catalyze Related Reactions in Two Different Metabolic Pathways." *Proceedings of the National Academy of Sciences* 97 (18): 9925 LP-9930. <http://www.pnas.org/content/97/18/9925.abstract>.
- Kabsch, W. 1976. "A Solution for the Best Rotation to Relate Two Sets of Vectors." *Acta*

- Crystallographica Section A* 32 (5): 922–23. <https://doi.org/10.1107/S0567739476001873>.
- Kabsch, W, and C Sander. 1983. “Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features.” *Biopolymers* 22 (12): 2577–2637. <https://doi.org/10.1002/bip.360221211>.
- Kahraman, Abdullah, Richard J. Morris, Roman A. Laskowski, and Janet M. Thornton. 2007. “Shape Variation in Protein Binding Pockets and Their Ligands.” *Journal of Molecular Biology* 368 (1): 283–301. <https://doi.org/10.1016/j.jmb.2007.01.086>.
- Kahraman, Abdullah, Richard J. Morris, Roman A. Laskowski, Angelo D. Favia, and Janet M. Thornton. 2010. “On the Diversity of Physicochemical Environments Experienced by Identical Ligands in Binding Pockets of Unrelated Proteins.” *Proteins* 78 (5): 1120–36. <https://doi.org/10.1002/prot.22633>.
- Kaltenbach, Miriam, Stephane Emond, Florian Hollfelder, and Nobuhiko Tokuriki. 2016. “Functional Trade-Offs in Promiscuous Enzymes Cannot Be Explained by Intrinsic Mutational Robustness of the Native Activity.” *PLoS Genetics* 12 (10): e1006305. <https://doi.org/10.1371/journal.pgen.1006305>.
- Kapoor, Manali, and Munishwar Nath Gupta. 2012. “Lipase Promiscuity and Its Biochemical Applications.” *Process Biochemistry* 47 (4): 555–69. <https://doi.org/https://doi.org/10.1016/j.procbio.2012.01.011>.
- Karplus, P. A., and G. E. Schulz. 1989. “Substrate Binding and Catalysis by Glutathione Reductase as Derived from Refined Enzyme: Substrate Crystal Structures at 2 Å Resolution.” *Journal of Molecular Biology* 210 (1): 163–80.
- Kato, Tsuyoshi, and Nozomi Nagano. 2011. “Discriminative Structural Approaches for Enzyme Active-Site Prediction.” *BMC Bioinformatics* 12 Suppl 1 (Suppl 1): S49. <https://doi.org/10.1186/1471-2105-12-S1-S49>.
- Kellogg, G. E., S. F. Semus, and D. J. Abraham. 1991. “HINT: A New Method of Empirical Hydrophobic Field Calculation for CoMFA.” *Journal of Computer-Aided Molecular Design* 5 (6): 545–52.
- Kelly, Brenda S., William E. Antholine, and Owen W. Griffith. 2002. “*Escherichia coli* γ -Glutamylcysteine Synthetase. Two Active Site Metal Ions Affect Substrate and Inhibitor Binding.” *Journal of Biological Chemistry* 277 (1): 50–58. <https://doi.org/10.1074/jbc.M107961200>.
- Khanal, Akhil, Sean Yu McLoughlin, Jamie P. Kershner, and Shelley D. Copley. 2015. “Differential Effects of a Mutation on the Normal and Promiscuous Activities of Orthologs: Implications for Natural and Directed Evolution.” *Molecular Biology and Evolution* 32 (1): 100–108. <http://dx.doi.org/10.1093/molbev/msu271>.
- Khare, Sagar D., Yakov Kipnis, Per Jr Greisen, Ryo Takeuchi, Yacov Ashani, Moshe Goldsmith, Yifan Song, et al. 2012. “Computational Redesign of a Mononuclear Zinc Metalloenzyme for Organophosphate Hydrolysis.” *Nature Chemical Biology* 8 (3): 294–300. <https://doi.org/10.1038/nchembio.777>.
- Khazanov, Nikolay A., and Heather A. Carlson. 2013. “Exploring the Composition of Protein-Ligand Binding Sites on a Large Scale.” *PLoS Computational Biology* 9 (11). <https://doi.org/10.1371/journal.pcbi.1003321>.
- Khersonsky, Olga, Cintia Roodveldt, and Dan S. Tawfik. 2006. “Enzyme Promiscuity: Evolutionary and Mechanistic Aspects.” *Current Opinion in Chemical Biology* 10 (5): 498–508. <https://doi.org/10.1016/j.cbpa.2006.08.011>.
- Khersonsky, Olga, Daniela Rothlisberger, Andrew M. Wollacott, Paul Murphy, Orly Dym, Shira Albeck, Gert Kiss, K. N. Houk, David Baker, and Dan S. Tawfik. 2011. “Optimization of the In-Silico-Designed Kemp Eliminate KE70 by Computational Design and Directed Evolution.” *Journal of Molecular Biology* 407 (3): 391–412. <https://doi.org/10.1016/j.jmb.2011.01.041>.

- Khersonsky, Olga, and Dan S Tawfik. 2005. "Structure–Reactivity Studies of Serum Paraoxonase PON1 Suggest That Its Native Activity Is Lactonase." *Biochemistry* 44 (16): 6371–82. <https://doi.org/10.1021/bi047440d>.
- Khersonsky, Olga, and Dan S Tawfik. 2006. "The Histidine 115-Histidine 134 Dyad Mediates the Lactonase Activity of Mammalian Serum Paraoxonases." *The Journal of Biological Chemistry* 281 (11): 7649–56. <https://doi.org/10.1074/jbc.M512594200>.
- Khersonsky, Olga, and Dan S Tawfik. 2010. "8.03 - Enzyme Promiscuity – Evolutionary and Mechanistic Aspects." In , edited by Hung-Wen (Ben) Liu and Lew B T - Comprehensive Natural Products I I Mander, 47–88. Oxford: Elsevier. <https://doi.org/https://doi.org/10.1016/B978-008045382-8.00155-6>.
- Kiefhaber, Thomas, Annett Bachmann, and Kristine Steen Jensen. 2012. "Dynamics and Mechanisms of Coupled Protein Folding and Binding Reactions." *Current Opinion in Structural Biology* 22 (1): 21–29. <https://doi.org/10.1016/j.sbi.2011.09.010>.
- Kim, Juhan, and Shelley D Copley. 2007. "Current Topics Why Metabolic Enzymes Are Essential or Nonessential for Growth of Escherichia Coli K12 on Glucose †" 46 (44). <https://doi.org/10.1021/bi7014629>.
- Kim, Juhan, and Shelley D Copley. 2012. "Inhibitory Cross-Talk upon Introduction of a New Metabolic Pathway into an Existing Metabolic Network." *Proceedings of the National Academy of Sciences* 109 (42): E2856 LP-E2864. <http://www.pnas.org/content/109/42/E2856.abstract>.
- Kim, Juhan, Jamie P Kershner, Yehor Novikov, Richard K Shoemaker, and Shelley D Copley. 2010. "Three Serendipitous Pathways in E. Coli Can Bypass a Block in Pyridoxal-5'-Phosphate Synthesis." *Molecular Systems Biology* 6 (November): 436. <https://doi.org/10.1038/msb.2010.88>.
- Kinoshita, Kengo, Jun'ichi Furui, and Haruki Nakamura. 2002. "Identification of Protein Functions from a Molecular Surface Database, EF-Site." *Journal of Structural and Functional Genomics* 2 (1): 9–22.
- Kirshner, Daniel a, Jerome P Nilmeier, and Felice C Lightstone. 2013. "Catalytic Site Identification--a Web Server to Identify Catalytic Site Structural Matches throughout PDB." *Nucleic Acids Research* 41 (Web Server issue): W256-65. <https://doi.org/10.1093/nar/gkt403>.
- Kochańczyk, Marek. 2011. "Prediction of Functionally Important Residues in Globular Proteins from Unusual Central Distances of Amino Acids." *BMC Structural Biology* 11 (1): 34. <https://doi.org/10.1186/1472-6807-11-34>.
- Kolb, Peter, Rafaela S Ferreira, John J Irwin, and Brian K Shoichet. 2009. "Docking and Chemoinformatic Screens for New Ligands and Targets." *Current Opinion in Biotechnology* 20 (4): 429–36. <https://doi.org/10.1016/j.copbio.2009.08.003>.
- Kondo, T, and T Iida. 1997. "[gamma-GCS and glutathione--new molecular targets in cancer treatment]." *Gan to kagaku ryoho. Cancer & chemotherapy* 24 (15): 2219–25.
- Koshland, D E. 1958. "Application of a Theory of Enzyme Specificity to Protein Synthesis." *Proceedings of the National Academy of Sciences of the United States of America* 44 (2): 98–104. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC335371/>.
- Koshland, D E, Emil Fischer, Nathan A Baker, David Sept, Simpson Joseph, Michael J Holst, J Andrew McCammon, et al. 2004. "Einfluss Der Configuration Auf Die Wirkung Der Enzyme." *Nucleic Acids Research* 44 (1): W344–50. <https://doi.org/10.1093/nar/gkw408>.
- Kraut, Daniel A, Paul A Sigala, Brandon Pybus, Corey W Liu, Dagmar Ringe, Gregory A Petsko, and Daniel Herschlag. 2006. "Testing Electrostatic Complementarity in Enzyme Catalysis: Hydrogen Bonding in the Ketosteroid Isomerase Oxyanion Hole." *PLOS Biology* 4 (4): e99. <https://doi.org/10.1371/journal.pbio.0040099>.

- Krishnan, Rajaraman, Franz Hefti, Haim Tsubery, Michal Lulu, Ming Proschitsky, and Richard Fisher. 2017. "Conformation as the Therapeutic Target for Neurodegenerative Diseases." *Current Alzheimer Research* 14 (4): 393–402. <https://doi.org/10.2174/1567205014666170116152622>.
- Kular, Baldeep, Nicola Leyland, Jaime Mejia-carranza, Helen Reynolds, Stanislaw Karpinski, and Philip M Mullineaux. 2004. "Evidence for a Direct Link between Glutathione Biosynthesis and Stress Defense Gene Expression in Arabidopsis." *The Plant Cell* 16 (September): 2448–62. <https://doi.org/10.1105/tpc.104.022608.1>.
- Kumar, Shailesh, Neha Kasturia, Amit Sharma, Manish Datt, and Anand K Bachhawat. 2013. "Redox-Dependent Stability of the γ -Glutamylcysteine Synthetase Enzyme of Escherichia Coli: A Novel Means of Redox Regulation." *The Biochemical Journal* 449 (3): 783–94. <https://doi.org/10.1042/BJ20120204>.
- Kumar, Sudhir, Glen Stecher, and Koichiro Tamura. 2016. "MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets." *Molecular Biology and Evolution* 33 (7): 1870–74. <https://doi.org/10.1093/molbev/msw054>.
- Kuznetsova, Ekaterina, Michael Proudfoot, Claudio F Gonzalez, Greg Brown, Marina V Omelchenko, Ivan Borozan, Liran Carmel, et al. 2006. "Genome-Wide Analysis of Substrate Specificities of the Escherichia Coli Haloacid Dehalogenase-like Phosphatase Family." *The Journal of Biological Chemistry* 281 (47): 36149–61. <https://doi.org/10.1074/jbc.M605449200>.
- La, David, and Dennis R Livesay. 2005. "Predicting Functional Sites with an Automated Algorithm Suitable for Heterogeneous Datasets." *BMC Bioinformatics* 6 (May): 116. <https://doi.org/10.1186/1471-2105-6-116>.
- Lamare, Sylvain, Marie-dominique Legoy, Marianne Graber, Laboratoire De Biotechnologies, De Chimie, Bâtiment Marie Curie, De La Rochelle, Avenue Michel Crépeau, and F- E-mail. 2004. "Solid / Gas Bioreactors : Powerful Tools for Fundamental Research and Efficient Technology for Industrial Applications." *Green Chemistry*, 445–58.
- Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. Mcgettigan, H. McWilliam, F. Valentin, et al. 2007. "Clustal W and Clustal X Version 2.0." *Bioinformatics* 23 (21): 2947–48. <https://doi.org/10.1093/bioinformatics/btm404>.
- Laskowski, R A. 1995. "SURFNET: A Program for Visualizing Molecular Surfaces, Cavities, and Intermolecular Interactions." *Journal of Molecular Graphics* 13 (5): 307-308,323-330.
- Laskowski, R A, N M Luscombe, M B Swindells, and J M Thornton. 1996. "Protein Clefts in Molecular Recognition and Function." *Protein Science : A Publication of the Protein Society* 5 (12): 2438–52. <https://doi.org/10.1002/pro.5560051206>.
- Laskowski, Roman A, Victor V Chistyakov, and Janet M Thornton. 2005. "PDBsum More : New Summaries and Analyses of the Known 3D Structures of Proteins and Nucleic Acids" 33: 266–68. <https://doi.org/10.1093/nar/gki001>.
- Laskowski, Roman a, James D Watson, and Janet M Thornton. 2005. "ProFunc: A Server for Predicting Protein Function from 3D Structure." *Nucleic Acids Research* 33 (Web Server issue): W89-93. <https://doi.org/10.1093/nar/gki414>.
- Lata, Kusum, and Kausik Chattopadhyay. 2015. "Helicobacter Pylori TlyA Forms Amyloid-like Aggregates with Potent Cytotoxic Activity." *Biochemistry* 54 (23): 3649–59. <https://doi.org/10.1021/acs.biochem.5b00423>.
- Laurie, Alasdair T R, and Richard M Jackson. 2005. "Q-SiteFinder: An Energy-Based Method for the Prediction of Protein-Ligand Binding Sites." *Bioinformatics (Oxford, England)* 21 (9): 1908–16. <https://doi.org/10.1093/bioinformatics/bti315>.
- Leesong, Minsun, Barry S. Henderson, James R. Gillig, John M. Schwab, and Janet L. Smith. 1996. "Structure of a Dehydratase-Isomerase from the Bacterial Pathway for Biosynthesis of Unsaturated

- Fatty Acids: Two Catalytic Activities in One Active Site.” *Structure* 4 (3): 253–64. [https://doi.org/10.1016/S0969-2126\(96\)00030-5](https://doi.org/10.1016/S0969-2126(96)00030-5).
- Lehmann, Christopher, Victoria Doseeva, Sadhana Pullalarevu, Wojciech Krajewski, Andrew Howard, and Osnat Herzberg. 2004. “YbdK Is a Carboxylate-Amine Ligase with a γ -Glutamyl: Cysteine Ligase Activity: Crystal Structure and Enzymatic Assays.” *Proteins: Structure, Function and Genetics* 56 (2): 376–83. <https://doi.org/10.1002/prot.20103>.
- Letunic, Ivica, and Peer Bork. 2016. “Interactive Tree of Life (ITOL) v3: An Online Tool for the Display and Annotation of Phylogenetic and Other Trees.” *Nucleic Acids Research* 44 (W1): W242–45. <https://doi.org/10.1093/nar/gkw290>.
- Li, Lin, Chuan Li, Subhra Sarkar, Jie Zhang, Shawn Witham, Zhe Zhang, Lin Wang, Nicholas Smith, Marharyta Petukh, and Emil Alexov. 2012. “DelPhi: A Comprehensive Suite for DelPhi Software and Associated Resources.” *BMC Biophysics* 5 (May): 9. <https://doi.org/10.1186/2046-1682-5-9>.
- Li, Shenhui, and Mei Hong. 2011. “Protonation, Tautomerization, and Rotameric Structure of Histidine: A Comprehensive Study by Magic-Angle-Spinning Solid-State NMR.” *Journal of the American Chemical Society* 133 (5): 1534–44. <https://doi.org/10.1021/ja108943n>.
- Li, Weizhong, and Adam Godzik. 2006. “Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences.” *Bioinformatics (Oxford, England)* 22 (13): 1658–59. <https://doi.org/10.1093/bioinformatics/btl158>.
- Lichtarge, O, H R Bourne, and F E Cohen. 1996. “An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families.” *Journal of Molecular Biology* 257 (2): 342–58. <https://doi.org/10.1006/jmbi.1996.0167>.
- Lin, Chih-Peng, Shao-Wei Huang, Yan-Long Lai, Shih-Chung Yen, Chien-Hua Shih, Chih-Hao Lu, Cuen-Chao Huang, and Jenn-Kang Hwang. 2008. “Deriving Protein Dynamical Properties from Weighted Protein Contact Number.” *Proteins* 72 (3): 929–35. <https://doi.org/10.1002/prot.21983>.
- Lindahl, Erik, and Berk Hess. 2001. “GROMACS 3 . 0 : A Package for Molecular Simulation and Trajectory Analysis,” 306–17. <https://doi.org/10.1007/s008940100045>.
- Linsky, Thomas, and Walter Fast. 2010. “Mechanistic Similarity and Diversity among the Guanidine-Modifying Members of the Penten Superfamily.” *Biochimica et Biophysica Acta* 1804 (10): 1943–53. <https://doi.org/10.1016/j.bbapap.2010.07.016>.
- Linster, Carole L, Emile Van Schaftingen, and Andrew D Hanson. 2013. “Metabolite Damage and Its Repair or Pre-Emption.” *Nature Chemical Biology* 9 (January): 72. <http://dx.doi.org/10.1038/nchembio.1141>.
- Liu, X.-S., and W.-L. Guo. 2008. “Robustness of the Residue Conservation Score Reflecting Both Frequencies and Physicochemistries.” *Amino Acids* 34 (4): 643–52. <https://doi.org/10.1007/s00726-007-0017-2>.
- Liu, Zhi-Qiang, Zi-Wei Xiang, Zhou Shen, Qi Wu, and Xian-Fu Lin. 2014. “Enzymatic Enantioselective Aldol Reactions of Isatin Derivatives with Cyclic Ketones under Solvent-Free Conditions.” *Biochimie* 101: 156–60. <https://doi.org/https://doi.org/10.1016/j.biochi.2014.01.006>.
- Livesay, Dennis R, Per Jambeck, Atipat Rojnuckarin, and Shankar Subramaniam. 2003. “Conservation of Electrostatic Properties within Enzyme Families and Superfamilies.” *Biochemistry* 42 (12): 3464–73. <https://doi.org/10.1021/bi026918f>.
- Llamas, Angel, Alejandro Chamizo-Ampudia, Manuel Tejada-Jimenez, Aurora Galvan, and Emilio Fernandez. 2017. “The Molybdenum Cofactor Enzyme MARC: Moonlighting or Promiscuous Enzyme?” *BioFactors (Oxford, England)* 43 (4): 486–94. <https://doi.org/10.1002/biof.1362>.
- Lo, Yu-chih, Su-chang Lin, Jei-fu Shaw, and Yen-chywan Liaw. 2005. “Substrate Specificities of Escherichia Coli Thioesterase I / Protease I / Lysophospholipase L 1 Are Governed by Its Switch

- Loop Movement †,” 1971–79.
- Lolis, E, and G A Petsko. 1990. “Crystallographic Analysis of the Complex between Triosephosphate Isomerase and 2-Phosphoglycolate at 2.5-Å Resolution: Implications for Catalysis.” *Biochemistry* 29 (28): 6619–25.
- London, Nir, Jeremiah D Farelli, Shoshana D Brown, Chunliang Liu, Hua Huang, Magdalena Korczynska, Nawar F Al-Obaidi, et al. 2015. “Covalent Docking Predicts Substrates for Haloalkanoate Dehalogenase Superfamily Phosphatases.” *Biochemistry* 54 (2): 528–37. <https://doi.org/10.1021/bi501140k>.
- Loo, Bert van, Stefanie Jonas, Ann C Babbie, Alhosna Benjdia, Olivier Berteau, Marko Hyvönen, and Florian Hollfelder. 2010. “An Efficient, Multiply Promiscuous Hydrolase in the Alkaline Phosphatase Superfamily.” *Proceedings of the National Academy of Sciences of the United States of America*. <https://doi.org/10.1073/pnas.0903951107>.
- López-Canut, Violeta, Maite Roca, Juan Bertrán, Vicent Moliner, and Iñaki Tuñón. 2011. “Promiscuity in Alkaline Phosphatase Superfamily. Unraveling Evolution through Molecular Simulations.” *Journal of the American Chemical Society* 133 (31): 12050–62. <https://doi.org/10.1021/ja2017575>.
- Lopez-Gallego, Fernando, Sean A Agger, Daniel Abate-Pella, Mark D Distefano, and Claudia Schmidt-Dannert. 2010. “Sesquiterpene Synthases Cop4 and Cop6 from *Coprinus Cinereus*: Catalytic Promiscuity and Cyclization of Farnesyl Pyrophosphate Geometric Isomers.” *Chembiochem: A European Journal of Chemical Biology* 11 (8): 1093–1106. <https://doi.org/10.1002/cbic.200900671>.
- Lopez-Gallego, Fernando, Grayson T Wawrzyn, and Claudia Schmidt-Dannert. 2010. “Selectivity of Fungal Sesquiterpene Synthases: Role of the Active Site’s H-1 Alpha Loop in Catalysis.” *Applied and Environmental Microbiology* 76 (23): 7723–33. <https://doi.org/10.1128/AEM.01811-10>.
- Lu, Chih-hao, Chin-sheng Yu, Yu-tung Chien, and Shao-wei Huang. 2014. “EXIA2: Web Server of Accurate and Rapid Protein Catalytic Residue Prediction” 2014.
- Lu, Shelly C. 2009. “Regulation of Glutathione Synthesis.” *Molecular Aspects of Medicine* 30 (1–2): 42–59. <https://doi.org/10.1016/j.mam.2008.05.005>.
- Lu, Shelly C. 2013. “Glutathione Synthesis.” *Biochimica et Biophysica Acta* 1830 (5): 3143–53. <https://doi.org/10.1016/j.bbagen.2012.09.008>.
- M., Villiers Benoit R, and Hollfelder Florian. 2009. “Mapping the Limits of Substrate Specificity of the Adenylation Domain of TycA.” *ChemBioChem* 10 (4): 671–82. <https://doi.org/10.1002/cbic.200800553>.
- Mackerell, Alexander D Jr, Michael Feig, and Charles L 3rd Brooks. 2004. “Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of Gas-Phase Quantum Mechanics in Reproducing Protein Conformational Distributions in Molecular Dynamics Simulations.” *Journal of Computational Chemistry* 25 (11): 1400–1415. <https://doi.org/10.1002/jcc.20065>.
- Mallinson, Sam J B, Melodie M Machovina, Rodrigo L Silveira, Marc Garcia-Borràs, Nathan Gallup, Christopher W Johnson, Mark D Allen, et al. 2018. “A Promiscuous Cytochrome P450 Aromatic O-Demethylase for Lignin Bioconversion.” *Nature Communications* 9 (1): 2487. <https://doi.org/10.1038/s41467-018-04878-2>.
- Manning, Christopher D, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.
- Martin, T W, and Z S Derewenda. 1999. “The Name Is Bond--H Bond.” *Nature Structural Biology*. United States. <https://doi.org/10.1038/8195>.
- Martínez-Núñez, Mario Alberto, Augusto Cesar Poot-Hernandez, Katya Rodríguez-Vázquez, and Ernesto Perez-Rueda. 2013. “Increments and Duplication Events of Enzymes and Transcription Factors Influence Metabolic and Regulatory Diversity in Prokaryotes.” *PLoS ONE* 8 (7).

<https://doi.org/10.1371/journal.pone.0069707>.

- Martinez-Nunez, Mario Alberto, Katya Rodriguez-Vazquez, and Ernesto Perez-Rueda. 2015. "The Lifestyle of Prokaryotic Organisms Influences the Repertoire of Promiscuous Enzymes." *Proteins* 83 (9): 1625–31. <https://doi.org/10.1002/prot.24847>.
- Martínez-Núñez, Mario, Zuemy Rodríguez-Escamilla, Katya Rodríguez-Vázquez, and Ernesto Pérez-Rueda. 2017. "Tracing the Repertoire of Promiscuous Enzymes along the Metabolic Pathways in Archaeal Organisms." *Life* 7 (3): 30. <https://doi.org/10.3390/life7030030>.
- Masip, Lluís, Karthik Veeravalli, and George Georgiou. 2006. "The Many Faces of Glutathione in Bacteria." *Antioxidants & Redox Signaling* 8 (5–6): 753–62. <https://doi.org/10.1089/ars.2006.8.753>.
- McDonald, Andrew G, Sinéad Boyce, and Keith F Tipton. 2009. "ExplorEnz: The Primary Source of the IUBMB Enzyme List." *Nucleic Acids Research* 37 (suppl_1): D593–97. <http://dx.doi.org/10.1093/nar/gkn582>.
- McDonald, I K, and J M Thornton. 1994. "Satisfying Hydrogen Bonding Potential in Proteins." *Journal of Molecular Biology* 238 (5): 777–93. <https://doi.org/10.1006/jmbi.1994.1334>.
- McGuffin, Liam J, Kevin Bryson, and David T Jones. 2000. "The PSIPRED Protein Structure Prediction Server ." *Bioinformatics* 16 (4): 404–5. <http://dx.doi.org/10.1093/bioinformatics/16.4.404>.
- McIntyre, T M, and N P Curthoys. 1980. "The Interorgan Metabolism of Glutathione." *International Journal of Biochemistry* 12 (4): 545–51. [https://doi.org/10.1016/0020-711X\(80\)90005-1](https://doi.org/10.1016/0020-711X(80)90005-1).
- McMillan, Andrew W, Mariana S Lopez, Mingzhao Zhu, Benjamin C Morse, In-Cheol Yeo, Jaleesia Amos, Ken Hull, Daniel Romo, and Margaret E Glasner. 2014. "Role of an Active Site Loop in the Promiscuous Activities of Amycolatopsis Sp. T-1-60 NSAR/OSBS." *Biochemistry* 53 (27): 4434–44. <https://doi.org/10.1021/bi500573v>.
- Meister, A, and M E Anderson. 1983. "Glutathione." *Annual Review of Biochemistry* 52: 711–60. <https://doi.org/10.1146/annurev.bi.52.070183.003431>.
- Mellor, Joseph, Ioana Grigoras, Pablo Carbonell, and Jean-Loup Faulon. 2016. "Semisupervised Gaussian Process for Automated Enzyme Search." *ACS Synthetic Biology* 5 (6): 518–28. <https://doi.org/10.1021/acssynbio.5b00294>.
- Meszaros, Balint, Zsuzsanna Dosztanyi, and Istvan Simon. 2012. "Disordered Binding Regions and Linear Motifs--Bridging the Gap between Two Models of Molecular Recognition." *PloS One* 7 (10): e46829. <https://doi.org/10.1371/journal.pone.0046829>.
- Miller, Brian G, and Ronald T Raines. 2004. "Identifying Latent Enzyme Activities: Substrate Ambiguity within Modern Bacterial Sugar Kinases." *Biochemistry* 43 (21): 6387–92. <https://doi.org/10.1021/bi049424m>.
- Milton, Ross D, Fei Wu, Koun Lim, Sofiene Abdellaoui, David P Hickey, and Shelley D Minter. 2015. "Promiscuous Glucose Oxidase: Electrical Energy Conversion of Multiple Polysaccharides Spanning Starch and Dairy Milk." *ACS Catalysis* 5 (12): 7218–25. <https://doi.org/10.1021/acscatal.5b01777>.
- Mistry, Jaina, Alex Bateman, and Robert D Finn. 2007. "Predicting Active Site Residue Annotations in the Pfam Database." *BMC Bioinformatics* 8 (January): 298. <https://doi.org/10.1186/1471-2105-8-298>.
- Mitternacht, Simon, and Igor N Berezovsky. 2011. "A Geometry-Based Generic Predictor for Catalytic and Allosteric Sites," 1–10.
- MIYAKE, Koichiro, and Shingo KAKITA. 2009. "A Novel Catalytic Ability of γ -Glutamylcysteine Synthetase of *Escherichia Coli* and Its Application in Theanine Production." *Bioscience, Biotechnology, and Biochemistry* 73 (12): 2677–83. <https://doi.org/10.1271/bbb.90538>.

- Moghe, Gaurav D, and Robert L Last. 2015. "Something Old , Something New : Conserved Enzymes and the Evolution of Novelty in Plant Specialized Metabolism 1" 169 (November): 1512–23. <https://doi.org/10.1104/pp.15.00994>.
- Mohamed, Mark F, and Florian Hollfelder. 2013. "Efficient, Crosswise Catalytic Promiscuity among Enzymes That Catalyze Phosphoryl Transfer." *Biochimica et Biophysica Acta* 1834 (1): 417–24. <https://doi.org/10.1016/j.bbapap.2012.07.015>.
- Monecke, Thomas, Juliane Buschmann, Piotr Neumann, Elmar Wahle, and Ralf Ficner. 2014. "Crystal Structures of the Novel Cytosolic 5'-Nucleotidase IIIB Explain Its Preference for M7GMP." *PLoS One* 9 (3): e90915. <https://doi.org/10.1371/journal.pone.0090915>.
- Morais, M C, W Zhang, A S Baker, G Zhang, D Dunaway-Mariano, and K N Allen. 2000. "The Crystal Structure of Bacillus Cereus Phosphonoacetaldehyde Hydrolase: Insight into Catalysis of Phosphorus Bond Cleavage and Catalytic Diversification within the HAD Enzyme Superfamily." *Biochemistry* 39 (34): 10385–96.
- Morris, Gm, and Ruth Huey. 2009. "AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility." *Journal of ...* 30 (16): 2785–91. <https://doi.org/10.1002/jcc.21256>.
- Morris, Richard J, Rafael J Najmanovich, Abdullah Kahraman, and Janet M Thornton. 2005. "Real Spherical Harmonic Expansion Coefficients as 3D Shape Descriptors for Protein Binding Pocket and Ligand Comparisons." *Bioinformatics (Oxford, England)* 21 (10): 2347–55. <https://doi.org/10.1093/bioinformatics/bti337>.
- Morrison, K L, and G A Weiss. 2001. "Combinatorial Alanine-Scanning." *Current Opinion in Chemical Biology* 5 (3): 302–7.
- Musgrave, William B., Hankuil Yi, Dustin Kline, Jeffrey C. Cameron, Jonathan Wignes, Sanghamitra Dey, Himadri B. Pakrasi, and Joseph M. Jez. 2013. "Probing the Origins of Glutathione Biosynthesis through Biochemical Analysis of Glutamate-Cysteine Ligase and Glutathione Synthetase from a Model Photosynthetic Prokaryote." *Biochemical Journal* 450 (1): 63–72. <https://doi.org/10.1042/BJ20121332>.
- Nagano, Nozomi. 2005. "EzCatDB: The Enzyme Catalytic-Mechanism Database." *Nucleic Acids Research* 33 (Database issue): D407-12. <https://doi.org/10.1093/nar/gki080>.
- Najmanovich, R, J Kuttner, V Sobolev, and M Edelman. 2000. "Side-Chain Flexibility in Proteins upon Ligand Binding." *Proteins* 39 (3): 261–68.
- Nakahigashi, Kenji, Yoshihiro Toya, Nobuyoshi Ishii, Tomoyoshi Soga, Miki Hasegawa, Hisami Watanabe, Yuki Takai, Masayuki Honma, Hirotada Mori, and Masaru Tomita. 2009. "Systematic Phenome Analysis of Escherichia Coli Multiple-Knockout Mutants Reveals Hidden Reactions in Central Carbon Metabolism." *Molecular Systems Biology* 5 (306): 1–14. <https://doi.org/10.1038/msb.2009.65>.
- Nakamura, Haruki, Katsuichiro Komatsu, Setsuko Nakagawa, and Hideaki Umeyama. 1985. "Visualization of Electrostatic Recognition by Enzymes for Their Ligands and Cofactors." *Journal of Molecular Graphics* 3 (1): 2–11. [https://doi.org/https://doi.org/10.1016/0263-7855\(85\)80007-2](https://doi.org/https://doi.org/10.1016/0263-7855(85)80007-2).
- Nam, Hojung, Nathan E Lewis, Joshua A Lerman, Dae-hee Lee, Roger L Chang, Donghyuk Kim, and Bernhard O Palsson. 2012a. "Evolution to Enzyme Specificity" 6185 (August).
- Nam, Hojung, Nathan E Lewis, Joshua A Lerman, Dae-Hee Lee, Roger L Chang, Donghyuk Kim, and Bernhard O Palsson. 2012b. "Network Context and Selection in the Evolution to Enzyme Specificity." *Science (New York, N.Y.)* 337 (6098): 1101–4. <https://doi.org/10.1126/science.1216861>.
- Narainsamy, Kinsley, Sandrine Farci, Emilie Braun, Christophe Junot, Corinne Cassier-Chauvat, and Franck Chauvat. 2016. "Oxidative-Stress Detoxification and Signalling in Cyanobacteria: The

- Crucial Glutathione Synthesis Pathway Supports the Production of Ergothioneine and Ophthalmate.” *Molecular Microbiology* 100 (1): 15–24. <https://doi.org/10.1111/mmi.13296>.
- Nath, Abhinav, and William M Atkins. 2008. “A Quantitative Index of Substrate Promiscuity.” *Biochemistry* 47 (1): 157–66. <https://doi.org/10.1021/bi701448p>.
- Nayal, Murad, and Barry Honig. 2006. “On the Nature of Cavities on Protein Surfaces: Application to the Identification of Drug-Binding Sites.” *Proteins* 63 (4): 892–906. <https://doi.org/10.1002/prot.20897>.
- Needleman, S B, and C D Wunsch. 1970. “A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins.” *Journal of Molecular Biology* 48 (3): 443–53.
- Nevin Gerek, Zeynep, Sudhir Kumar, and Sefika Banu Ozkan. 2013. “Structural Dynamics Flexibility Informs Function and Evolution at a Proteome Scale.” *Evolutionary Applications* 6 (3): 423–33. <https://doi.org/10.1111/eva.12052>.
- Newton, G. L., and B. Javor. 1985. “ γ -Glutamylcysteine and Thiosulfate Are the Major Low-Molecular-Weight Thiols in Halobacteria.” *Journal of Bacteriology* 161 (1): 438–41.
- Newton, Gerald L., Karen Arnold, Mitchel S. Price, Christopher Sherrill, Stephen B. Delcardayre, Yair Aharonowitz, Gerald Cohen, Julian Davies, Robert C. Fahey, and Charles Davis. 1996. “Distribution of Thiols in Microorganisms: Mycothiol Is a Major Thiol in Most Actinomycetes.” *Journal of Bacteriology* 178 (7): 1990–95. <https://doi.org/10.1128/jb.178.7.1990-1995.1996>.
- Newton, Matilda S, Vickery L Arcus, Monica L Gerth, and Wayne M Patrick. 2018. “Enzyme Evolution: Innovation Is Easy, Optimization Is Complicated.” *Current Opinion in Structural Biology* 48: 110–16. <https://doi.org/https://doi.org/10.1016/j.sbi.2017.11.007>.
- Nilmeier, Jerome P., Daniel a. Kirshner, Sergio E. Wong, and Felice C. Lightstone. 2013a. “Rapid Catalytic Template Searching as an Enzyme Function Prediction Procedure.” *PLoS ONE* 8 (5). <https://doi.org/10.1371/journal.pone.0062535>.
- Nilmeier, Jerome P, Daniel a Kirshner, Sergio E Wong, and Felice C Lightstone. 2013b. “Rapid Catalytic Template Searching as an Enzyme Function Prediction Procedure.” *PloS One* 8 (5): e62535. <https://doi.org/10.1371/journal.pone.0062535>.
- Nobeli, Irene, Angelo D. Favia, and Janet M. Thornton. 2009. “Protein Promiscuity and Its Implications for Biotechnology.” *Nature Biotechnology* 27 (2): 157–67. <https://doi.org/10.1038/nbt1519>.
- Notebaart, Richard A., Bálint Kintsés, Adam M. Feist, and Balázs Papp. 2018. “Underground Metabolism: Network-Level Perspective and Biotechnological Potential.” *Current Opinion in Biotechnology* 49: 108–14. <https://doi.org/10.1016/j.copbio.2017.07.015>.
- Notebaart, Richard A, Balázs Szappanos, Bálint Kintsés, Ferenc Pál, Ádám Györkei, Balázs Bogos, Viktória Lázár, et al. 2014. “Network-Level Architecture and the Evolutionary Potential of Underground Metabolism.” *Proceedings of the National Academy of Sciences* 111 (32): 11762–67. <https://doi.org/10.1073/pnas.1406102111>.
- O’Boyle, Noel M, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. 2011. “Open Babel: An Open Chemical Toolbox.” *Journal of Cheminformatics* 3 (1): 33. <https://doi.org/10.1186/1758-2946-3-33>.
- O’Brien, P J, and D Herschlag. 1999. “Catalytic Promiscuity and the Evolution of New Enzymatic Activities.” *Chemistry & Biology* 6 (4): R91–105. [https://doi.org/10.1016/S1074-5521\(99\)80033-7](https://doi.org/10.1016/S1074-5521(99)80033-7).
- Oberhardt, Matthew A., Raphy Zarecki, Leah Reshef, Fangfang Xia, Miquel Duran-Frigola, Rachel Schreiber, Christopher S. Henry, et al. 2016. “Systems-Wide Prediction of Enzyme Promiscuity Reveals a New Underground Alternative Route for Pyridoxal 5’-Phosphate Production in *E. Coli*.” *PLoS Computational Biology* 12 (1): 1–19. <https://doi.org/10.1371/journal.pcbi.1004705>.
- Olguin, Luis F, Sarah E Askew, AnnMarie C O’Donoghue, and Florian Hollfelder. 2008. “Efficient

- Catalytic Promiscuity in an Enzyme Superfamily: An Arylsulfatase Shows a Rate Acceleration of 1013 for Phosphate Monoester Hydrolysis.” *Journal of the American Chemical Society* 130 (49): 16547–55. <https://doi.org/10.1021/ja8047943>.
- Ondrechen, Mary Jo, James G Clifton, and Dagmar Ringe. 2001. “THEMATICS: A Simple Computational Predictor of Enzyme Function from Structure.”
- Onesti, S, G Desogus, a Brevet, J Chen, P Plateau, S Blanquet, and P Brick. 2000. “Structural Studies of Lysyl-TRNA Synthetase: Conformational Changes Induced by Substrate Binding [In Process Citation].” *Biochemistry* 39 (42): 12853–61. <https://doi.org/10.1021/bi001487r>.
- Pabis, Anna, Fernanda Duarte, and Shina C L Kamerlin. 2016. “Promiscuity in the Enzymatic Catalysis of Phosphate and Sulfate Transfer.” *Biochemistry* 55 (22): 3061–81. <https://doi.org/10.1021/acs.biochem.6b00297>.
- Pabis, Anna, and Shina Caroline Lynn Kamerlin. 2016. “Promiscuity and Electrostatic Flexibility in the Alkaline Phosphatase Superfamily.” *Current Opinion in Structural Biology* 37 (April): 14–21. <https://doi.org/10.1016/j.sbi.2015.11.008>.
- Pai, E F, and G E Schulz. 1983. “The Catalytic Mechanism of Glutathione Reductase as Derived from X-Ray Diffraction Analyses of Reaction Intermediates.” *The Journal of Biological Chemistry* 258 (3): 1752–57.
- Pande, Jyoti, Magdalena M Szewczyk, and Ashok K Grover. 2010. “Phage Display: Concept, Innovations, Applications and Future.” *Biotechnology Advances* 28 (6): 849–58. <https://doi.org/10.1016/j.biotechadv.2010.07.004>.
- Pandya, Chetanya, Jeremiah D. Farelli, Debra Dunaway-Mariano, and Karen N. Allen. 2014. “Enzyme Promiscuity: Engine of Evolutionary Innovation.” *Journal of Biological Chemistry* 289 (44): 30229–36. <https://doi.org/10.1074/jbc.R114.572990>.
- Panigrahi, Sunil K, and Gautam R Desiraju. 2007. “Strong and Weak Hydrogen Bonds in the Protein-Ligand Interface.” *Proteins* 67 (1): 128–41. <https://doi.org/10.1002/prot.21253>.
- Park, Jooyoung, Ann M Guggisberg, Audrey R Odom, and Niraj H Tolia. 2015. “Cap-Domain Closure Enables Diverse Substrate Recognition by the C2-Type Haloacid Dehalogenase-like Sugar Phosphatase Plasmodium Falciparum HAD1.” *Acta Crystallographica Section D: Biological Crystallography*. <https://doi.org/10.1107/S1399004715012067>.
- Parkison, C, K Ashizawa, P McPhie, K H Lin, and S Y Cheng. 1991. “The Monomer of Pyruvate Kinase, Subtype M1, Is Both a Kinase and a Cytosolic Thyroid Hormone Binding Protein.” *Biochemical and Biophysical Research Communications* 179 (1): 668–74.
- Pastore, A, F Piemonte, M Locatelli, A Lo Russo, L M Gaeta, G Tozzi, and G Federici. 2001. “Determination of Blood Total, Reduced, and Oxidized Glutathione in Pediatric Subjects.” *Clinical Chemistry* 47 (8): 1467–69.
- Patrick, Wayne M, and Ichiro Matsumura. 2008. “A Study in Molecular Contingency: Glutamine Phosphoribosylpyrophosphate Amidotransferase Is a Promiscuous and Evolvable Phosphoribosylanthranilate Isomerase.” *Journal of Molecular Biology* 377 (2): 323–36. <https://doi.org/10.1016/j.jmb.2008.01.043>.
- Patrick, Wayne M, Erik M Quandt, Dan B Swartzlander, and Ichiro Matsumura. 2007. “Multicopy Suppression Underpins Metabolic Evolvability.” *Molecular Biology and Evolution* 24 (12): 2716–22. <https://doi.org/10.1093/molbev/msm204>.
- Peracchi, A. 2001. “Enzyme Catalysis: Removing Chemically ‘essential’ Residues by Site-Directed Mutagenesis.” *Trends in Biochemical Sciences* 26 (8): 497–503.
- Petrova, Natalia V, and Cathy H Wu. 2006. “Prediction of Catalytic Residues Using Support Vector Machine with Selected Protein Sequence and Structural Properties.” *BMC Bioinformatics* 7

(January): 312. <https://doi.org/10.1186/1471-2105-7-312>.

- Pettersen, Eric F, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. 2004. "UCSF Chimera--a Visualization System for Exploratory Research and Analysis." *Journal of Computational Chemistry* 25 (13): 1605–12. <https://doi.org/10.1002/jcc.20084>.
- Pham, Thi Thanh My, and Michel Sylvestre. 2013. "Has the Bacterial Biphenyl Catabolic Pathway Evolved Primarily to Degrade Biphenyl? The Diphenylmethane Case." *Journal of Bacteriology* 195 (16): 3563–74. <https://doi.org/10.1128/JB.00161-13>.
- Pham, Thi Thanh My, Youbin Tu, and Michel Sylvestre. 2012. "Remarkable Ability of *Pandoraea Pnomenusa* B356 Biphenyl Dioxygenase to Metabolize Simple Flavonoids." *Applied and Environmental Microbiology* 78 (10): 3560–70. <https://doi.org/10.1128/AEM.00225-12>.
- Phillips, James C., Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. 2005. "Scalable Molecular Dynamics with NAMD." *Journal of Computational Chemistry* 26 (16): 1781–1802. <https://doi.org/10.1002/jcc.20289>.
- Poelarends, G J, V Puthan Veetil, and C P Whitman. 2008. "The Chemical Versatility of the Beta-Alpha-Beta Fold: Catalytic Promiscuity and Divergent Evolution in the Tautomerase Superfamily." *Cellular and Molecular Life Sciences : CMLS* 65 (22): 3606–18. <https://doi.org/10.1007/s00018-008-8285-x>.
- Poelarends, Gerrit J, Hector Serrano, William H Jr Johnson, David W Hoffman, and Christian P Whitman. 2004. "The Hydratase Activity of Malonate Semialdehyde Decarboxylase: Mechanistic and Evolutionary Implications." *Journal of the American Chemical Society* 126 (48): 15658–59. <https://doi.org/10.1021/ja044304n>.
- Pompella, Alfonso, and Alessandro Corti. 2015. "Editorial: The Changing Faces of Glutathione, a Cellular Protagonist." *Frontiers in Pharmacology* 6 (May): 98. <https://doi.org/10.3389/fphar.2015.00098>.
- Pompella, Alfonso, Athanase Visvikis, Aldo Paolicchi, Vincenzo De Tata, and Alessandro F Casini. 2003. "The Changing Faces of Glutathione, a Cellular Protagonist." *Biochemical Pharmacology* 66 (8): 1499–1503.
- Porter, Craig T, Gail J Bartlett, and Janet M Thornton. 2004. "The Catalytic Site Atlas: A Resource of Catalytic Sites and Residues Identified in Enzymes Using Structural Data." *Nucleic Acids Research* 32 (Database issue): D129–33. <https://doi.org/10.1093/nar/gkh028>.
- Prilusky, Jaime, Clifford E Felder, Tzviya Zeev-Ben-Mordehai, Edwin H Rydberg, Orna Man, Jacques S Beckmann, Israel Silman, and Joel L Sussman. 2005. "FoldIndex©: A Simple Tool to Predict Whether a given Protein Sequence Is Intrinsically Unfolded." *Bioinformatics* 21 (16): 3435–38. <http://dx.doi.org/10.1093/bioinformatics/bti537>.
- Qiu, Huan, Dana C Price, Eun Chan Yang, Hwan Su Yoon, and Debashish Bhattacharya. 2015. "Evidence of Ancient Genome Reduction in Red Algae (Rhodophyta)." *Journal of Phycology* 51 (4): 624–36. <https://doi.org/10.1111/jpy.12294>.
- Radzicka, A, and R Wolfenden. 1995. "A Proficient Enzyme." *Science (New York, N.Y.)* 267 (5194): 90–93.
- Rahimi, Mehran, Jan-Ytzen van der Meer, Edzard M Geertsema, Harshwardhan Poddar, Bert-Jan Baas, and Gerrit J Poelarends. 2016. "Mutations Closer to the Active Site Improve the Promiscuous Aldolase Activity of 4-Oxalocrotonate Tautomerase More Effectively than Distant Mutations." *Chembiochem : A European Journal of Chemical Biology* 17 (13): 1225–28. <https://doi.org/10.1002/cbic.201600149>.
- Rahman, Syed Asad, Sergio Martinez Cuesta, Nicholas Furnham, Gemma L Holliday, and Janet M Thornton. 2014. "EC-BLAST: A Tool to Automatically Search and Compare Enzyme Reactions."

- Nature Methods* 11 (2): 171–74. <https://doi.org/10.1038/nmeth.2803>.
- Redinbo, M R, S Bencharit, and P M Potter. 2003. “Human Carboxylesterase 1: From Drug Metabolism to Drug Discovery.” *Biochemical Society Transactions* 31 (Pt 3): 620–24. <https://doi.org/10.1042/>.
- Redinbo, Matthew R. 2004. “Promiscuity: What Protects Us, Perplexes Us.” *Drug Discovery Today*. England. [https://doi.org/10.1016/S1359-6446\(04\)03087-9](https://doi.org/10.1016/S1359-6446(04)03087-9).
- Renata, Hans, Z. Jane Wang, and Frances H. Arnold. 2015. “Expanding the Enzyme Universe: Accessing Non-Natural Reactions by Mechanism-Guided Directed Evolution.” *Angewandte Chemie International Edition*, no. 150: n/a-n/a. <https://doi.org/10.1002/anie.201409470>.
- Ribeiro, Antonio J M, Gemma L Holliday, Nicholas Furnham, Jonathan D Tyzack, Katherine Ferris, and Janet M Thornton. 2018. “Mechanism and Catalytic Site Atlas (M-CSA): A Database of Enzyme Reaction Mechanisms and Active Sites.” *Nucleic Acids Research* 46 (D1): D618–23. <https://doi.org/10.1093/nar/gkx1012>.
- Rice, P, I Longden, and A Bleasby. 2000. “EMBOSS: The European Molecular Biology Open Software Suite.” *Trends in Genetics : TIG* 16 (6): 276–77.
- Ringe, D. 1995. “What Makes a Binding Site a Binding Site?” *Current Opinion in Structural Biology* 5 (6): 825–29.
- Rothman, Steven C, and Jack F Kirsch. 2003. “How Does an Enzyme Evolved in Vitro Compare to Naturally Occurring Homologs Possessing the Targeted Function? Tyrosine Aminotransferase from Aspartate Aminotransferase.” *Journal of Molecular Biology* 327 (3): 593–608.
- Rzem, R, M-F Vincent, E Van Schaftingen, and M Veiga-da-Cunha. 2007. “L-2-Hydroxyglutaric Aciduria, a Defect of Metabolite Repair.” *Journal of Inherited Metabolic Disease* 30 (5): 681–89. <https://doi.org/10.1007/s10545-007-0487-0>.
- Safo, M K, F N Musayev, M L di Salvo, and V Schirch. 2001. “X-Ray Structure of Escherichia Coli Pyridoxine 5'-Phosphate Oxidase Complexed with Pyridoxal 5'-Phosphate at 2.0 Å Resolution.” *Journal of Molecular Biology* 310 (4): 817–26. <https://doi.org/10.1006/jmbi.2001.4734>.
- Saito, Takaya, and Marc Rehmsmeier. 2015. “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.” *PLoS One* 10 (3): e0118432. <https://doi.org/10.1371/journal.pone.0118432>.
- Sankararaman, Sriram, Fei Sha, Jack F. Kirsch, Michael I. Jordan, and Kimmen Sjölander. 2010. “Active Site Prediction Using Evolutionary and Structural Information.” *Bioinformatics* 26 (5): 617–24. <https://doi.org/10.1093/bioinformatics/btq008>.
- Sankararaman, Sriram, and Kimmen Sjölander. 2008a. “INTREPID - Information-Theoretic Tree Traversal for Protein Functional Site Identification.” *Bioinformatics* 24 (21): 2445–52. <https://doi.org/10.1093/bioinformatics/btn474>.
- Schäfer, Holger J., Steffen Greiner, Thomas Rausch, and Angela Haag-Kerwer. 1997. “In Seedlings of the Heavy Metal Accumulator Brassica Juncea Cu²⁺-differentially Affects Transcript Amounts for γ -Glutamylcysteine Synthetase (γ -ECS) and Metallothionein (MT2).” *FEBS Letters* 404 (2–3): 216–20. [https://doi.org/10.1016/S0014-5793\(97\)00132-4](https://doi.org/10.1016/S0014-5793(97)00132-4).
- Schmid, A, J S Dordick, B Hauer, A Kiener, M Wubbolts, and B Witholt. 2001. “Industrial Biocatalysis Today and Tomorrow.” *Nature* 409 (6817): 258–68. <https://doi.org/10.1038/35051736>.
- Schmidt, Dawn M Z, Emily C Mundorff, Michael Dojka, Ericka Bermudez, Jon E Ness, Sridhar Govindarajan, Patricia C Babbitt, Jeremy Minshull, and John A Gerlt. 2003. “Evolutionary Potential of (Beta/Alpha)₈-Barrels: Functional Promiscuity Produced by Single Substitutions in the Enolase Superfamily.” *Biochemistry* 42 (28): 8387–93. <https://doi.org/10.1021/bi034769a>.
- Scott, Colin, Colin J Jackson, Chris W Coppin, Roslyn G Mourant, Margaret E Hilton, Tara D Sutherland,

- Robyn J Russell, and John G Oakeshott. 2009. "Catalytic Improvement and Evolution of Atrazine Chlorohydrolase ." *Applied and Environmental Microbiology*. <https://doi.org/10.1128/AEM.02634-08>.
- Scott, D L, Z Otwinowski, M H Gelb, and P B Sigler. 1990. "Crystal Structure of Bee-Venom Phospholipase A2 in a Complex with a Transition-State Analogue." *Science (New York, N.Y.)* 250 (4987): 1563–66.
- Seebeck, Florian P. 2010. "In Vitro Reconstitution of Mycobacterial Ergothioneine Biosynthesis." *Journal of the American Chemical Society* 132 (19): 6632–33. <https://doi.org/10.1021/ja101721e>.
- Seelig, Gail Foure, and Alton Meister. 1984. "Y-Glutamylcysteine Synthetase" 259 (6): 3534–38.
- Sekine, Shun-ichi, Osamu Nureki, Daniel Y Dubois, Stéphane Bernier, Robert Chênevert, Jacques Lapointe, Dmitry G Vassylyev, and Shigeyuki Yokoyama. 2003. "ATP Binding by Glutamyl-TRNA Synthetase Is Switched to the Productive Mode by TRNA Binding." *The EMBO Journal* 22 (3): 676–88. <https://doi.org/10.1093/emboj/cdg053>.
- Sharma, Ashok K, Shubham K Jaiswal, Nikhil Chaudhary, and Vineet K Sharma. 2017. "A Novel Approach for the Prediction of Species-Specific Biotransformation of Xenobiotic/Drug Molecules by the Human Gut Microbiota." *Scientific Reports* 7 (1): 9751. <https://doi.org/10.1038/s41598-017-10203-6>.
- Sharma, Vivek, Sujata Sharma, Kerstin Hoener Zu Bentrup, John D. McKinney, David G. Russell, William R. Jacobs, and James C. Sacchettini. 2000. "Structure of Isocitrate Lyase, a Persistence Factor of Mycobacterium Tuberculosis." *Nature Structural Biology* 7 (8): 663–68. <https://doi.org/10.1038/77964>.
- Shehadi, Ihsan A, and Alper Uzun. 2004. "THEMATICS Is Effective for Active Site Prediction in Comparative Model Structures."
- Shin, J S, and B G Kim. 2001. "Comparison of the Omega-Transaminases from Different Microorganisms and Application to Production of Chiral Amines." *Bioscience, Biotechnology, and Biochemistry* 65 (8): 1782–88.
- Sillitoe, Ian, Tony E. Lewis, Alison Cuff, Sayoni Das, Paul Ashford, Natalie L. Dawson, Nicholas Furnham, et al. 2015. "CATH: Comprehensive Structural and Functional Annotations for Genome Sequences." *Nucleic Acids Research* 43 (D1): D376–81. <https://doi.org/10.1093/nar/gku947>.
- Silva, Giordano F Z da, and Li-June Ming. 2005. "Catechol Oxidase Activity of Di-Cu²⁺-Substituted Aminopeptidase from Streptomyces Griseus." *Journal of the American Chemical Society* 127 (47): 16380–81. <https://doi.org/10.1021/ja056034u>.
- Sivakumar, Tadi Venkata, Anirban Bhaduri, Rajasekhara Reddy Duvvuru Muni, Jin Hwan Park, and Tae Yong Kim. 2018. "SimCAL: A Flexible Tool to Compute Biochemical Reaction Similarity." *BMC Bioinformatics* 19 (1): 254. <https://doi.org/10.1186/s12859-018-2248-5>.
- Skolnick, Jeffrey, and Mu Gao. 2013. "Interplay of Physics and Evolution in the Likely Origin of Protein Biochemical Function." *Proceedings of the National Academy of Sciences* 110 (23): 9344 LP-9349. <http://www.pnas.org/content/110/23/9344.abstract>.
- Sobolev, V, A Sorokine, J Prilusky, E E Abola, and M Edelman. 1999. "Automated Analysis of Interatomic Contacts in Proteins." *Bioinformatics (Oxford, England)* 15 (4): 327–32.
- Somers, W S, M L Stahl, and F X Sullivan. 1998. "GDP-Fucose Synthetase from Escherichia Coli: Structure of a Unique Member of the Short-Chain Dehydrogenase/Reductase Family That Catalyzes Two Distinct Reactions at the Same Active Site." *Structure (London, England : 1993)* 6 (12): 1601–12.
- Soo, Valerie W C, Yuliana Yosaatmadja, Christopher J Squire, and X Wayne M Patrick. 2016. "Mechanistic and Evolutionary Insights from the Reciprocal Promiscuity of Two Pyridoxal

- Phosphate-Dependent” 291 (38): 19873–87. <https://doi.org/10.1074/jbc.M116.739557>.
- Soteras Gutiérrez, Ignacio, Fang Yu Lin, Kenno Vanommeslaeghe, Justin A. Lemkul, Kira A. Armacost, Charles L. Brooks, and Alexander D. MacKerell. 2016. “Parametrization of Halogen Bonds in the CHARMM General Force Field: Improved Treatment of Ligand–protein Interactions.” *Bioorganic and Medicinal Chemistry* 24 (20): 4812–25. <https://doi.org/10.1016/j.bmc.2016.06.034>.
- Steinkellner, Georg, Christian C Gruber, Tea Pavkov-Keller, Alexandra Binter, Kerstin Steiner, Christoph Winkler, Andrzej Lyskowski, et al. 2014. “Identification of Promiscuous Ene-Reductase Activity by Mining Structural Databases Using Active Site Constellations.” *Nature Communications* 5 (May): 4150. <https://doi.org/10.1038/ncomms5150>.
- Steinmetz, Andrea, Maria Vyazmensky, Danilo Meyer, Ze Ev Barak, Ralph Golbik, David M Chipman, and Kai Tittmann. 2010. “Valine 375 and Phenylalanine 109 Confer Affinity and Specificity for Pyruvate as Donor Substrate in Acetohydroxy Acid Synthase Isozyme II from *Escherichia Coli*.” *Biochemistry* 49 (25): 5188–99. <https://doi.org/10.1021/bi100555q>.
- Stout, Jan, Dirk De Vos, Bjorn Vergauwen, and Savvas N. Savvides. 2012. “Glutathione Biosynthesis in Bacteria by Bifunctional GshF Is Driven by a Modular Structure Featuring a Novel Hybrid ATP-Grasp Fold.” *Journal of Molecular Biology* 416 (4): 486–94. <https://doi.org/10.1016/j.jmb.2011.12.046>.
- Strait, Bonnie J, and T Gregory Dewey. 1996. “The Shannon Information Entropy.” *Biophysical Journal* 71 (July): 148–55.
- Sugase, Kenji, H Jane Dyson, and Peter E Wright. 2007. “Mechanism of Coupled Folding and Binding of an Intrinsically Disordered Protein.” *Nature* 447 (7147): 1021–25. <https://doi.org/10.1038/nature05858>.
- Sunden, Fanny, Ishraq AlSadhan, Artem Lyubimov, Tzanko Doukov, Jeffrey Swan, and Daniel Herschlag. 2017. “Differential Catalytic Promiscuity of the Alkaline Phosphatase Superfamily Bimetallo Core Reveals Mechanistic Features Underlying Enzyme Evolution.” *The Journal of Biological Chemistry* 292 (51): 20960–74. <https://doi.org/10.1074/jbc.M117.788240>.
- Taglieber, Andreas, Horst Höbenreich, J. Daniel Carballeira, Régis J.G. Mondière, and Manfred T. Reetz. 2007. “Alternate-Site Enzyme Promiscuity.” *Angewandte Chemie - International Edition* 46 (45): 8597–8600. <https://doi.org/10.1002/anie.200702751>.
- Tang, Yu-Rong, Zhi-Ya Sheng, Yong-Zi Chen, and Ziding Zhang. 2008. “An Improved Prediction of Catalytic Residues in Enzyme Structures.” *Protein Engineering, Design and Selection* 21 (5): 295–302. <http://dx.doi.org/10.1093/protein/gzn003>.
- Tawfik, Dan S. 2010. “Messy Biology and the Origins of Evolutionary Innovations.” *Nature Chemical Biology* 6 (10): 692–96. <https://doi.org/10.1038/nchembio.441>.
- Tawfik, Olga Khersonsky and Dan S. 2010. “Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective.” *Annual Review of Biochemistry* 79 (1): 471–505. <https://doi.org/10.1146/annurev-biochem-030409-143718>.
- Taylor, Alexander B, Robert M Czerwinski, William H Johnson, Christian P Whitman, and Marvin L Hackert. 1998. “Crystal Structure of 4-Oxalocrotonate Tautomerase Inactivated by 2-Oxo-3-Pentynoate at 2.4 Å Resolution: Analysis and Implications for the Mechanism of Inactivation and Catalysis.” *Biochemistry* 37 (42): 14692–700. <https://doi.org/10.1021/bi981607j>.
- The UniProt Consortium. 2017. “UniProt: The Universal Protein Knowledgebase.” *Nucleic Acids Research* 45 (D1): D158–69. <http://dx.doi.org/10.1093/nar/gkw1099>.
- Theodossis, Alex, Helen Walden, Elaine J Westwick, Helen Connaris, Henry J Lambie, David W Hough, Michael J Danson, and Garry L Taylor. 2004. “The Structural Basis for Substrate Promiscuity in 2-Keto-3-Deoxygluconate Aldolase from the Entner-Doudoroff Pathway in *Sulfolobus Solfataricus*.” *Journal of Biological Chemistry* 279 (42): 43886–92. <https://doi.org/10.1074/jbc.M407702200>.

- Thompson, Matthew K, Mary E Keithly, Joel Harp, Paul D Cook, Kevin L Jagessar, Gary A Sulikowski, and Richard N Armstrong. 2013. "Structural and Chemical Aspects of Resistance to the Antibiotic Fosfomycin Conferred by FosB from *Bacillus Cereus*." *Biochemistry* 52 (41): 7350–62. <https://doi.org/10.1021/bi4009648>.
- Tobi, Dror, and Ivett Bahar. 2005. "Structural Changes Involved in Protein Binding Correlate with Intrinsic Motions of Proteins in the Unbound State." *Proceedings of the National Academy of Sciences of the United States of America* 102 (52): 18908–13. <https://doi.org/10.1073/pnas.0507603102>.
- Tokuriki, Nobuhiko, Colin J Jackson, Livnat Afriat-Jurnou, Kirsten T Wyganowski, Renmei Tang, and Dan S Tawfik. 2012. "Diminishing Returns and Tradeoffs Constrain the Laboratory Optimization of an Enzyme." *Nature Communications* 3: 1257. <https://doi.org/10.1038/ncomms2246>.
- Tokutake, N, J Hiratake, M Katoh, T Irie, H Kato, and J Oda. 1998. "Design, Synthesis and Evaluation of Transition-State Analogue Inhibitors of Escherichia Coli Gamma-Glutamylcysteine Synthetase." *Bioorganic & Medicinal Chemistry* 6 (10): 1935–53.
- Tommaso, Paolo Di, Sebastien Moretti, Ioannis Xenarios, Miquel Orobitg, Alberto Montanyola, Jia Ming Chang, Jean Francois Taly, and Cedric Notredame. 2011. "T-Coffee: A Web Server for the Multiple Sequence Alignment of Protein and RNA Sequences Using Structural Information and Homology Extension." *Nucleic Acids Research* 39 (SUPPL. 2): 13–17. <https://doi.org/10.1093/nar/gkr245>.
- Tong, Wenxu, Ying Wei, Leonel F. Murga, Mary Jo Ondrechen, and Ronald J. Williams. 2009. "Partial Order Optimum Likelihood (POOL): Maximum Likelihood Prediction of Protein Active Site Residues Using 3D Structure and Sequence Properties." *PLoS Computational Biology* 5 (1). <https://doi.org/10.1371/journal.pcbi.1000266>.
- Torre, Oliver, Ignacio Alfonso, and Vicente Gotor. 2004. "Lipase Catalysed Michael Addition of Secondary Amines to Acrylonitrile." *Chemical Communications (Cambridge, England)*, no. 15 (August): 1724–25. <https://doi.org/10.1039/b402244k>.
- Traverso, Nicola, Roberta Ricciarelli, Mariapaola Nitti, Barbara Marengo, Anna Lisa Furfaro, Maria Adelaide Pronzato, Umberto Maria Marinari, and Cinzia Domenicotti. 2013. "Role of Glutathione in Cancer Progression and Chemoresistance." *Oxidative Medicine and Cellular Longevity* 2013: 972913. <https://doi.org/10.1155/2013/972913>.
- Tremblay, Lee W, Debra Dunaway-Mariano, and Karen N Allen. 2006. "Structure and Activity Analyses of Escherichia Coli K-12 NagD Provide Insight into the Evolution of Biochemical Function in the Haloalkanoic Acid Dehalogenase Superfamily." *Biochemistry* 45 (4): 1183–93. <https://doi.org/10.1021/bi051842j>.
- Tseng, Yan Yuan, and Wen-Hsiung Li. 2011. "Evolutionary Approach to Predicting the Binding Site Residues of a Protein from Its Primary Sequence." *Proceedings of the National Academy of Sciences* 108 (13): 5313–18. <https://doi.org/10.1073/pnas.1102210108>.
- Tsuchiya, Yuko, Kengo Kinoshita, and Haruki Nakamura. 2004. "Structure-Based Prediction of DNA-Binding Sites on Proteins Using the Empirical Preference of Electrostatic Potential and the Shape of Molecular Surfaces." *Proteins* 55 (4): 885–94. <https://doi.org/10.1002/prot.20111>.
- Tyzack, Jonathan D., Nicholas Furnham, Ian Sillitoe, Christine M. Orengo, and Janet M. Thornton. 2017. "Understanding Enzyme Function Evolution from a Computational Perspective." *Current Opinion in Structural Biology* 47: 131–39. <https://doi.org/10.1016/j.sbi.2017.08.003>.
- Ufarte, Lisa, Elisabeth Laville, Sophie Duquesne, Diego Morgavi, Patrick Robe, Christophe Klopp, Angeline Rizzo, Sandra Pizzut-Serin, and Gabrielle Potocki-Veronese. 2017. "Discovery of Carbamate Degrading Enzymes by Functional Metagenomics." *PloS One* 12 (12): e0189201. <https://doi.org/10.1371/journal.pone.0189201>.
- Uppenberg, J, M T Hansen, S Patkar, and T A Jones. 1994. "The Sequence, Crystal Structure Determination and Refinement of Two Crystal Forms of Lipase B from *Candida Antarctica*." *Structure (London, England : 1993)* 2 (4): 293–308.

- Veeravalli, Karthik, Dana Boyd, Brent L. Iverson, Jon Beckwith, and George Georgiou. 2011a. "Laboratory Evolution of Glutathione Biosynthesis Reveals Natural Compensatory Pathways." *Nature Chemical Biology* 7 (2): 101–5. <https://doi.org/10.1038/nchembio.499>.
- Vergauwen, Bjorn, Dirk De Vos, and Jozef J. Van Beeumen. 2006. "Characterization of the Bifunctional γ -Glutamate-Cysteine Ligase/Glutathione Synthetase (GshF) of *Pasteurella Multocida*." *Journal of Biological Chemistry* 281 (7): 4380–94. <https://doi.org/10.1074/jbc.M509517200>.
- Villali, Janice, and Dorothee Kern. 2010. "Choreographing an Enzyme's Dance." *Current Opinion in Chemical Biology* 14 (5): 636–43. <https://doi.org/https://doi.org/10.1016/j.cbpa.2010.08.007>.
- Wallden, Karin, and Par Nordlund. 2011. "Structural Basis for the Allosteric Regulation and Substrate Recognition of Human Cytosolic 5'-Nucleotidase II." *Journal of Molecular Biology* 408 (4): 684–96. <https://doi.org/10.1016/j.jmb.2011.02.059>.
- Wang, Guoli, and Roland L. Dunbrack. 2003. "PISCES: A Protein Sequence Culling Server." *Bioinformatics* 19 (12): 1589–91. <https://doi.org/10.1093/bioinformatics/btg224>.
- Wang, Kai, Jeremy a Horst, Gong Cheng, David C Nickle, and Ram Samudrala. 2008. "Protein Meta-Functional Signatures from Combining Sequence, Structure, Evolution, and Amino Acid Property Information." *PLoS Computational Biology* 4 (9): e1000181. <https://doi.org/10.1371/journal.pcbi.1000181>.
- Wang, Peng, Mingming Jin, and Guoping Zhu. 2012. "Biochemical and Molecular Characterization of NAD(+)-Dependent Isocitrate Dehydrogenase from the Ethanologenic Bacterium *Zymomonas Mobilis*." *FEMS Microbiology Letters* 327 (2): 134–41. <https://doi.org/10.1111/j.1574-6968.2011.02467.x>.
- Wang, Susan C, William H Jr Johnson, and Christian P Whitman. 2003. "The 4-Oxalocrotonate Tautomerase- and YwhB-Catalyzed Hydration of 3E-Haloacrylates: Implications for the Evolution of New Enzymatic Activities." *Journal of the American Chemical Society* 125 (47): 14282–83. <https://doi.org/10.1021/ja0370948>.
- Waterhouse, Andrew M, James B Procter, David M A Martin, Michèle Clamp, and Geoffrey J Barton. 2009. "Jalview Version 2—a Multiple Sequence Alignment Editor and Analysis Workbench." *Bioinformatics* 25 (9): 1189–91. <http://dx.doi.org/10.1093/bioinformatics/btp033>.
- Webb, Benjamin, and Andrej Sali. 2016. "Comparative Protein Structure Modeling Using MODELLER." *Current Protocols in Protein Science* 86 (November): 2.9.1-2.9.37. <https://doi.org/10.1002/cpp.20>.
- Weber, K. 1968. "New Structural Model of E. Coli Aspartate Transcarbamylase and the Amino-Acid Sequence of the Regulatory Polypeptide Chain." *Nature* 218 (5147): 1116–19.
- Whisstock, James C, and Arthur M Lesk. 2003. "Prediction of Protein Function from Protein Sequence and Structure." *Quarterly Reviews of Biophysics* 36 (3): 307–40.
- Wieman, Heather, Kristin Tøndel, Endre Anderssen, and Finn Drabløs. 2004. "Homology-Based Modelling of Targets for Rational Drug Design." *Mini Reviews in Medicinal Chemistry* 4 (7): 793–804. <https://doi.org/10.2174/1389557043403639>.
- Wilfried, Blokzijl, and Engberts Jan B F N. 2018. "Hydrophobic Effects. Opinions and Facts." *Angewandte Chemie International Edition in English* 32 (11): 1545–79. <https://doi.org/10.1002/anie.199315451>.
- Wilkinson, Bonney, and Hiram F Gilbert. 2004. "Protein Disulfide Isomerase." *Biochimica et Biophysica Acta* 1699 (1–2): 35–44. <https://doi.org/10.1016/j.bbapap.2004.02.017>.
- Witek, Marta A, Emily G Kuiper, Elizabeth Minten, Emily K Crispell, X Graeme L Conn, and Norma Allewell. 2017. "A Novel Motif for S -Adenosyl- L -Methionine Binding by the Ribosomal RNA Methyltransferase TlyA from *Mycobacterium*" 292 (5): 1977–87. <https://doi.org/10.1074/jbc.M116.752659>.

- Wright, Peter E, and H Jane Dyson. 2009. "Linking Folding and Binding." *Current Opinion in Structural Biology* 19 (1): 31–38. <https://doi.org/10.1016/j.sbi.2008.12.003>.
- Wu, Rui, John A Latham, Danqi Chen, Jeremiah Farelli, Hong Zhao, Kaila Matthews, Karen N Allen, and Debra Dunaway-Mariano. 2014. "Structure and Catalysis in the Escherichia Coli Hotdog-Fold Thioesterase Paralogs YdiI and YbdB." *Biochemistry* 53 (29): 4788–4805. <https://doi.org/10.1021/bi500334v>.
- Yamamoto, Kohji, Akifumi Higashiura, Md Tofazzal Hossain, Naotaka Yamada, Takahiro Shiotsuki, and Atsushi Nakagawa. 2015. "Structural Characterization of the Catalytic Site of a Nilaparvata Lugens Delta-Class Glutathione Transferase." *Archives of Biochemistry and Biophysics* 566: 36–42. <https://doi.org/10.1016/j.abb.2014.12.001>.
- Yang, Gloria, Nansook Hong, Florian Baier, Colin J Jackson, and Nobuhiko Tokuriki. 2016. "Conformational Tinkering Drives Evolution of a Promiscuous Activity through Indirect Mutational Effects." *Biochemistry* 55 (32): 4583–93. <https://doi.org/10.1021/acs.biochem.6b00561>.
- Yang, Lee-Wei, and Ivet Bahar. 2005. "Coupling between Catalytic Site and Collective Dynamics: A Requirement for Mechanochemical Activity of Enzymes." *Structure (London, England : 1993)* 13 (6): 893–904. <https://doi.org/10.1016/j.str.2005.03.015>.
- Yang, Yi, Ying Chen, Elisabet Johansson, Scott N Schneider, Howard G Shertzer, Daniel W Nebert, and Timothy P Dalton. 2007. "Interaction between the Catalytic and Modifier Subunits of Glutamate-Cysteine Ligase." *Biochemical Pharmacology* 74 (2): 372–81. <https://doi.org/https://doi.org/10.1016/j.bcp.2007.02.003>.
- Yao, Jianzhuang, Haobo Guo, Minta Chairprasongsuk, Nan Zhao, Feng Chen, Xiaohan Yang, and Hong Guo. 2015. "Substrate-Assisted Catalysis in the Reaction Catalyzed by Salicylic Acid Binding Protein 2 (SABP2), a Potential Mechanism of Substrate Discrimination for Some Promiscuous Enzymes" 2. <https://doi.org/10.1021/acs.biochem.5b00638>.
- Ycas, M. 1974. "On Earlier States of the Biochemical System." *Journal of Theoretical Biology* 44 (1): 145–60.
- Yeung, David T, David E Lenz, and Douglas M Cerasoli. 2005. "Analysis of Active-Site Amino-Acid Residues of Human Serum Paraoxonase Using Competitive Substrates." *The FEBS Journal* 272 (9): 2225–30. <https://doi.org/10.1111/j.1742-4658.2005.04646.x>.
- Yon, J. M., D. Perahia, and C. Ghélis. 1998. "Conformational Dynamics and Enzyme Activity." *Biochimie* 80 (1): 33–42. [https://doi.org/10.1016/S0300-9084\(98\)80054-0](https://doi.org/10.1016/S0300-9084(98)80054-0).
- Yoshikuni, Yasuo, Thomas E Ferrin, and Jay D Keasling. 2006. "Designed Divergent Evolution of Enzyme Function." *Nature* 440 (7087): 1078–82. <https://doi.org/10.1038/nature04607>.
- Youn, Eunseog, Brandon Peters, Predrag Radivojac, and Sean D Mooney. 2007. "Evaluation of Features for Catalytic Residue Prediction in Novel Folds." *Protein Science: A Publication of the Protein Society* 16 (2): 216–26. <https://doi.org/10.1110/ps.062523907>.
- Yuan, Zheng, Ju Zhao, and Zhi-Xin Wang. 2003. "Flexibility Analysis of Enzyme Active Sites by Crystallographic Temperature Factors." *Protein Engineering* 16 (2): 109–14. <https://doi.org/10.1093/proeng/gzg014>.
- Zandvoort, Ellen, Bert-Jan Baas, Wim J Quax, and Gerrit J Poelarends. 2011. "Systematic Screening for Catalytic Promiscuity in 4-Oxalocrotonate Tautomerase: Enamine Formation and Aldolase Activity." *Chembiochem: A European Journal of Chemical Biology* 12 (4): 602–9. <https://doi.org/10.1002/cbic.201000633>.
- Zhang, Fan, Yong-Heng Wang, Xiaowen Tang, and Ruibo Wu. 2018. "Catalytic Promiscuity of the Non-Native FPP Substrate in the TEAS Enzyme: Non-Negligible Flexibility of the Carbocation Intermediate." *Physical Chemistry Chemical Physics* 20 (22): 15061–73. <https://doi.org/10.1039/C8CP02262C>.

- Zhang, Tuo, Hua Zhang, Ke Chen, Shiyi Shen, Jishou Ruan, and Lukasz Kurgan. 2008. "Accurate Sequence-Based Prediction of Catalytic Residues." *Bioinformatics* 24 (20): 2329–38. <https://doi.org/10.1093/bioinformatics/btn433>.
- Zhang, Yang. 2008. "I-TASSER Server for Protein 3D Structure Prediction." *BMC Bioinformatics* 9 (1): 40. <https://doi.org/10.1186/1471-2105-9-40>.
- Zhang, Yang, and Jeffrey Skolnick. 2004. "Scoring Function for Automated Assessment of Protein Structure Template Quality." *Proteins* 57 (4): 702–10. <https://doi.org/10.1002/prot.20264>.
- Zhang, Yang, and Jeffrey Skolnick. 2005. "TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score." *Nucleic Acids Research* 33 (7): 2302–9. <https://doi.org/10.1093/nar/gki524>.
- Zhang, Yu, Jiao An, Guang-Yu Yang, Aixi Bai, Baisong Zheng, Zhiyong Lou, Geng Wu, et al. 2015. "Active Site Loop Conformation Regulates Promiscuous Activity in a Lactonase from *Geobacillus Kaustophilus* HTA426." *PLoS One* 10 (2): e0115130. <https://doi.org/10.1371/journal.pone.0115130>.
- Zhang, Zhong-Yin. 2003. "Chemical and Mechanistic Approaches to the Study of Protein Tyrosine Phosphatases." *Accounts of Chemical Research* 36 (6): 385–92. <https://doi.org/10.1021/ar020122r>.
- Zhao, Bin, Li Lei, Dmitry G Vassilyev, Xin Lin, David E Cane, Steven L Kelly, Hang Yuan, David C Lamb, and Michael R Waterman. 2009. "Crystal Structure of Albaflavenone Monooxygenase Containing a Moonlighting Terpene Synthase Active Site." *The Journal of Biological Chemistry* 284 (52): 36711–19. <https://doi.org/10.1074/jbc.M109.064683>.
- Zhou, Xianzhi, and Michael D Toney. 1999. "PH Studies on the Mechanism of the Pyridoxal Phosphate-Dependent Dialkylglycine Decarboxylase." *Biochemistry* 38 (1): 311–20. <https://doi.org/10.1021/bi981455s>.
- Zhuang, Zhihao, Feng Song, Hong Zhao, Ling Li, Jian Cao, Edward Eisenstein, Osnat Herzberg, and Debra Dunaway-Mariano. 2008. "Divergence of Function in the Hot Dog Fold Enzyme Superfamily: The Bacterial Thioesterase YciA." *Biochemistry* 47 (9): 2789–96. <https://doi.org/10.1021/bi702334h>.
- Zvelebil, M.J.J.M., and M J E Sternberg. 1988. "Analysis and Prediction of the Location of Catalytic Residues in Enzymes." *Protein Engineering, Design and Selection* 2 (2): 127–38. <https://doi.org/10.1093/protein/2.2.127>.