Deciphering Structural Dynamics of Proteins with Disease Causing Mutations

Thesis submitted to

Manipal University
Manipal, INDIA

for the Degree of

DOCTOR OF PHILOSOPHY

Malkaram Sridhar Achary
Registration Number: 040100006

Centre for DNA Fingerprinting and Diagnostics
Hyderabad, INDIA
March 2008

CDFD सी डी एफ डी
CENTRE FOR DNA FINGERPRINTING AND DIAGNOSTICS
(An Autonomous Centre of the Department of Biotechnology, Ministry of Science & Technology, Govt. of India)
डीएनए फिंगरप्रिंटिंग एवं निदान केन्द्र
(जैव प्रौद्योगिकी विभाग, विज्ञान एवं तकनीकी मंत्रालय, भारत सरकार की स्वायत्त संस्था)
Nacharam नाचारम, Hyderabad हैदराबाद - 500 076, India भारत

CERTIFICATE

This is to certify that this thesis entitled, **Deciphering Structural Dynamics of Proteins with Disease Causing Mutations**, submitted by **Mr. Malkaram Sridhar Achary** for the degree of **Doctor of Philosophy** to the **Manipal University** is based on the work carried out by him at the Centre for DNA Fingerprinting and Diagnostics, Hyderabad. This work is original and has not been submitted in part or full for any other diploma or degree of any other Institution or University.

Dr. H.A. Nagarajaram
Thesis Supervisor
Centre for DNA Fingerprinting and Diagnostics, Hyderabad.

Dr. Shekhar C. Mande
Dean, Academic Affairs
Centre for DNA Fingerprinting and Diagnostics, Hyderabad.

Tel. दूर भाष : (040) 27151344 -46-47-56, 27150727
: 27155607-08-12, 27171491, 27171503
Fax फैक्स : (91-040) 27155479 & 27155610
Website : http://www.cdfd.org.in

Centre for DNA Fingerprinting and Diagnostics
Laboratory of Computational Biology
Gandipet Campus, Hyderabad- 500076. INDIA

Declaration

The research work embodied in this thesis entitled, "Deciphering Structural Dynamics of Proteins with Disease Causing Mutations", has been carried out by me at the Centre for DNA Fingerprinting and Diagnostics, Hyderabad, under the supervision of Dr. H. A. Nagarajaram. I hereby declare that this work is original and has not been submitted in part or full for any other Degree or Diploma of any other Institution or University.

M. Sridhar Achary

Malkaram Sridhar Achary

Affectionately dedicated to

my parents

Sathyanarayana & Bharathi bai

and my wife and son

Sandhya & Shrihan

Acknowledgements

I would like to take this opportunity to thank all the people who have helped me in their various capacities during the tenure of my Ph.D. My sincere thanks to Dr. H.A. Nagarajaram, my Academic Supervisor, for his meticulous suggessions and subtle guidance, which had helped me to acquire the knowledge and aptitude for Computational Biology research. I am grateful to him for his constant support throughout the course of my PhD research. I am also thankful to Dr. J. Gowrishankar, the Director of CDFD, and Dr. Seyed E. Hasnain, the former Director of CDFD, for their encouragement and motivation throughout my stay at CDFD.

A few sentences are not sufficient to rightly describe the good time I had with my lab-mates. The scientific and funful debates with sreenu, the former student were sources of a healthy debate. I cherish all the friendly and warm conversations with Pankaj, my immediate junior. We have been mutually supportive and helpful, in many other things apart from the matters of research. The conversations with Tabrez, a helpful person, have always been interesting. During my final years of my PhD, my association with Vishal, another junior, who had a sportive outlook even in research, is memorable. Though I interacted quite less with Anupam and Rachita, the newly joined students, their names too are now part of my address book.

Apart from the PhD students, I had a chance to work with other bright people in the lab. My interaction with Anwar, a modest person with good reasoning, was memorable. I had a very friendly association with Gyan, and cherish all the fun and friendly chats with him. I should also acknowledge my friendly association with Deepak, who always had a positive answer during all our conversations. Several sharp people like Arun Siddharth, Aravind Thiagarajan, Sumeet Chadha, Venu Madhav, Suresh, had worked as Project Assistants in the lab during my PhD, and my conversations with them had been very stimulating and encouraging. I should mention that some of my initial trials with computer programming happened during my association with Naresh. We used to have good discussions at the road-side cafe. I also take the opportunity to thank my good old friend Saikat for being in touch and for extending his help whenever needed. I am thankful to the staff of LCB, Geeta, Swaminathan, Ranjit, Pavan and Prashanthi, who have also contributed indirectly to my learning experience.

Few staff and students from other labs, have also been very helpful during my research. I am highly thankful to Dr. Gayatri Ramakrishna, staff scientist and warden of the CDFD hostel, for her many informal suggessions. I am also grateful to her for her guidance in my search for postdoctoral position. My sincere thanks to Dr. Shekhar C. Mande, staff scientist and Dean of Academic affairs, for his support and advice at several times during my PhD. I have to express my thanks to Prof. Balasubhramanian, Director of L.V. Prasad Eye Institute, for indirectly being an inspiration in my research

and also for his enthusiastic support for my thesis work which is related to the disease PCG. I had the opportunity to listen to his charismatic speeches. I am also inpired by the simplistic and focussed nature of Prof. C. Ramakrishnan, when I had the opportunity to listen to his lectures at CDFD. During my PhD, I had also the chance to listen to the extempore lectures by Prof. T. Ramasarma, the Chairperson of CDFD, which taught me the importance of basics in science. I also had the chance to interact and work with Dr. Chandra Verma, Scientist at Bioinformatics Institute, Singapore, which helped me in the initial stages of my research in Molecular Dynamics. I shall be grateful to Dr. Dhananjaya Bhattacharya, staff scientist at Saha Institute of Nuclear Physics, Kolkata, from whom I got the first ideas of Molecular Dynamics technique. I am happy to have developed lasting friendship with Vaibhav, a junior who had come to my rescue The friendly and courteous association with Dhananjay Joshi, a physicist is has been educative. I am happy to have associated with Subbu, Sailu, Swetha, Prabhat, Aravind, Pavani, Anoop and many other students at CDFD, who My undergraduate junior, K. Srinivasa Rao, who worked as a horticulturist, had been very encouraging. I am also thankful to other seniors and staff at CDFD, Hari Narayan, Abhijeet, Dr. Mahalingam, Dr. J Gowrishankar and Dr. Prabhakar whose speeches and conversations have indirectly inpired me in research. I am grateful for having associated with Mr. K. Srinivas Rao, the security officer, for his help and useful suggessions at the time of completion of my PhD, especially when I was in dearth of funds.

At last, I should mention my indebtedness to my parents- Sathyanarayana and Bharathi bai, family members and especially to my wife Sandhya, who had always been helpful and encouraged me with hopeful words at difficult times. My son Shrihan and his kid-colleagues in our home- Sashank, Himanshu, Mihir, Hridya and Shriya have been the buffering agents for me, in whose presence I would get over the difficult situations that I had to face.

Contents

List of Figures

List of Tables

List of Abbreviations

Å	Angstrom
χ	Chi side-chain torsion angle
ω	Omega peptide torsion angle
ϕ	Phi peptide torsion angle
ψ	Psi peptide torsion angle
2D	Two Dimensional
3D	Three Dimensional
aa	Amino Acid
BLAST	Basic Local Alignment Search Tool
C$^{\alpha}$	C-alpha
CYP	Cytochrome p450
E2	17β-Estradiol
ED	Essential Dynamics
EL	Eigen Value
EM	Energy Minimization
EV	Eigen Vector
FIR	Functionally Important Region
GA	Genetic Algorithm
HB	Hydrogen Bond
HBL	Heme Binding Loop
HBR	Heme Binding Region
IP	Inner Product
LSF	Least Squares Fit
MD	Molecular Dynamics
MM	Molecular Mechanics
MSA	Multiple Sequence Alignment
MT	Mutant
NMR	Nuclear Magnetic Resonance
ns	nano second

Continued...

PC	Principal Component
PCG	Primary Congenital Glaucoma
PDB	Protein Data Bank
ps	pico second
r	Correlation Coefficient
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
SAC	Substrate Access Channel
SASA	Solvent Accessible Surface Area
SBR	Substrate Binding Region
SNP	Single Nucleotide Polymorphism
SO	Subspace Overlap
SST	Secondary Structure
TPF	Total Positional Fluctuation
WT	Wild type

Preface

Proteins are the bio-molecules having a wide variety of structural and catalytic roles in almost all the life processes. The information stored in the nucleotide sequence of DNA is expressed through the proteins which determine the structure and function of living systems. Spontaneous changes can happen in the DNA sequence, which can be passed to future generations without being corrected. These changes can be synonymous or non-synonymous depending on whether the translated amino acid is altered or not respectively. Further, the non-synonymous changes can be phenotypically neutral or can have marked phenotypic effects or lethality.

The versatility of the proteins to perform all the structural and functional roles comes from the specificity of their three-dimensional structures. Genetic diseases are caused by mutations that alter the structural and dynamic properties of the proteins involved in critical metabolic reactions/pathways. A comparative analysis of the native and mutant proteins at the sequence level and at the structural level can reveal the deleterious effects of mutations. This thesis describes an in-depth structural analysis in atomistic detail, of the nature of effects of disease-causing mutations on the protein structure and dynamics, using the example of mutations in Human CYP1b1 protein. Molecular modeling and molecular dynamics simulations were employed for comparative analysis, focusing on the structural and dynamic properties of the Wild type and some Mutant forms CYP1b1 protein implicated in the disease Primary Congenital Glaucoma (PCG).

PCG is a severe eye disorder occurring at birth or early childhood leading to loss

of vision. The disease which is mostly found to be autosomally recessive in inheritance, is characterized by a developmental abnormality of the trabecular meshwork, in the anterior chamber angle of the eye leading to increased intra-ocular pressure, resulting in optic-nerve damage and permanent loss of vision. Genetic linkage studies have indicated that PCG is, genetically, a heterogeneous disease, mapping on to at least three different loci. However, in majority (50%) of the PCG instances, the candidate locus has been found to be GLC3A on chromosome 2 that codes for a cytochrome p450 protein called CYP1b1. Sequence analyses have so far revealed several mutations in the CYP1b1 gene, of which some are found only in the PCG affected individuals. The prevalence of the disease varies geographically, and despite lower CYP1b1 mutation frequency, the frequency of different types of mutations was found to be highest in the Indian populations. To-date the mechanism or the pathway by which CYP1b1 functions in the development of trabecular meshwork is not known. The exact effects of mutations on CYP1b1 protein leading to PCG are not known, and it has been proposed that the mutations in CYP1b1 gene produce a defective enzyme, affecting its yet unknown functionality. Thus, it is important to know how certain mutations would lead to a non-functional protein.

Chapter 1 forms the Introductory review of all the relevant literature pertaining to the work presented in the thesis. Hence it begins with an overall introduction to protein structure, the methods for structure determination followed by computational approaches to protein structure prediction. This is followed by a background on the techniques used in this work, i.e., Sequence Analysis, Homology modeling, Energy minimization, Molecular Dynamics, Essential Dynamics analysis and Molecular Docking. It also includes a brief review on cytochrome p450 proteins in general and specifically about the Human CYP1b1 protein and the disease Primary Congenital Glaucoma. Finally the chapter introduces the objectives of current work.

Any protein structural analysis investigation, starts with the availability of 3 dimensional structures. The human CYP1b1 is a 543 amino acids long protein made up of three regions: the 53 residue long membrane bound N-terminal region, a 10 residue long proline-rich region called the hinge and the 480 residue long cytosolic globular domain. The structure of CYP1b1 has not been solved experimentally, and thus modeling of the protein by comparative modeling became a prerequisite. Comparative modeling is now a standard technique to get a good approximation of the structure of the protein. Moreover, a detailed structural analysis would require a sufficiently accurate model. The comparative modeling procedure to generate models for CYP1b1 proteins was a rigorous one, which also included a rigorous evaluation of the model. Chapter 2 details the procedures and techniques employed to get the models of CYP1b1 and the mutant structures corresponding to the 8 PCG mutations; A115P, M132R, Q144P, P193L, E229K, S239R, R368H and G466D.

CYPs, in general, are oxido-reductase enzymes catalyzing the oxidation of substrates (mainly toxic xenobiotics or endobiotics) and detoxifying them. The rate of enzyme activity may be limited by few checkpoints, like the accessibility of substrate, the exit of the product, substrate-binding, heme-binding, the interaction with the reductase, charge transfer, etc. The mutations may affect any of these processes, thus leading to loss of function. Chapter 3 describes the details of the molecular dynamics simulations of CYP1b1 and its mutants, which were done with the intention of comparative study of the time evolution as well as time averaged values of structural properties, especially of the functionally important regions. The mutant structures show differential structural properties during the simulation period compared to the wild type. Very importantly structural characteristics of functionally important regions are quite different in the mutant structures compared to the wild type and these altered structural properties may not to be conducive for the enzymatic function.

Essential motions in proteins are large, collective and anharmonic motions that are functionally important, in contrast to the near-constraint Gaussian motions that are harmonic in nature. Essential dynamics analysis or Principal component analysis identifies these motions which are also known as Principal modes. Chapter 4 describes the details of the procedures and inferences of the Essential Dynamics analysis of CYP1b1 and its mutant forms. Molecular dynamics simulations for 50 nanoseconds were used to study the Essential motions in Wild type and Mutant proteins. Thus, any negative effect in the essential motions in the Mutants are a direct indication of the deleteriousness to function. The study here is again a comprehensive comparative analysis of the essential motions in the Wild type and Mutants, to get insights into the mechanism of loss of function in Mutants. This study revealed qualitative and quantitative differences in collective properties between Wild type and Mutants, many of which are associated with the functionally important regions of the protein.

Chapter 5 describes the details of docking studies on the wild type and Mutant proteins. Estradiol was used as the ligand in these studies as it is known to bind to all known CYPs. The purpose of docking studies was to investigate into the binding properties of the WT vis-a-vis Mutants, and the necessity for this stemmed from the fact that the SBR showed differential structural properties in Mutants as compared to the Wildtype. The receptor structures used for docking were obtained from the MD simulations after trajectory stabilization. Finally, Chapter 6 provides concluding remarks of the results obtained in the investigations with pertinent discussion referring relevant literature.

Apart from the aforementioned main investigations, the author was also involved in molecular modeling studies and the details of these are given in the appendices. Appendix A describes a molecular dynamics study of the DNA-ligand(PBD) complexes. Appendix B describes a pilot work on modeling the interactions of N protein with the

elongation complex of Ecoli RNA polymerase. Details of the mapping of some anti-termination resistant mutations were presented.

Most of the work presented in the thesis including that given in appendices has been published or communicated. The list of Publications is provided on Page xxiii. The tables, figures and equations referred throughout this thesis are numbered chapter-wise. Modeling was done using MODELLER (Sali & Blundell, 1993) package. The simulations and trajectory analysis was performed using GROMACS (Van Der Spoel *et al.*, 2005) program package. Docking calculations were done using GOLD (Morris *et al.*, 1999) software. XMGRACE (Turner, 2005) was used to plot most of the graphs. The references cited in various chapters have been collectively given at the end and are arranged in the alphebetical order. The typesetting of this thesis including the table and figures was done using LaTeX 2_ε.

List of Publications

1. **Achary MS**, Reddy AB, Chakrabarti S, Panicker SG, Mandal AK, Ahmed N, Balasubramanian D, Hasnain SE, Nagarajaram HA. Disease-causing mutations in proteins: structural analysis of the CYP1B1 mutations causing primary congenital glaucoma in humans. *Biophys J* (2006) 91(12), 4329-39.

2. Ahmed Kamal, G. Ramesh, O. Srinivas, P. Ramulu, N. Laxman,Tasneem Rehana, M.Deepak, **M.S. Acharya** and H. A. Nagarajaram. Design, Synthesis, and Evaluation of Mixed Imine-Amine Pyrrolobenzodiazepine Dimers with Efficient DNA-Binding Affinity and Potent Cytotoxicity. *Bioorganic & Medicinal Chemistry* (2004) 12, 5427-5436.

3. Frontiers in Bioinformatics Research: The Biodiversity Issues. R.K.Gundu, B.Bose, S.Swamynathan, V.B.Sreenu, N.Pavan, **S.Acharya** and H.A.Nagarajaram *"Biodiversity: Status and Prospects"* Editors: Pramod Tandon, Manju Sharma and Renu Swarup, Narosa Publishing House, New Delhi. (2005).

4. EMBnet India Node (EIN) at the Centre for DNA Fingerprinting and Diagnostics: Serving the Indian sub-continent in Bioinformatics. G Ranjit Kumar, M Narendar Pavan, B. Bose, S Swaminathan, Geetha Thanu, P Prashanthi, B Phani Prasad, V B Sreenu, **M Sridhar Acharya** and H A Nagarajaram. *Bioinformatics India* (2003) 1 (2), 67-77.

5. Anoop Cheeran, Rajan Babu Suganthan, G. Swapna, Irfan Bandey, **M. Sridhar Acharya**, H. A. Nagarajaram and Ranjan Sen. Escherichia coli RNA Polymerase Mutations Located Near the Upstream Edge of an RNA:DNA Hybrid and the Beginning of the RNA-exit Channel are Defective for Transcription Antitermination by the N Protein from Lambdoid Phage H-19B. *J. Mol. Biol* (2005) 352, 28-43.

6. The NMITLI-BioSuite Team. (**M.S. Achary**, Part of the team). BioSuite: A comprehensive bioinformatics software package (A unique industryacademia collaboration). *Curr. Sci* (2007) 92 (1), 29-38.

Manuscripts Under Preparation

1. Shedding Light on the Effects of Disease Causing Mutations on the Essential Motions in Proteins. Achary, M.S. and Nagarajaram, H.A. (Manuscript under preperation).

2. Molecular Docking Studies of Estradiol Binding with Wildtype and PCG Mutant models of Human CYP1b1 enzyme. Achary, M.S. and Nagarajaram, H.A. (Manuscript under preperation).

Symposia Attended

1. "**International Conference in Bioinformatics-INCOB** (2006)", organized by Indian Institute of Technology, New Delhi, INDIA. Presented the poster entitled, "Dynamics-Function Correlation Studies on the Wild type and PCG-Associated Mutant Forms of Human CYP1b1".

2. "**Bioinfosummer** (2005)", organized by Australian National University, Canberra, AUSTRALIA. Presented the poster entitled, "Molecular dynamics simulation analysis of Human CYP1B1 mutations associated with Primary Congenital Glaucoma".

3. "**National Workshop on Bioinformatics** (2004)", organized by Osmania University, Hyderabad, INDIA. Worked as a resource person.

4. "**Supercomputing Applications** (2003)", organized by Centre for Development in Advanced Computing (CDAC), Bangalore, INDIA.

5. "**Molecular Dynamics** (2002)", organized by International Institute for Information Technology (IIIT), Hyderabad, INDIA.

1

Introductory Review

1.1 Protein Structure

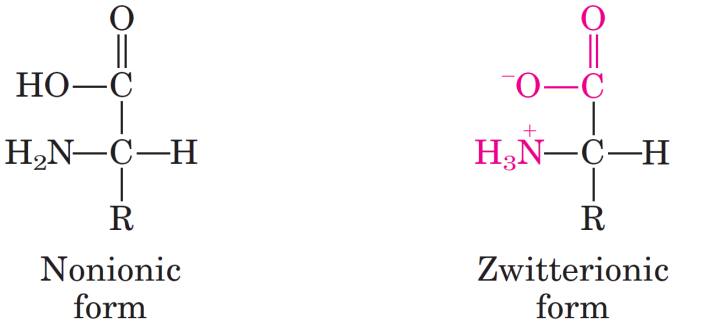
Proteins are the bio-molecules having a wide variety of structural and catalytic roles in almost all the life processes. The information stored in the nucleotide sequence of DNA is expressed through the proteins, which determine the structure and function of living systems. Proteins exhibit the most variation in structure among all the biological macro-molecules. The versatility of the proteins to perform all the structural and functional roles comes from the specificity of their three-dimensional structures. The structure of proteins can be studied at four hierarchical levels of organization.

1.1.1 Primary Structure

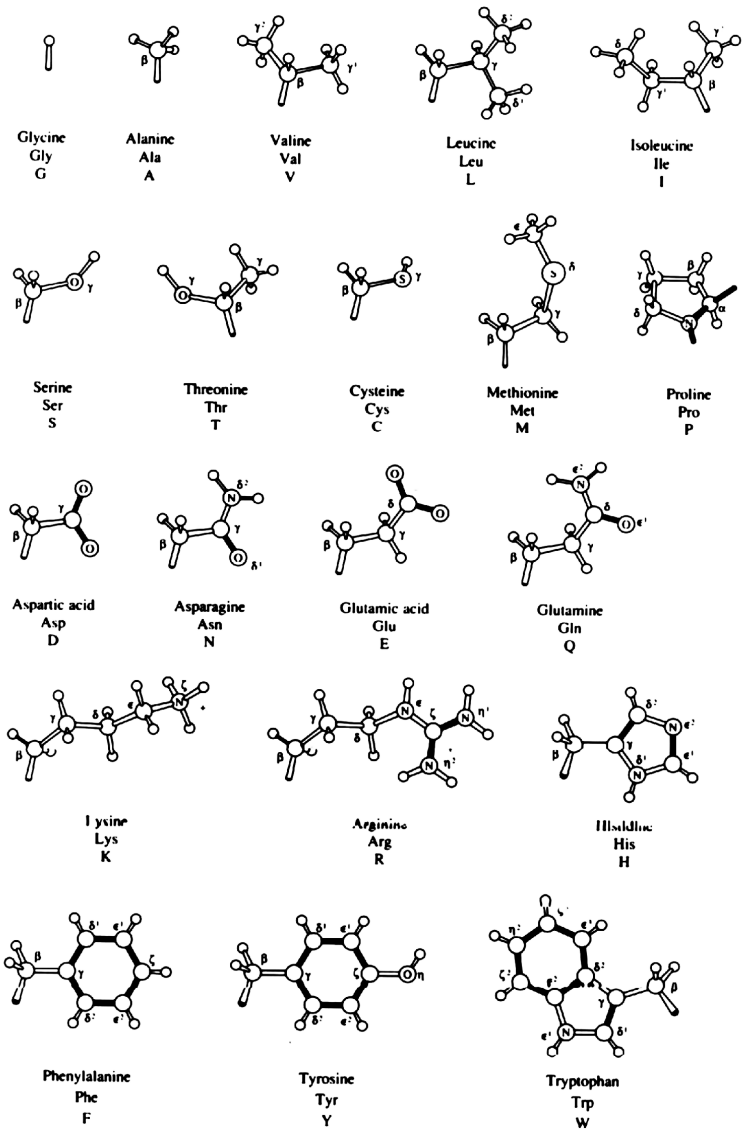
Primary structure is the term used for the linear sequence of amino acids (AA) that make up the protein. The primary structure determines the three-dimensional structure of the protein (Anfinsen, 1973).

There are 20 naturally occurring AAs in proteins. They are usually represented in one letter or three letter codes. The general structure of AA is shown in Figure 1.1(A). Schematic representations of the side chain structures of the 20 amino acids are also shown (Figure 1.1(B)). All the naturally occurring AAs (except Glycine) have an asymmetric α-carbon atom in 'L' geometrical configuration, to which an amino group, a carboxylic group and an 'R' group representing the side chain are attached. The α-carbon of Glycine has 'H' as the side chain and therefore is not asymmetric. The AAs mainly differ by the type of side chain present, which determines their specific properties. Thus, by the virtue of the type of side chain, the AAs exhibit different physico-chemical properties, which are shown in Figure 1.2 and Table 1.1. The adjacent AAs are linked together by peptide bonds as shown in Figure 1.3. The four groups attached to the α-carbon are in tetrahedral geometry.

The polypeptide chain formed by the peptide bonds is not a rigid structure and has certain limited conformational flexibility due to rotations about the single bonds N-C^{α} and C^{α}-C as shown in Figure 1.4. The four substituent groups are in a single plane, either in *cis* ($\omega = \tilde{0}^{o}$) or *trans* ($\omega = \tilde{1}80^{o}$) configuration (Figure 1.5). The *cis* configuration is rare in proteins, found in only about 5-10% of cases in prolines (Ramachandran & Mitra, 1976). The geometrical details of the *trans* peptide unit are shown in Figure 1.6. The twists and turns in polypeptide chain that give rise to its globular or compact folded structure can be represented in the form of a set of ϕ, ψ angles at every amino acid residue. The ϕ, ψ values that are 'allowed' for peptides and proteins are very well summarised by the famous Ramachandran Map (Figure 1.7) (Ramachandran & Ramakrishnan, 1963). In addition to the backbone torsion angles, the side chain conformations too can be described by means of side chain torsion angles, which have optimal and sub-optimal values. The torsion angles of the side chain are named $\chi 1$, $\chi 2$... etc, as shown in Figure 1.8. The steric hindrance



A



B

Figure 1.1: A. The general structure of Amino acid. The amino acid exists as zwitterions at physiological pH. B. The structures of the naturally occurring amino-acids. Figure obtained from (Creighton, 1992)

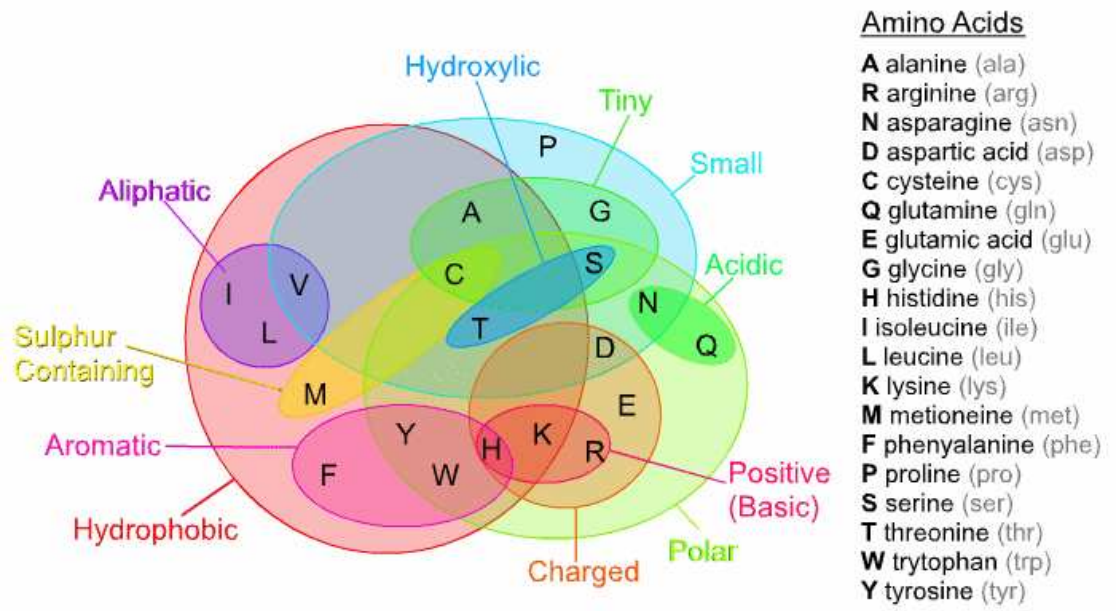


Figure 1.2: The 20 Naturally occurring AAs, shown in a Venn diagram, indicating their physico-chemical properties. The 1 letter and 3 letter codes of the AAs are given on the right side.

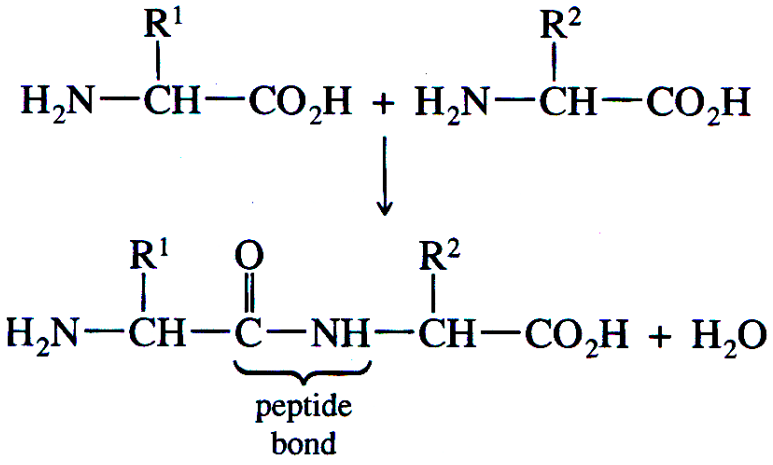


Figure 1.3: The formation of peptide bond (Creighton, 1992)

Table 1.1: Physico-chemical properties of AAs

Name	3-letter code	1-letter code	Side Chain	Occurence %	[a]Mass -H2O	[b]surface	[c]volume	[d]pKa	[e]pI	[e]density	[e]solubility	[f]Residue non-polar surface area^2	[g]Estimated hydrophobic effect for residue burial (kcal/mol)	[g]Estimated hydrophobic effect for side chain burial (kcal/mol)	UV Abs Log_e1 max pH 7.0
Alanine	ALA	A	CH3-	7.49	71.079	115	88.6	-	6.107	1.401	16.65	86	2.15	1	-
Arginine	ARG	R	HN=C(NH2)-NH-(CH2)3-	5.22	156.188	225	173.4	ĩ2	10.76	1.1	15	89	2.23	1.1	-
Asparagine	ASP	D	H2N-CO-CH2-	4.53	114.104	150	111.1	4.5	2.98	1.66	0.778	42	1.05	-0.1	-
Aspartic acid	ASN	N	HOOC-CH2-	5.22	115.089	160	114.1	-	-	1.54	3.53	45	1.13	-0.1	-
Cysteine	CYS	C	HS-CH2-	1.82	103.145	135	108.5	9.1-9.5	5.02	-	>>	48	1.2	0	2.45(250nm)
Glutamine	GLU	E	H2N-CO-(CH2)2-	4.11	128.131	190	138.4	4.6	3.08	1.46	0.864	66	1.65	0.5	-
Glutamic acid	GLN	Q	HOOC-(CH2)2-	6.26	129.116	180	143.8	-	-	-	2.5	69	1.73	0.5	-
Glycine	GLY	G	H-	7.1	57.052	75	60.1	-	6.064	1.607	24.99	47	1.18	0	-
Histidine	HIS	H	N=CH-NH-CH=C-CH2- ‖——————‖	2.23	137.141	195	153.2	6.2	7.64	-	4.19	43+86	2.45	1.3	3.77(211nm)
Isoleucine	ILE	I	CH3-CH2-CH(CH3)-	5.45	113.16	175	166.7	-	6.038	-	4.117	155	3.88	2.7	-
Leucine	LEU	L	(CH3)2-CH-CH2-	9.06	113.16	170	166.7	-	6.036	1.191	2.426	164	4.1	2.9	-
Lysine	LYS	K	H2N-(CH2)4-	5.82	128.17	200	168.6	10.4	9.47	-	>>	122	3.05	1.9	-
Methionine	MET	M	CH3-S-(CH2)2-	2.27	131.199	185	162.9	-	5.74	1.34	3.381	137	3.43	2.3	-
Phenylalanine	PHE	F	Phenyl-CH2-	3.91	147.177	210	189.9	-	5.91	-	2.965	39+155	3.46	2.3	3.97(206nm) 2.30(257nm)
Proline	PRO	P	-N-(CH2)3-CH- ‖————‖	5.12	97.117	145	112.7	-	6.3	-	162.3	124	3.1	1.9	-
Serine	SER	S	HO-CH2-	7.34	87.078	115	89	-	5.68	1.537	5.023	56	1.4	0.2	-
Threonine	THR	T	CH3-CH(OH)-	5.96	101.105	140	116.1	-	-	-	>>	90	2.25	1.1	-
Tryptophan	TRP	W	Phenyl-NH-CH=C-CH2- ‖————————‖	1.32	186.213	255	227.8	-	5.88	-	1.136	37+199	4.11	2.9	4.67(219nm) 3.75(280nm)
Tyrosine	TYR	Y	4-OH-Phenyl-CH2-	3.25	163.176	230	193.6	9.7	5.63	1.456	0.0453	38+116	2.81	1.6	3.90(222nm) 3.15(274nm)
Valine	VAL	V	CH3-CH(CH2)-	6.48	99.133	155	140	-	6.002	1.23	8.85	135	3.38	2.2	-

[a]mass [dalton], surface [Å^{22}], volume [Å^{33}], pK_a [side chain], pI [at 25°C], solubility [g/100g, 25°C], density [crystal density, g/ml], name: information from NIST Chemistry WebBook, three letter code: GIF, one letter code: VRML

[b]C.Chothia, J. Mol. Biol., 105(1975)1-14

[c]A.A. Zamyatin, Prog. Biophys. Mol. Biol., 24(1972)107-123

[d]C. Tanford, Adv. Prot. Chem., 17(1962)69-165

[e]The Merck Index, Merck & Co. Inc., Nahway, N.J., 11(1989); CRC Handbook of Chem.& Phys., Cleveland, Ohio, 58(1977)

[f]All surfaces associated with main- and side-chain carbon atoms were included except for amide, carb- oxylate and guanidino carbons. For aromatic side chains, the aliphatic and aromatic surface areas are reported seperately

[g]The values are obtained from the previous column by substracting the value for Gly (1.18 kcal/mol) from each residue

[†]Hydrophobic scales P.A.Karplus, Protein Science 6(1997)1302-1307

[†]sidechains from http://swift.cmbi.kun.nl/swift/future/aainfo/struct.htm

[†]properties from http://www.imb-jena.de/IMAGE_AA.html

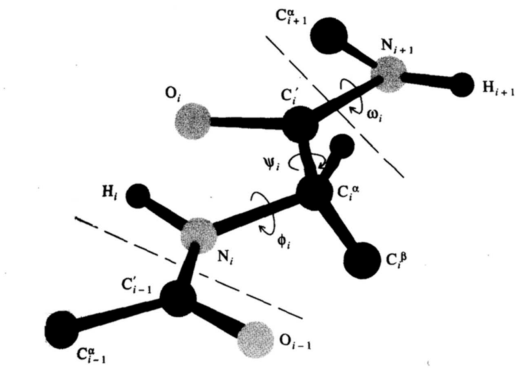


Figure 1.4: Perspective drawing of a segment of polypeptide chain comprising two peptide units. Only the C^{β} atom of each side chain is shown. The limits of a single residue (number i of the chain) are indicated by the dashed lines. The recommended notations for atoms and torsion angles are indicated. The polypeptide chain is shown in the fully extended conformation, where ϕ=ψ=ω = 180°. (Creighton, 1992)

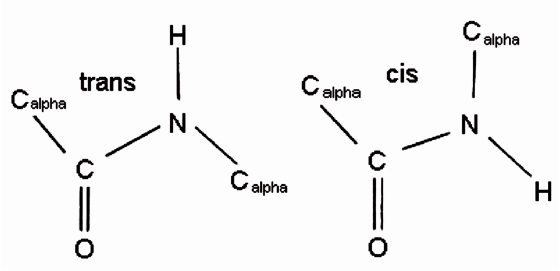


Figure 1.5: The *cis* and *trans* peptide configurations

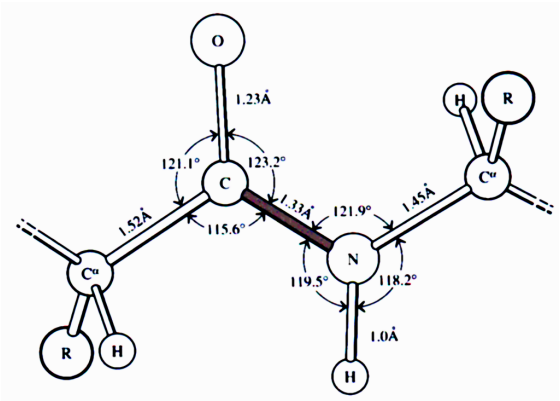


Figure 1.6: The geometry of the peptide backbone, with a trans peptide bond, showing all the atoms between two C^{α} atoms of adjacent residues. The peptide bond is stippled. The dimensions given are the averages observed crystallographically in amino acids and small peptides (Ramachandran & Kolaskar, 1974)

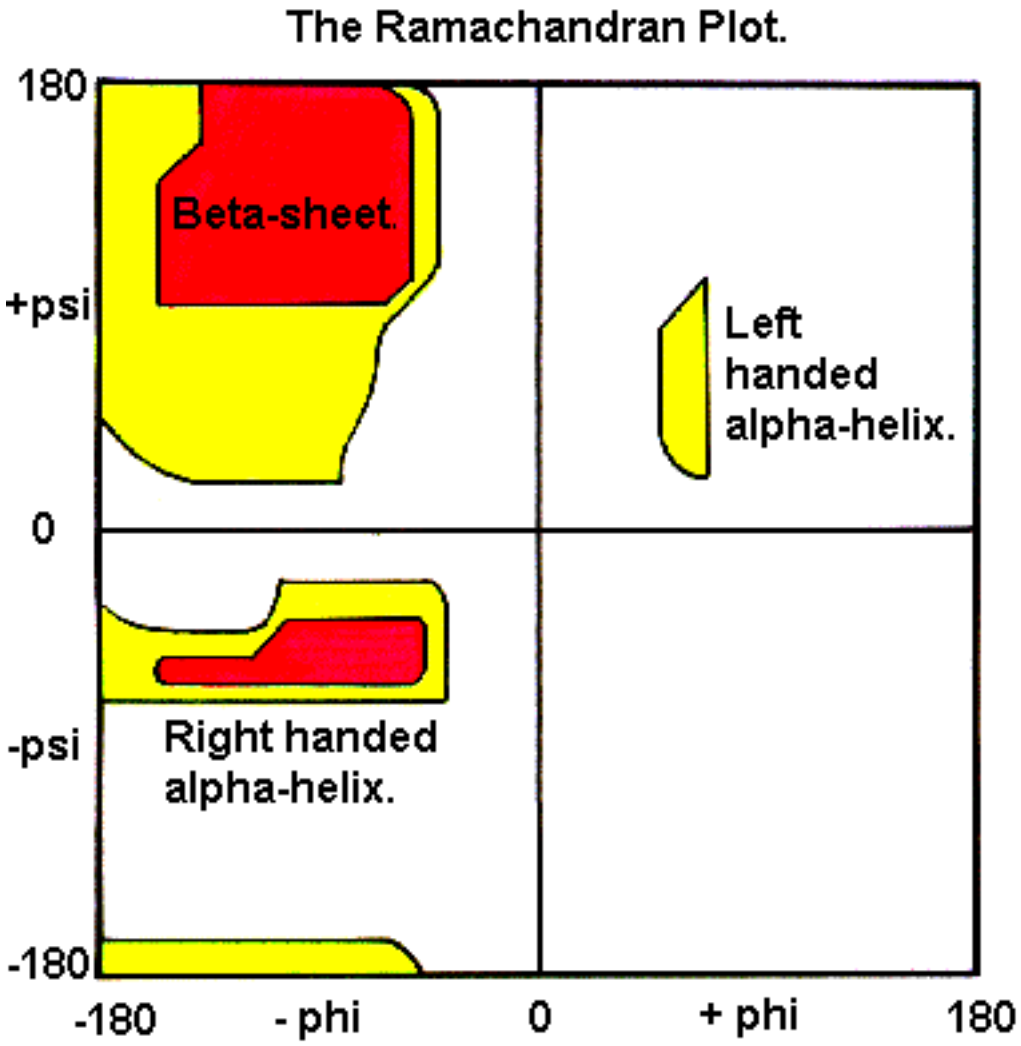


Figure 1.7: The Ramachandran Plot. The shaded regions (Red, followed by yellow) represent the most favorable ϕ,ψ combinations.

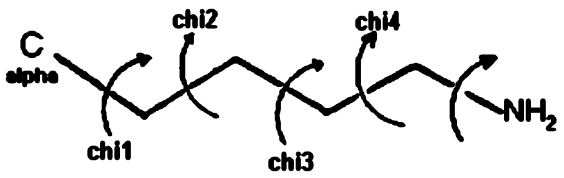


Figure 1.8: The side chain torsion angles

between the atoms in the side chain and the main chain atoms pose certain restrictions on the χ angles, giving rise to conformations called *gauche(+)*, *trans* and *gauche(-)* as shown in Figure 1.9. *Gauche(+)* and trans conformations are the most abundant conformations, in which the γ side chain atom is opposite to the main chain carbonyl group and the main chain Nitrogen groups respectively. The side chain conformation for all the 20 residues is usually described in terms of the rotameric state. The limits of possible side chain dihedral angles for every AA is given in Table 1.2. Known protein structures have been analysed to calculate the probable rotameric states for AAs (Ponder & Richards, 1987), into backbone dependent and backbone independent rotameric libraries.

1.1.2 Secondary Structure

Secondary structure, is the next hierarchical level of structural organization. All proteins have four major classes of secondary structures namely; helices, β-strands, turns and loops (often turns and loops are grouped as random coils). The primary forces for the formation of secondary structures is believed to be the intrinsic local property of the peptide segments. Other forces that drive the formation and stabilization of secondary structures are Hydrogen Bonds (HBs), Van der Waals forces and salt bridges (Dill, 1990).

1.1.2.1 Helices

Helices are the common forms of protein secondary structure in which the polypeptide twists repeatedly with ϕ and ψ values roughly -60^o, in the same direction giving rise to a helical conformation. α-helix is the most common type of helix found in proteins, which has 3.6 AAs per turn of the helix and with charecteristic Hydrogen bonding occurring

Table 1.2: Limits for rotamer library *chi* angles[†]. ((Dunbrack & Karplus, 1993))

A. Ser, Thr, Cys, Val, Phe, His, Tyr		
	χ_1 **limits**	
1	0°→120°	
2	120°→240°	
3	-120°→0°	
B. Lys, Arg, Met, Gln, Glu, Ile, Leu		
	χ_1 **limits**	χ_2 **limits**
1	0°→120°	0°→120°
2	0°→120°	120°→240°
3	0°→120°	-120°→0°
4	120°→240°	0°→120°
5	120°→240°	120°→240°
6	120°→240°	-120°→0°
7	-120°→0°	0°→120°
8	-120°→0°	120°→240°
9	-120°→0°	-120°→0°
C. Trp		
	χ_1 **limits**	χ_2 **limits**
1	0°→120°	0°→180°
3	0°→120°	-180°→0°
4	120°→240°	0°→180°
6	120°→240°	-180°→0°
7	-120°→0°	0°→180°
9	-120°→0°	-180°→0°
D. Asp, Asn		
	χ_1 **limits**	χ_2 **limits**
1	0°→120°	-90°→-30°
2	0°→120°	-30°→30°
3	0°→120°	30°→90°
4	120°→240°	-90°→-30°
5	120°→240°	-30°→30°
6	120°→240°	30°→90°
7	-120°→0°	-90°→-30°
8	-120°→0°	-30°→30°
9	-120°→0°	30°→90°
E. Pro		
	χ_1 **limits**	χ_2 **limits**
1	0°→60°	-60°→0°
3	-60°→0°	0°→60°

[†]The amino acids with flexible χ_3 and χ_4 dihedral angles (Lys, Arg, Glu, Gln), has the same limits as described for χ_1, except for χ_3 of Glu and Gln, which has the limits described for Asp and Asn χ_2.

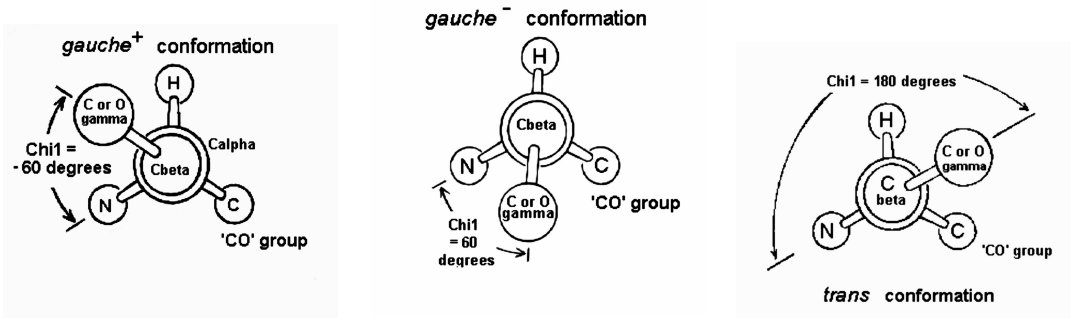
between C=O group of AA i and H-N group of AA i+4. The propensities of amino acids for helix expressed as $\Delta(\Delta G)$ values relative to Ala, the most helix favoring AA are given in Table 1.3 (Pace & Scholtz, 1998). Mostly, right-handed α-helices are observed although left handed helices also occur very rarely, as the left handed arrangement (ϕ and ψ $\tilde{+}60^o$) leads to over crowded arrangement of atoms. Other types of helices found to lesser extent in proteins are the 3_{10}-helix and the π-helix. The structural properties of the three types of helices are given in Table 1.4.

1.1.2.2 β-Strand

β-strands are the extended state conformations, in which the N-H and C=O groups are perpendicular to the direction of the strand. The AA propensities for the formation of β-sheet are given in Table 1.5 (Street & Mayo, 1999). β-strands lined up side-by-side form β sheets in which Hydrogen bonding occurs between the N-H and C=O of the adjacent strands. The adjacent β strands can have either parallel or perpendicular arrangements based on the direction of the polypeptide chain, as shown in Figure 1.10. Due the optimal alignment of HBs, anti parallel β strands are more stable than parallel β strands.

1.1.2.3 Turns

Turns are one form of the secondary structures that change the direction of the polypeptide chain. Turns are proposed to play a role in protein folding, by bringing together regular secondary structures. They are responsible for the polypeptide to fold into compact structure. Since they occur mostly in the exposed surface of the proteins, they may also represent antigenic sites or involve molecular recognition. These secondary structures often have specific AA preferences (Hutchinson & Thornton, 1994) and commonly contain proline and/or glycine residues. Turns can be classified based



gauche + gauche − trans

Figure 1.9: The side chain conformations

Table 1.3: A helix propensity scale based on experimental studies of proteins and peptides. (Pace & Scholtz, 1998)

Amino acid	**Helix propensity**[†] **(Kcal.mol^{-1})**
Ala	0.00
Glu^{0}	0.16
Leu	0.21
Met	0.24
Arg^{+}	0.21
Lys^{+}	0.26
Gln	0.39
Glu^{-}	0.40
Ile	0.41
Asp^{0}	0.43
Ser	0.50
Trp	0.49
Tyr	0.53
Phe	0.54
Val	0.61
Thr	0.66
His^{0}	0.56
His^{+}	0.66
Cys	0.68
Asn	0.65
Asp^{-}	0.69
Gly	1.00
Pro	3.16

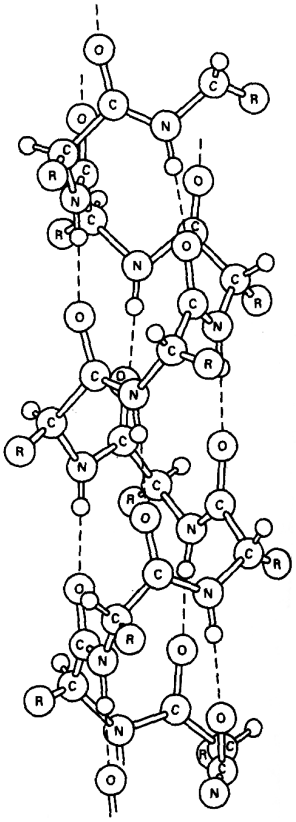
[†]The amino acids are grouped to highlight structural similarities, and they are not ranked in strict order of decreasing helix propensity

Table 1.4: The average structural properties of the three common forms of helices. ϕ and ψ are the backbone torsion angles, n is the average number of residues per turn, r is the rise in height per residue, p is the pitch of the helix. (source: http://www.cryst.bbk.ac.uk/PPS2/course/section8)

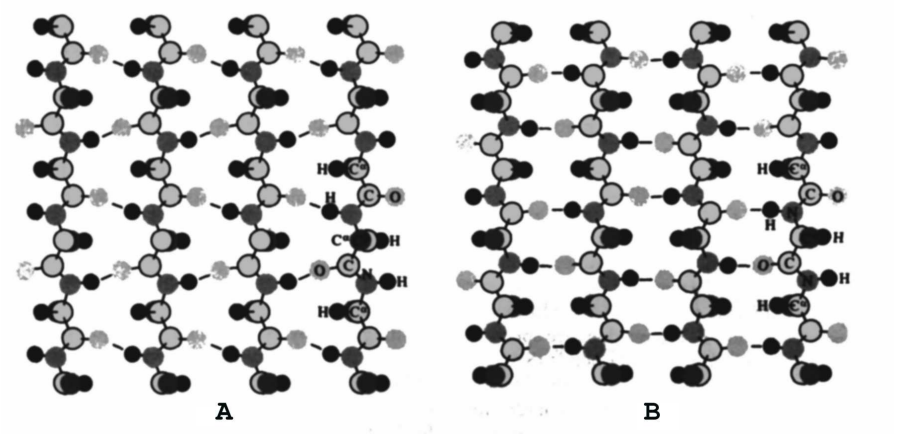
helix-type	**frequency**	ϕ	ψ	**n**	**r(Å)**	**p(Å)**	**HB (CO: HN)**	**atoms in H-bonded loop**	**radius(Å) (back-bone)**
α	abundant	-57.8	-47.0	3.6	1.5	5.5	i,i+4	13	2.3
3_{10}	infrequent	-74.0	-4.0	3.0	2.0	6.0	i,i+3	10	1.9
π	rare	-57.1	-69.7	4.4	1.1.	5.0	i,i+5	16	2.8

Table 1.5: Calculated change in entropy (ΔS) and Helmholtz free energy (ΔA) on folding into a β-sheet and the average normalized experimental propensity of the naturally occurring amino acids (Street & Mayo, 1999)

Amino acid	**ΔS, cal.mol^{-1}.K^{-1}**	**ΔA, Kcal.mol^{-1}.K^{-1}**	**Average normalized experimental propensity**
Ile	-1.59	6.58	0.10
Val	-1.69	6.88	0.13
Thr	-1.70	6.79	0.06
Phe	-1.73	7.14	0.13
Tyr	-1.74	7.15	0.11
Glu	-1.80	7.47	0.35
Gln	-1.80	7.47	0.34
Cys	-1.81	7.50	0.25
Leu	-1.82	7.56	0.32
Lys	-1.84	7.60	0.34
Ser	-1.84	7.58	0.30
Arg	-1.85	7.66	0.35
Met	-1.86	7.70	0.26
His	-1.88	7.81	0.37
Trp	-1.89	7.66	0.24
Ala	-1.99	8.30	0.47
Asp	-2.19	8.95	0.72
Asn	-2.19	8.95	0.40



α-helix



β-strand

Figure 1.10: Schematic representation of α-helix and β-sheets (A. Parallel, B. Anti-parallel)(Creighton, 1992)

on the pattern of Hydrogen bonding and their backbone dihedral angles (Richardson, 1981). Based on the number of atoms involved in a turn there are 5 classes of turns (Chou & Fasman, 1977).

The properties of common types of turns are given in Table 1.6. β-turns are the most common types of turns and were first recognized by Venkatachalam when he analyzed for conformations that are stabilized by backbone HBs and proposed 3 types of conformations based on the ϕ and ψ values (Venkatachalam & Ramachandran, 1969). Later Lewis *et al* (Lewis & Momany, 1973) has given a more general classification of the turns which classified the β turns into 10 types. Richardson (Richardson, 1981) further evaluated the β turns and reclassified them into 6 distinct types based on the ϕ, ψ angles. The major types of γ turns and β turns are shown in Figure 1.11 (Rose *et al.*, 1985).

1.1.3 Tertiary Structure

Tertiary structure is the three dimensional arrangement of all the atoms of the protein. In other words, it is the overall topology of the polypeptide (Figure 1.12). Proteins fold into specific tertiary structures, so as to take the minimum energy conformations. The tertiary structure of a protein determines its function. The tertiary structure is mainly stabilized by the hydrophobic interactions among the non-polar side chains (Anfinsen, 1972). In addition to this, stabilizing forces are contributed by HBs, ionic bonds formed between the acidic and basic AAs and covalent interactions between the side chains of cysteine residues (Figure 1.13). The tertiary structure of the protein can be divided into structural units called domains. There are several definitions proposed to describe protein domains. They are identified based on their independence in protein folding (Wetlaufer, 1973), compact structure (Richardson, 1981) and function and

Table 1.6: Properties of common types of turns

Type of Turn	Residues involved	Charecteristic Hydrogen bond
δ turn	2	NH(i) == CO(i+1)
γ turn	3	CO(i) == NH(i+2)
β turn	4	CO(i) == NH(i+3)
α turn	5	CO(i) == NH(i+4)
π turn	6	CO(i) == NH(i+5)

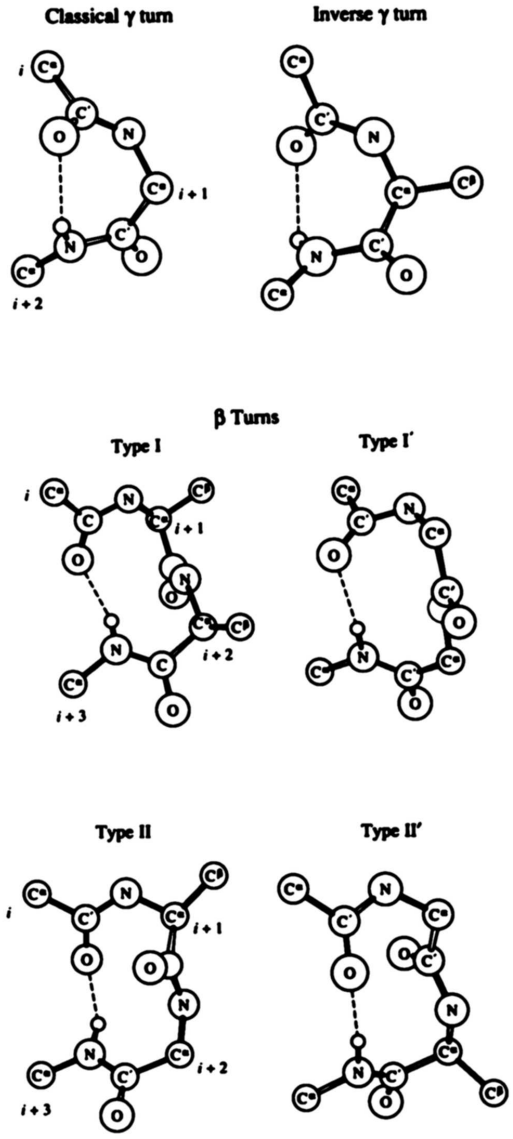


Figure 1.11: The most common γ turns and β turns connecting adjacent strands of an antiparallel β-sheet. The three and four residues, respctively, that are considered to define the turns are shown, with the first residue designated i. C^{β} atoms are shown only in positions where non-Gly residues occur frequently. The charecteristic HB is shown as dashed line (Rose *et al.*, 1985)

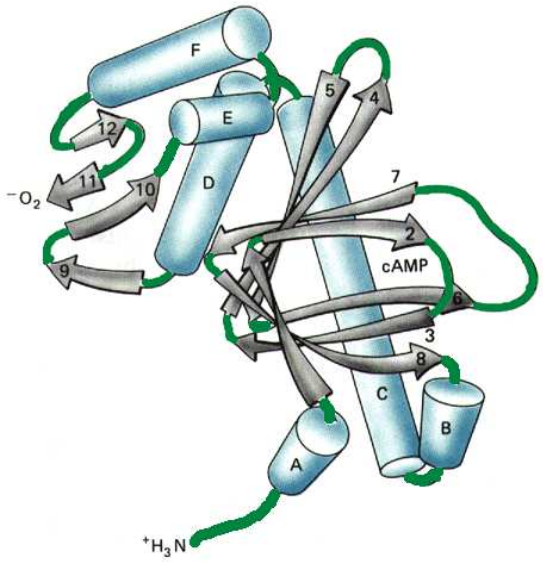


Figure 1.12: Projection diagram with ribbon representation of the Tertiary structure of a protein

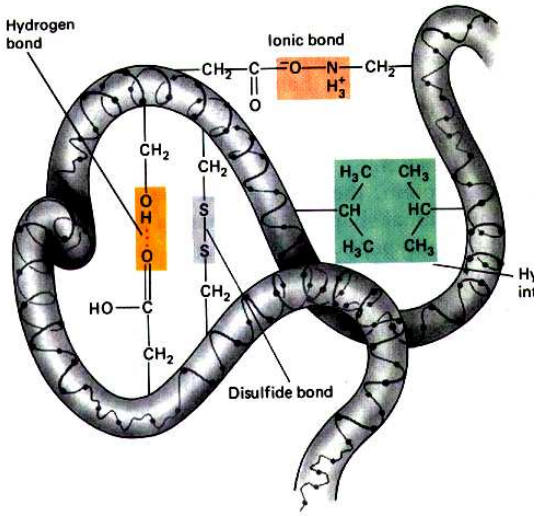


Figure 1.13: The interactions that stabilize the Tertiary structure of a protein. (Figures 1.12 and 1.13 are from http://academic.brooklyn.cuny.edu/biology/bio4fv/page/terti.htm)

evolution (Bork, 1991). These definitions, which are valid for different contexts often overlap. The domains and domain superfamilies are defined and described in the Structural Classification of Proteins database, SCOP (Murzin & Brenner, 1995). The domain superfamilies are annotated with respect to their function or usual role in a protein, in a particular pathway or in the cell/organism. The functions are categorized into into 7 general categories.

Domains can vary in size from 36 residues as in E-selectin to 692 residues as in lipoxygenase-1. Majority of the domains have less than 200 residues. Small domains less than 40 residues are stabilized by metal ions or disulphide bonds and larger domains greater than 300 residues may have multiple hydrophobic cores (Jones *et al.*, 1998). A domain can be defined as a structural unit that can exist and function independently of the rest of the protein. Thus, domain is considered as the basic unit of function rather than the whole protein. Evolutionarily related proteins can have the common domains. This is evident in the case of multi-domain proteins in which different domains with different functional properties are found in the same protein (Chothia, 1992). In Pyruvate Kinase, an enzyme which regulates the conversion of fructose-1,6-bi-phosphate to pyruvate, there are 3 domains. The all-β regulatory domain, α/β substrate binding domain and an α/β nucleotide binding domain (George & Heringa, 2002). These three domains have different functions and each of the domains occur in diverse sets of protein families. Domains also serve as modules for building up large molecular assemblies, like the virus particles or muscle fibers and also serve for regulatory functions.

Based on the secondary structural content of the domains, the protein tertiary structures are classified into four main classes. The all-α domains have a domain core built exclusively from α-helices. The all-β domains have a core comprising of anti-parallel β-sheets. The $\alpha + \beta$ domains are a mixture of all-α and all-β motifs. The

α/β domains are made from a combination of β-α-β motifs that predominantly form a parallel β-sheet surrounded by amphipathic α-helices. Each class is further classified into distinct hierarchical topological classes, as found in SCOP (Structural Classification of Proteins) (Murzin & Brenner, 1995) and CATH (Class Architecture Topology and Homologous super family) (Orengo *et al.*, 1997) databases.

1.1.4 Quaternary Structure

Quaternary structure is the next level of organization in proteins. Quaternary structures are actually formed by the association of protein subunits that can have independent existence (Jones & Thornton, 2001). The association or aggregation into quaternary structures is necessary in some cases to get the functional form of the protein. Hemoglobin is a quaternary structure of four chains, that together form the active site responsible for oxygen binding. In some cases, the catalytic activity may be regulated by cooperative interactions among protein subunits in the Quarternary structure. Quarternary structures are stabilized by the same kind of stabilizing forces that are found in tertiary structure.

1.2 Determination of Protein Structure

The availability of structure greatly improves the understanding of molecular interactions. Since the structure of a protein determines its function, the availability of proteins structure is the starting point for many functional genomics research and is the key step in the structure based drug discovery (Buchanan & Sauder, 2002).

The approximate secondary structural content of a protein can be estimated experimentally using circular dichroism (Whitmore & Wallace, 2007). A marked minimum at 208nm and 222nm indicates α-helical structure and a minimum at 204nm

or 217nm indicates a random-coil or a β-strand. The less commonly used experimental methods to determine the secondary structure are infrared spectroscopy (Manning, 2005) and Nuclear Magnetic Resonance (NMR) spectroscopy (Wuthrich *et al.*, 1991). The tertiary structure of protein can be determined experimentally by X-ray crystallography and NMR spectroscopy. The Protein Data Bank (PDB) (Sussman & Lin, 1998) stores the three-dimensional protein structural data determined by these experimental methods.

1.2.1 X-ray Crystallography

When molecules are kept in a beam of X-rays, specific diffraction patterns, characteristic of atomic arrangements in the molecules are generated (Smyth & Martin, 2000). Unlike visible light, X-rays are suited for diffraction since the wavelength of the X-rays is of the order of spacing between atoms in proteins. The diffraction pattern can be interpreted mathematically using Fourier transform to convert into electron density maps, so as to deduce the 3D structure of proteins. The first structure to be solved by X-ray crystallography was that of sperm whale myoglobin (Kendrew *et al.*, 1958) and since then structures of over 35000 proteins have been solved. Crystallization of the protein is a prerequisite for this technique, which is the major bottleneck of this method (DeLucas & Bray, 2003). Crystals are made using the conditions of supersaturated solution. Since the proteins are complex molecules their crystallization is rather difficult. Many a times, the amount of protein that is available for crystallization is a constraint. However, the 3D model that is built using the data is quite accurate, as the final deduced structure is approximately equivalent to a snapshot of the real molecule.

1.2.2 NMR Spectroscopy

NMR or Nuclear Magnetic Resonance spectroscopy is another widely used experimental method for determining the structures of proteins (Rule & Hitchens, 2006). The phenomenon of magnetic resonance results from the interaction of the magnetic moment of an atomic nucleus with an external magnetic field. In this method, the sample which is placed in a magnetic field is bombarded with radio waves. The spin of the positively charged nuclei in the atoms of the sample creates a magnetic moment. When radio waves hit the nuclei they experience a torque which gives rise to resonance. The signals emitted by the resonating nuclei is decoded using a Fourier Transform algorithm that is used to deduce the structure.

In NMR terms, chemical shift is the variation of the nuclear magnetic resonance frequencies of the same kind of nucleus due to variations in the electron distribution. It describes the dependence of nuclear magnetic energy levels on the electronic environment in a molecule. The electron distribution of the same type of nucleus varies according to the local geometry (bond lengths, angles...) and with the local magnetic field at each nucleus. Ideally each distinct nucleus in the molecule experiences a distinct chemical environment and thus has a distinct chemical shift by which it can be recognized. By understanding the different chemical environments, the chemical shift can be used to obtain structural information about the molecule in the sample (Wuthrich, 1990). The advantages of NMR spectroscopy are that apart from structure, information on the dynamics of the macromolecules is also obtained. However, NMR suffers the limitations of inherently lower sensitivity and relative complexity of the interpretation of the data and a limitation on size of proteins of $<$30kD (Wishart, 2005).

1.3 Prediction of Protein Structure

Apart from experimental methods, there are theoretical methods that use available information to predict protein structures. The structures predicted by these methods are called models. Examples of databases holding such theoretically determined models are PMDB (Protein Model DataBase) (Castrignano *et al.*, 2006), and ModBase (Pieper *et al.*, 2006). The theoretical methods use certain empirical knowledge based procedures, which gives a best approximation of structure, within the constraints of the available information. Theoretical methods exist for prediction of structure at the secondary and tertiary levels.

1.3.1 Secondary Structure Prediction

As the name suggests secondary structure prediction methods give information about the secondary structural composition, i.e, the location and lengths of α-helices, β-strands and other secondary structures.

Knowledge based prediction methods are most commonly used to assign secondary structure to protein sequences. Protein secondary structure prediction is usually the useful preliminary step in many tertiary structure prediction methods (Rost, 1997). Accurate secondary structure prediction helps to identify the fold of the protein. For example, the pattern of $\beta\alpha\beta\beta\alpha\beta$ is the signature for ferredoxin fold.

The early methods of protein secondary structure prediction were single sequence based methods, based on the helix or sheet forming propensities of individual amino-acids. The Chou-Fasman method (Chou & Fasman, 1974) used the conformational propensities for each of the AAs to be found in α-helix β-sheet, β-turns or coil. In this method, the propensity values are used to find the nucleation sites for secondary structures, where either 4 of 6 residues are α-helix formers and 3 of 5 residues are β-

sheet formers. The nucleation sites are extended on either side till the propensity for the conformations remained. Another method uses the log values of the AA propensities and computes what is called the information difference, which is the difference of the probability of the conformational state of the residue and the probability of all other conformational states (Garnier *et al.*, 1978). Both methods were found to be less than 55% accurate, primarily due to limited data availability and because they are purely focused on local features (Nishikawa, 1983).

Later, methods employing neural networks (Qian & Sejnowski, 1988; Rost & Sander, 1993), Hidden Markov models (Martin *et al.*, 2006) and Support Vector Machines (Ward *et al.*, 2003) have been developed. These methods improved the prediction accuracies up to 80% by taking into account multiple sequence alignments and information on neighbouring residues. These methods also have an additional confidence level for the predictions at every position, thus making them more useful. These methods are accurate in prediction, which take into account the specific aspects of the secondary structure. Among the recently developed prediction methods, special mention may be made on APSSP2 developed by Raghava (Raghava, 2002), which predicted all the targets of CASP5 competition with high accuracy and further it also ranked 2^{nd} and 4^{th} in the two categories of the CAFASP3 (Critical Assessment of Fully Automated Structure Prediction) competition. This method predicts the secondary structure in three steps. In the first step, the multiple sequence alignment from PSI-BLAST and neural network is used to predict the secondary structure. In the second step, it uses a modified form of example based learning technique to predict the secondary structure. Based on the reliability scores, the first two steps are combined to predict the final structure.

1.3.2 Tertiary structure prediction

The diverse biological functions of the proteins are a virtue of the specific three-dimensional structures, or the tertiary structure that they adopt. Thus, the knowledge of the tertiary structure is invaluable in the understanding of their function. Experimental methods are the best way to get detailed tertiary structure information. However, not all the times it is possible to get the tertiary structure of proteins experimentally, owing to the technical difficulties involved. In these cases, the tertiary structure prediction methods come to our rescue.

Further, with the availability of huge amounts of protein sequence information, resulting from large-scale sequencing of the genomes, the experimental determination of the protein structure for all of them would be too expensive and time consuming, thus making only the theoretical methods practically feasible (Baker & Sali, 2001). Since the protein tertiary structure is determined by its primary structure, it is theoretically possible to deduce the tertiary structure through a conformational search using suitable energy function. However, due to the enormous number of possible conformations, it is computationally too difficult and prohibitively time consuming to arrive at the native structure. Fortunately, methods exist that can predict the tertiary structure with reasonable accuracies. An overview of the most common approach taken for protein structure prediction is given in Figure 1.14. There are three important classes of tertiary structure prediction methods, as detailed below. The Homology modeling method which is used in the current investigation is described in detail.

1.3.3 Homology Modeling

Homology modeling is a method of structure prediction, which is based on the fact that sequences that are related by homology (common origin) tend to have similar structures.

It is based on the premise that, similar sequences adopt similar structures (Marti-Renom *et al.*, 2000). Structural information from a protein which is similar to the target sequence is used as a guide to build a model for the target. In the absence of experimental data, model building based on structure of a homologous protein is the only reliable method of structure prediction (Hilbert *et al.*, 1993). In the strict sense of the definition, homology requires an evolutionary relationship between the target and the template. However, in practise, the templates selected are based on sequence similarity alone. Though success through homology modeling is limited by the availability of template structure of sufficient sequence identity, it is by far the most frequently used and most accurate of structure prediction methods available to date (Forrest *et al.*, 2006; Tramontano *et al.*, 2001). The accuracy of homology modeling thus depends on the sequence similarity between the target and the template structure (Chothia & Lesk, 1986). Based on the sequence identity of the target with the template, the accuracy of modeling can be classified into 3 zones. A sequence identity of $\geq$ 40% is considered as safe zone, since this sequence identity implies structural similarity. Sequence identity between 20% to 40% is called Twilight zone as this squence identity may or may not imply structural similarity. Sequence identity less than 20% is called Midnight zone, because of considerable differences in the structures and the structure cannot be used as a template (Rost, 1997; Rost, 1999). The methodology of homology modeling can be broadly divided into four stages.

1.3.3.1 Template Search

Templates which fall in the safe zone of modeling could be easily detected in a database of sequences, using BLAST (Altschul *et al.*, 1990), or FASTA (Pearson, 1990), which are fast similarity searching programs based on certain heuristics. FASTA uses sequence patterns or words of length k or greater (k-tuples) to find exact matches in the

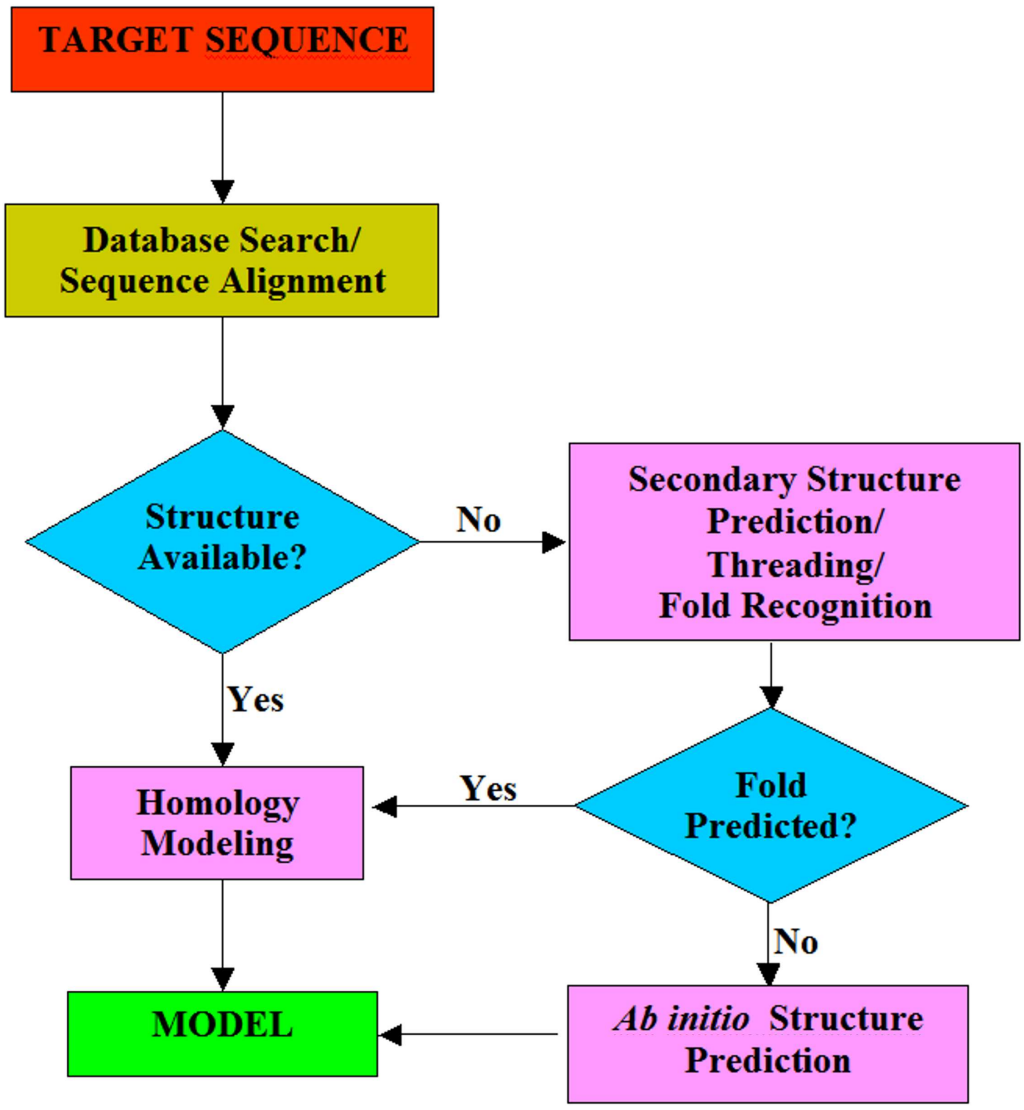


Figure 1.14: Flowchart showing the steps involved in Protein structure prediction

query sequence. The sequences that have the highest density of common words are identified. The partially aligned sequences are then re-scored using a similarity matrix, while trimming the ends that do not contribute to high score. In the high scoring sequences, the aligned segments are then joined together including penalties for gaps. Then the dynamic programming method is used in a window of 32 residues to get the optimal alignment. The disadvantage of FASTA is that it can miss significantly similar sequences, as it is based on finding exact matches. The BLAST method on the other hand uses similarity measure (obtained from substitution matrices). In the first stage, it finds matches of word length w and having a similarity score greater than T. In the next stage, it extends the matches in both directions, without consideration for insertions and deletions, such that the alignment score increases. In the third stage, the highest scoring sequences are aligned with consideration of gaps, using a variation of Smith-Waterman algorithm.

Search using BLAST or FASTA against PDB data bank can identify structures having similar sequences. If there are no similar sequences in the database, i.e. for sequences falling in the twilight zone, it is possible to find remote homologous using PSI-BLAST (Altschul *et al.*, 1997), an iterative method, in which a profile generated from a multiple sequence alignment of similar sequences identified in each run, is used to search the database in the next iteration. This improves the sensitivity of finding remotely related protein sequences. Alternatively, fold recognition software like FUGUE and 3DPSSM could be used to find proteins having similar fold. FUGUE (Shi *et al.*, 2001) uses environment based fold profiles that are created from structural alignments to make sequence structure comparison. 3DPSSM (Kelley *et al.*, 2000) is a threading method based on sequence profiles, solvation potentials and secondary structural information. Recently, a Support Vector Machine based method, using the secondary structural state and solvent accessibility state frequencies of amino acids

and amino acid pairs, could predict the fold of unknown sequences with the highest accuracy (Shamim *et al.*, 2007). Since the fold recognition methods are not solely dependent on sequence information, they are helpful to identify suitable templates, in the twilight and even in the midnight zones.

1.3.3.2 Target-template Alignment

The sequence of template thus identified, is aligned with the target sequence such that the number of aligned amino-acid pairs are maximized and the number of gaps minimized. The dynamic programming algorithm (Needleman & Wunsch, 1970) can find the optimal global alignment for the sequences. In many instances, the alignment of sequences becomes difficult due to presence of highly variable regions. Information from Multiple sequence alignment could be used in such cases. Multiple sequence alignment can be used to derive position specific scoring matrices that aid in producing correct alignments. CLUSTAL is a popular tool for sequence alignment including multiple sequence alignment (Thompson *et al.*, 1994). In the case of template identified by fold recognition, the structural alignments can be used to guide the alignment of the target. The quality of the sequence alignment is of utmost importance in modeling, as alignment errors lead to wrongly built models and cannot be rectified in the subsequent stages of the modeling. Thus, careful inspection and manual adjustment of the initial alignment can greatly improve the quality of the model.

1.3.3.3 Model Building

After having an optimal alignment, the model can be built using the methods employing template based fragment assembly, as used in COMPOSER (Sutcliffe *et al.*, 1987; Johnson *et al.*, 1994) or SWISS-MODEL (Guex & Peitsch, 1997) or satisfaction of spatial restraints, as used in MODELLER (Sali & Blundell, 1993). In the template based

fragment assembly, the structurally conserved core regions from homologous proteins with known structures are identified. An averaged backbone of all the templates is used to build a model core. The alignment is used to copy the backbone structure for the core residues of the target that align with the template. The side chains are also copied if the residues are identical. This generates the backbone for the core regions. The next step is to model the loop regions that usually correspond to the gaps in the alignment. Modeling of the loop regions requires a conformational change in the backbone due to the presence of gaps either in the target or the template. This is the difficult part of modeling. Loop modeling is done using two main methods. In the knowledge based method, a database of loop structures is searched for loops with matching residues and the loop conformation is copied into the model (Sali & Blundell, 1993). In energy based method, an energy function is used to judge the quality of the loop. Molecular dynamics (MD) or Monte Carlo techniques are applied to minimize the energy function to arrive at the best conformation (Marti-Renom *et al.*, 2000). Once all the backbone atoms are modeled, the side chains are modeled in the next step. Optimal side chain conformations are found using information from homologous structures, backbone dependent side chain rotamer libraries and energetic and packing criteria (Dunbrack & Karplus, 1994).

The second method of homology modeling is based on satisfaction of spatial restraints. In this method, for the corresponding residues in the target, spatial restraints are extracted from the template. Restraints are extracted for distances, angles, torsional angles, etc. The residues in the target structure are modeled to satisfy the spatial restraints. MODELLER, a popular homology modeling software uses this method for homology modeling (Sali & Blundell, 1993). This methodology is discussed in more detail in the section 1.3.6.

1.3.3.4 Model Quality Assessment

The quality of the model obtained in homology modeling primarily depends on the choice and quality of template structure used. If the sequence identity between the target and template is $\geq$ 90% the accuracy of the model will be comparable to that of crystallographically determined structures (Sippl, 1993). As the sequence identity decreases, the errors in the model increase. The quality (or stereochemical quality as it is often referred) of the model can be checked using certain criteria. These criteria are the normality indices that describe how well a given feature of the model resembles the same feature in the real structures. The features that are suitable for such analysis include, the main chain dihedral angle conformations, the planarity of the peptide bonds, the rotameric states of the side chains, the Hydrogen bonding in buried polar atoms, the environments for hydrophobic and hydrophilic residues and atom-atom bad contacts or clashes, etc. A useful program which evaluates most of these features is PROCHECK (Laskowski *et al.*, 1993).

Another program, VERIFY-3D (Luthy *et al.*, 1992) compares the model to its sequence, using a 3D profile. The profile is based on the statistical preferences of each of the 20 AAs for particular environments within the protein. The preferred environments for the AAs are derived from known 3D structures and are defined by three parameters, namely, the buried surface area, the fraction of side chain covered by polar atoms and the local secondary structure. The 1D profile constructed from the 3D structure based on these environment variables, can reveal problem areas in the model.

The PROSA method uses the knowledge based mean fields to judge the quality of the models (Sippl, 1993). Using the Boltzmann's principle, information is extracted on the forces which maintain the native folds in the experimentally solved 3D structures.

The extracted information is used to derive potentials of mean force. The force field for a particular protein is then obtained by a recombination of these potentials as a function of the AA sequence. This method has been shown to distinguish native folds from misfolded decoys (Hendlich *et al.*, 1990; Sippl & Weitckus, 1992). There is also a method called HORMONY by Blundell and Co-workers (Pugalenthi & Shameer, 2006), in which the structure-sequence compatibility in a model is checked by a scoring function which is based on environment based substitution tables (Overington & Johnson, 1990).

1.3.3.5 Model Refinement

The many approximations used in the process of homology modeling may result in considerable geometrical errors in the final model. These errors can be minimized using the Energy minimization (EM) or MD techniques (Kairys & Gilson, 2006). EM is used to an extent to refine the geometry of the molecule. EM could also be performed with initial restraints on the structurally conserved regions. The restraints may be removed in the subsequent steps of EM. MD, which mimics the protein folding process is also used for model refinement (Fan & Mark, 2004).

1.3.4 Threading and Fold Recognition

Threading and protein fold recognition are synonymously used in the computational methods of protein structure prediction. Threading methods could be useful in the case of absence of template information with sufficient sequence similarities. Proteins that are evolutionarily remotely related, may have very low sequence similarity, but could still have significant structural similarity. Some of the proteins with sequence similarities of $\leq$ 10 % were found to be homologous, identified through structural alignments (Brenner *et al.*, 1998; Gerstein & Levitt, 1998; Rost, 1997). Threading

methods are thus based on the concept that only a limited number of distinct protein folds exist in nature, and all the proteins should belong to any one of these folds (Chothia, 1992). The target sequence is threaded through the backbones of a library of template proteins to search for structural analogues. Then a scoring/fitness function is employed to choose the optimum fitting structure. The scoring functions employ terms that take into account the interactions between paris of AAs and also the solvation energy terms.

Threading methods vary in the algorithm used to fit the sequence to the entries in the library and the subsequent evaluation of the fit. In the simplest methods, scoring function calculated from the environment of each residue in the structure is used in combination with the dynamic programming to align the sequences (Bowie *et al.*, 1991). In advanced methods, the statistically derived pair-wise interaction potentials between residue pairs are used to evaluate the best alignments (Jones *et al.*, 1992). In other methods, the predicted secondary structure or the solvent accessibility information are used to align the target and the library sequences (Jones *et al.*, 1997). Recently, a Support Vector Machine based method has been developed for protein fold recognition (Shamim *et al.*, 2007). From a set of several sequence and structure based features used to train the SVM, a combination of the secondary structural state frequencies and solvent accessibility state frequencies of amino acids and amino acid pairs, was shown to discriminate between protein folds with $>$70% accuracy. This is the most accurate fold classification method till-date.

1.3.5 Ab initio structure prediction

The third important class of structure prediction methods is the *Ab initio* method. As the name indicates, these methods use little or no prior structural information for

predicting the structure of a given sequence. *Ab initio* methods are employed when no suitable templates could be found and when other methods of structure prediction fail (Hardin *et al.*, 2002). The protein is defined in a suitable representation with certain energy functions. Then suitable algorithms are used to minimize the energy function through a thorough search in conformational space. Since, the 3D protein models are built from scratch based on physical principles, an all atom representation of the protein in the energy functions will provide accurate results. However, evidently such a representation would make the solution too complex to be computed within reasonable time. Hence, for practical purposes, reduced representations of the protein that are amenable for computation of solutions within manageable time limits are used (Bonneau & Baker, 2001). For example in the ROSETTA method, the protein is divided into short sequence segments and their conformations are assumed to be those found in all the known protein structures (Simons *et al.*, 1997). Using *Ab initio* methods, long contiguous protein segments could be predicted to an accuracy within RMSD of 6Å (Bonneau & Baker, 2001). The reduced representation used in *Ab initio* methods often leads to over-simplification of the empirical potential, making the energy function unable to differentiate the native state from the conformations that are close to the native state, thus reducing the accuracy of the method. As such, owing to the high complexity of conformational space in proteins, these methods require high computational power and at present could be applied only for the smallest of proteins.

1.3.6 The MODELLER Package

MODELLER predicts the most probable structure for a sequence given its alignment with related structues (Sali & Blundell, 1993). It uses the approach of satisfying spatial restraints that are obtained from the alignment, to build the 3D model. The spatial restraints on the unknown sequence are obtained from the statistical analysis

of the relationships between various features of protein structure, using a database of 17 family alignments containing 80 proteins. These relationships are described as probability density functions (ρ) for the features to be predicted. The value of ρ for a feature x, is non-negative and integrates to 1 over the range of all possible values for x. The probability of an event $x_1 \leq x < x_2$ is obtained by integration of ρ (Equation 1.1).

$$\rho(x_1 \leq x < x_2) = \int_{x_1}^{x_2} \rho(x)dx \tag{1.1}$$

The calculation of the 3D model by satisfaction of spatial restraints is achieved by optimization of the molecular ρ, which is a combination of ρs restraining individual spatial features of the whole molecule. In the following subsections, the procedures used for derivation of spatial restraints and the subsequent model building by satisfaction of these spatial restraints are described.

1.3.6.1 Derivation of Spatial Restraints

The probability density functions (ρ) can be derived either analytically by using statistical and classical mechanics or empirically using the database of known structures. The ρ suitable for a certain feature x can be written as conditional ρ, which gives the probability density of x when other variables like $a, b, ...c$ are specified (Equation 1.2).

$$\rho(x/a, b, ..., c) \tag{1.2}$$

This takes into consideration the influence of other variables on x. For example the side chain dihedral angle χ_1 can be predicted using the conditional probability $\rho(\chi_1/residuetype, \phi, \psi)$, which takes into consideration, the residue type and ϕ and ψ main-chain dihedral angles. To be useful for modeling, the features in the conditional

probability density function of x should be known at the prediction stage and x has to be a spatial feature of the sequence to be modeled. In reality only approximate conditional probability functions could be obtained as in Equation 1.3, where $W_{x,a,b,...c}$ is a table spanned by $x, a, b, ..., c$ that contains as its elements the observed relative frequencies for the occurrence of x given $a, b, ..c$, and f is an analytic function fitted to the observed W.

$$\rho(x/a, b, ...c) \approx W_{x,a,b,....c} \approx f(x, a, b, ...c, q) \tag{1.3}$$

W and f must satisfy the integration and non-negativity criteria as mentioned before. The optimum number of parameters q for the function f is defined as the q that minimizes the Equation 1.4.

$$r.m.s = \sqrt{\sum_{x,a,b,...,c} [W_{x,a,b,...,c} - f(x, a, b, .., c, q)]^2} \tag{1.4}$$

A multidimensional table of relative frequencies W is calculted from the absolute frequencies W' using the Equation 1.5, where W' are obtained directly by counting the number of occurrences of each combination of $x, a, b, .., c$ values in the sample.

$$W_{x,a,b,..,c} = \frac{W'_{x,a,b,...,c}}{\sum_x W'_{x,a,b,...,c}} \tag{1.5}$$

1.3.6.2 Satisfaction of Spatial Restraints

The restraints from homologous structure, which are expressed as probability density functions(ρ) are used to derive a 3D model. The 3D model is obtained by optimization of the molecular ρ which depends on the model and on the restraints, such that the violations of the given restraints by the model are minimized. The basis ρs described

above are used to derive the feature ρs. Feature ρs are those that describe any quantity with a particular set of atoms. It combines all the information about the possible values that the feature can assume. The following example illustrates the concept of feature ρ. Consider a feature ρ for a particular C^{α}-C^{α} distance in a given sequence. If there are two related structures with equivalent distances, then two corresponding basis ρs are possible. In addition, these two basis ρs have to comply with a third basis ρ, the van der Waals criterion. A feature ρ is made combining these 3 basis ρs. Sequence alignments for the three proteins (2 templates and the sequence to be modeled) with the proteins in database are made. The feature ρ, $\rho(d/d',d'')$ can be modeled as a weighted sum of the individual ρs $\rho(d/d')$ and $\rho(d/d'')$ (Equation 1.6).

$$\rho(d/d',d'',s',s'') = \omega(s').\rho(d/d') + \omega(s'').\rho(d/d'') \tag{1.6}$$

In the final step the van der Waals restraint is included in the feature ρ (Equation 1.7).

$$\rho^{D}(d) = [\omega_1\rho_1^{d}(d) + \omega_2\rho_2^{d}(d)]\rho^{v}(d) \tag{1.7}$$

This approach of combining basis ρs is used for any number of basis ρs of the same type that were derived from related structures. Using this approach the feature ρs for C^{α}-C^{α} distances, N-O distances, and stereo-chemical restraints like, bond length, bond angle and dihedral angles, van der Waals contact, angles and distances of disulphide bonds and the side chain dihedral angles are determined. Subsequently, all the feature ρs are combined to give a molecular probability density function (P) 1.8.

$$P = \prod_i \rho^{F}(f_i) \tag{1.8}$$

P is the probability for occurrence of any combination of the features simultaneously. Thus, optimising (maximizing) the function P gives the most probable model for the 3D structure of the unknown sequence, given the alignment with the known structures. In MODELLER, the function that is actually optimized is a transformation of the P (Equation 1.9), where all the features are expressed in terms of atomic Cartesian co-ordinates.

$$F = -ln(P) \tag{1.9}$$

This function, F is referred to as the Objective function and is computationally better suited for optimization (minimization), rather than P (maximization) (Equation 1.8). The final model is obtained by optimizing the objective function in Cartesian space. Optimization of objective function is done using a variable target function procedure (Braun & Go, 1985) in Cartesian space, which employs the methods of conjugate gradients and molecular dynamics with simulated annealing.

1.4 Molecular Mechanics

Molecular mechanics deals with the application of the laws of classical mechanics to the molecular systems. It is the term given to the empirical potential energy functions that are used to describe atomic interactions. The potential energy of a molecule depends on the Cartesian coordinates of all atoms. The potential energy can be empirically calculated from the bonded and non-bonded energy terms (Equation 1.10).

$$V = [V_{bond} + V_{angle} + V_{torsion}] \; + \; [V_{vdw} + V_{ele}] \tag{1.10}$$

The bonded energy term includes energy terms associated with bond-length stretch, bond-angle bend and torsion-angle rotations in the molecule. The non-bonded energy term include the contributions from Van der Waals and electrostatic energy terms. A schematic representation of the key contributions to the molecular mechanics force field is given in Figure 1.15. Several software packages have been developed to perform Molecular mechanics calculations on bio-molecules. Among them GROMACS stands very attractive, as it is free, flexible and very fast (Van Der Spoel *et al.*, 2005).

1.4.1 GROMACS Force Field

GROMACS dosen't have its own implementation of forcefield. Its default forcefield is the GROMOS-96 and it also supports the OPLS and AMBER forcefields. The forces and energies in GROMACS are computed based on three types of interactions. In some cases, modified form of the basic representation of forcefield equations have been implemented for computational efficiency.

1.4.1.1 Bonded Interactions

Bonded interactions are between two, three or four atoms pertaining to bond stretching, angle bending and torsional rotations respectively. The bond stretching term between two covalently bonded atoms i and j is given by the basic Equation 1.11 and for the reason of computational efficiency implemented in GROMOS96 using the Equation 1.12

$$V_b(r_{ij}) = (\frac{k_{ij}^b}{2})(r_{ij} - b_{ij})^2 \tag{1.11}$$

$$V_b(r_{ij}) = (\frac{k_{ij}^b}{4})(r_{ij}^2 - b_{ij}^2)^2 \tag{1.12}$$

For systems requiring an-harmonic bond stretching potentials, the package also includes, the Morse potential, Cubic bond stretching potential and FENE (finitely extensible nonlinear elastic) potentials. The angle bending interaction is harmonic and given by the basic Equation 1.13. For computational efficiency in GROMOS96 an alternate Equation in terms of cosine of the angle is used (equation 1.14).

$$V_a(\theta_{ijk}) = (\frac{k^{\theta}_{ijk}}{2})(\theta_{ijk} - \theta^0_{ijk})^2 \tag{1.13}$$

$$V_a(\theta_{ijk}) = (\frac{k^{\theta}_{ijk}}{2})(cos(\theta_{ijk}) - cos(\theta^0_{ijk}))^2 \tag{1.14}$$

The four particle dihedral interaction term is a periodic function (Equation 1.15) of the dihedral angle (ϕ) in conjunction with a special 1-4 Lennard-Jones interaction. In the case of alkanes, a power series (using Ryckaert-Bellemans potential, with up to the 5^{th} power) of cos ϕ without 1-4 interaction is used (Equation 1.16).

$$V_d(\phi_{ijkl}) = k_{\phi}(1 + cos(n\phi - \phi_s)) \tag{1.15}$$

$$V_{rb}(\phi_{ijkl}) = \sum_{n=0}^{5} C_n(cos(\psi))^n \tag{1.16}$$

There is also a harmonic improper dihedral term to keep the planarity of the planar groups and to prevent molecules from flipping over to their mirror images (Equation 1.17).

$$V_{id}(\epsilon_{ijkl}) = (\frac{k^{ijkl}_{\epsilon}}{2})(\epsilon_{ijkl} - \epsilon^{ijkl}_0)^2 \tag{1.17}$$

1.4.1.2 Non bonded interactions

The non-bonded interactions are pair-additive and centro-symmetric and consists of a repulsion term, a dispersion term and a Coulomb term. The repulsion and dispersion term are calculated using either Lennard-Jones (6-12 potential) (Equation 1.18) or the Buckingham (exp-6 potential) (Equation 1.19).

$$V_{LJ}(r_{ij}) = 4\epsilon_{ij}((\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6}) \tag{1.18}$$

$$V_{bh}(r_{ij}) = A_{ij}exp(-B_{ij}r_{ij}) - \frac{C_{ij}}{r_{ij}^6} \tag{1.19}$$

The coulomb interaction between two charge particles is given by equation 1.20.

$$V_c(r_{ij}) = \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{\epsilon_r r_{ij}} \tag{1.20}$$

The Coulomb interaction has a fixed dielectric constant and can be modified by a reaction field which mimics the effect of a homogeneous dielectric environment beyond the cutoff radius, including the effect of ionic strength. The Long-range interactions can be chosen using single or twin range cutoff or a lattice sum evaluation of the Coulomb forces like in Ewald summation or Particle-Mesh-Ewald method (Darden, 1993).

1.4.1.3 Special interactions

In GROMACS, special interactions can be defined to impose restraints on positions of atoms, bond angles or atomic distances. The position restraints are harmonic interactions of specified atoms with fixed positions (Equation 1.21).

$$V_{pr}(r_i) = \frac{k_{pr}}{2}|r_i - R_i|^2 \tag{1.21}$$

In solvent simulations these are helpful to allow solvent to find reasonable configurations and prevents drastic rearrangements within the macro molecule. Angle restraints are used to incorporate any available experimental data on orientation (Equation 1.22).

$$\begin{aligned} V_{ar}(r_i, r_j, r_k, r_l) &= k_{ar}(1 - cos(n(\theta - \theta_0))) \\ where\ \theta &= arccos(\frac{r_j - r_i}{||r_j - r_i||}.\frac{r_l - r_k}{||r_l - r_k||}) \end{aligned} \tag{1.22}$$

Distance restraints are used to specify the experimentally determined restraints as obtained from Nuclear Overhauser Effect data in high-resolution NMR (Equation 1.23).

$$\begin{aligned} V_{dr}(r_{ij}) &= \frac{k_{dr}}{2}(r_{ij} - r_0)^2 \quad (for\ r_{ij} < r_0) \\ V_{dr}(r_{ij}) &= 0 \quad (for\ r_0 \leq r_{ij} < r_1) \\ V_{dr}(r_{ij}) &= \frac{k_{dr}}{2}(r_{ij} - r_1)^2 \quad (for\ r_1 \leq r_{ij} < r_2) \\ V_{dr}(r_{ij}) &= \frac{k_{dr}}{2}(r_2 - r_1)(2r_{ij} - r_2 - r_1) \quad (for\ r_2 \leq r_{ij}) \end{aligned} \tag{1.23}$$

1.4.2 Energy Minimization

The free energy of the system depends on the strain from the optimum conformations of atoms in molecule, as discussed in the preceding section. Molecules at their native state are at their free energy minimum. Structures predicted by the Modeling methods invariably contain errors that increase their free energy. In order to bring the structures to their native conformations, which are presumably at the global energy minimum certain EM methods have to be employed. The landscape of the potential energy function of large molecules like proteins is highly complex (Tavernelli *et al.*, 2003). Moreover, the existence of saddle points between the numerous local minima and the

global minimum poses a major problem to compute the global minima by numerical methods. In practise, using the EM methods it is possible to find the nearest local energy minimum. The EM problem can be formaly stated as follows: given a function f, which depends on one or more independent variables $x_1, x_2, ..., x_i$, the task is to find the values of these variables, where f has a minimum value. At the minimum point the first derivative of the function with respect to each of the variables is zero and the second derivatives are all positive (Equation 1.24) (Leach, 2001).

$$\frac{\delta f}{\delta x_i} = 0; \frac{\delta^f}{\delta x_i^2} > 0 \tag{1.24}$$

The common algorithms of EM are described below. The steepest descents method which is used in the current work is explained in detail.

1.4.2.1 Steepest Descents

The steepest descents algorithm proceeds in the direction of the negative gradient of potential, without considering the history built up in the previous steps. For 3N Cartesian coordinates this direction is most conveniently represented by a 3N-dimensional unit vector, s_k (Equation 1.25).

$$s_k = -\frac{g_k}{|g_k|} \tag{1.25}$$

Once the direction is determined, the move can be made towards the minimum by performing a line search or by taking arbitrary size step (Figure 1.16). In line search moves are made in the direction of the minimum such that, the minimum is bracketed within the start and end points. That is, energy of the middle point is lower than start and the end points. Thus, if three such points are determined, at least one minimum should lie between the start and end points. The process is iterated, such that the

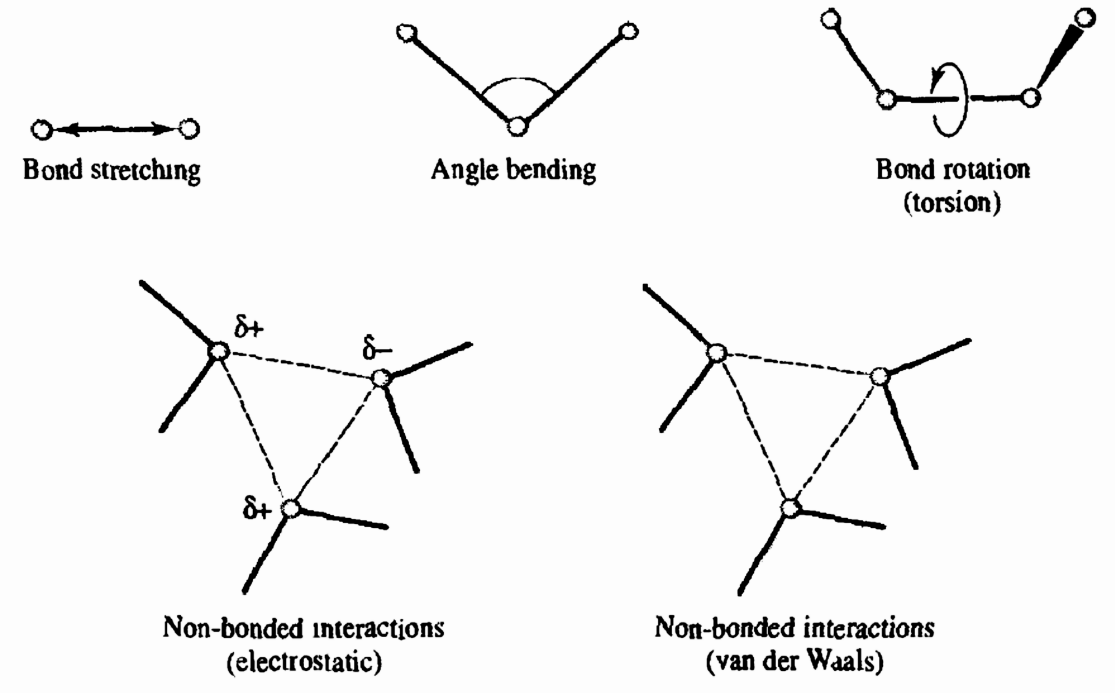


Figure 1.15: Schematic representation of the key contributions to a molecular mechanics force field (Leach, 2001)

distance between the three points is decreased gradually restricting the minimum to an even smaller region. Though this procedure is simple conceptually, it requires a considerable number of functin evaluations, making it computationally expensive. As an alternative, a quadratic function could be fit to the three points, and differntiating the function can give an approximation to the minimum along the line to be identified. Since line search is computationally demanding, the new coordinates could be obtained by taking an arbitrary length, as given by the Equation 1.26, along the gradient unit vector s_k, where λ_k is the step size (Figure 1.17).

$$x_{k+1} = x_k + \lambda_k s_k \tag{1.26}$$

A predetermined value is set for step size in most implementations of the algorithm. If an iteration leads to reduction in energy, the step size is increased by a multiplicative factor for the next iteration. The process is repeated until each iteration leads to reduction in energy. When the energy actually increases, it is assumed that the algorithm has leapt across the valley containing the minimum. In that case, the step size is reduced by a multiplicative factor. The arbitrary step size method, may require more number of steps to reach the minimum but often requires fewer functional evaluations and thus is computationally less intensive. The algorithm can be stopped after a specified number of force evaluations or when maximum of the absolute values of the force(gradient) components is smaller than a specified value. Thus, the Steepest Descent algorithm always takes the molecule downhill the energy landscape, and method is rather fast when the molecule is far away from the local minimum. However, in the vicinity of the local minimum, the convergence becomes quite slow.

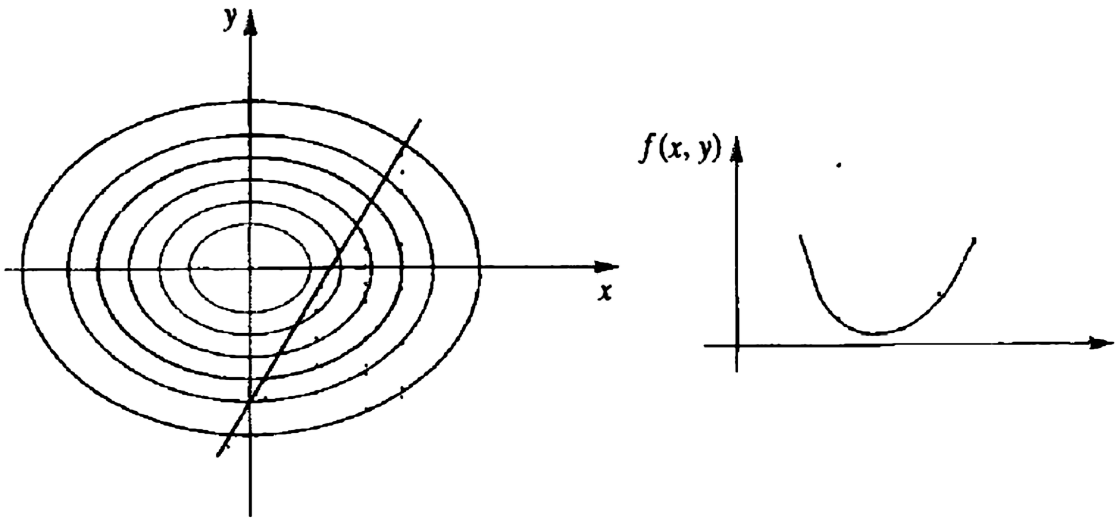


Figure 1.16: Illustration of a line search to locate the minimum in the function, in steepest descents algorithm

1.4.2.2 Conjugate Gradient

In this method, gradient information form previous steps is used to get a fast convergence. The algorithm is slower than steepest descent method but more efficient when closer to the energy minimum. The parameters used and the stop criterion are the same as for the steepest descent method. The conjugate gradient method brings the system very close to the local minimum, but in contrast to the steepest descent method, it performs badly far away from the minimum.

1.4.2.3 Newton Raphson

This belongs to the third class of EM methods which use the information of second derivative of the energy function, to locate the energy minimum. The method has a faster convergence but computationally time consuming as the computation of the second derivative and handling of its matrix becomes time consuming for large systems. Another disadvantage of the method is that the method is not globally convergent. Since the method converges rapidly in the end, which is quite opposite to that of Steepest descent method, a combination of both the methods results in a more efficient hybrid method. That is Starting with steepest descent method, one can switch over to the Newton-Raphson method when the progress by the former gets slow.

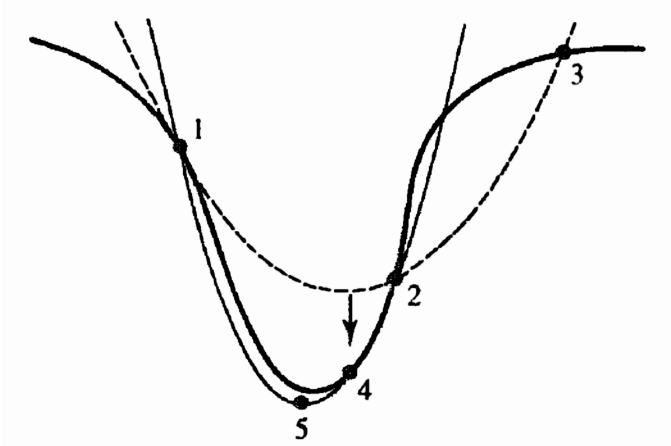


Figure 1.17: The minimum in a line search may be found more effectively by fitting a function such as a quadratic to the initial set of three points (1,2 and 3). A better estimate of the minimum can then be found by fitting a new function to the points 1,2 and 4 and finding the minimum. Figure adapted from ...

1.5 Molecular Dynamics

MD is a computational procedure that calculates the time dependent behavior of molecular system. The forces between the atoms are explicitly calculated and a suitable numerical integration method is used to compute the motions of the atoms. In MD simulation basically the Newton's Equations of motion are solved for a system of N interacting atoms to find their new positions of atoms after a time step, according to the Equation 1.27, where the forces are the negative derivatives of the potential function V (Equation 1.28).

$$m_i \frac{\delta^2 r_i}{\delta t^2} = F_i \quad i = 1, 2, ..., N \tag{1.27}$$

$$F_i = -\frac{\delta V}{\delta r_i} \tag{1.28}$$

At the new positions the forces are recalculated and new positions are found after another time step. This process is repeated indefinitely. The positions and occasionally the velocities of the atoms are stored at regular time steps, which constitutes the trajectory of the system. The most expensive part of the simulation is the calculation of the non-bonded forces, which is proportional to square of the number of atoms in the system. They are calculated using an empirical potential such as the commonly used Lennard-Jones potential (Equation 1.29).

$$V_{ij}^{LJ}(r_{ij}) = \epsilon_{ij}[(\frac{r_{ij}^{min}}{r_{ij}})^{12} - 2(\frac{r_{ij}^{min}}{r_{ij}})^{6}] \tag{1.29}$$

When the two interacting atoms are close, the force acting between atoms is repulsive and at longer ranges it is attractive and at the energy minimum the force is zero. Since the LJ potential has an infinite range, practically it is advantageous to ignore the interactions between atoms separated by distance greater than a certain cutoff,

resulting in speeding up of the calculations. However, this is also associated with disadvantages, because imposition of the cutoff gives rise to break in the continuity of the function, which causes jumps in the energy function across the cutoff. To avoid this the potential is often is shifted in order to vanish at the cutoff radius.

1.5.1 Periodic Boundary Conditions

In MD, the technique of Periodic boundary conditions is used to eliminate the surface effects. Using the periodic boundary conditions, the system of only a few hundred atoms behaves equivalent to a system of infinite size, without much increase in the cost of computation. The atoms enclosed in a periodic box are replicated to infinity by translation in all the three Cartesian directions, completely filling the space. Each particle has interaction with other particles within the box but also with its images in nearby boxes. This virtually eliminates the surface effects from the system. The complexity of calculations resulting from the use of periodic boundary conditions is reduced by allowing the particles to interact only with the closest image, using the convention called the minimum image criterion. This requires that the 'rcut' (cutoff-radius) cannot be greater than half the width of the cell.

1.5.2 Integration Algorithms

In an MD simulation, the numerical methods used to calculate the new positions and velocities for the atoms, after each time step are called integration algorithms. The algorithms basically solve the Newton's Equations of motion for the atoms. The methods are devised to cater for the accuracy, stability, speed and economy of carrying out the simulations.

The Verlet algorithm is the most commonly used in MD. Equations 1.30 and 1.31

represent the Taylor expansions for the position $r(t)$, one in the forward and one in backward in time. In the Equations, v stands for velocities, a for accelerations and b for the third derivatives of r with respect to t. Adding the two Equations give Equation 1.32, which is the basic form of Verlet algorithm.

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \frac{1}{2}a(t)\Delta t^2 + \frac{1}{6}b(t)\Delta t^3 + O(\Delta t^4) \tag{1.30}$$

$$r(t-\Delta t) = r(t) - v(t)\Delta t + \frac{1}{2}a(t)\Delta t^2 - \frac{1}{6}b(t)\Delta t^3 + O(\Delta t^4) \tag{1.31}$$

$$r(t+\Delta t) = 2r(t) - r(t-\Delta t) + a(t)\Delta t^2 + O(\Delta t^4) \tag{1.32}$$

The position after time Δt could be found using information from the current position, previous position and acceleration. According the Newtons Equations, the acceleration is given by the Equation 1.33.

$$a(t) = -(\frac{1}{m})\Delta V(r(t)) \tag{1.33}$$

It could be seen that the truncation error of the algorithm is of the order of fourth power of Δt. The Verlet algorithm is simple to implement, stable and accurate. The disadvantage is that the velocities are not explicitly generated. The velocity which is required to compute the kinetic energy could be calculated using Equation 1.34.

$$v(t) = \frac{r(t+\Delta t) - r(t-\Delta t)}{2\Delta t} \tag{1.34}$$

As seen in Equation atom positions for three consecutive time steps have to be stored, in order to calculate the velocities. The error associated with this expression is of the order Δt^2. To over come this, variants of the Verlet algorithms have been developed. In leap frog method the velocities obtained are more accurate. The velocities are calculated

using Equation 1.35 and then the new positions are calculated using Equation 1.36. The velocity at time t can be calculated using Equation 1.37.

$$v(t+\frac{\Delta t}{2}) = v(t=\frac{\Delta t}{2}) + a(t) \tag{1.35}$$

$$r(t+\Delta t) = r(t) + v(t+\frac{\Delta t}{2})\Delta t \tag{1.36}$$

$$v(t) = \frac{1}{2}v(t+\frac{\Delta t}{2}) + v(t-\frac{\Delta t}{2})) \tag{1.37}$$

In the Velocity Verlet algorithm, the new atomic positions after time Δt are calculated as per the Equations 1.38 to 1.41. The is a requirement of 9N memory locations for 3N positions, velocities and accelerations, but none of these quantities need to be simultaneously stored at two different times.

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \frac{1}{2}a(t)\Delta t^2 \tag{1.38}$$

$$v(t+\frac{\Delta t}{2}) = v(t) + \frac{1}{2}a(t)\Delta t \tag{1.39}$$

$$a(t+\Delta t) = -(\frac{1}{m})\Delta V(r(t+\Delta t)) \tag{1.40}$$

$$v(t+\Delta t) = v(t+\frac{\Delta t}{2}) + \frac{1}{2}a(t+\Delta t)\Delta t \tag{1.41}$$

1.5.3 Constraint Algorithms

In MD, constraints are applied to reset the bond lengths and angles to their correct values. Two popular constraint solving algorithms are described below. The SHAKE (Ryckaert *et al.*, 1977) algorithm uses an iterative procedure to reset the coordinates to new coordinates that fulfill a list of distance constraints. The algorithm continues until all the constraints are satisfied within a relative tolerance limit. The

algorithm stops with a warning when the bond length deviations are too high or when the maximum number of iterations are reached. The LINCS (Hess *et al.*, 1997) algorithm which is specifically used with bond constraints and isolated angle constraints, is more stable and faster than SHAKE. The algorithm resets bonds to their correct lengths after an unconstrained update. It is a non-iterative, 2-step method. In the first step, the projections of the new bonds on the old bonds are set to zero. In the second step, a correction is applied for the lengthening of the bonds due to rotation. Unlike SHAKE, the LINCS algorithm is designed to function even when it is impossible to reset the constraints. A warning message is generated when a bond rotates over more than a predefined angle.

1.5.4 Temperature and Pressure Coupling

In MD simulation the temperature and pressure of the system could deviate abnormally from the set values due to reasons like, drift during equilibration, frictional forces, integration errors etc. It is necessary to control the temperature of the system using a temperature coupling algorithm (Berendsen *et al.*, 1984). In the Berendsen method of temperature coupling, the system is weakly coupled with first order kinetics to an external heat bath at a given temperature T_0. The deviation in temperature of the system is corrected slowly such that the temperature deviation decays exponentially with a time constant τ. Similarly to correct the pressure the system is coupled to a pressure bath. The Berendsen algorithm achieves this by scaling the coordinates and box vectors at every step.

1.5.5 Analysis of MD trajectories

The integration of Equations of motion results in a trajectory which has information about the positions, velocities and accelerations of the particles throughout the simulation time. The trajectories contain information at the atomic level. The conformations of the molecule generated in the simulation is equivalent to an ensemble of molecules. Statistical mechanics is used to convert the information from the trajectory at the atomic level to the macroscopic properties like, pressure, energy, etc at the ensemble level. This is based on the Ergodic hypothesis, which states that the time average equals the ensemble average. To satisfy the Ergodic hypothesis and to get correct estimates of the average properties of the system, the MD simulation should be of considerably long duration.

1.5.6 The GROMACS Package

GROMACS is a public domain simulation software suite, which stands for GROningen MAchine for Chemical Simulation (Lindahl *et al.*, 2001; Van Der Spoel *et al.*, 2005). The software provides tools for micro-canonical Hamiltonian mechanics, stochastic dynamics including Langevin and Brownian dynamics and energy minimization. The package also includes a variety of analysis tools aiding in trajectory analysis and normal mode and principal component analysis of structural fluctuations.

In GROMACS, several adjustments can be made to obtain the longest possible time step without much loss of accuracy. The united atom instead of all-atom force field can reduce the number of interactions roughly by a factor of 4. Constraining the bond lengths allows a time step of up to 2 fs. Usually the Hydrogen atom is defined as a virtual particle by adding its mass to the heavy atom it is connected to, and when required, reconstructing the hydrogen position from the position of the three

nearby atoms. Similarly, increasing the mass of the hydrogen atoms at the expense of the mass of the neighboring heavy atoms, in the case of planar ring systems, can extend the time step up to 4 fs. The Global MD algorithm of GROMACS is shown in Figure 1.18 (Spoel *et al.*, 2006). The input for MD or EM run is an initial set of coordinates with initial velocities for all particles involved. The system is bounded by a box whose size is determined by three vectors representing the three basis vectors of the periodic box. If the initial velocities are not available, the program generates initial atomic velocities with a Maxwellian distribution at a given absolute temperature. The forces are computed using the Equations 1.11 to 1.23. The Potential energy of each interaction term is computed as a sum of Lennard-Jones, Coulomb and Bonded terms. The total Kinetic energy is given by the Equation 1.42.

$$E_{kin} = \frac{1}{2}\sum_{i=1} N m_i v_i^2 \tag{1.42}$$

From this, the Temperature T is computed using Equation 1.43, where k is Boltzmann's constant and N_{df} is the number of degrees of freedom which can be computed from $N_{df} = 3N - N_c - N_{com}$, where N_c is number of constraints imposed on the system, and N_{com} is the number of degrees of freedom for center of mass.

$$\frac{1}{2}N_{df}kT = E_{kin} \tag{1.43}$$

The Pressure is calculated from the difference between Kinetic energy E_{kin} and the Virial (Equation 1.44).

$$P = \frac{2}{V}(E_{kin} - \Xi) \tag{1.44}$$

The Equations of motion are integrated using the Leap-frog algorithm. The positions r at time t and velocities v at time $t - \frac{\Delta t}{2}$ are used to update positions and velocities

THE GLOBAL MD ALGORITHM

1. Input initial conditions

Potential interaction V as a function of atom positions

Positions r of all atoms in the system

Velocities v of all atoms in the system

$\Downarrow$

repeat 2,3,4 for the required number of steps:

2. Compute forces

The force on any atom

$$F_i = -\frac{\delta V}{\delta r_i}$$

is computed by calculating the force between non-bonded atom pairs

$$F_i = \sum_j F_{ij}$$

plus the forces due to bonded interactions (which may depend on 1,2,3 or 4 atoms), plus restraining and/or external forces.

The potential and kinetic energies and pressure tensor are computed.

$\Downarrow$

3. Update configuration

The movement of the atoms is simulated by numerically solving Newton's equation of motion

$$\frac{d^2 r_i}{dt^2} = \frac{F_i}{m_i}$$

or

$$\frac{dr_i}{dt} = v_i \; ; \; \frac{dv_i}{dt} = \frac{F_i}{m_i}$$

$\Downarrow$

4. if required: **Output step**

write positions, velocities, energies, temperature, pressure, etc.

Figure 1.18: Flowchart of the global MD algorithm of GROMACS.

using the force $F(t)$ determined by the positions at time t (Equations 1.45).

$$v(t+\frac{\Delta t}{2}) = v(t-\frac{\Delta t}{2}) + \frac{F(t)}{m}\Delta t$$
$$r(t+\Delta t) = r(t) + v(t+\frac{\Delta t}{2})\Delta t \tag{1.45}$$

The Equations of motion are modified for temperature and pressure couplings, and extended to include the conservation of constraints. The flow chart for update of velocities and coordinates is given in Figure 1.19 (Spoel *et al.*, 2006). The main output of MD run is the trajectory file, which contains particle coordinates and optional velocities at selected regular intervals. When analyzing the MD trajectories, the averages $< x >$ and fluctuations $< \sigma >$of a quantity are computed using Equations 1.46 and 1.47.

$$< x >= \frac{1}{N_x}\sum_{i=1}^{N_x} x_i \tag{1.46}$$

$$\sigma_x = \sum_{i=1} N_x[x_i - < x >]^2 \tag{1.47}$$

1.6 Essential Dynamics Analysis

Essential Dynamics analysis helps to filter the large collective motions from random motions in the MD data. It is based on the technique of Principal component analysis, which provides a way to reduce a complex data set to a lower dimension. This can reveal the hidden properties, which are otherwise difficult to perceive due to the complexity of high dimensional data (Alakent & Doruker, 2004). The ED method was first introduced by Amadei *et al* to study the biologically relevant large collective motions filtered from random constraint like high frequency motions in a MD simulation (Amadei

et al., 1993). The ED method consists of diagonalization of the covariance matrix of atomic fluctuations, to get a set of Eigen vectors and corresponding Eigen values. The covariance matrix is calculated by the Equation 1.48, where M is a diagonal matrix containing the masses of the atoms. C is a symmetric 3N X 3N matrix (Eigen vectors), which can be diagonalized with an orthonormal transformation matrix R (Eigen values) (Equation 1.49). The Eigen vectors arranged in the decreasing order of magnitude of the Eigen values, indicate the directions of collective motion. The trajectory can be projected on to principal modes to give the principal components $p_i(t)$ (Equation 1.50). ED analysis is useful in revealing low frequency motions of the molecules which are functionally important, even with short simulation data.

$$C_{ij} =< M_{ii}^{\frac{1}{2}}(x_i - < x_i >)M_{jj}^{\frac{1}{2}}(x_i - < x_i >) > \tag{1.48}$$

$$R = diag(\lambda_1, \lambda_2, ..., \lambda_{3N}) where \lambda_1 \geq \lambda_2 \geq .... \geq \lambda_{3N} \tag{1.49}$$

$$p_t = R^T M_{\frac{1}{2}}(x(t) - < x >) \tag{1.50}$$

Since ED analysis is used to obtain functionally relevant collective motions in MD simulations, it is essential that the simulations are sufficiently long and equilibrated. The reliability of the calculated principal modes could be estimated, by studying the similarity of Eigen vector sets calculated for the two halves of the trajectory. This can be done by calculating the subspace overlap of the top ranking Eigen vectors, between the two Eigen vector sets (de Groot *et al.*, 1996b). Equation 1.51 gives the overlap between the two Eigen vector sets, where $v_1, ...v_n$ are the m orthonormal vectors of the first half of simulation, $w_1, ...w_m$ are the m orthonormal vectors of the second half of simulation.

$$overlap(v, w) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} (v_i . w_j)^2 \tag{1.51}$$

THE UPDATE ALGORITHM

Given:
Positions r of all atoms at time t
Velocities v of all atoms at time $t - \frac{\Delta t}{2}$
Accelerations $\frac{F}{m}$ on all atoms at time t.
(Forces are computed disregarding any constraints)
Total kinetic energy and virial

$\Downarrow$

1. Compute the scaling factors for temperature and pressure

$\Downarrow$

2. Update and scale velocities: $v' = \Lambda(v + a\Delta t)$

$\Downarrow$

3. Compute new unconstrained coordinates: $r' = r + v'\Delta t$

$\Downarrow$

4. Apply constraint algorithm to coordinates: constrain$(r' \rightarrow r''; r)$

$\Downarrow$

5. Correct velocities for constraints: $v = (r'' - r)/\Delta t$

$\Downarrow$

6. Scale coordinates and box: $r = \mu r''; \; b = \mu b$

Figure 1.19: Flowchart of the MD update algorithm of GROMACS

The cosine content is another useful check for the reliability of the simulations. It has been shown that the principal components of random diffusion are cosines with the number of periods equal to half the principal component index (Hess, 2000; Hess, 2002). The cosine content is calculated by the Equation 1.52.

$$\frac{2}{T}\left(\int_0^T cos(i\pi t)p_i(t)dt\right)^2 \left(\int_0^T p_i^2(t)dt\right)^{-1} \tag{1.52}$$

High cosine content indicate that the largest fluctuations are not connected with the potential, but with random diffusion. ED analysis of MD trajectories has been used in the past to reveal functional motions. The hinge-bending motion of thermolysisn was revealed by ED analysis, which is responsible for the opening and closing of the active site (van Aalten & Amadei, 1995). ED analysis revealed the possible mechanism of ligand binding to beta-secretase protein. A large conformational change resulting from the swing motion of the flap structure structure facilitates the substrate binding (Xiong & Huang, 2004). The exit site of the retinol in the retinol binding protein could be revealed by strong correlations of the motions of three surface loops of the protein with the retinol motions (van Aalten & Findlay, 1995).

1.7 Molecular Docking

Molecular Docking is the process of computational identification of the optimum binding orientation of a substrate or ligand at the binding site of a receptor. It predicts the preferred orientation of one molecule to a second when bound to each other, to form a stable complex (Lengauer & Rarey, 1996). Docking is useful to in the field of signal transduction, to predict the strength and type of signal produced. Signal transduction is an important biological process effected by the association between the bio-molecules; proteins, nucleic acids, carbohydrates and lipids. The relative orientation

of the interacting molecules determines both, the type of signal and the strength of signal produced. It is also widely used in the field of structure based drug design, wherein drugs of therapeutic value against specific receptors (usually proteins) are discovered (Kitchen *et al.*, 2004).

The receptor and ligand structures in the complex are analogous to a lock-and-key model. However, since both the protein and ligand molecules are flexible, and not rigid like the lock and key, the hand-in-glove is a more appropriate analogy for the ligand-receptor complex (Jorgensen, 1991), in which conformational adjustments are induced during the binding process. This process of receptor ligand binding is referred to as induced-fit mechanism (Wei *et al.*, 2004).

1.7.1 Docking Algorithms

The ligand and the receptors in their optimum binding conformations, have geometric and chemical complementarity which is energetically favorable. The main purpose of Molecular Docking algorithms is to identify the energetically favorable binding orientations. Based on the methodology of identifying optimum receptor-ligand orientations, molecular docking algorithms can be classified broadly into two categories. Those which rely on the protein and ligand shape complementarity and those which simulate the process of complex formation, using the ligand-protein pairwise interaction energies.

In the shape complementarity methods, a set of features that describe the ligand and protein are used. These features include descriptors like the solvent-accessible surface area, hydrophobicity and Fourier descriptors. Shape complementarity methods have the advantage that they are fast and robust but pose limitations when the dynamic changes occurring due to the flexibility of the receptor and the ligand are taken into

account (Shoichet *et al.*, 2004). In the methods of second class, a simulation approach is used in which ligand finds its position into the receptor's binding site after a certain number of moves in its conformational space. The moves incorporate rigid body transformations such as translations and rotations as well as changes in the ligand including torsion angle rotations. The binding free energy of the complex is calculated at every step. The advantage of these methods is that the flexibility of ligand is accounted in these methods and the process is physically closer to what happens in reality. The main disadvantage of these methods is the time required to find the optimum binding orientation, since a large energy landscape of the complex needs to be explored (Fieg *et al.*, 2004). The success of docking depends on the conformational search algorithm and the scoring function.

1.7.2 Conformational Search

The basic requirement in docking calculation is the structure of the receptor and the ligand. The receptor structure is usually determined experimentally (X-ray or NMR). The structure of ligands which are usually small molecules can be obtained from a database of ligand structures, or can be built using molecular modeling software. The entire conformational search space in any typical docking experiment is too large to completely sample using current computational resources. To address this problem, some effective strategies are employed to sample the conformational space, which use MD, shape complementarity methods and genetic algorithms.

In MD simulation methods, the ligand is normally allowed to freely explore within a fixed receptor structure. It is followed by simulated annealing and EM steps of some of the generated conformations. The conformations thus generated are ranked using the energies determined in the MD run. The MD method is advantageous

in that it needs no special energy function and MD force fields could be used for the purpose (Goodsell & OLson, 1990). Shape complementarity methods are the most common methods, which look for match between the receptor and the ligand to arrive at the optimal orientation. Several popular docking programs, DOCK, FRED, GLIDE, SLIDE etc are based on this approach. In shape complementarity methods, the molecules are assigned with some descriptors like solvent accessible area, shape and geometry constraints, Hydrogen bonding interactions, hydrophobic interactions, etc (Gabb *et al.*, 1997). Compared to MD simulation methods, these methods are more efficient in finding the optimum binding orientations. The Genetic algorithms encode each spatial conformation of protein and ligand as genes with a particular energy. The complete genome can be considered equivalent to the entire energy landscape that has to be explored. The process of biological evolution is mimicked by cross-over of fragments of the genome, including occasional incorporation of mutations in the offspring (Morris *et al.*, 1999). Although Genetic algorithms are useful for a more efficient sampling of a large conformational space, they may not be as efficient as the shape complementarity based methods in screening large databases of compounds. Further, they need to be run multiple times in order to obtain reliable results.

1.7.3 Scoring Functions

In any docking procedure, the orientations of the ligand and receptors need to be evaluated and ranked to find the optimal or best conformations. This is done using some form of scoring function. Scoring functions are the mathematical methods to predict the binding affinity between the two interacting molecules. The functions are parametrized or trained using experimentally determined binding affinities between the molecular species similar to the species that one wishes to predict (Rajamani & Good, 2007). Scoring functions are of three types, 1. The force-field based: The binding affinities are

determined by the strength of the van der Waals and electrostatic interactions between the ligand and receptor atoms, and the intra-molecular strain energy associated within the two binding partners. 2. Empirical methods: These methods score based on the number of molecular interactions between the binding partners. The common types of interactions that are counted are, HBs, hydrophobic interactions and solvent interactions. The number of rotatable bonds which are immobilized during the complex formation are also counted (Bohm, 1998). 3. Knowledge based: These methods are based on the assumption that close inter-molecular interactions between certain types of atoms or functional groups that occur more frequently are likely to be energetically favorable than other possible types of interactions. These are statistical methods based on information from large structure databases like Protein Data Bank or Cambridge Structural Database. Information on the inter molecular close contacts are used to derive 'potentials of mean force' (Muegge, 2006).

1.7.4 The GOLD Package

GOLD or Genetic Optimization for Ligand Docking, is an automated ligand docking program that uses a Genetic algorithm (GA) to explore the full range of ligand conformational flexibility with partial flexibility of the protein (Jones *et al.*, 1997). An evolutionary strategy is employed to explore the conformational variability of a flexible ligand while simultaneously sampling available binding modes into a partially flexible protein active site. A GA is a computer program that mimics the process of evolution by manipulating a collection of data structures called chromosomes. Each of these structures encode possible solutions (ie. a possible ligand orientation within the protein binding site) to the docking problem and may be assigned a fitness score based on the relative merit of that solution. A steady state operator based GA was used to explore conformation space and ligand binding modes. The algorithm works in

the following steps,

1. A set of operators, crossover, mutation etc., is chosen with weights assigned to them.
2. An initial population is randomly created and the fitness of its members determined.
3. An operator is chosen using roulette wheel selection based on their weights.
4. The parents required by the operator are chosen using roulette wheel selection based on scaled fitness.
5. The operator is applied and to produce the child chromosomes. The fitness is evaluated for the child chromosomes.
6. If not already present in the population, the children replace the least fit members of the population.
7. If a maximum of 100000 operators have been applied then the interactions are stopped, otherwise the cycle is repeated from step 3.

1.7.4.1 Preparation of protein and ligand

The protein molecule is prepared for docking as follows. Any water molecules in the structures are removed and hydrogen atoms added appropriately considering the protonation states. The Ligand binding site is defined by the user by specifying two parameters; a location or origin in structure and a radius to define space about this location. The location can be specified by a point, atom or a list of atoms. A flood-fill algorithm (Ho & Marshall, 1990) is used to locate solvent accessible surface within the specified radius of the point origin. Then a cavity detection algorithm (Delaney1992, 1992) is employed to locate all solvent-accessible cavities that are less than 7.5Å wide. A second pass through the algorithm filters those cavities that are less than 2Å wide. Then a second pass through flood-fill algorithm is again made to locate the solvent-accessible protein surface. Within the detected cavity, all the donor and acceptor atoms

for HBs are identified using SYBYL atom-type characterization. Lone pairs are added to acceptors at a distance of 1Å. The ligand used for docking is fully minimized using the MAXIMIN2 module of SYBYL (Clark *et al.*, 1989). All acyclic and non-terminal single bonds are marked as rotatable.

1.7.4.2 The chromosome representation

The conformation information is encoded using two binary strings; one for protein and one for ligand, where each byte in the string encodes an angle of rotation about a rotatable bond. Each torsion is allowed to vary between -180°to 180°in step-sizes of 1.4°. Two integer strings encode mappings, suggesting possible HBs between the protein and the ligand. The first of these strings encodes a mapping form acceptors in the ligand to donor hydrogen atoms in the protein, such that if V was the integer value at position P on the string, then the P^{th} acceptor in the ligand was mapped to the V^{th} donor hydrogen in the protein. V could also be a null value, indicating that the acceptor was not mapped to any protein hydrogen. In a similar manner the second string encodes a mapping from donor hydrogen atoms in the ligand to acceptors within the ligand. On decoding a chromosome, GOLD uses the least-squares fitting method to form as many of these HBs as possible.

1.7.4.3 Fitness function

In GOLD the best docking solutions are generated and scored using the fitness function. The fitness function is evaluated at six stages as follows.

1. Generation of the conformations for ligand and protein. The binary coding is used to generate the conformations for the active site and ligand.

2. Placing of the ligand within the active site using a least squares fitting procedure. For every donor hydrogen atom a virtual fitting point was created collinear with the bond at a distance of 2.9Å from the donor. Fitting points are created at the center of each acceptor. A Least squares fitting technique is used to dock the ligand within the active site. In the second pass of least squares fitting, the points that are less than 1.5Å apart are used.

3. For each possible combination of donor hydrogen atom and acceptor the energy is examined against the geometrical criteria and a weight is assigned between 0 and 1 for potential bond. The weight is combination of 2 weights, the distance weight and angle weight. If the distance is less than 0.25Å, then the distance weight assigned is 1. If the distance is more than a cutoff distance, the weight is 0. The cutoff distance which is 4Å at the start, is decreased linearly after application of 75000 GAs. This was done to initially allow long range contacts between donors and acceptors to fitness score. Similarly angle weights are considered. Angle is calculated between the Donor-H and lone-pair-Acceptor bonds. If the angle is greater than 60°then the weight is 0. If it is less than 20°the weight is 1. Between angles of 20°to 60°the weight is between 0 and 1. The HB energies are scaled by the weights. The total HB energy is the sum of all individual HB energies.

4. The steric energy of the complex, which is the energy of interaction between the protein and the ligand is calculated pairwise using a 4-8 potential with linear cut-off, as per the Equation 1.53), where E_{ij} is the energy of interaction between two atoms and d_{ij} is the distance between them. This form of potential is much softer than the standard 6-12 potential. Adjustments are made to E_{ij} if the two atoms are involved in a HB. All pairwise interaction energies of ligand and protein

atoms in close contact (<1.5Å) are added to give Complex energy.

$$E_{ij} = \frac{A}{d_{ij}^8} - \frac{B}{d_{ij}^4} \tag{1.53}$$

5. The internal energy of the ligand is a sum of ligand steric and torsional energies. The steric energy is determined using 6-12 potential as in Equation 1.54.

$$E_{ij} = \frac{C}{d_{ij}^{12}} - \frac{D}{d_{ij}^6} \tag{1.54}$$

The torsional energy is calculated using the Equation 1.55, where E_{ijkl} is the torsional energy associated with four consecutively bonded atoms i, j, k, l, ω is the torsional angle, n is the periodicity and V the barrier to rotation.

$$E_{ijkl} = \frac{1}{2}V_{ijkl}[1 + \frac{n_{ijkl}}{|n_{ijkl}|}cos(|n_{ijkl}|.\omega_{ijkl})] \tag{1.55}$$

6. In the final stage the above energy terms are added together to give a final fitness score.

The final fitness score is given by; HB_Energy - (Internal_Energy + Complex_Energy). For efficiency of the algorithm, all the encoded chromosomes are divided into sub populations. Three genetic operators, cross-over, mutation and migration are used. They are chosen using roulette-wheel selection, based on the operator weights. The weights are chosen such that cross-over and mutation are applied with equal probability and migration was applied 5% of the time. After the application of 100,000 genetic operations, the algorithm terminates giving the highest scoring docking.

1.8 Disease causing mutations in Proteins

SNPs or Single Nucleotide Polymorphisms are single base pair changes occurring in the DNA. The single base pair changes, also called point mutations can be of non-synonymous or synonymous types depending on whether it leads to AA change in the translated protein or not respectively. The non-synonymous SNPs which alters the AA in the translated protein, can affect the native function of the protein (Halushka *et al.*, 1999). However certain non-synonymous SNPs do not lead to marked changes in function of the translated protein. These SNPs are silent with respect to function and are well accommodated into the protein structure. SNPs are the most common type of genetic variation in humans accounting for about 90% of differences (Collins *et al.*, 1998) The SNP database (www.ncbi.nlm.nih.gov/SNP) at NCBI harbors all the discovered SNPs in humans. Non-synonymous SNPs that have deleterious effects on the proteins' function are also called Mutations. According to the Human Gene Mutation Database, the mis-sense mutations account for almost half of all DNA mutations that are known to cause genetic disease. Since the SNPs can have varied effects on the protein function, a knowledge of the effects of SNPs on protein structure will be helpful in the assessment of susceptibility to diseases and individualizing the therapeutic measures (Masood, 1999).

The effects of SNPs can be estimated at the DNA level by genetic linkage studies. But for polygenic diseases the linkage studies are difficult (Pratt & Dzau, 1999). At the protein structure level many structural properties could be used for prediction of changes due to SNPs. For example, changes in properties of the hydrophobic interactions (Eriksson *et al.*, 1992), HBs (Shirley *et al.*, 1992), Van der Waals interactions (Xu *et al.*, 1998), electrostatic interactions (Horovitz *et al.*, 1990) or disulphide bonds (Betz, 1993) could be used for estimating the effect of SNPs.

At the AA sequence level, multiple sequence alignments of homologous proteins can be used to calculate sequence profiles and measures of AA conservation could be used to predict the effect of SNPs (Sunyaev *et al.*, 2000). Amino acid substitution matrix BLOSUM62 was used to classify the SNPs in the coding regions as conservative or non-conservative (Cargill *et al.*, 1999). However the use of substitution scoring matrices gives general information and does not incorporate protein specific information for specific SNPs. The aligned sequences at the protein family level give position specific information unlike substitution matrices. SIFT is based on this principle which predicts whether the substitution at a particular position in a protein will have a phenotypic effect (Ng & Henikoff, 2001). SIFT generates a multiple alignment of homologous sequences after searching a database. Then normalized probabilities for all possible substitutions at each position from the alignment are calculated. Substitution at each position with normalized probabilities less than a chosen cutoff are predicted to be deleterious, and those greater than or equal to the cutoff are predicted to be tolerated. The predictions based on SIFT method were better than those based on the substitution scoring matrices.

Adams and Chasman have used structure and sequence based features to derive a selection of structure and sequence based features that serve as indicators that are used to predict the effects of AAs substitutions (Chasman & Adams, 2001). It was estimated from the predictions that 26 to 32% of the natural non-synonymous substitutions in the human population are likely to affect protein function.

Sunyaev et al (Sunyaev *et al.*, 2001), have devised certain rules based on sequence and structural features. An AA variant is predicted to affect function or structure of the protein, if it satisfies certain empirical conditions derived from known protein structures, interactions and evolution. These conditions are, 1. The variant is located in any one of the annotated sites defined in the SWISS-PROT database. 2. The variant

is not compatible with the context of AA substitutions at the position in the family of homologous proteins. 3. The variant is likely to destroy the hydrophobic core of the protein, based on the solvent accessible areas. 4. The variant in the buried site displays a change in the electrostatic charge. 5. The variant is predicted to affect solubility of the protein. 6. The variant involves a proline residue in the α helix. and 7. The variant is predicted to affect protein-ligand interactions. Using these rules, it was predicted that about 20% of the common human non-synonymous SNPs damage the protein.

Wang and Moult used the data on *in vitro* site directed mutagenesis, together with the data on structural context of known disease-causing mutations to develop a model for the functional impact of mutations. Each mutation is associated with an effect on one or more roles of the residue considered. The roles that may be affected are, protein stability or folding, ligand binding, catalysis, regulation by allosteric and other mechanisms, post-translational modifications etc. The changes in the types of interactions are also identified. 90% of the known disease causing missense mutations could be explained by the model. Majority of the disease causing mutations were found to affect protein stability predicted due to over-packing, loss of HBs, strain in backbone and due to buried charged residues.

PolyPhen server uses the structural and evolutionary information in the prediction process (Ramensky *et al.*, 2002). The annotated structural information is obtained from SWALL database (Johnson & Todd, 2000) to find if the substitution site is involved in disulphide bond formation, or a part of active site, signal peptide or trans membrane regions. It uses DSSP (Kabsch & Sander, 2004) program to calculate the secondary structure and the solvent accessibilities. In addition, it also uses the evolutionary information from the alignment of homologous proteins of 30 to 90% identities. In the third stage, it maps the substitution sites to 3D structure of the protein or homologoues having greater than 50% identity, to find whether the substitution will affect the local

structural properties like the disruption of hydrophobic core, change in electrostatic interactions, change in interactions with ligand or other proteins. The predictions are of two categories, 1. nsSNPs possibly damaging protein function and 2. nsSNPs probably damaging protein function. From the analysis of predictions it was found that various effects on protein stability are responsible for accumulation of slightly deleterious nsSNPs in human genes.

1.8.1 Human CYP1b1 in PCG and other diseases

Primary Congenital Glaucoma is a severe eye disorder occurring at birth (Akarsu *et al.*, 1996) or early childhood (Bejjani *et al.*, 1998) up to the age 3 years, and is a major cause of blindness in infancy. This disease is characterized by a developmental abnormality of the trabecular mesh work, in the anterior chamber angle of the eye leading to increased intra-ocular pressure, resulting in optic-nerve damage and permanent loss of vision (deLuise & Anderson, 1983). PCG has been found to be mainly autosomal recessive in inheritance. The prevalence of the disease has been reported to vary geographically, with incidence ratios as low as 1:10,000 in the Western countries to as high as 1:1,250 among the Slovak gypsies (Gencik *et al.*, 1982). High incidence ratios have usually been found in populations that practise inbreeding and consanguinity.

Genetic linkage studies have indicated that PCG is, genetically, a heterogeneous disease, mapping on to at least three different loci. The first locus is GLC3A located in the region 2p21 of chromosome 2 (Sarfarazi *et al.*, 1995); the second is GLC3B, located in the region 1p36 of chromosome 1 (Akarsu *et al.*, 1996); and the third is GLC3C, located in the region 14q24.3 on chromosome 14 (Stoilov & Sarfaraze, 2002). In majority of the PCG instances the candidate locus has been found to be GLC3A that codes for a cytochrome p450 protein called CYP1b1. It is thought that CYP1b1

participates in the development of trabecular mesh work of eye which serves as a filter for drainage of anterior chamber fluid (Stoilov *et al.*, 2001). To-date the mechanism or the pathway by which CYP1b1 functions in the development of trabecular mesh work is not known. However, it is suspected that CYP1b1 may be involved in the elimination of a metabolite, the presence of which may have toxic effect on eye development. It has also been thought that the protein may be involved in the generation of a regulatory molecule which controls the expression of genes involved in the development of the anterior chamber angle of the eye (Stoilov *et al.*, 2001). It is certain that mutations having deleterious effect on protein function do hamper the normal development of the trabecular mesh work. Sequence analysis have so far revealed several mutations in the CYP1b1 gene, of which some are found only in the PCG affected individuals (Bejjani *et al.*, 1998; Bejjani *et al.*, 2000; Kakiuchi *et al.*, 1999; Martin *et al.*, 2000; Mashima *et al.*, 2001; Ohtake *et al.*, 2000; Plasilova *et al.*, 1998a; Plasilova *et al.*, 1998b; Stoilov *et al.*, 2001; Stoilov *et al.*, 1998; Stoilov *et al.*, 2002).

The Human CYP1b1 is a relatively recently identified p450 which has distinct properties compared to the other members of the family. It is a single copy gene which maps to human chromosome 2 at 2p21-22, having three exons and two introns (Sutter *et al.*, 1994). CYP1b1 is expressed in several normal tissue and cell types including liver, lymphocytes, breast tissue, lung epithelial cells, endometrium etc (Hakkola *et al.*, 1997; Hellmold *et al.*, 1998; Vadlamuri *et al.*, 1998). The constitutive expression of CYP1b1 is low in the normal tissues (Murray *et al.*, 2001). However, CYP1b1 was found expressed at high levels in various tumor tissues than the corresponding normal tissues (Murray *et al.*, 1997; Cheung *et al.*, 1999). Liehr and Ricci (Liehr & Ricci, 1996) found that it is expressed at high levels in breast cancer tissues. This was confirmed by increased estradiol 4-hydroxylse activity in microsomes prepared from breast cancer. The activity is not detected in normal tissue, where 2-hydroxylation of estradiol occurred.

Polymorphisms in CYP1b1 in various ethnic populations have been discovered (Inoue *et al.*, 2000). The polymorphic CYP1b1 showed altered catalytic activities towards the steroid hormones, etradiol testosterone and progesterone and are found associated with incidence breast cancer (Li *et al.*, 2000; Stoilov *et al.*, 1998). The association of CYP1b1 with the disease Primary Congenital Glaucoma (PCG) has been fist established by Sarfarazi in which mutations within the CYP1b1 gene where associated with the disease (Sarfarazi *et al.*, 1995).

1.9 Cytochrome p450 super family

The cytochrome p450 family is a large group of heme containing proteins found in almost all organisms. The name is derived from the fact that they have an absorbance maximum at 450nm when the enzyme is bound to carbon monoxide. They are involved in a wide variety of enzymatic reactions using both exogenous and endogenous compounds as substrates.

1.9.1 Mechanism of Catalysis

The p450s, are involved in many diverse metabolic reactions inside the cells of almost all of the living organisms (Oesterheld, 1998). p450 belong to the class of enzymes called oxygenases, and they are specifically monoxygenases or mixed function oxidases. The common form of reaction catalyzed by p450s can be illustrated by the Equation 1.56, in which one atom of oxygen is incorporated into organic substrate while the other oxygen atom is reduced to water (Coon, 2004).

$$RH + O_2 + 2H^+ + 2e^- - - > ROH + H_2O \tag{1.56}$$

All the CYPs have an absolutely conserved motif at the active site, RXCXG, in which the Cysteine is bound to Iron (Fe) of heme by the thiolate bond. In the ligand free form, the Fe is at $+3$ oxidation state. Binding the substrate initiates the transfer of electrons and oxygen binding. The electrons are supplied by an accessory protein called p450 reductase. The reduced Fe binds molecular oxygen splitting it into nascent oxygen. The reactive oxygen oxidizes the substrate to either an epoxide or an alcohol thereby regenerating the resting state of the enzyme.

1.9.2 Structural features of p450s

The 3D structure among Cytochrome p450s is highly conserved despite low sequence similarities (can be less than 20%). This high structural conservation among p450s is perhaps due to their common mechanism of electron and proton transfer and oxygen activation (Werck-Reichhart & Feyereisen, 2000). The active site of the typical enzyme is deep inside the protein at a location above the plane of the heme. The substrates enter and the products exit from the active site via a channel that opens on the protein surface (Scott *et al.*, 2003). The general structural topology of p450s is schematically represented in Figure 1.20 (Werck-Reichhart & Feyereisen, 2000). The structure can be divided into the α-rich and β-rich domains (Cupp-Vickery & Poulos, 1995). The conserved core of the protein consists of four helix bundle (containing helices D, E, I and L), the helices J and K, two sets of β-sheets and a loop region called meander. The conserved core has three characteristic p450 signature sequences. The first sequence, F-X-X-G-X-R-X-C-X-G with an absolutely conserved cysteine serving as the fifth ligand to the Fe of heme, occurs in the heme binding loop, just before the L helix. The second absolutely conserved signature sequence E-X-X-R is in the K helix. The first and second sequences are located on the proximal side of the heme and probably needed to stabilize the core structure. The third sequence occurs in the I helix with a signature, A/G-G-

X-D/E-T-T/S. This corresponds to the proton transfer groove on the distal side of the heme.

There are regions which show considerable sequence variability. For example, at the region of substrate recognition and substrate access channel and the catalytic sites. This is probably to cater for the diverse substrates that p450s catalyze.

1.9.3 Classification of Cytochrome p450s

Cytochrome p450s can be broadly classified into bacterial forms and microsomal forms. The bacterial forms are soluble enzymes while the microsomal or eukaryote forms are bound to the endoplasmic reticulum. p450s can also be classified based on their type of interaction with redox partners. Class I proteins require both FAD-containing NADPH reductase and an iron-sulfur redoxin as electron donors. While the proteins are soluble in prokaryotes, they are attached to mitochondrial inner membrane in eukaryotes. Class II proteins require FAD/FMN containing NADPH p450 reductase. In prokaryotes the reductase protein is fused with the p450 protein, while in eukaryotes the p450 and reductase proteins are separate and found attached to outer surface of the endoplasmic reticulum by N-terminal hydrophobic domain. The class II proteins are the most abundant in eukaryotes. The class I and class II enzymes catalyze the most diverse functions, from activation of xenobiotics, to biosynthesis and catabolism of signaling molecules, steroid hormones (Hasler *et al.*, 1999). The class III p450s unlike class I and II are self sufficient enzymes, and do not require molecular oxygen or an external electron source. They catalyze rearrangement or dehydration of alkyl hydroperoxides or alkylperoxides that are generated by dioxygenases. They are involved in the synthesis of signaling molecules such as prostaglandins in mammals and jasmonate in plants (Mansuy, 1998). Finally in class IV p450s the electrons are directly received

from NADH. For example, the soluble fungal p450 reduces NO to N2O (Mansuy, 1998). p450s belonging to Class III and IV are not so abundant and can be considered as the remains of the most ancestral form of p450.

Apart from the above general classification, systematic classification of p450s is based on the sequence identity, phylogenetic criteria and gene organization (Nelson *et al.*, 1996). The super family or root is denoted as 'CYP'. The families are denoted with numerals 1,2,... etc. Members with more than 40% sequnce identities belong to the same family. The families are further divided into sub-families denoted by alphabets, a,b,c..etc. The members of the same sub-family share more than 55% sequence identity. Finally individual members of each sub-family are numbered 1,2...etc.

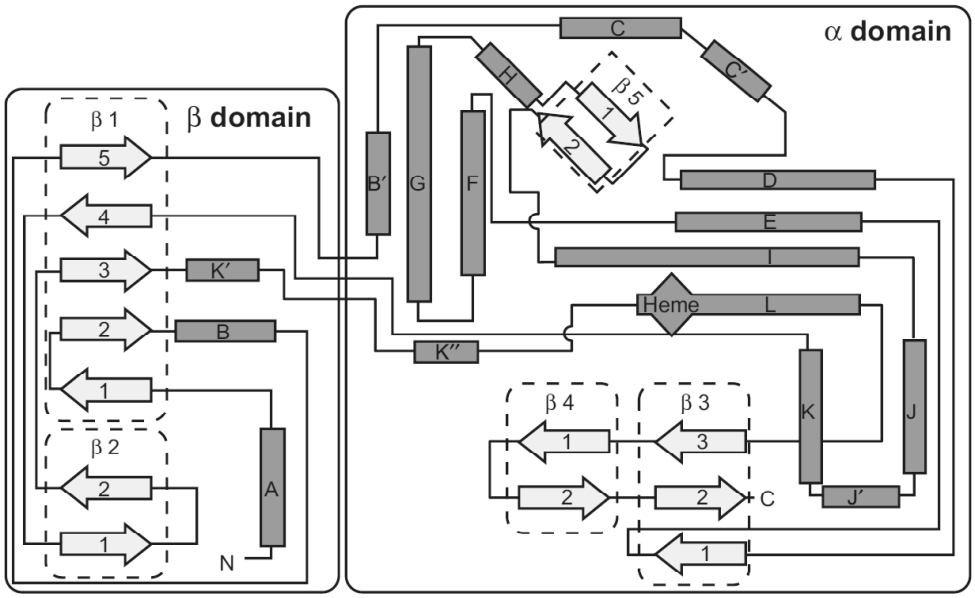


Figure 1.20: The arrangement of secondary structural elements of a typical p450 protein. The α helices are named alphabetically A,B,... and the βsheets are numbered 1,2,..

1.10 The Present work

In the above review, the utility of predictive methods for assessing the effects of disease causing mutations was discussed. However, the question of the detailed mechanisms by which disease causing mutations affect the function of the protein has not been adequately studied. This aspect needs a detailed exploration of the structure and dynamic properties of the Wild type (WT) and the Mutant (MT) form of the protein.

The present work on Deciphering the structural dynamics of disease causing mutations in Proteins, aims to investigate how the effects of disease causing mutations are propogated into the structure, resulting in altered properties. The structure and dynamics of the Human CYP1b1 and its 8 PCG causing MTs (A115P, M132R, Q144P, P193L, E229K, S239R, R368H and G466D) are investigated. These Mutations which were reported by Panicker and coworkers (Panicker *et al.*, 2002; Reddy *et al.*, 2004) in the Indian PCG patients, are studied by both sequence and structure based analysis.

1.11 Summary and Conclusion

In this Chapter, a brief review on the organization of protein structure and the methods used for the determination of protein structure is presented. A detailed background on the methods used in the current investigation, namely Homology Modeling, Energy Minimization, Molecular Dynamics and Molecular Docking is given. Further, an overview of the software used in the investigation namely, MODELLER, GROMACS and GOLD was made. A comprehensive overview was provided on Cytochrome p450 proteins, specifically Human CYP1b1 and disease causing mutants, which is the focus of the current investigation. A brief account on Primary Congenital Glaucoma, caused due to mutations in CYP1b1 is also made. Finally the objectives of the current work are indroduced.

2

Sequence Analysis and Modeling Studies

2.1 Introduction

In the previous chapter, the general importance of CYPs in the detoxification system in the body, and especially the importance of CYP1b1 in Humans, which is implicated in some of the Human diseases including PCG has been discussed. It was also seen that current methods of computational sequence and structural analysis have been helpful in a better understanding of deleterious mutations in proteins. Comparative sequence analysis provides a way to understand the significance of mutations in the protein sequence in the context of the AA variation observed in the related protein sequences. It provides with a background knowledge of protein's similarity and uniqueness to the generally observed trends in AA substituions in the family, thereby useful in predicting the effects of mutations. Hence in the present chapter an account of the studies pertaining to sequence analysis of Human CYP1b1 is presented.

Apart from the sequence information, the investigation on many functional aspects of the protein cannot be accurately carried out without having structural information.

Structure provides atomistic detail of the protein that helps in understanding its functional aspects. Usually the structure determined by X-ray crystallography or NMR is used in the structural analysis. However, for Human CYP1b1 the experimentally determined structure is not available. A comparative model or homology model of the protein based on a template, is the best alternative to the experimentally determined structure (Forrest *et al.*, 2006; Tramontano *et al.*, 2001). Thus one of the main aspects of the current investigation was to build a comparative model for Human CYP1b1.

In the comparative modeling procedure, the reliability of the model depends largely on its sequence similarity, and accuracy of sequence alignment with the template used for modeling (Chothia & Lesk, 1986). Selection of the appropriate template structure is thus an important step in modeling. As discussed earlier, most of the eukaryotic CYPs are membrane bound proteins and hence difficult to crystallize, which has resulted in only a limited number of them having the structure determined experimentally. As on September, 2007, there were about 5544 CYP sequences in the trEMBL of the SWISSPROT data bank, out of which for only about 189 proteins the structures have been solved. Out of these, 27 are Human CYP structures (Consisting of the unbound and ligand bound forms of CYPs; 1A2, 2C8,2C9,3A4,2A6, 2D6, 2R1 and 2A13).

As already mentioned, experimentally determined structure (X ray or NMR) is not available for the Human CYP1b1, for that matter, for none of the proteins belonging to CYP1b1 family. However, homology based models have been reported earlier for Human CYP1b1 using both the bacterial and the microsomal p450 templates. Stoilov and coworkers, used the bacterial CYP102 to build a model of CYP1b1 and found some PCG causing mutations to be located in the hinge or the conserved core of the protein (Stoilov *et al.*, 1998). CYP1b1 models built using CYP1a1 and CYP1a2 as the templates were used to find the solvent exposed regions of protein to synthesize specific antibodies for the protein. The antibodies have been successfully used to detect

CYP1b1 expression in cases of breast cancer (McFadyen *et al.*, 1999). Two CYP1b1 models built using the templates 2hpd and 2c17 when superimposed, were found to have high structural similarity in the conserved regions. This indicates the reliability of comparative modeling procedure to get consistent models. The better model was further used to develop models for four missense mutations D192V, A330F, V364M, R444Q which are implicated in PCG, to find the effects of mutations on the structure.

The structure of the first mammalian p450, CYP2c5 was solved in the year 2000 (Williams *et al.*, 2000) and this paved way for the modeling of a number of microsomal p450s including CYP1b1 (Belkina *et al.*, 1998; Tsigelny *et al.*, 2004). CYP1b1 models built using CYP2c5 were used to explain the probable nature of the effects of mutations in terms of changes in the interactions and folding of the enzyme (Mashima *et al.*, 2001). In 2003 Shimada and co-workers (Lewis *et al.*, 2003) used CYP2c5 as template to model CYP1b1 to investigate the binding modes of selected CYP1b1 substrates and the effects of allelic variants on metabolism. Using the model the substrate selectivity of the enzyme could be inferred, based on the specific contacts of the substrates with the AA residues in the putative active site (Lewis *et al.*, 2003).

In CYPs the substrates are recruited into the active site through an opening formed between the F/G loop and the B' helix (Ludemann *et al.*, 2000a). The F/G loop is also involved in substrate recognition (Raucy & Allen, 2001) and membrane interaction (Storbeck *et al.*, 2007). The templates used for modeling CYP1b1 in modeling studies mentioned above, lacked proper structural information for this region. During the time of the current investigation, the structure of Human cytochrome p450 (CYP2c9) (PDB-CODE: 1OG2) (Williams *et al.*, 2003) was reported. CYP2c9 obviously was the template of choice for modeling CYP1b1, given the fact that CYPs within a species are expected to have more structural similarity. Importantly in CYP2c9,

the loop region between F and G helices contains short helices F and G, a Human specific CYP structural feature, not observed in bacterial and other mammalian CYP structures (Williams *et al.*, 2003). In following sections, the procedures used for comparative sequence analysis and comparative modeling of CYP1b1 and its PCG MT forms, and the subsequent analysis of the models are described.

2.2 Material and Methods

The Human CYP1b1 has 543 AAs (see Figure 2.1). The PFAM domain description for the protein (PfamID: PF00067) is shown in Figure 2.2. The segment starting from residue number 51 till 520 maps to the p450 domain, residues 20 to 40 is trans membrane domain and the rest are low-complexity domains. The P450 domain is the cytosolic domain and is mainly responsible for the catalytic function.

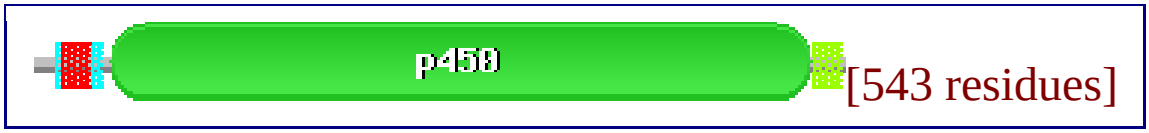
2.2.1 Comparative Sequence Analysis of CYPs

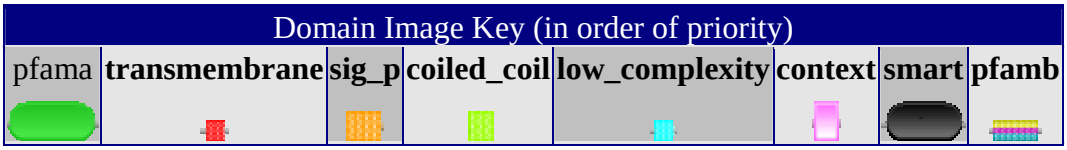
In order to aid comparative sequence analysis, a multiple sequence alignment(MSA) of all the CYP sequences was obtained from the PFAM database (Finn & Tate, 2008). The MSA corresponding to only the Human sequences, only CYP1 sequences and only the CYP1b sequences were separately analyzed, to get the conservation patterns specific to Human CYP sequences, CYP1 family and CYP1b subfamily respectively. The conservation pattern of AAs in CYPs was determined by the frequency of occurrence of each of the 20 AAs sometimes referred to profile and the entropy at each residue position.

The frequencies (in %) of occurrence of AAs at each of the alignment positions were computed using the Equation 2.1 where, $F_{i,j}$ denotes the frequency of AA i at position j, $AA_{i,j}$ is the number of occurrences of AA i at position j and N is the total

```
>Q16678|CP1B1_HUMAN Cytochrome P4501B1 - Homo sapiens(Human)
MGTSLSPNDPWPLNPLSIQQTTLLLLLSVLATVHVGQRLLRQRRRQLRSAPPGPFAWPLI
GNAAAVGQAAHLSFARLARRYGDVFQIRLGSCPIVVLNGERAIHQALVQQGSAFADRPAF
ASFRVVSGGRSMAFGHYSEHWKVQRRAAHSMMRNFFTRQPRSRQVLEGHVLSEARELVAL
LVRGSADGAFLDPRPLTVVAVANVMSAVCFGCRYSHDDPEFRELLSHNEEFGRTVGAGSL
VDVMPWLQYFPNPVRTVFREFEQLNRNFSNFILDKFLRHCESLRPGAAPRDMMDAFILSA
EKKAAGDSHGGGARLDLENVPATITDIFGASQDTLSTALQWLLLLFTRYPDVQTRVQAEL
DQVVGRDRLPCMGDQPNLPYVLAFLYEAMRFSSFVPVTIPHATTANTSVLGYHIPKDTVV
FVNQWSVNHDPLKWPNPENFDPARFLDKDGLINKDLTSRVMIFSVGKRRCIGEELSKMQL
FLFISILAHQCDFRANPNEPAKMNFSYGLTIKPKSFKVNVTLRESMELLDSAVQNLQAKE
TCQ
```

Figure 2.1: Primary Structure of Human CYP1b1





Source	Domain	Start	End
seg	Low Complexity	16	30
Phobius	Transmembrane	20	40
seg	Low Complexity	37	48
PfamA	p450	51	520
seg	Low Complexity (hidden)	334	347
ncoils	Coiled Coil (hidden)	519	539

Figure 2.2: Functional domains in CYP1b1

number of sequences in the alignment.

$$F_{i,j} = \frac{AA_{i,j}}{N} * 100 \tag{2.1}$$

The sequence entropy was calculated with a modified equation for Shannon's negative entropy (Shannon, 1948) using the Equation 2.2 where, S_j denotes the entropy at position j; $P_{i,j}$ is the probability of AA i and at position j; Pg_j is the probability of gap at position j. Thus, a penalty was applied for the occurrence of gaps. Highly conserved positions have low entropy values and *vice versa*, with the maximum entropy value depending on the number of sequences in the alignment.

$$S_j = -\sum_{i=1}^{20} P_{i,j} \ log_2 \frac{P_{i,j}}{1 - Pg_j} \tag{2.2}$$

The sensitivity of the residue positions for mutations along the CYP1b1 sequence is determined as follows. A list of all the known Human CYP mutations was obtained from the Human Gene Mutation Database. The equivalent residue positions corresponding to the mutations were found in the Human CYP1b1 from the multiple sequence alignment. Further, the Entropy and AA profile information corresponding to these positions was used to make the comparative analysis.

2.2.2 Modeling of Human CYP1b1 and the PCG associated Mutants

Modeling of CYP1b1 protein was done using the comparative modeling procedure. The AA sequence of Human CYP1b1 was obtained from SWISS-PROT data bank (ID: Q16678). The information about the domain organization was retrieved from the PFAM database server. BLAST (Altschul *et al.*, 1990) search was made against the

Protein Data bank (Sussman & Lin, 1998) to find the homologous proteins with known structures. The sequence of the selected Template protein was aligned with CYP1b1 sequence using CLUSTAL (Thompson *et al.*, 1994). The secondary structure of the sequence was predicted using PSIPRED (McGuffin *et al.*, 2000) to get the secondary structural preferences for the residues. This information was used to manually adjust the gaps in the alignment. The academic version of MODELLER.v.4.0 (Sali & Blundell, 1993) (http://salilab.org/modeller) installed on SGI-O2 system was used for model building using the options as given in Table 2.1. The models generated were checked for their stereo-chemical quality using the PROCHECK software (Laskowski *et al.*, 1993). The 3D-1D sequence-structure compatibility was examined by computing profile scores using the VERIFY-3D software (Luthy *et al.*, 1992).

The WT structure was used to generate the disease MT structures by replacing wild type residues with disease associated residues. Overall, eight disease MT structures were generated corresponding to the eight PCG mutations; A115P, M132R, Q144P, P193L, E229K, S239R, R368H and G466D. The residue replacements were carried out using the MUTATE_MODEL command of MODELLER. In this function, the side chain of WT residue is replaced with the MT residue and the replaced side chain is subjected to energy minimization. During Energy Minimization, the dihedral angle restraints are applied using a 'homologue and dihedral' library. For modeling the side chain, the coordinates from the template native structure are transferred and the remaining unknown coordinates are built using the internal coordinates. The conformation of the MT side chain is then optimized by 200 steps of Conjugate Gradient algorithm and refined using short Molecular Dynamics.

Table 2.1: Options used for comparative modeling in MODELLER

INCLUDE		
SET MAXRES	=	650
SET OUTPUT_CONTROL	=	1 1 1 1 2
SET ALNFILE	=	'mod3CYP_1OG2.pir'
SET KNOWNS	=	'1og2'
SET SEQUENCE	=	'cyp1'
SET ATOM_FILES_DIRECTORY	=	'./:../atom_files'
SET OUTPUT_DIRECTORY	=	'./'
SET PDB_EXT	=	'.atm'
SET ALIGNMENT_FORMAT	=	'PIR'
SET STARTING_MODEL	=	1
SET HETATM_IO	=	on
SET WATER_IO	=	on
SET ENDING_MODEL	=	20
SET TOPOLOGY_MODEL	=	1
SET TOPLIB	=	'/usr/local/bin/modeller4/modlib/top.lib'
SET PARLIB	=	'/usr/local/bin/modeller4/modlib/par.lib'
SET MD_LEVEL	=	'refine_2'
SET DEVIATION	=	4.000000
CALL ROUTINE	=	'model'

2.2.3 Identification of Functionally Important Regions (FIRs)

All the CYPs are characterized by a highly conserved structural core, which is divisible into an α-rich region, comprising of about 14 α-helices and a β-rich region comprising of 4 to 6 β-sheets (Graham-Lorence & Peterson, 1996). Functionally, CYP structure can also be characterized by certain functionally important regions (FIRs) like the Heme binding region (HBR), substrate binding region (SBR), substrate access channel (SAC), the residues involved in charge relay and the surface region that binds to CYP reductase protein, which donates the electrons required for catalysis. For the purpose of the present study, three FIRs namely the HBR, SBR and SAC regions were identified in the model as follows:

Heme Binding Region (HBR)

The HBR is defined by a set of residues that are in contact with the heme co-factor. The solvent accessible surface area (SASA) of all residues in the protein with and without the heme were calculated and the residues that showed a decrease in SASA in the presence of heme were denoted as those comprising the HBR. PSA (Mizuguchi *et al.*, 1998) was used to calculate SASA values.

Substrate Binding Region (SBR)

To identify the SBR, the crystal structures of various CYP-ligand complexes (PDB-codes: 2CPP, 1PHA, 6CP4, 1FAG, 1E9X, 1EGY, 1EA1, 1F4T) were used. The SBR residues in each structure were identified as those showing an increase in SASA upon the removal of the bound ligand from the complex. The putative SBR residues in CYP1b1 were then identified, as those at the structurally equivalent positions in CYP1b1 in the multiple structural alignment of CYP1b1 with the CYP-ligand complex structures.

Substrate Access Channel (SAC)

Earlier studies indicated that the substrate access into the active site of the enzyme occurs through an opening between the F/G loop and the B'-helix (Graham-Lorence *et al.*, 1995; Hasemann *et al.*, 1995; Dai *et al.*, 1998; Williams *et al.*, 2000). The pathway (SAC) from outside to the active site of the protein is bounded by the F and G helices and the B/C loop. Wade and coworkers (Ludemann *et al.*, 2000a), have studied the substrate access pathway in P450cam and P450BM-3. They computationally demonstrated that the F/G loop and the B'-helix are very likely to be involved in substrate entry and exit. For our analysis, the residues in F/G loop and B'-helix are considered to denote the SAC.

2.3 Results and Discussion

2.3.1 Sequence Analysis of CYPs

Sequence analysis was carried out to get information on residue conservation in p450's in general and in the context of PCG mutations of CYP1b1. For this purpose, the frequency of occurrence of each of the 20 AAs and entropy at each of the CYP1b1 sequence positions was calculated from the multiple sequence alignment to determine the conserved and non-conserved regions with respect to the CYP super family. The PFAM sequence alignment consisting of a total of 8735 CYP sequences were used for the calculation. The sequence entropy information was earlier used to determine the AA conservation in proteins (Valdar, 2002). While calculating entropy, weightage for occurrence of gaps (Equation 2.2) was given such that presence of gaps contributes to an increase in entropy at the position. This method gives better estimates of entropy as compared to the basic Shannon's entropy method (Ramazzotti *et al.*, 2004).

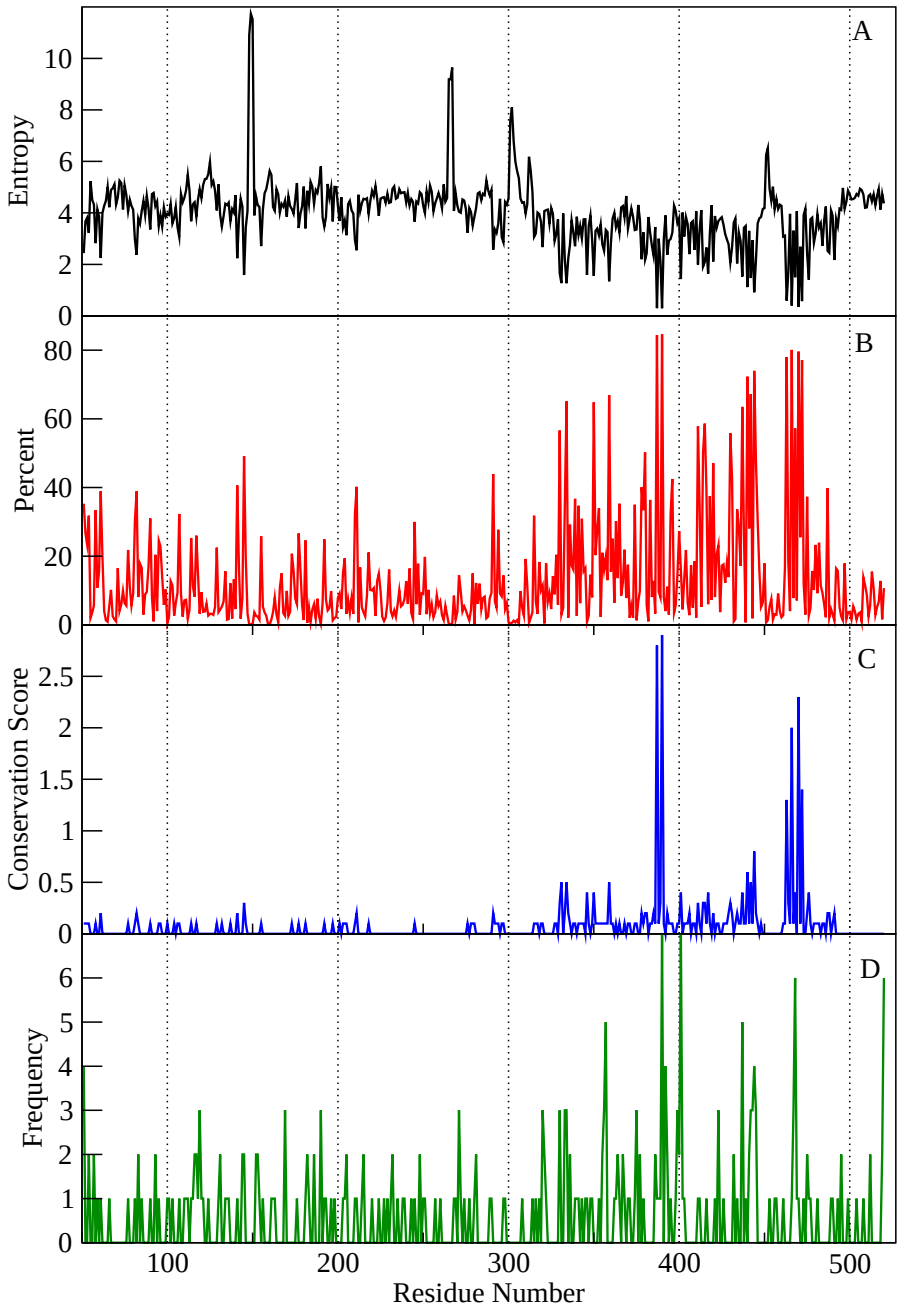


Figure 2.3: (A) Sequence entropy found in CYPs. (B) The percentage occurrance of CYP1b1 specific AAs in all p450s. (C) The conservation score along the sequence. (D) The frequencies of occurrence of Disease causing mutations in Human CYPs. In (A) to (D) the positions are equivalent to the Human CYP1b1 AA positions from 50 to 527.

Figure 2.3(A) shows the entropy along the CYP1b1 sequence. There is a clear variation of the entropy along the sequence. There are positions with high or low entropies. The average entropy for the entire sequence is about 4.0. The region spanning residues 325 and 485 has an average entropy of 3.1, which is considerably lower than the overall average. This region having lesser entropy corresponds to the I, J and K helices and the Heme binding loop, thus emphasizing conservation of residues in this region. Apart from this, there are several narrow peaks and drops in the graph representing respectively, the specific conserved and variable regions. To find the usefulness of entropy values in analysis of residue conservation, the entropy at the highly conserved positions among all p450s, was examined.

All the CYPs are characterized by some absolutely conserved residues. These residues belong to the three highly conserved motifs called 'CYP signatures'. Three important CYP signature sequences exist. They are the 'F-X-X-G-X-R-X-C-X-G' motif in the heme binding loop, the 'E-X-X-R' motif in the K' helix, and the 'A/G-G-X-D/E-T-T/S' motif in the middle of the I helix (Werck-Reichhart & Feyereisen, 2000). The sequence entropy values for these motifs are listed in Table 2.2. It can be seen that the conserved residues possess low entropy values compared to the residues indicated by 'X', within these motifs. In fact, the entropy values for these highly conserved residues were among the lowest along the whole sequence. The representation (in %) of the CYP1b1 AAs in the alignment of all p450s is given in Figure 2.3(B). It can be observed that the region that was observed to have lesser entropy, also has a high representation of CYP1b1 residues. This indicates that the AA composition in CYP1b1, especially in the conserved regions (low entropy), is similar to that found in p450s in general. Further, to accurately identify the most conserved residues in the family, a measure of conservation called Conservation score, was calcuated using the Equation 2.3 where in the maximum of AA frequencies of the 20 AAs at each position is divided by the

sequence entropy at that position.

$$Conservation\ score_j = \frac{MAX(F_{i,j})}{S_j} \quad i = 1\ to\ 20 \tag{2.3}$$

Figure 2.3(C) gives the plot of Conservation score along the CYP1b1 sequence. The height of the peaks represent the AA conservation at that position. The prominent peaks correspond to the highly conserved p450 signature sequences, that were mentioned earlier. Evidently, this method could identify the characteristic p450 signature sequences, with a better discrimination from the rest of the sequence, than the normal entropy method. It can thus be useful in the identification of highly conserved regions in proteins.

Further, the disease causing mutations in all Human CYP proteins were obtained from the HGMD database (Krawczak & Cooper, 1997) and their position wise frequency of occurrence was calculated (Figure 2.3(D)). This was done to study the frequency distribution of the disease causing mutations at positions equivalent to Human CYP1b1.

In general, the disease causing mutations are more frequent in regions of lesser entropy. In other words, disease causing mutations happen in regions of highest residue conservation. Figure 2.4 shows the relationship between the average entropy values and the mutation frequency. The frequency of mutations leading to disease increases with decreasing entropy. From the Figure 2.3(C and D) it could also be seen that the frequency of disease mutations is highest in the regions of maximum conservation scores. These data indicate that many of the p450 related diseases are a result of mutations in the highly conserved positions of the proteins. Subsequently the specific mutations of Human CYP1b1 that are implicated in PCG were studied. Table 2.3 gives the entropy values at the PCG mutation positions in CYP1b1. As seen from the table, the mutation sites are at positions of both high and low entropy .

Table 2.2: The Entropy values for the CYP signature sequences

Motif 1	Entropy	Motif 2	Entropy	Motif 3	Entropy
F	0.59	E	0.30	A/G	1.66
X	1.70	X	3.00	G	1.27
X	3.96	X	2.13	X	3.97
G	0.40	R	0.30	D/E	2.21
X	3.31			T	1.26
R	1.50			T/S	2.23
X	4.06				
C	0.35				
X	2.68				
G	0.57				

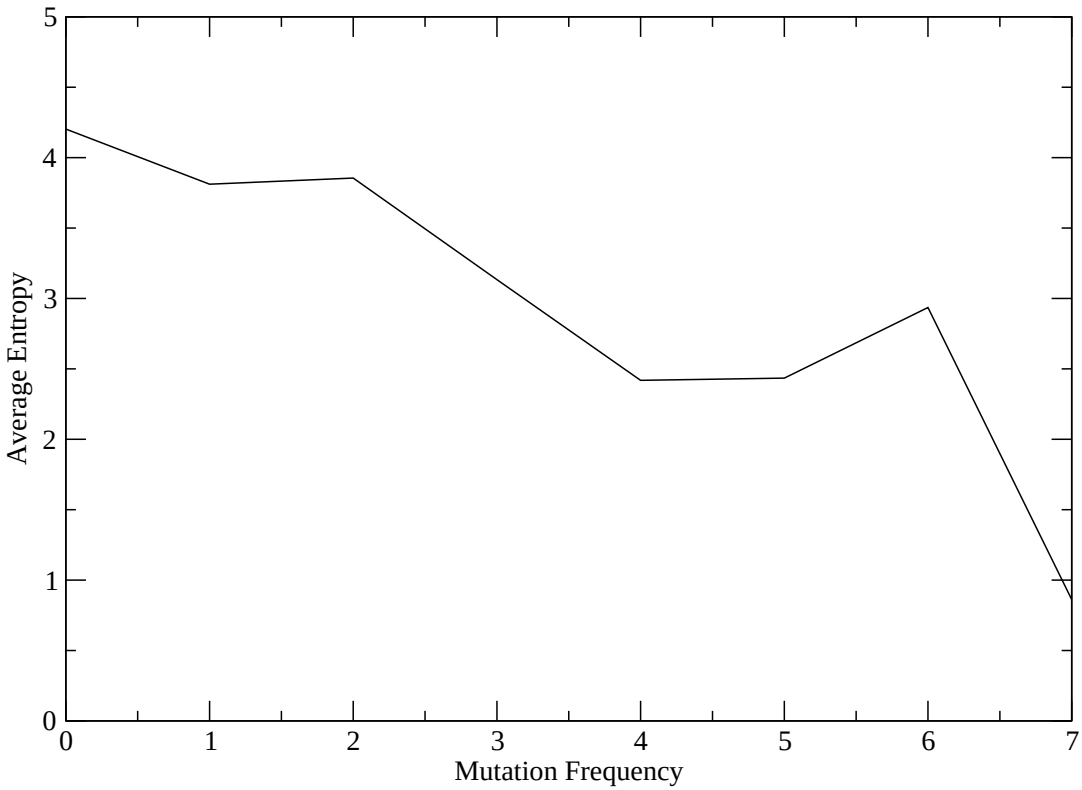


Figure 2.4: Relationship between Average entropy values and the Mutation frequency

Table 2.3: Entropy values at PCG Mutation positions in CYP1b1

Mutation site	Entropy
115	4.33
132	3.46
144	4.24
193	3.82
229	4.74
239	4.80
368	3.83
466	0.40

Table 2.4: Average Entropy and Mutation frequency at the FIRs

FIR	Entropy	Mutation Frequency
HBR	3.66	1.08
SBR	3.98	0.88
SAC	4.13	0.64
RBR	4.04	0.62
CHR	3.91	0.73
Non-FIR	4.07	0.56

Thus, some of the mutations which are implicated in PCG also occur in positions of high entropy. This is in contrast to the general trend seen above, where mostly the disease causing mutations occur in the regions of low entropy. This indicates that the entropy information obtained from a consideration of all p450 sequences, might not accurately describe the specificity associated with CYP1b1 sequence. To examine the variation in the AA preferences caused because of the data set used, the AA profiles at the PCG mutation sites were analyzed at four levels; considering all p450s, only the Human p450 sequences, only the CYP1 sequences and only the CYP1b1 sequences respectively. Table 2.5 gives the frequency of occurrence of each of the 20 AAs at the mutation sites. The frequencies calculated from alignments of all p450s, only Human p450s, only p450 of family 1 and only p450 belonging to 1b1 subfamily respectively, are indicated in the sub tables, A, B, C and D. The WT AA is colored in Blue, while the MT AA is colored in Red. In Table 2.5(A) corresponding to all CYPs, it could be seen that the mutation positions are not highly conserved for the WT residues. Only the mutation site 466 seem to be conserved among all the p450s. In Human p450s (Table 2.5(B)) however, it could be seen that the specificities have increased for the WT residues.

Further, in the p450s belonging to family1 (Table 2.5(C)), it could be observed that, many of the mutation sites have the WT allele occurring to a greater frequency than the MT residue. The other mutation positions though not highly conserved seem to be more specific to CYP1b1s where the WT residue has the highest frequency of occurrence (Table 2.5(D)). Furthermore, at 193 position the residue L occurs with considerable frequency but less than P in CYPs, but in CYP1b1s it is never found. This might mean that CYP1b1 has adapted more specifically to P rather than L, and P might be important for its structural integrity. It is clear that none of the PCG MT AAs occur among the three CYP1b1 members and as well as other p450 proteins except

Table 2.5: Profile of Amino acid Frequencies at the mutation sites. The WT AA is colored Blue, while the MT AA is colored Red. Values that are between 0 and 1 are shown as 0.

(A) All CYPs 8735 sequences in MSA

	A	G	C	S	P	T	N	D	E	K	R	H	Q	Y	W	F	I	L	V	M	-
115	**11**	2	1	14	**0**	1	2	0	1	4	1	1	0	1	0	0	0	1	3	0	48
132	0	0	0	1	1	1	0	0	0	0	**0**	0	0	0	0	1	14	17	9	**5**	45
144	2	0	0	0	**0**	1	0	0	0	0	4	6	**7**	1	0	1	3	13	3	6	44
193	2	0	0	0	**7**	1	0	0	0	0	0	0	0	0	0	2	9	**13**	10	5	45
229	3	2	0	2	0	1	4	6	**6**	**5**	5	3	5	1	0	1	0	1	2	0	44
239	3	0	0	**7**	0	3	6	1	1	0	**4**	0	0	0	3	5	1	3	3	2	47
368	3	8	0	4	1	1	1	6	4	4	**21**	**0**	4	0	0	0	0	1	0	0	34
466	0	**80**	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	19

(B) Only Human CYPs 258 sequences in MSA

	A	G	C	S	P	T	N	D	E	K	R	H	Q	Y	W	F	I	L	V	M	-
115	**22**	1	0	18	**0**	2	1	0	1	12	0	1	0	0	0	0	0	0	5	0	29
132	0	0	1	1	0	0	0	0	0	0	**0**	0	0	0	0	0	18	25	15	**4**	31
144	0	0	0	0	**0**	3	2	0	0	0	3	24	**12**	0	0	0	11	8	3	1	27
193	0	0	0	0	**27**	0	0	0	0	0	0	0	0	0	0	1	15	**8**	14	5	25
229	0	1	0	6	0	0	11	3	**5**	**6**	1	3	18	1	0	3	0	6	0	1	27
239	2	1	0	**16**	0	6	4	1	0	0	**9**	0	1	0	9	7	0	9	0	0	27
368	9	2	0	15	0	0	0	5	3	0	**29**	**0**	6	0	0	0	0	0	0	1	25
466	0	**72**	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	27

(C) Only CYP1 (47 sequences in MSA)

	A	G	C	S	P	T	N	D	E	K	R	H	Q	Y	W	F	I	L	V	M	-
115	**29**	0	0	6	**0**	0	0	0	0	53	0	0	0	0	0	0	0	0	0	4	6
132	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	46	0	**48**	4
144	0	2	0	0	**0**	0	0	0	0	0	80	0	**6**	0	0	0	0	4	0	0	6
193	0	0	0	0	**95**	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	4
229	0	0	0	6	0	2	25	29	**10**	**6**	2	12	0	0	0	0	0	0	0	0	4
239	0	0	0	**25**	0	0	63	0	0	0	**0**	0	0	6	0	0	0	0	0	0	4
368	0	0	0	0	0	0	0	0	0	0	**97**	**0**	0	0	0	0	0	0	0	0	2
466	0	**93**	0	0	0	0	0	**2**	0	0	0	0	0	0	0	0	0	0	0	0	4

(D) Only CYP1b1 (3 sequences in MSA)

	A	G	C	S	P	T	N	D	E	K	R	H	Q	Y	W	F	I	L	V	M	-
115	**100**	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
132	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	66	0	**33**	0
144	0	0	0	0	**0**	0	0	0	0	0	33	0	**66**	0	0	0	0	0	0	0	0
193	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0
229	0	0	0	0	0	0	0	0	**100**	**0**	0	0	0	0	0	0	0	0	0	0	0
239	0	0	0	**100**	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0
368	0	0	0	0	0	0	0	0	0	0	**100**	**0**	0	0	0	0	0	0	0	0	0
466	0	**100**	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0

for MT AAs at positions 193, 229 and 239, indicating the incompatibility of the MT AAs for these positions in all CYPs in general. In the case of positions 193, 229 and 239 though some p450s do have the MT AAs, in the case of CYP1b1s, the MT AAs are not found indicating the structural specificity of CYP1b1 for the WT AAs.

In addition to this, the residue conservation in the FIRs was examined. Table 2.4 gives the average entropy associated with the FIRs. It could be observed that the average entropy in Non-FIR regions is higher compared to the FIR regions indicating the importance of the FIRs. Among the FIRs, the HBR region followed by the SBR and the CHR regions have the least average entropy, indicating their relative importance in protein structure. The SAC region has more entropy than other FIRs, even higher than the NON-FIR regions, probably with a significance for such property. Table 2.4 also shows the average frequency of occurrence of all disease causing mutations in Human CYPs, at the equivalent FIR regions of CYP1b1. It could be noted that the disease causing mutations are more in FIRs than Non-FIRs. The HBR region followed by SBR which have the lowest average entropy, has the highest average occurrance of disease causing mutations.

2.3.2 Modeling Wild type and PCG Mutant structures of CYP1b1

2.3.2.1 Selection of Template

BLAST search was made against the PDB data bank to find homologous proteins whose structures have been solved. By the time of this modeling work, the structures of only two Mammalian p450 structures had been solved. The Rabbit cytochrome p450, namely CYP2c5 (PDBID:1dt6) and later the first Human CYP structure, CYP2c9 (PDBID:1og2). Human CYP2c9 was the template of choice as it was solved at a better

resolution of 2.6Å than CYP2c5 (3.0Å) and moreover it belongs to the same species.

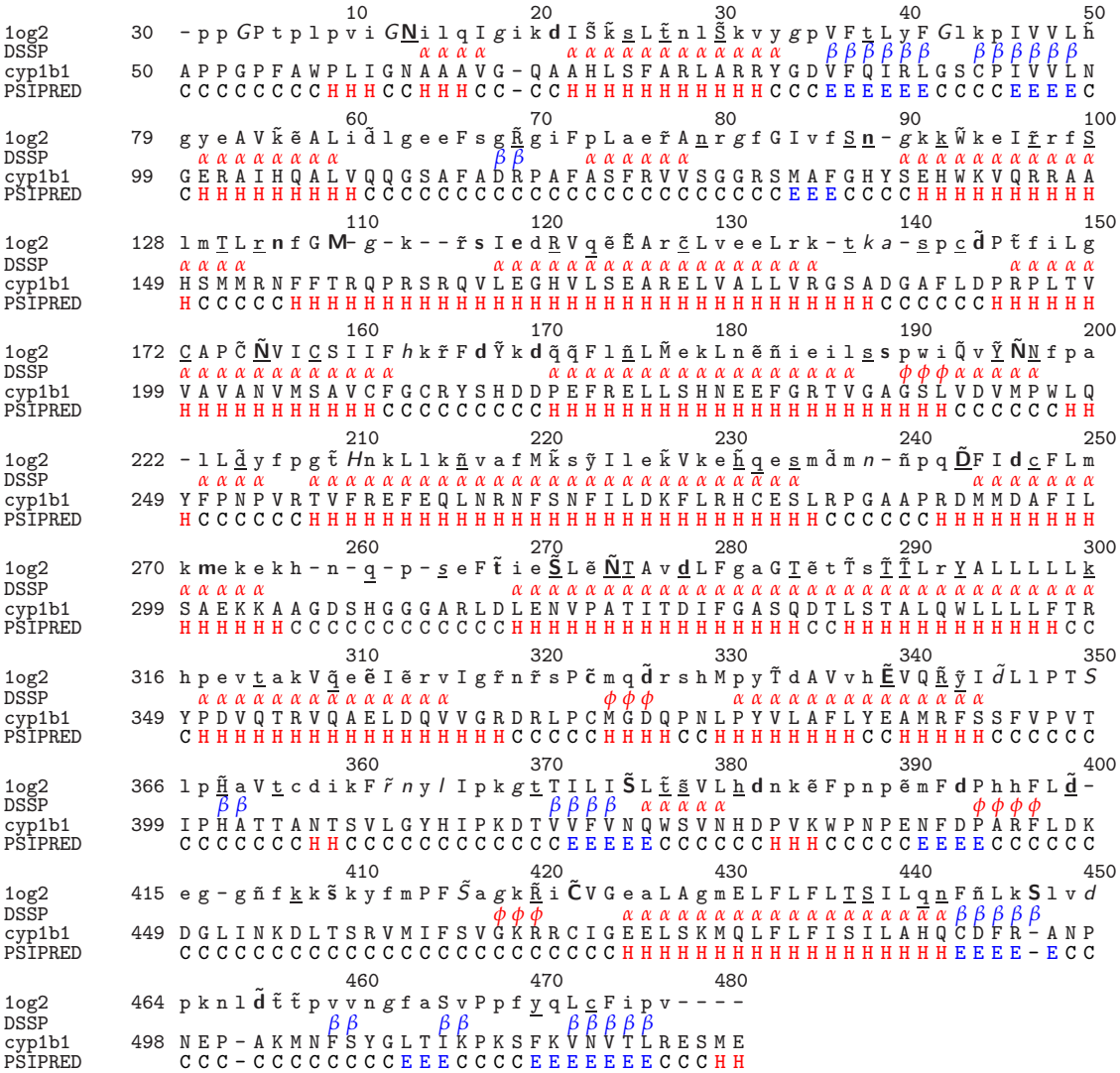
2.3.2.2 Target-Template Sequence Alignment

The CLUSTALW (Thompson *et al.*, 1994) sequence alignment program was used for aligning the target and template sequences. The sequence alignment in comparative modeling is the most crucial step, since the accuracy of the model largely depends on this step. Misalignment in the structurally conserved regions invariably leads to wrongly modeled protein. Further, the accuracy of the conformation of functionally important regions like the active site region depends largely on the accuracy of alignment at those regions. Thus, in many instances manual adjustment of the sequence alignment becomes necessary. The predicted secondary structural information for the sequence will be very useful in improving the sequence alignment.

The secondary structural information for CYP1b1, predicted using PSIPRED (McGuffin *et al.*, 2000), was used to manually adjust the alignment, such that care was taken to avoid gaps in the regular secondary structural regions. Care was taken to check that the predicted secondary structural regions of the target and the secondary structural regions of the template are largely aligned. The sequence of CYP1b1 corresponding to residues 1-49 and 528-543, which has no template information was excluded form the alignment. The final alignment that was used for modeling is shown in Figure 2.5.

2.3.2.3 Comparative Modeling

MODELLER generated about 20 models during the modeling process. The best model which corresponds to the least Molecular Probability Density function was selected for further studies. Figure 2.6 gives a projection diagram of the model in ribbon representation. The locations of 8 PCG causing mutations are indicated in ball-and-stick form.



Key to the alignment

solvent inaccessible	UPPER CASE	X
solvent accessible	lower case	x
positive ϕ	*italic*	*x*
cis-peptide	breve	x̆
HB to other sidechain	tilde	x̃
HB to mainchain amide	**bold**	**x**
HB to mainchain carbonyl	underline	x̲
disulphide bond	cedilla	x̧

Figure 2.5: Sequence alignment of CYP1b1 (from residue 50-527) with 1OG2 template (from residue 30-490). The secondary structure assignment calculated by 'DSSP' for the template and predicted by 'PSIPRED' for the target is also indicated. This alignment was produced using JOY software. The key for residue structural environments for the template is given below the alignment.

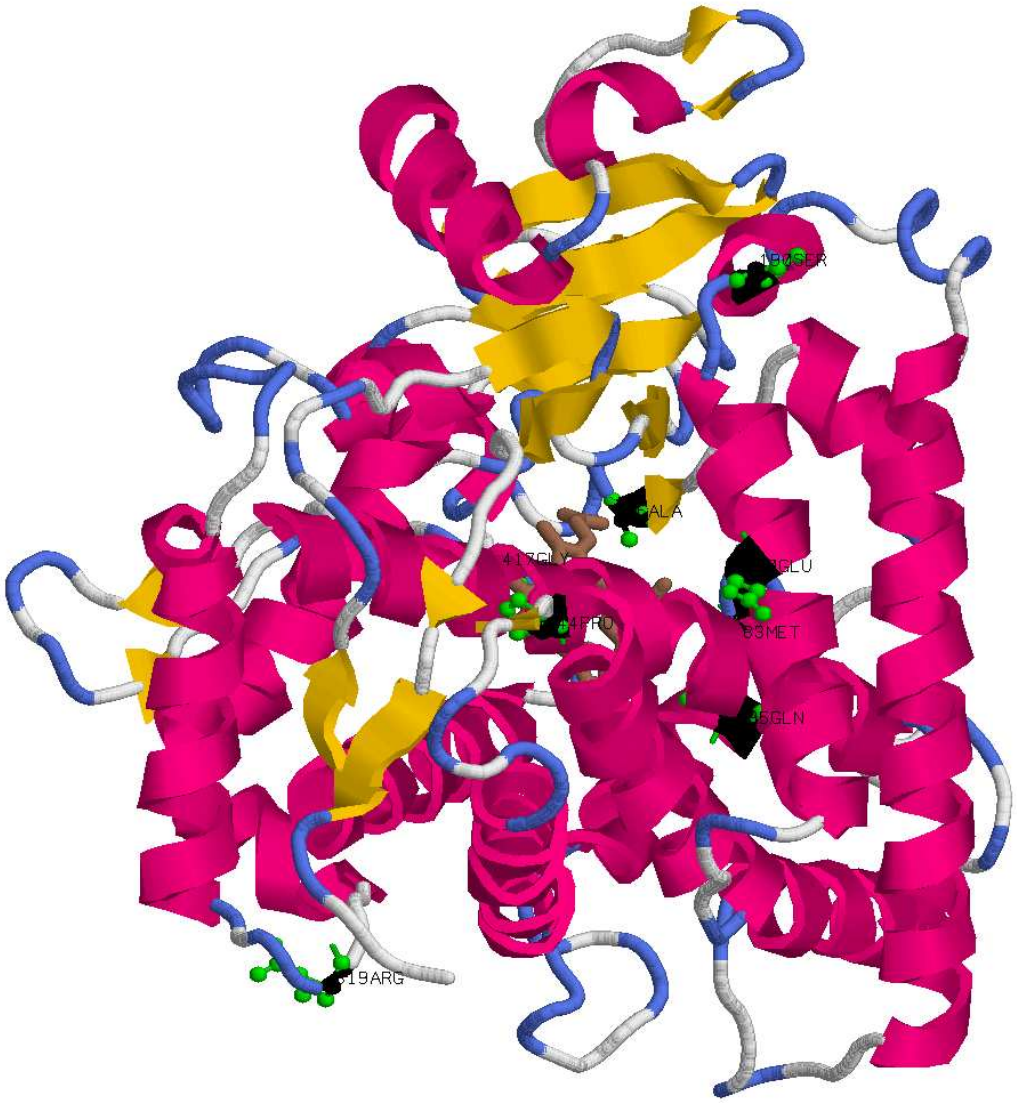


Figure 2.6: Projection diagram of the ribbon representation of CYP1b1 model. The PCG mutations are shown in ball-and-stick representation.

2.3.2.4 Quality Assessment of the Model

The results of PROCHECK analysis for this model is given in Table 2.10. The overall G-score for the model is -0.21 indicating a satisfactory stereo-chemical quality. It is worthwile to mention that a G-score higher than -0.5 indicates a good model. Furthermore, about 92% of the non-Glycine and non-Proline residues occur in the most favored regions of the Ramachandran Map (Table 2.10, Figure 2.7 and Table 2.11).

The Verify-3D plot for the model is shown in Figure 2.8. It could be seen that none of the residues in the model have negative scores, which means that the residue environments in the model are similar to the environment expected for this sequence in CYP fold. In modeling attempts using templates other than CYP2c9 including CYP2c5, the F/G loop region could not be modeled accurately. This is due to the fact that in CYP2c5 the structural information for F/G loop region is absent. On the other hand in CYP2c9 this region is well defined and furthermore, characterized by the presence of two unique short helices, F' and G' within the loop region, which is unique to Human CYPs. Hence, this region (residues 245 to 255) could be built correctly in the model as evident from the positive environment scores of VERIFY 3D. Furthermore, the 3D model obtained from MODELLER was further subjected to refinement by a combination of EM and short MD to relieve stereo-chemical strains.

2.3.3 Mapping of PCG mutations onto CYP1b1 Model

The model thus built was used for the identification of the three FIRs as discussed in the Materials and Methods section. The projection diagrams highlighting the three FIR regions are shown in Figure 2.9. The residues belonging to these FIRs are expected to be directly associated with the functions of heme binding, substrate binding and the access of the substrate to the active site respectively. Thus, mutations in any of

Table 2.10: Summary of PROCHECK G-Scores

Parameter	**Score**	**Average Score**
Dihedral angles:-		
Phi-psi distribution	0.09	
Chi1-chi2 distribution	-0.39	
Chi1 only	-0.14	
Chi3 & chi4	0.52	
Omega	-0.30	——
		-0.09
		——
Main-chain covalent forces:-		
Main-chain bond lengths	-0.13	
Main-chain bond angles	-0.67	——
		-0.44
		——
OVERALL AVERAGE		-0.21
		——

Table 2.11: Summary of the Ramachandran plot

Residues in most favoured regions	[A,B,L]	381	91.6%
Residues in additional allowed regions	[a,b,l,p]	31	7.5%
Residues in generously allowed regions	[a, b, l, p]	4	1.0%
Residues in disallowed regions	[XX]	0	0.0%
		———	———-
Number of non-glycine and non-proline residues		416	100.0%
Number of end-residues (excl. Gly and Pro)		3	
Number of glycine residues		30	
Number of proline residues		30	
		———-	
Total number of residues		479	

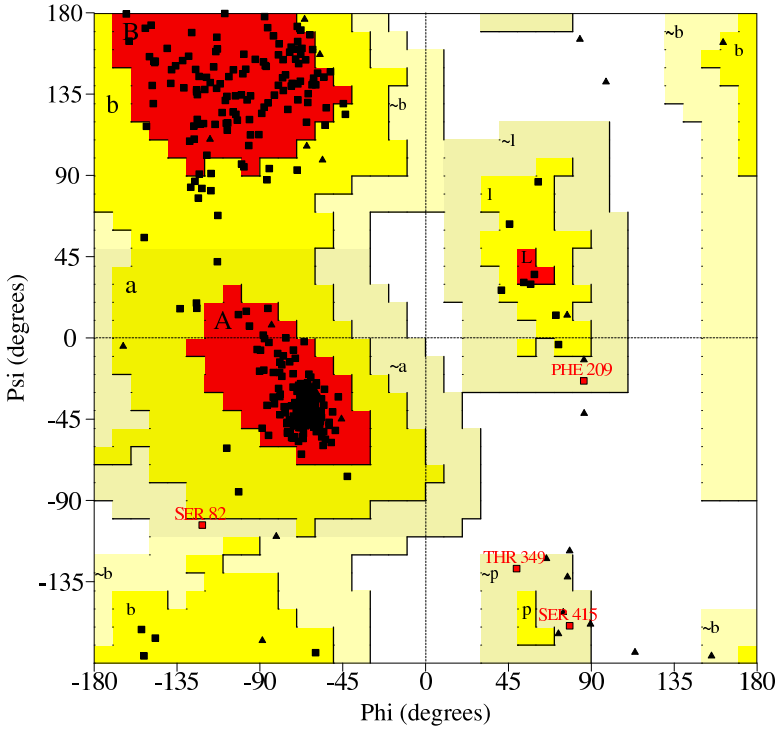


Figure 2.7: Ramachandran Plot

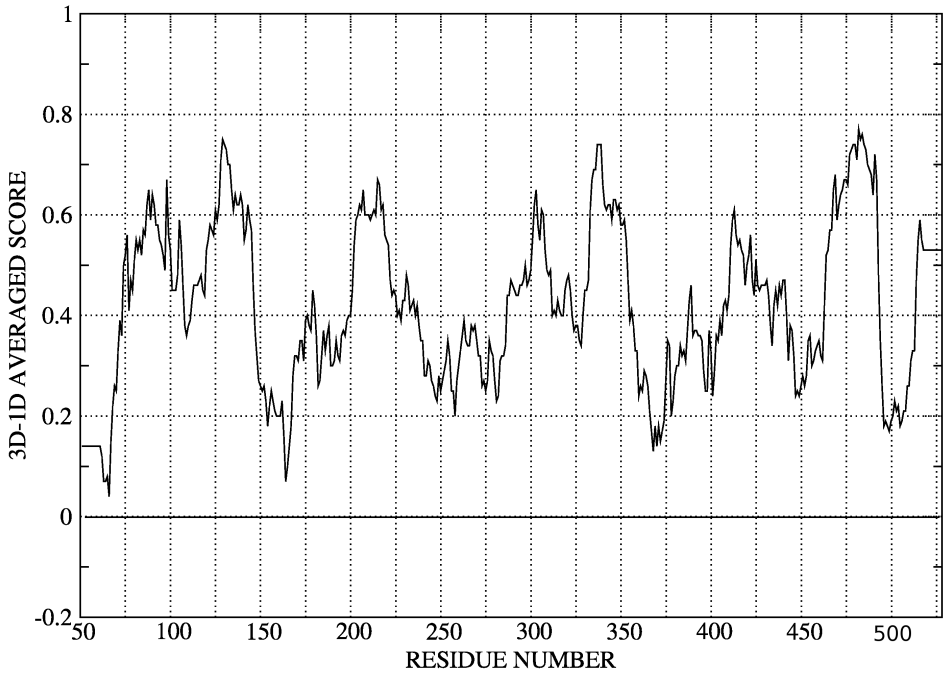


Figure 2.8: Verify-3D Environment plot

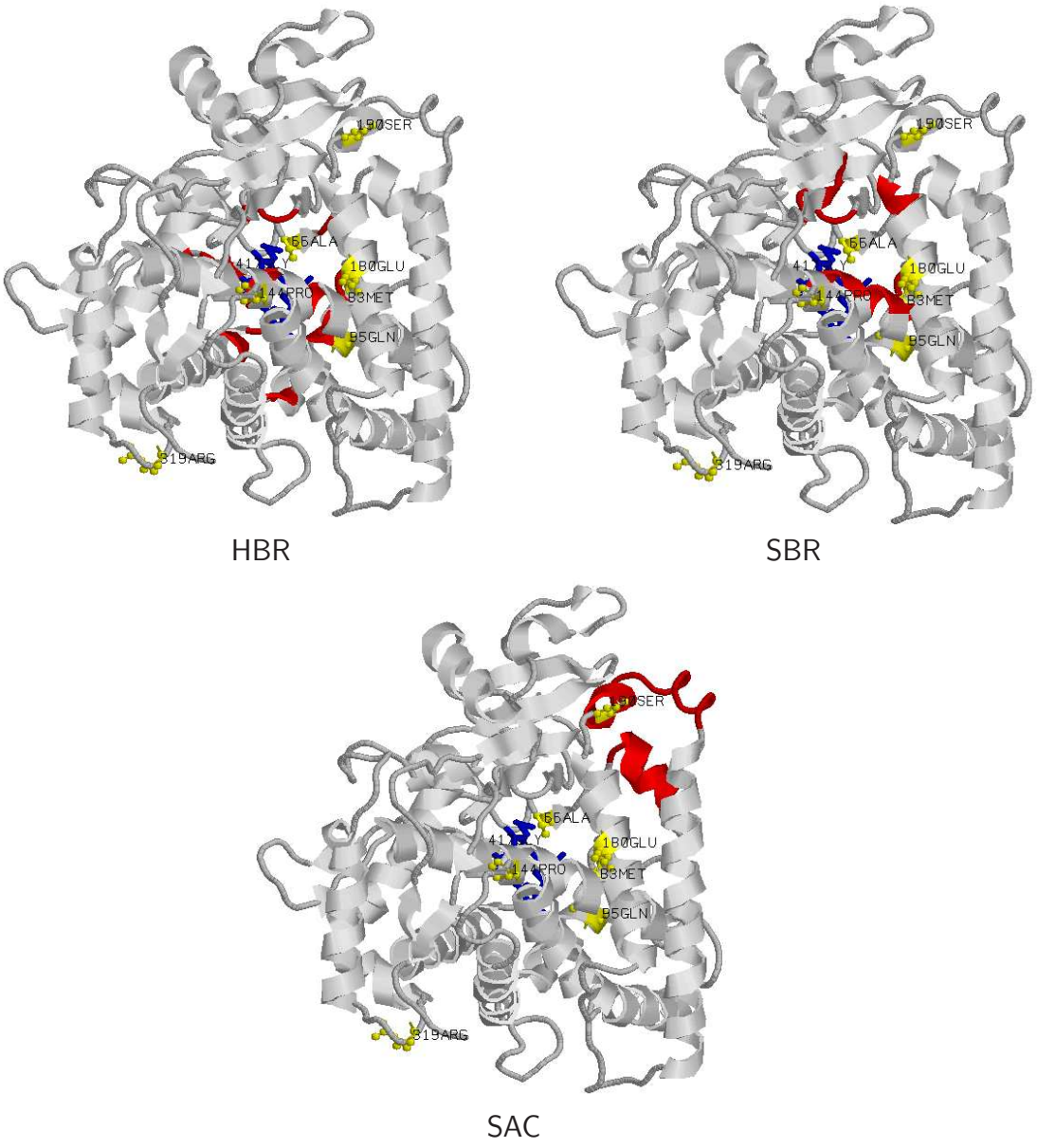


Figure 2.9: The three FIRs. A. Heme binding region B. Substrate binding region C. Substrate access channel. The mutation sites are represented in Yellow, in ball and stick. The FIR regions are colored in red.

these residues can affect the native functionality associated with these regions. The mutation sites were studied in the WT with respect to the three FIRs viz., HBR, SBR and SAC. Table 2.12 gives the list of residues that comprise the FIRs. The structures in which these residues occur are also indicated. The minimum distances of any of the FIR residues from the mutation sites are also given. It is observed that some of the mutation sites are either among the residues comprising the FIRs or at a distance close to them, while the others are located far from any of the FIRs (Table 2.12). Mutations in spatially close sites can be expected to exert an influence on FIRs. A comparison of the WT and the MTs structures in the vicinity of the mutation sites gave rise to the following observations.

A115P

This mutation occurs within the B/B' loop at the C-terminal end of B-helix and is in close proximity to the HBR. 'A' to 'P' substitution not only restricts conformational freedom at the site but also precludes Hydrogen bonding interaction capability due to the absence of the amide hydrogen.

M132R

This mutation site is in the HBR, in the loop connecting B and C helices. In WT the side chain of 'M' extends into the interior and packs well between I-helix (residue N319 to Y349) and one of the heme propionate groups. The 'N' of M132 donates a HB to D326 of I helix and accepts HB from G135, W141 and R468, thus interconnecting I helix, C helix and heme binding loop. Mutation to 'R' leads to congestion in the packing in the immediate vicinity of the mutation site thereby potentially harming the HB interactions.

Table 2.12: The range of distances (in Å) at which the residues in the three FIRs are located from the mutation sites. The minimum distances of the mutation sites from the FIR are given. The residues and the corresponding secondary structures in which they occur in each FIR are also given.

FIR	Structures involved	Comprising Residues	Mut Site	Min dist
HBR	**B/B' loop**	R117	115	6.17
	B'/C loop	M132, A133	132	0.00
	C helix	R145, M152	144	3.77
	I helix	I327, A330, S331, T334, L335, A338	193	12.06
	K helix	M389		
	B-sheet S1	V395, T398, I399	239	17.83
	HBL	I462, F463, S464, R468, C470, I471, G472	466	5.58
	L helix	L475, S476	368	21.23
SBR	**B/C loop**	S131, A133, F134	115	11.53
	F helix	V235, G236	229	9.75
			239	8.59
	I helix	D326, G329, A330, D333, T334	193	13.69
	B-sheet S1	V395, P396, T398, I399	466	12.32
	B-sheet S3	G508, L509, T510		
SAC	**B' helix**	F120, A121, S122, F123, R124, V125, V126	115	16.34
	F/G loop	L240, V241, D242, V243, M244, P245, W246, L247, Q248, Y249, F250, P251, N252, P253, V254	193	26.17

Q144P

This mutation occurs in the middle of C-helix and is also very close to the HBR. Mutation to 'P', a helix-breaking residue, can potentially break the continuity of the C-helix. The side chain of WT residue 'Q' is in a polar environment surrounded by the side chains of residues S131, H140, R314 and N319 and makes HBs with W141 and A147. Apart from this, 'Q' makes interaction with R314, in the N term of I helix, N319, T323 of I helix and S131 of B'/C loop, thus this site is located at the crucial junction interconnecting helix I, helix C, and the B'/C helix.

P193L

This site is far away from any of the FIRs and is in the N-cap position of E-helix. Replacement of 'P' which is a better N-cap residue as compared to 'L' can affect stability of the E-helix, thereby affecting the packing in this region.

E229K

This mutation occurs in the C terminal of the F-helix in the vicinity of the SBR. Substitution of 'E' to 'K' leads to a change from negatively charged residue to positively charged side chain and this in turn affects the local charge distribution. The mutation disrupts an important cluster of salt-bridges. In WT, the ionic bonds, R194:E229, R194:D333 and D333:K512 form a triangle of interactions, holding I helix with F helix and β-strand $S^{3.2}$. Upon mutation, the R194:E229 interaction is lost and this has potential to destabilize the other ionic interactions.

S239R

The position 239 is located in the F/G loop region and is an exposed site and thus may not have severe consequences. Nevertheless, since the region is close to the SAC, mutation of 'S' to 'R' could affect the structure and dynamics of this region. Nevertheless, since the region is close to the SAC, mutation of 'S' to 'R' could affect the structure and dynamics of this region. 'S' makes main chain HBs with T234, A237 at the C term of F helix. While its side chain makes a HB with T256, and T234.

R368H

This site is in between the J and K helices, in an exposed loop and the consequence of this change is not immediately apparent, except that positively charged 'R' is replaced by 'H' whose charge state depends on its protonation state. In WT, R368 interacts with G365, D367, V363, Q362 and D374. Upon mutation, the bonds to D367 and D374 are weakened. How this affects the conformation and functionality of the protein is not clear.

G466D

This is located in the middle of the heme-binding loop (HBL) and is a site involved in the interaction with reductase protein. Presence of 'G' is completely conserved in all CYP proteins at this position. From the conformational point of view 'G' is found in a turn and is accommodated well in the limited volume between HBL and B-helix. This site being a major docking area of the reductase protein, the mutation from 'G' to an acidic residue 'D' may be undesirable for the reductase binding since neutral/basic residues are preferred in this reductase binding region. Presence of 'G' is completely conserved in all CYP proteins at this position.

The above observations reveal the specific effects of mutations on the local interactions in the protein. This supports the observations made in an earlier study, wherein the variable prognosis of some of the PCG mutations (A115P, M132R, Q144P, S239R and G466D) was hypothesized to be due to the specific effects of each mutation on CYP1b1 protein (Reddy *et al.*, 2004). From the mapping studies it is also clear that, some of the mutations are actually part of FIRs or spatially close to the residues comprising the FIRs, indicating their possible role in affecting function.

2.4 Conclusion

In this chapter, the comparative sequence analysis of CYPs including the residue conservation pattern and position wise entropies is presented. Analysis was done with respect to the disease causing mutaions reported for all CYPs and more specifically to PCG mutations. From this sequence analysis it became clear that the disease mutations are least represented, indicating their incompatibility with the structure. Some regions of the protein have lower entropy (for example, Heme binding region) than the rest, indicating their relative importance in the structure and function. Among the three Functionally important regions studied, the frequency of occurrence of deleterious mutations is highest in the HBR. This indicates that in many cases, deleterious mutations may affect the protein function by affecting Heme binding function.

As the structure for Human CYP1b1 is not yet available and also that the reported homology models were obtained using non Human templates, a 3D model of Human CYP1b1 was built using the best available template. The model was seen to be structurally complete and reasoned to be better compared to the models developed earlier using bacterial templates. The model was used to locate and study the mutation sites for their structural contexts. Three regions in the model were identified that are

associated with the functions of Heme binding, Substrate binding and Substrate access, into the enzyme's active site. Mapping of the mutations onto the structure revealed that some of the mutations are actually part of FIRs or spatially close to the residues comprising the FIRs, indicating their possible role in affecting function. A comparative study of the MT models indicated that the mutations lead to changes in the interactions with the local residues. Although the sequency analysis and mapping of the mutations does reveal potential effects of the mutations, it was not clear how they bring out deleterious changes to the structure. For this reason, detailed MD studies on the WT and the MTs were carried out, which is described in the next chapter.

3

Structural Analysis of the Wild type and Mutant CYP1b1 Proteins

3.1 Introduction

Studies presented in Chapter 2 pertaining to sequence analysis of CYP1b1 and its PCG associated MTs revealed that many of the disease mutations occur in the structurally conserved regions of the protein. Furthermore, the disease causing mutations are either rare or completely absent in the CYP super family, indicating that the disease mutations are not among the accepted mutations during the evolution. This indicates that the disease mutations are perhaps not conducive to structural and functional integrity of the protein.

Modeling of the CYP1b1 structure revealed that some of the mutations are located close to the FIRs of the protein, while other mutations are spatially far from the FIRs. The FIR regions are shown to be structurally important with average entropy lower (except for SAC) than other regions. Among the FIRs, the HBR was found to be more

conserved. The SAC region with entropy higher than even non-FIR regions, indicated higher sequence and structural flexibility for this region. These observations could reveal qualitatively, the possible deleterious nature of the specific mutations, especially those which are close to the FIRs. It is intriguing however, as how certain mutations which are spatially far from the FIRs can affect the enzyme's function. The quantitative aspects pertaining to the mechanism of effects of mutations that propagate into the protein structure, and which could be functionally disruptive, can be obtained through Molecular Dynamics (MD) simulation studies.

MD simulation has become an indispensable tool to address the structural and dynamic properties of proteins and other bio-molecules (Karplus & Kuriyan, 2005). Through MD simulations structural information could be obtained at atomistic detail, which is often difficult to get by experimental methods. It could be used to explain the observed structural or functional properties, or to predict the possible effects on structure and function, upon changes in the native protein sequence or its environment.

It was therefore tempting to carry out detailed studies, to explore possible impact of mutations on the structure and function of CYP1b1. The WT and MT models obtained by comparative modeling were subjected to long MD simulations with the intention of studying the time evolution as well as time averaged structural properties, especially of the FIRs to know the possible impact of these mutations on the protein structure and hence its function. In this chapter, the structure based analysis of CYP1b1 and the 8 PCG associated MTs using MD simulations is reported.

3.2 Material and Methods

3.2.1 MD Simulation setup

GROMACS (Lindahl *et al.*, 2001) program package adopting the GROMOS96 force field parameters was used for EM and MD simulations. All the simulations were done using explicit water as solvent. The initial structures were solvated (spc216 water representation) in a triclinic bounding box using periodic boundary conditions. The least distance from the box edge and the protein atoms was kept at 0.8 nm. The systems were neutralized with the addition of counterions (Cl^-) (3,3,4,3,3,5,4,2 and 2 nos., for WT, A115P, M132R, Q144P, P193L, E229K, S239R, R368H and G466D respectively). Finally there were 41820, 41821, 41823, 41824, 41825, 41828, 41836, 41820 and 41832 atoms respectively in WT, A115P, M132R, Q144P, P193L, E229K, S239R, R368H and G466D.

The solvated structures were energy minimized using the steepest descent method, terminating when maximum force is found smaller than 10 $KJ.M^{-1}.nm^{-1}$. The electrostatic interactions were calculated using the Particle-Mesh-Ewald summation method (Darden, 1993). The energy-minimized structures were subjected to position restrained dynamics for 25 ps, keeping the whole protein molecule fixed and allowing only the water molecules to move, so as to equilibrate water molecules with respect to the protein structure. The position restraints in GROMACS are applied using a variable potential as given in equation 3.1,

$$V_{pr}(r_i) = \frac{1}{2} k_{pr} |r_i - R_i|^2 \tag{3.1}$$

where V is the potential, r_i is the position of atom i, k is the force constant and R_i is the fixed reference position of atom i. This was followed by MD simulation of the

full system (protein and water) without any positional restraints. All the simulations were performed at constant temperature (300K) and pressure (1 bar), with a time step of 2 femto seconds. The non-bonded pair list was updated every 5 steps. During the simulation, constraints were applied on all bonds lengths and angles using the LINCS (Hess *et al.*, 1997) algorithm with parameters; LINCS-order-4 and LINCS-warn-angle=30^o. Coordinates were saved at regular time interval of every 1 ps.

3.2.2 Analysis of MD trajectories

3.2.2.1 RMSD

The Root mean square deviation of the C^{α} atoms is calculated using the function g_rms, with reference to the initial structure after a least squares fit as per the equation 3.2,

$$RMSD(t_1) = \left[\frac{1}{M}\sum_{i=1}^{N} m_i |r_i(t_1) - r_i(t_r)|^2\right]^{\frac{1}{2}} \tag{3.2}$$

where M is the total mass, m_i is the mass of atom i, $r_i(t_1)$ is position of atom i at time t and $r_i(t_r)$ is the position of atom i in the reference structure.

3.2.2.2 RMSF

Root mean square fluctuation of C^{α} atoms with reference to the average structure is calculated using the g_rmsf function, as per Equation 3.3, where, T is the total time of simulation x_{it} is the position of atom i at time t, x_{i0} is the average position of atom i.

$$RMSF(x_1) = \left[\frac{1}{T}\sum_{t=1}^{T} (x_{it} - x_{i0})^2\right]^{\frac{1}{2}} \tag{3.3}$$

3.2.2.3 Hydrogen Bonds

The HBs were computed using the g_hbond function. The criteria used for determining the existence of HB are; Donor-Acceptor distance $\leq$ 0.35nm and Hydrogen-Donor-Acceptor angle $\leq 30^{\circ}$. The g_hbond gives information on the number and distribution of all the HBs in the molecule, but not the individual HB details. Hence, the details of individual HB were calculated using the HBPLUS (McDonald & Thornton, 1994) (McDonald & Thornton, 1993) program. The geometric criteria used in the program for finding HBs are as follows. A minimum 90° angle between; Donor-Hydrogen-Acceptor, Hydrogen-Acceptor-Acceptor Antecedent and Donor-Acceptor-Acceptor Antecedent. A maximum distances of 3.9Å between the Donor-Acceptor atoms, 2.5Å between Hydrogen-Acceptor atoms and 3.0Å for salt bridges. Further a minimum covalent separation of 3 covalent bonds was adopted.

3.2.2.4 Secondary Structures and Other Structural properties

The secondary structural information was calculated using DSSP program (Kabsch & Sander, 2004). The volume of SBR was computed using the program POCKET (Levitt & Banaszak, 1992) with a water probe radius of 1.4Å. The solvent accessible areas and radii of gyration were computed using the g_sas and g_gyrate modules. All the average properties including HB interactions were computed after the time of equilibration of the systems. In addition to the inbuilt GROMACS modules, the SWISS-PDB VIEWER (Guex & Peitsch, 1997) and PSA (Mizuguchi *et al.*, 1998) programs were used for structural superimposition and solvent accessibility calculations respectively. The figures were produced using XMGRACE and inhouse developed programs.

3.3 Results and Discussion

The WT and the MT structures were subjected to MD simulations for 30 nano seconds (ns). The instantaneous structures (snapshots) saved at every successive 1 pico second (ps) from the start of the simulation, totaling 30,000 for each structure, were used for analysis of time evolution of various structural properties analyzed in this study.

3.3.1 Trajectories of various structural properties

Figures 3.1 and 3.2 give the trajectories of Potential and Kinetic energies in the WT and MT simulations. It could be seen that the Potential energy of the system decreases and becomes stable after about 2ns. The average Potential energy remains stable throughout the simulation. Similarly, the Kinetic energy of the system becomes stable after about 2ns, indicating the stability of the simulations. The trajectories of the HBs are shown in Figure 3.3. The trajectories take longer times to attain stability. The average value is also varying during the simulation. The radius of gyration (Figure 3.4) of the protein in WT and MT simulations decrease steeply to a stable value in about 5ns. Thus, the structures become more compact during the simulation. The trajectories of the solvent accessible surface areas (SASA) (Figure 3.5) reveal that the SASA values also decrease from initial high values to stable values after 5ns. The decrease in SASA values correlate with decrease in the radius of gyration and hence further supports that the structures become compact.

3.3.1.1 RMSD trajectories

The overall structural deviation in terms of RMSD, of the WT and the MTs from their starting structures during the entire course of simulations is shown in Figure 3.6. It should be noted that the starting structures of WT and MTs were identical except for

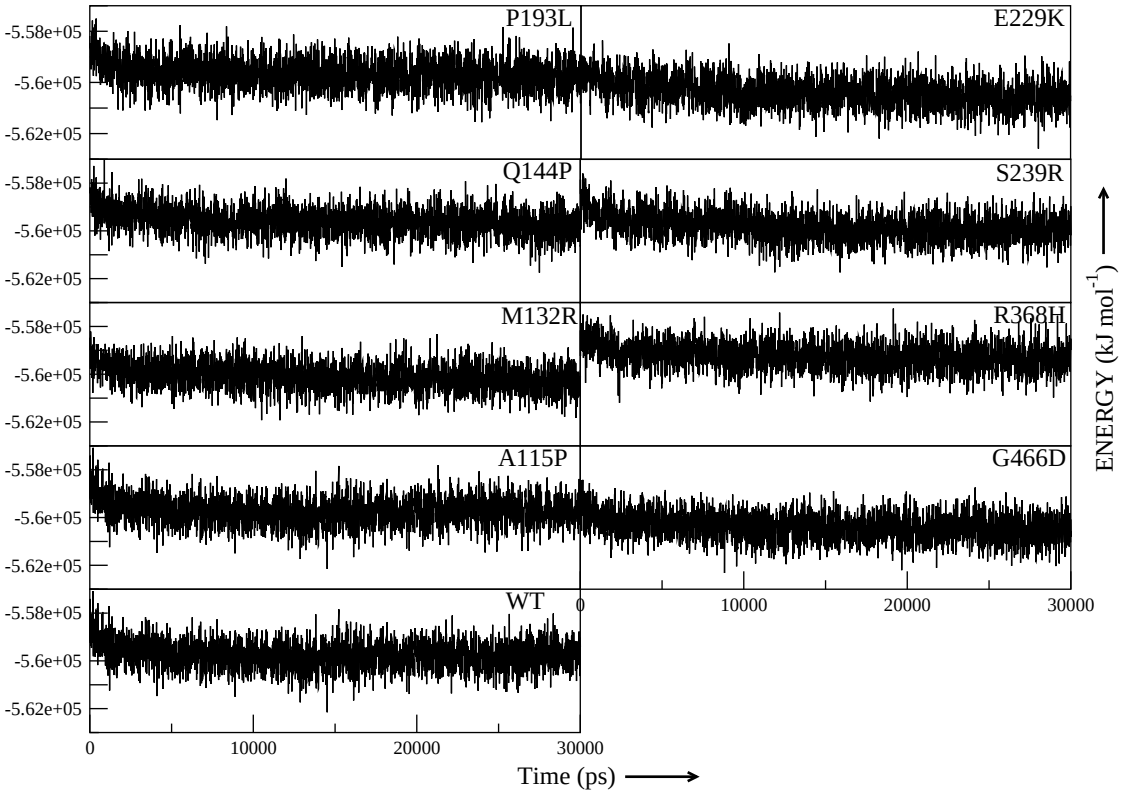


Figure 3.1: Trajectories of Potential Energy

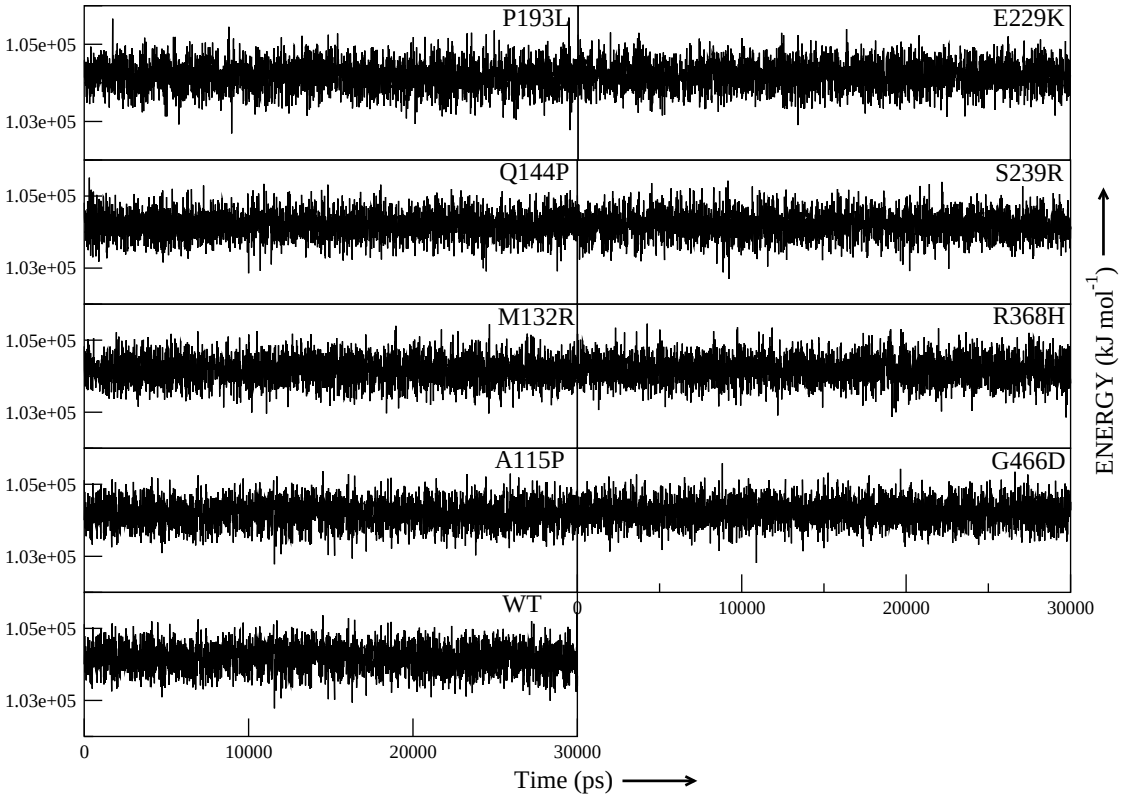


Figure 3.2: Trajectories of Kinetic Energy

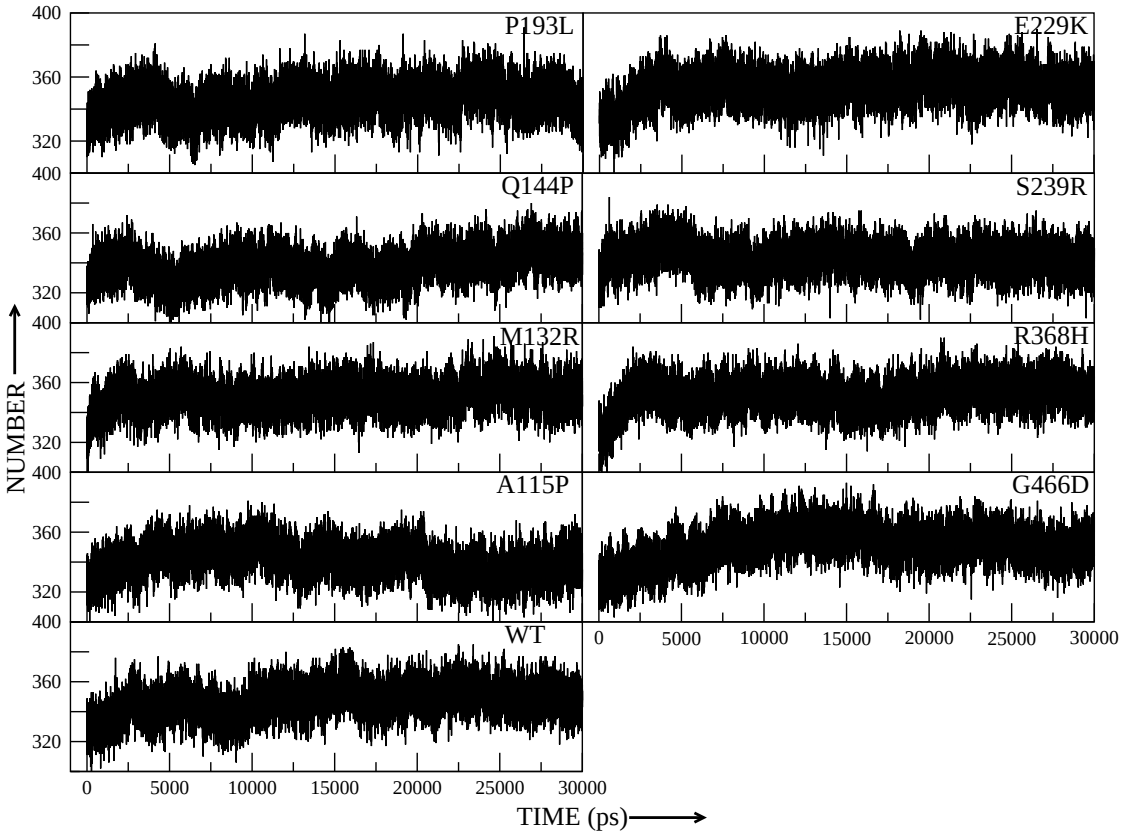


Figure 3.3: Trajectories of Number of Hydrogen Bonds

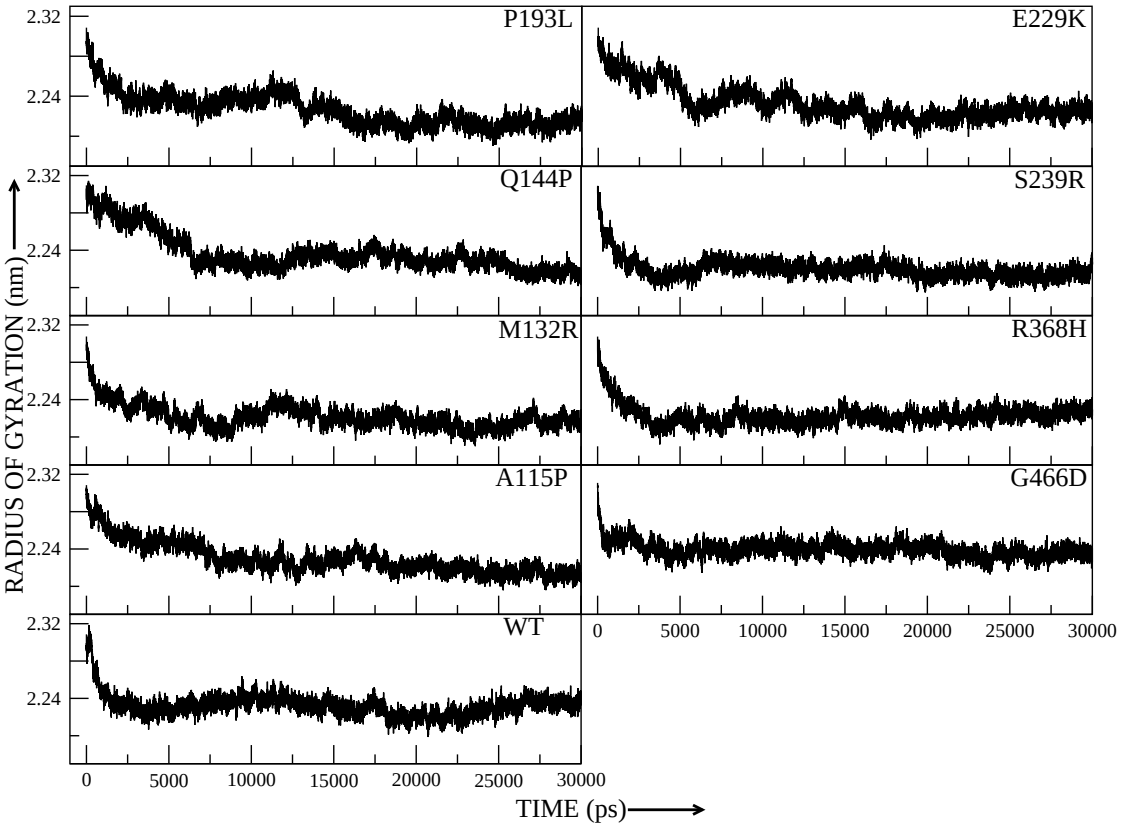


Figure 3.4: Trajectories of Radius of Gyration

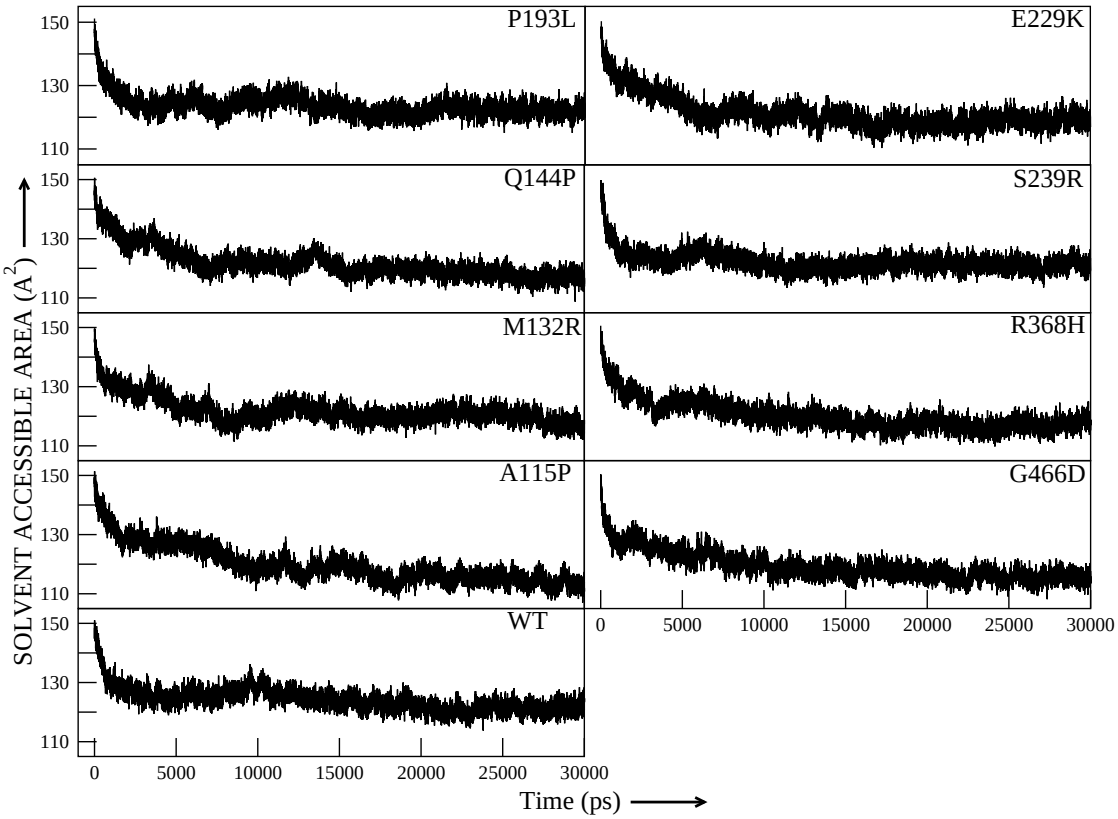


Figure 3.5: Trajectories of Solvent accessible Surface Area

the side chains of the residues at the mutation sites. The all-C^{α} trajectories shown in Figure 3.6 do not take precisely the same course; all the structures initially evolve rapidly during the first 1.5 to 2 ns and seem to stabilize at different times. While the all-C^{α} RMSD trajectories take long times for stabilization, the trajectories of the secondary structures (shown in gray in Figure 3.6) quickly stabilize in all the structures. It may be noted that some of the all-C^{α} RMSD trajectories seem to stabilize for short periods before drifting to stable structures. For example in the case of A115P and WT, after an initial increase, the RMSDs seem to stabilize for some time ($\approx$5 ns), followed by a gradual drift into another state. This behaviour is apperently due to some loop regions that undergo slow conformational transitions, during which time the RMSD trajectories show slow drift-like variation. This was revealed when the RMSD trajectory of the WT simulation is studied in detail.

The initial 20ns portion of the WT trajectory was examined during which time there is a marked increase in the RMSD. The graph marked A is the RMSD trajectory calculated for all-C^{α} atoms. A quick look at the figure gives an impression of gradual drift of the RMSD value, indicating that the system has not attained equilibrium until 13 ns. To check whether the drift in RMSD trajectory is due to gradual changes occurring globally or due to changes occurring in certain local regions of the protein, the contribution to total RMSD from individual structures in the protein was examined. Thus initially the C^{α} RMSD trajectory was calculated just for the protein core comprising only the secondary structures (SSTs). Figure 3.7B shows the RMSD trajectory of the protein core. It can be seen from the figure that the trajectory reaches a plateau at about 3000 ps and remains stable until the end of the simulation indicating that the protein core reaches equilibrated state at about 3000ns.

Further, the variation in the C^{α} trajectory, occurring upon inclusion of the loops in the RMSD calculation, was examined. It was found that, of the 24 loops only one

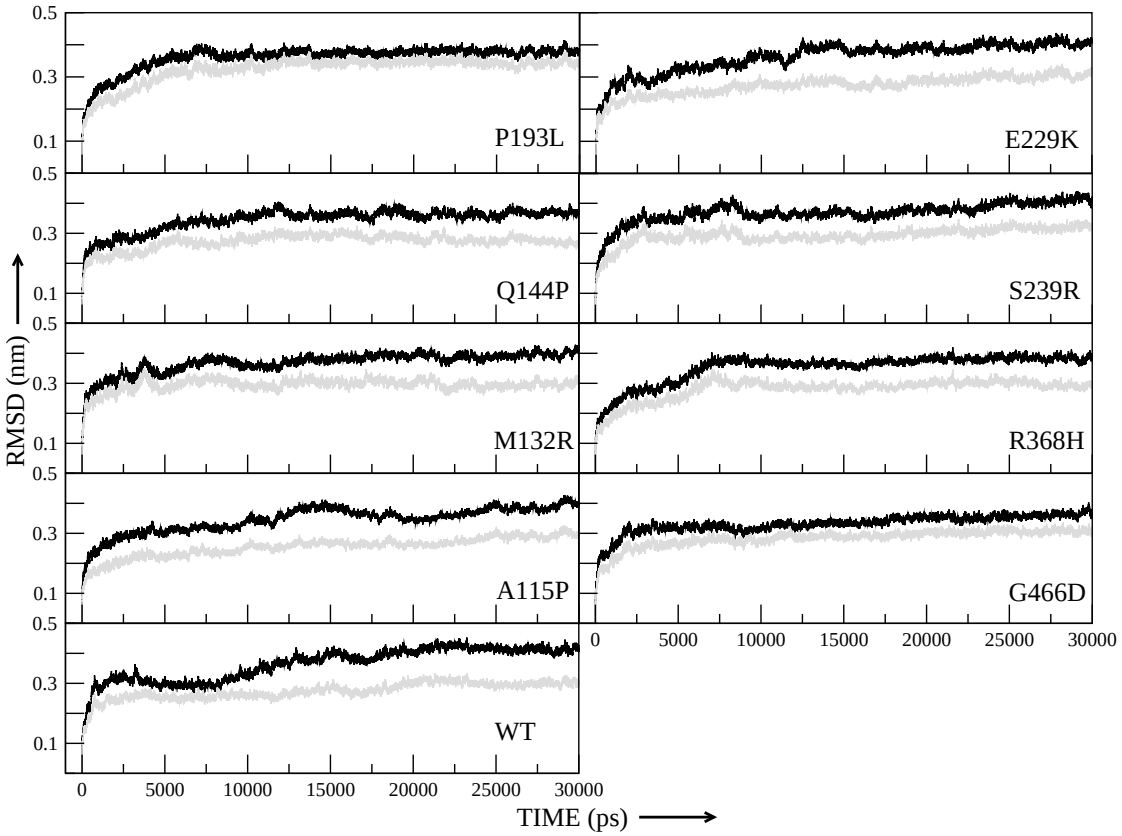


Figure 3.6: Overall RMSD trajecoties of C^{α} atoms from their starting structures. Gray lines indicate the C^{α} RMSD trajectories calculated for only the secondary structures.

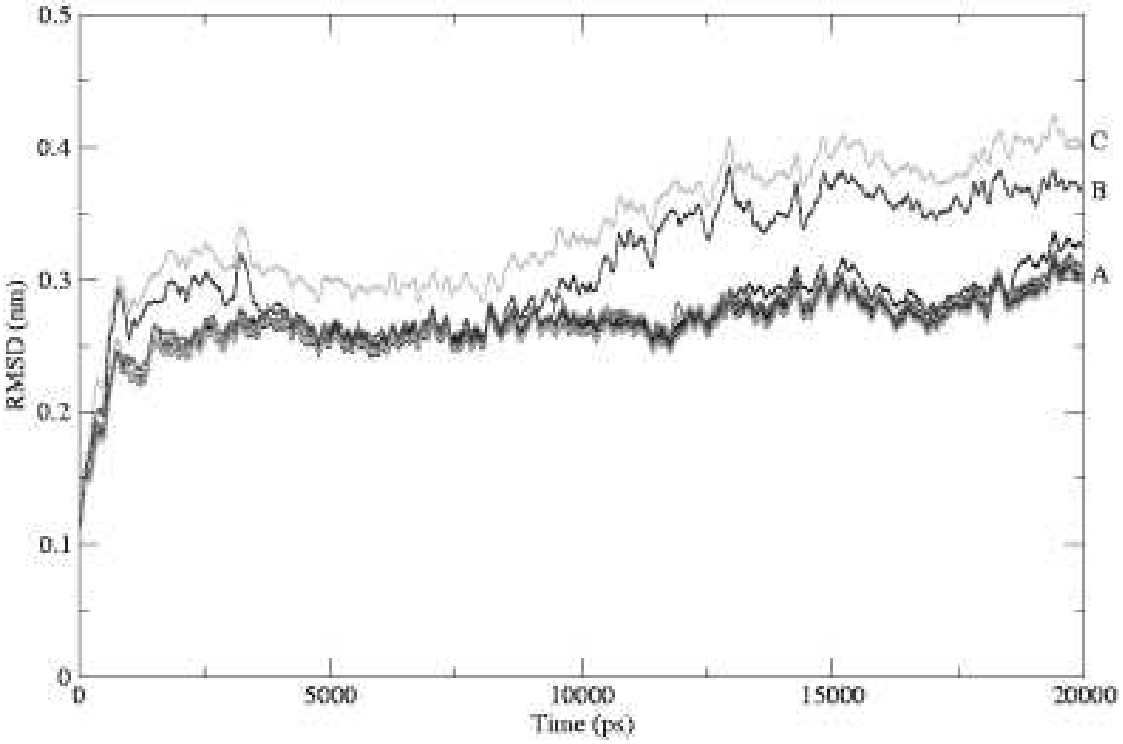


Figure 3.7: Trajectories of (a) all-C^{α} RMSD (b) SST's RMSD and (c) Trajectories of SST's RMSD (marked as A), the combined RMSD trajectories of SSTs and H/I loop (marked as B) and the all-C^{α} RMSD trajectory (marked as C). The combined trajectories of SSTs together with various loops other than H/I loop are similar to the SST's RMSD trajectory and are thus not distinct. All the plots are averaged over a running window of 100

loop between the H and I helices (loop H/I) spanning the residues L303-E318 gives rise to a trajectory very similar to the all-C^{α} RMSD trajectory (Figure 3.7C). When the correlation between each of these trajectories and the all-C^{α} RMSD trajectory were computed, only the trajectory upon addition of H/I loop, distinctly gave the highest correlation (r=0.96) with the all-C^{α} RMSD. The correlations were significantly far less in the case of other loops. Figure 3.8 shows the snapshots of the conformation of H/I loop taken after every nanosecond. Three distinct phases of the loop conformations become evident from this figure; the conformation, which is similar to the initial structure, occurs until 8^{th} ns, then the loop opens up into an elongated structure in the next 2ns. This conformation is maintained till 13^{th}ns. Again the loop folds rapidly in the 14^{th}ns and remains in the same conformation till the end of the simulation.

The loop residues were examined for their dihedral angle transitions happening throughout the simulation. L303 and A304, which were in α-R conformation during the first 10ns change to β-R conformation. A305 changes from β-R conformation to α-L after 13ns. Residues D307 and H309 gradually change from β-R to α-R conformations between 9 to 13ns. Frequent dihedral transitions occur in the three 'G' residues (310-312), in which the residues takes conformations in all the four quadrants of the Ramachandran plot. The C-terminal residues L315 to E318, which are adjoining I helix, are stable with respect to their dihedral angles. The flanking helices are maintained fairly stable. However, as seen earlier (Figure 3.1), the Potential energy of the system decreases and reaches a stable state by about 2 ns, indicating that the conformational transitions occurring in the loop after 2 ns do not drastically change the total potential energy and the protein transits from one local minimum to another having similar energies. Thus, the apparent non-stabilization of the all-C^{α} RMSD trajectories in MD simulations do not necessarily mean that the system has not reached the equilibrium state, rather they may reflect slow iso-energy conformational changes happening in

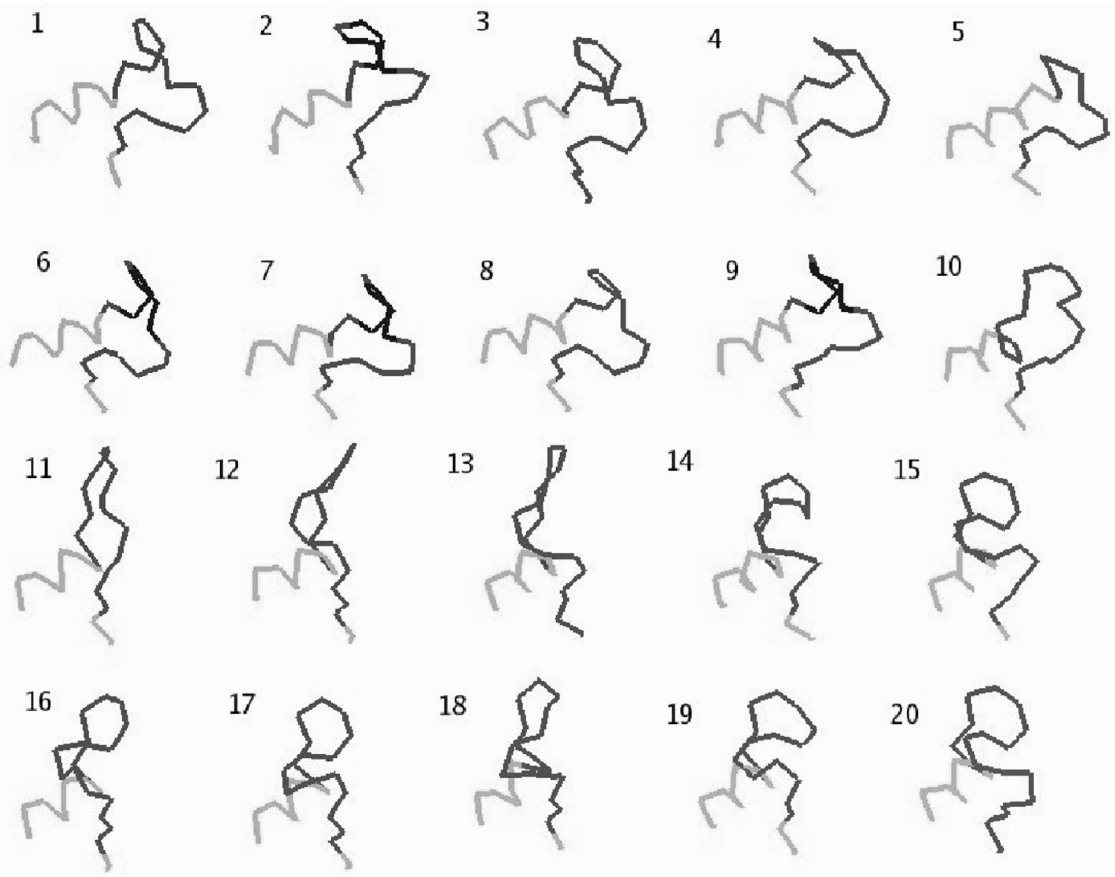


Figure 3.8: Influence of loops on the all-C^{α} RMSD. Conformation of the H-I loop structure during successive 1ns time intervals of the simulation. The figure shows one stable conformation occurring between 2 to 9 ns and another between 14 to 20 ns. The flanking H and I helices are partly shown

some of the loop structures, as it is shown in this case.

Owing to the fact that structures simulated are from comparative modeling, the early parts of simulation were discarded and only the data after RMSD stabilization was considered for further analysis. The average value of the all-C^{α} RMSD after stabilization is about 4.1Å in WT, whereas in MTs it ranges from 3.4Å to 3.9Å (Table 3.1). The average values of RMSD after stabilization for the secondary structures is about 2.9Å in WT and between 2.8Å to 3.4Å in the MTs. Higher RMSDs of the secondary structures in some MTs as compared to the WT reflect relative structural deviations occurring in the core of the MTs.

3.3.1.2 RMSF profile and Structural flexibility

The distribution of C^{α} root mean square fluctuation (RMSF) values is shown in Figure 3.9. In the WT, 54% of the residues have RMSF values less than 1Å (see Table 3.1). This could be considered as the basal level fluctuation for the WT protein. All the MTs, other than A115P have larger percentage of residues having RMSF values greater than 1Å as compared to the WT. Thus, on the whole, the MTs are characterized by increased protein flexibilities.

Similarly, the percentage of residues having RMSF greater than 1Å in the protein core, comprising of secondary structural elements are more in MTs (except A115P) as compared to WT, thus indicating that these MTs are associated with increase in flexibilities in some of the residues in the core region. The RMSF profiles in Figure 3.9 further illustrate residue-wise differences between WT and MTs, which are conspicuously evident in terms of peak heights that represent magnitudes of regional flexibilities. Comparison of these regional flexibilities further revealed characteristic decrease and increase in flexibilities in each of the MTs in certain loops, α-helices and β-sheets as given in Table 3.2.

Table 3.1: Average structural properties of WT and MT structures computed after RMSD stabilization.

	[t]TIME (ns)	[a]RMSD (Å)						[b]RMSF (Å)				[d]CYS-Fe (Å)		[e]SBR ($Å^3$)		[f]SAC (Å)		[g]HBs	
		All C^α		[c]SST C^α		heme		All C^α		SST C^α									
		μ	σ	μ	σ	μ	σ	$\leq$1Å	>1Å	$\leq$1Å	>1Å	μ	σ	μ	σ	μ	σ	μ	σ
WT	15	4.1	0.2	2.9	0.1	2.9	0.4	54	46	66	34	2.1	0.1	403	79	7.3	0.3	351	9.5
A115P	15	3.7	0.2	2.8	0.2	4.5	0.3	60	40	73	27	2.9	0.1	392	110	8.3	0.3	342	9.5
M132R	8	3.8	0.2	3.0	0.1	3.6	0.3	43	57	54	48	2.3	0.1	306	86	10.7	0.7	352	9.9
Q144P	12	3.6	0.1	2.8	0.1	3.6	0.4	49	51	60	40	2.5	0.1	308	69	9.6	0.8	341	10.8
P193L	7	3.8	0.1	3.4	0.1	2.3	0.3	41	59	53	47	3.3	0.4	610	172	7.9	1.2	347	10.0
E229K	15	3.9	0.1	2.9	0.2	2.6	0.4	53	47	61	39	2.1	0.1	377	152	9.9	1.0	358	9.6
S239R	8	3.8	0.2	3.0	0.2	2.6	0.4	35	65	43	57	1.9	0.1	355	106	7.7	0.4	355	9.7
R368H	8	3.8	0.1	3.0	0.1	3.6	1.1	41	59	54	46	2.3	0.3	254	83	10.5	0.6	356	10.1
G466D	5	3.4	0.2	2.9	0.1	2.3	0.3	51	49	61	39	2.2	0.1	747	295	9.2	0.4	342	9.5

[t]Time of RMSD stabilization. [a]Root Mean Square Deviation from the starting structure. [b]Root Mean Square fluctuation. [c]Secondary structural region. [d]Distance between CYS470-SG and Heme-FE. [e]Volume of the putative substrate binding region. [f]Size of the SAC opening. [g]Average number of Hydrogen bonds. For each property the mean (μ) and the standard deviations (σ) are given.

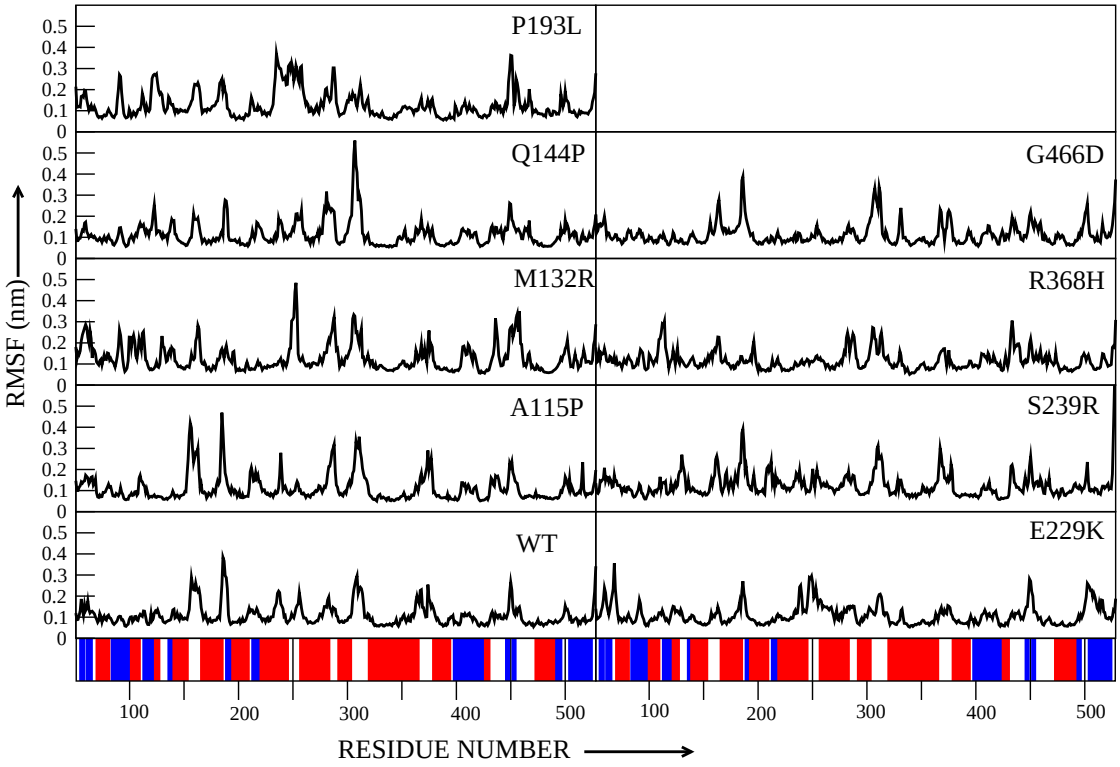


Figure 3.9: The residue wise RMSF profiles of C^{α} atoms of the WT and MT structures computed after stabilization of the RMSD trajectories. The bars at the bottom of the figure indicate the location of C^{α} -helices (red) and C^{β}-strands (blue).

Table 3.2: Regions (α-helices, β-sheets and loops) showing an average increase or decrease of RMSF in the MTs compared to the WT.

	INCREASE	DECREASE
A115P	D; $S^{4.2}$, $S^{3.2}$; C/D, B/B',E/F, G/H, H/I, J/K, K'/L'	B', F; D/E, K/K'
M132R	B, H; $S^{1.1}$, $S^{1.2}$, $S^{2.1}$, $S^{2.2}$, $S^{4.2}$; B/B', B'/C, G/H, H/I, K'/L', CHR, HBL, F/G	F; C/D, D/E, E/F, J/K, K/K'
Q144P	B', G, H; $S^{4.1}$; B/B', G/H, H/I, HBL	C/D, D/E, E/F, J/K
P193L	B', D, F, G; $S^{1.1}$, $S^{1.2}$, $S^{4.1}$, $S^{4.2}$; B/B', B'/C, F/G, G/H, L'/L, CHR, HBL	C/D, D/E, H/I, J/K, K/K'
E229K	G; $S^{4.1}$, $S^{3.1}$, $S^{3.2}$; F/G	J; C/D, D/E, E/F, H/I, J/K
S239R	D, E; $S^{5.1}$; B'/C, E/F, J/K, K'/L', HBL	C/D, D/E
R368H	B, B', E, K'; $S^{1.2}$; B/B', G/H, K'/L', HBL	F; C/D, D/E, E/F, J/K, L'/L
G466D	D, I, K'; H/I, K'/L'	B', F; C/D, D/E, E/F, L'/L

See Figure 3.13 for the definitions of α-helices and β-sheets. Loops are denoted by their flanking helices. HBL denotes Heme binding loop; CHR denotes Charge Relay Region. RMSF of a particular structure is taken to be increased or decreased if there is an average change in RMSF of greater than 0.3Å in at least 50% of its residues.

3.3.2 Structural changes at the mutation sites

As mentioned earlier substitutions of residues at the mutation sites are expected to bring out certain structural changes at the mutation sites. In order to know the changes brought out by the mutations, the average ϕ, ψ dihedral angles were computed and number of times a given HB (HB) interaction is observed during simulations (referred as the occupancy of HB interaction) involving the residues at the mutation site in WT and MTs, and the results are tabulated in Table 3.3. As can be seen from the table, some mutations are characterized by major conformational changes at the mutation site as compared to WT, while in some others not so severe changes have been observed. The MTs P193L, S239R, R368H and G466D are marked by major conformational changes as compared to WT, where ψ angles have changed by about 130°to 180°. Concomitantly these changes have brought out structural changes to the secondary structures associated with the mutation sites. Of these four sites only P193L occurs at the N-cap of the E-helix in WT, while the rest occur in loops. As a consequence of mutation at position 193, the N-cap is destabilized in MT leading to a shorter helix.

It is also worthwhile to mention that in the case of position Q144, which occurs in the middle of the C-helix in the WT, the mutation to P has led to deviation of ϕ, ψ angles from typical α(-R) helical to β(-R) region of Ramachandran map (Figure 3.3.2), thus destabilizing the N-terminal of the C-helix, leading to a shorter helix in MT (Figure 3.3.2). With regard to the HBing, absence or presence of HBs in the MTs has been determined by the absence or presence of functional groups capable of forming HB interactions. For example, in A115P, 'A' in WT is involved in two HBs, of which one is formed by the amide Hydrogen. 'P' in the MT is not involved in any such HBs. The C=O group of 'P' instead has HB with water molecule as it is on the

Table 3.3: The occupancies (in %) of HBs between heme and protein residues from structures surrounding heme, in the WT and MTs. Only occupancies $\geq$5% are shown.

	B/B'	B'/C		C		$S^{1.3}$			HBL							
	117	**135**	**137**	**141**	**145**	**398**	**399**	**401**	**464**	**465**	**466**	**467**	**468**	**469**	**470**	**471**
WT	48	-	-	84	73	31	48	-	41	-	17	10	94	22	-	-
A115P	24	-	-	-	25	-	-	-	-	-	97	-	-	46	-	-
M132R	-	-	-	72	18	-	-	55	-	-	48	10	67	46	-	-
Q144P	-	-	54	-	32	-	-	-	61	36	61	37	74	74	-	-
P193L	27	-	-	-	88	-	-	-	89	-	-	-	71	-	-	18
E229K	99	-	-	79	16	-	-	96	61	-	96	83	99	20	-	-
S239R	82	13	-	-	19	-	-	28	94	-	37	-	93	71	-	-
R368H	47	-	87	-	-	-	-	-	-	27	70	53	79	30	-	-
G466D	98	-	10	-	77	-	-	66	91	-	-	-	-	59	-	-

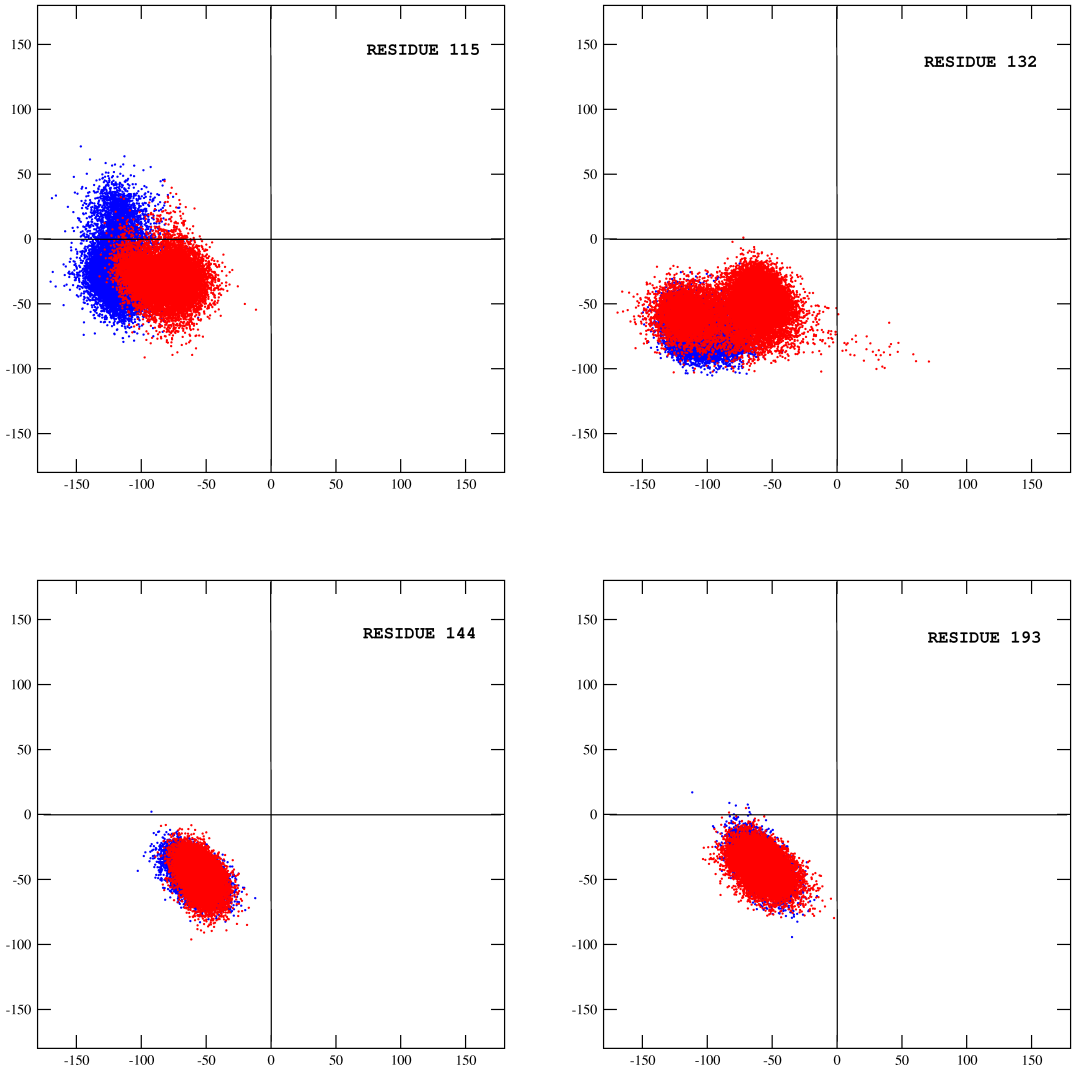


Figure 3.10: Variation of the ϕ, ψ dihedral angles at mutation sites during the simulation, in WT(blue) and MTs(red)

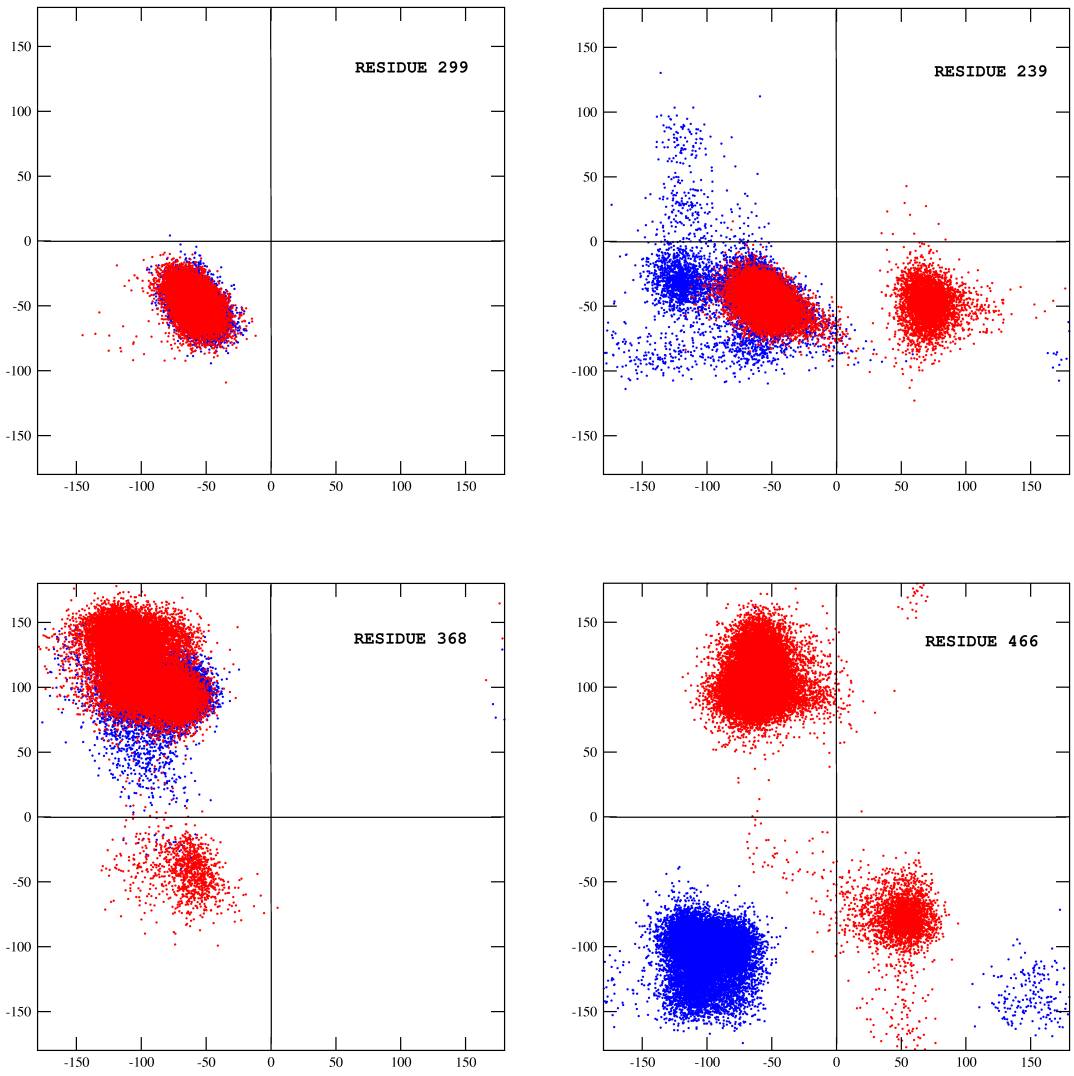


Figure 3.3.2 . . . continued

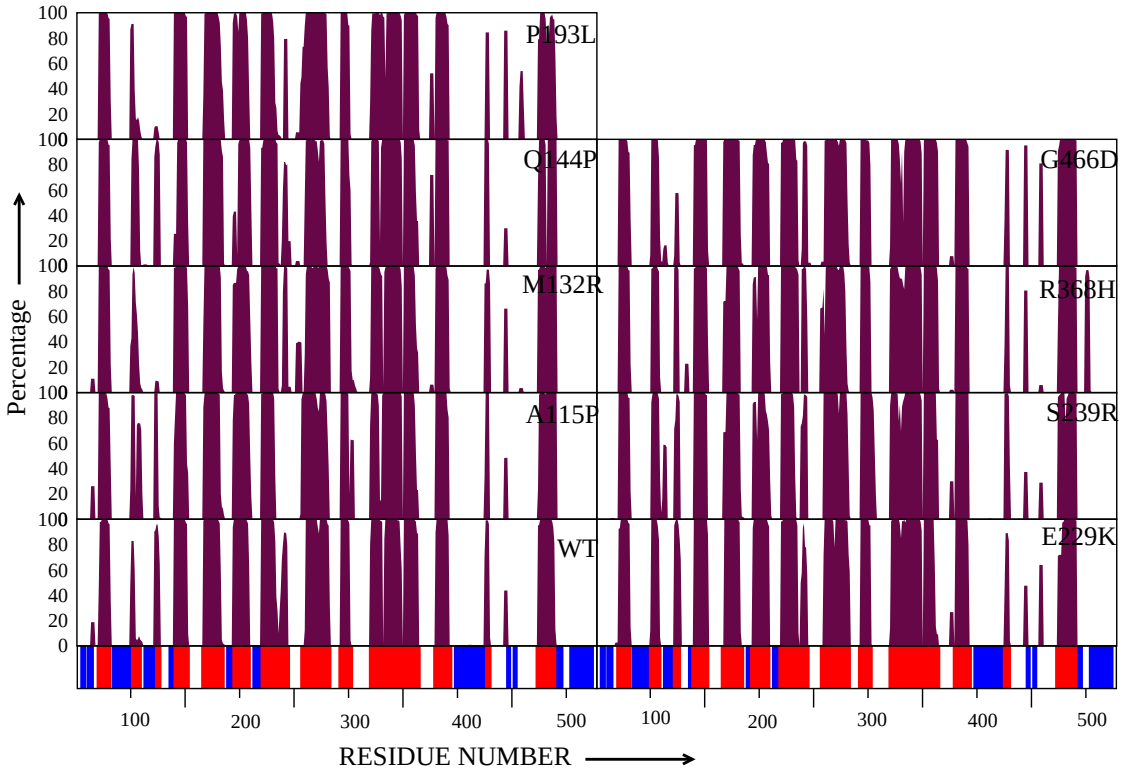


Figure 3.11: The Percentage of time of total simulation period during which the residues take α helical conformation. The shaded regions below the graphs indicate the helix and sheet regions in the WT structure before simulation.

surface. In some cases, MTs are characterized by the formation of more number of HB interactions. However, it should be noted that these additional HBs have low occupancies (see Table 3.3).

3.3.3 Structural properties of the FIRs

In the previous sections, it is was seen that MTs are characterized by changes in local interactions as well as global structural properties. It would be further interesting to study the changes in the structural properties with respect to the FIR regions. As discussed in Chapter 2, the FIR regions are directly associated with the function of the molecule and Mutations in these regions have a greater disease causing potential than the other regions. Thus the structural properties were analyzed in the context of the three FIRs.

3.3.3.1 Heme binding region

The heme moiety is held in place by hydrophobic and HB interactions, formed by the heme binding loop (HBL), C-helix, β-strand $S^{1,3}$, B/B and B/C loops (see Figure 3.13 for the definitions of secondary structures). The list of HB interactions with heme, and their occupancies in the WT and MTs during simulations is given in Table 3.5. In WT, the HBs between heme and its surrounding structures are present with about 50% or more occupancy. Comparatively, all the MTs show a reduction or complete absence of HB interactions from one or many of these structures, implying altered binding of the heme in all the MTs as compared to WT.

The decreased HB interactions in the case of MTs with the four structures surrounding the propionates correlate with their increased RMSF fluctuations in those structures (see Table 3.7). The positional alteration of the heme in MTs is also

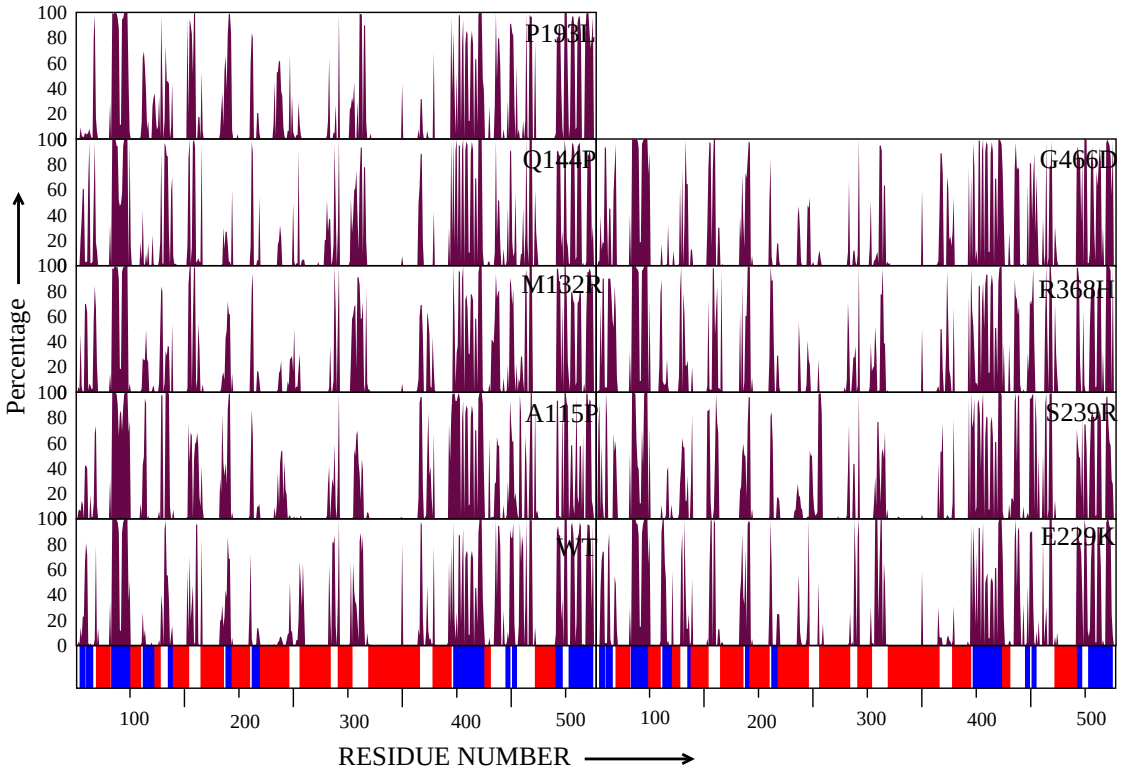


Figure 3.12: The Percentage of time of total simulation period during which the residues take β strand conformation. The shaded regions below the graphs indicate the helix and sheet regions in the WT structure before simulation.

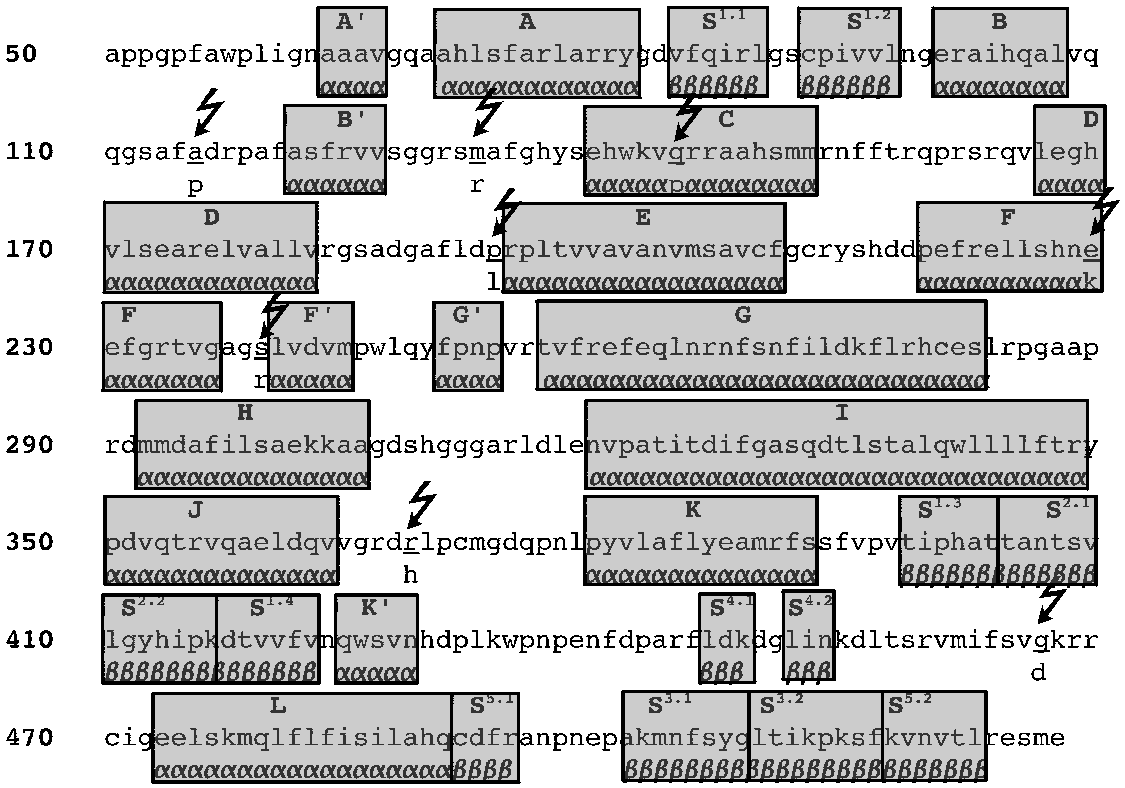


Figure 3.13: Location of Secondary structural regions in CYP1b1 sequence. The α-helices are indicated by upper case letters, A,B....L. The β-sheets are indicated by S with a superscript indicating the sheet and strand numbers. The locations of the loop structures are between the corresponding terminal secondary structure. For example the F/G loop comprises the region between F and G helices. The locations of mutation sites are indicated by the 'Lightning bolt' symbols.

Table 3.5: The average ϕ, ψ dihedral angle conformations, and the occupancies of the HB interactions, involving the mutation positions, in the WT and the MTs.

	WILDTYPE				MUTANT			
	ϕ	ψ	Hydrogen bonds		ϕ	ψ	Hydrogen bonds	
			Interaction	†Occ			Interaction	†Occ
A115P	-117	-23.5	[HBL]468ARG(NH1):::115ALA(O)[B/B]	91	-81.2	-30.8		
			[B/B]115ALA(N):::404THR(OG1)[S1.3]	75				
M132R	-106.7	-71.7	[B/C]184GLY(N):::132MET(O)[B/C]	57	-96.1	-47.2	[HBL]468ARG(NH1):::132ARG(O)[B/C]	74
			[HBL]468ARG(NH2):::132MET(O)[B/C]	31			[B/C]132ARG(N):::326ASP(OD1)[I]	34
			[B/C]132MET(N):::326ASP(OD1)[I]	15			[HBL]468ARG(NH2):::132ARG(O)[B/C]	31
			[B/C]132MET(N):::326ASP(OD2)[I]	10			[B/C]132ARG(N):::326ASP(OD2)[I]	22
							[B/C]132ARG(NE):::326ASP(OD1)[I]	16
							[B/C]132ARG(NH2):::144GLN(OE1)[C]	16
							[B/C]132ARG(NH2):::323THR(OG1)[I]	13
							[B/C]132ARG(NE):::322ALA(O)[I]	11
							[B/C]132ARG(NH1):::322ALA(O)[I]	11
							[B/C]132ARG(NE):::322ASP(OD2)[I]	9
							[B/C]132ARG(NH1):::323THR(OG1)[I]	9
Q144P	-58.5	-46.6	[C]148ALA(N):::144GLN(O)[B/C]	98	-75.8	-40.9	[E]197ALA(N):::144PRO(O)[C]	81
			[C]140GLN(N):::91-HIS(O)[C]	92				
			[I]323THR(OG1):::144GLN(OE1)[C]	26				
			[C]144GLN(NE2):::131SER(OG)[B/C]	15				
			[I]319ASN(ND2):::144GLN(OE1)[C]	15				
			[C]144GLN(NE2):::140HIS(O)[C]	14				
P193L	-58.2	-37.2	[E]197THR(N):::193PRO(O)[E]	96	-97.7	43.8	[E]197THR(N):::193LEU(O)[E]	91
			[E]196LEU(N):::193PRO(O)[E]	40			[E]193LEU(N):::516PHE(O)[S5.2]	72
							[E]196LEU(N):::193LEU(O)[E]	32
							[E]197THR(OG1):::193LEU(O)[E]	12
							[E]193LEU(N):::515SER(OG)[S3.2]	10
E229K	-57.6	-49.7	[F]229GLU(N):::225LEU(O)[F]	93	-83.8	4.4	[F]229LYS(N):::225LEU(O)[F]	95
			[F]233ARG(N):::229GLU(O)[F]	60			[F]233ARG(N):::229LYS(O)[F]	77
			[F]234THR(N):::229GLU(O)[F]	16			[F]229LYS(N):::226SER(O)[F]	14
			[F]229GLU(N):::226SER(O)[F]	15			[F]232GLY(N):::229LYS(O)[F]	9
S239R	-64.9	-40.5	[F]243VAL(N):::239SER(O)[F/G]	79	-77.9	103.4	[F]243VAL(N):::239ARG(O)[F/G]	94
			[F/G]239SER(N):::234THR(O)[F]	23			[F/G]239ARG(NH2):::234THR(O)[F]	40
			[F/G]239SER(N):::235VAL(O)[F]	20			[F/G]239ARG(NH2):::233ARG(O)[F]	26
			[F/G]242ASP(N):::239SER(O)[F/G]	14			[F/G]239ARG(NH2):::256THR(O)[G]	21
							[F/G]239ARG(NH1):::233ARG(O)[F]	16
							[F/G]239ARG(NE):::255ARG(O)[F/G]	15
							[F/G]242ASP(N):::239ARG(O)[F/G]	12
							[F/G]239ARG(NE):::234THR(O)[F]	11
							[F/G]239ARG(N):::236GLY(O)[F]	9
							[F/G]239ARG(NH1):::255ARG(O)[F/G]	9
R368H	-79.2	96	[J/K]365GLY(N):::368ARG(O)[J/K]	83	-113.3	-134.7	[J/K]368HIS(N):::365GLY(O)[J/K]	43
			[J/K]368ARG(NH1):::362GLN(O)[J]	14			[L]489HIS(NE2):::368HIS(O)[J/K]	40
							[J/K]365GLY(N):::368HIS(O)[J/K]	32
							[J/K]368HIS(NE2):::374ASP(OD1)[J/K]	9
							[J/K]368HIS(NE2):::374ASP(OD2)[J/K]	9
G466D	-98.1	-106	[HBL]469ARG(NE):::466GLY(O)[HBL]	30	-74.9	114.2	[HBL]469ARG(NH1):::466ASP(OD2)[HBL]	9
			[HBL]469ARG(NH1):::466GLY(O)[HBL]	17				
			[HBL]469ARG(NH2):::466GLY(O)[HBL]	15				

†Occ denotes the Hb occupancy in percentage. Only Hbs with occupancies of more than 10% are shown. For each Hb interaction, the secondary structures or loop regions involved in the interaction are indicated within square brackets.

indicated by the Fe-Cys co-ordination. Only in E229K, the co-ordination bond length is similar to that found in WT, whereas in all other MTs the co-ordination bond length is either shorter or longer than that of the WT, with P193L showing the highest deviation (Table 3.1). Further, the RMSD for heme was computed with respect to its starting position. The RMSD values (Table 3.1) indicate that the position of the heme undergoes more deviation in all the MTs except P193L and G466D, as compared to the WT.

3.3.3.2 Substrate binding region

The volume of the continuous void over the heme moiety and bounded by helices B', I, F and G, β-strand $S^{1.3}$ and B/C loop, was calcualted to represent the volume of the substrate binding region. Although the volume measured this way may be more than the actual volume occupied by any hypothetical substrate, it nevertheless gives a convenient representation of SBR for comparison purposes. In WT, on average, the volume is about 403Å^3 (see Table 3.1) and the value hardly fluctuates. The MTs R368H, M132R, Q144P, S239R, E229K and A115P show a reduction in the size of their SBRs whose volumes are respectively 63%, 76%, 76%, 88%, 94% and 97% of that of WT. In the case of P193L and G466D the volumes of SBR have increased to 151% and 185% respectively of that of the WT. This result indicates that the MTs are characterized by a distorted substrate-binding pocket, either smaller or larger than that found in the WT.

3.3.3.3 Substrate access channel

The catalytic site of CYP1b1 or the SBR is not directly accessible from the surface of the protein. Thermal motion pathway analysis (Ludemann & Wade, 1997) in p450cam had revealed the existence of three potential substrate entry/exit pathways

Table 3.7: Correlation coefficients (C_{ij}) between WT and MTs, for the HB occupancies and residue wise C^{α} RMSF profiles

WT vs. MT	HBs (C_{ij})	RMSF (C_{ij})
A115P	0.66	0.76
M132R	0.71	0.42
Q144P	0.55	0.66
P193L	0.73	0.56
E229K	0.61	0.50
S239R	0.57	0.72
R368H	0.58	0.50
G466D	0.76	0.68

(denoted by pw1, pw2 and pw3). Random expulsion MD ligand egress (Ludemann *et al.*, 2000a; Ludemann *et al.*, 2000b) trajectories further revealed three sub pathways (a, b and c) in the pathway pw2. Pathway pw2a, comprising of F/G loop and the B'-helix has been found to be common among the p450s (Hung *et al.*, 2004; Li & Poulos, 2004; Ludemann *et al.*, 2000a; Ludemann *et al.*, 2000b; Oprea *et al.*, 1997; Poulos, 2003; Scott *et al.*, 2003; Winn *et al.*, 2002) and hence the size of the opening between F/G loop and the B'-helix was monitored during the entire period of simulation (Figure 3.14).

In the WT, the average distance is about 7Å and is seemed to be stabilized by the HB between residues, R124:L247 occurring between the F/G loop and B'-helix. An earlier study involving REMD simulations (Ludemann *et al.*, 2000a) had revealed the breakage and formation of similar HB interactions between F/G loop and B'-helix. The MTs, A115P and S239R also show HBing pattern similar to WT. In correlation, the SAC size of A115P and S239R is similar to WT. While the MTs M132R, E229K, R368H and G466D show complete absence of HBs between F/G loop and B'-helix, the MTs Q144P and P193L have irregular pattern of HBs. Accordingly, in these MTs the conformation of SAC is wider and associated with noticeably large fluctuations. Apart from the above mentioned structural properties, the trajectories of two important structural features pertaining to the FIRs; the Cysteine-heme co-ordination bond distance (Figure 3.15) and the volume of the substrate binding region (Figure 3.16) are also computed. The co-ordination bond distance in WT has a stable trajectory during the simulation, which is also observed in some MTs. In some MTs the co-ordination bond distance is more fluctuating, as could be seen from the figure. The volume of the substrate binding region which is relatively stable in the WT is associated with fluctuations in some MTs. The average values of some of these structural properties are given in Table 3.1.

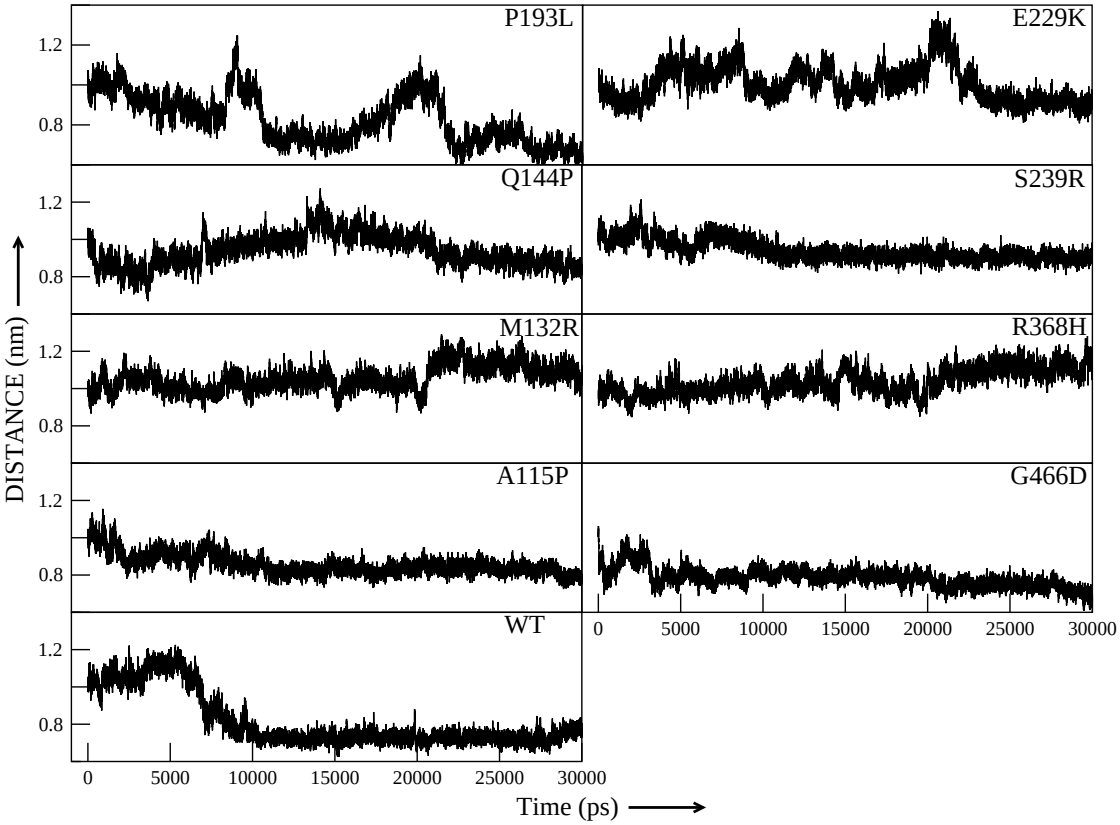


Figure 3.14: Trajectories of the size of SAC

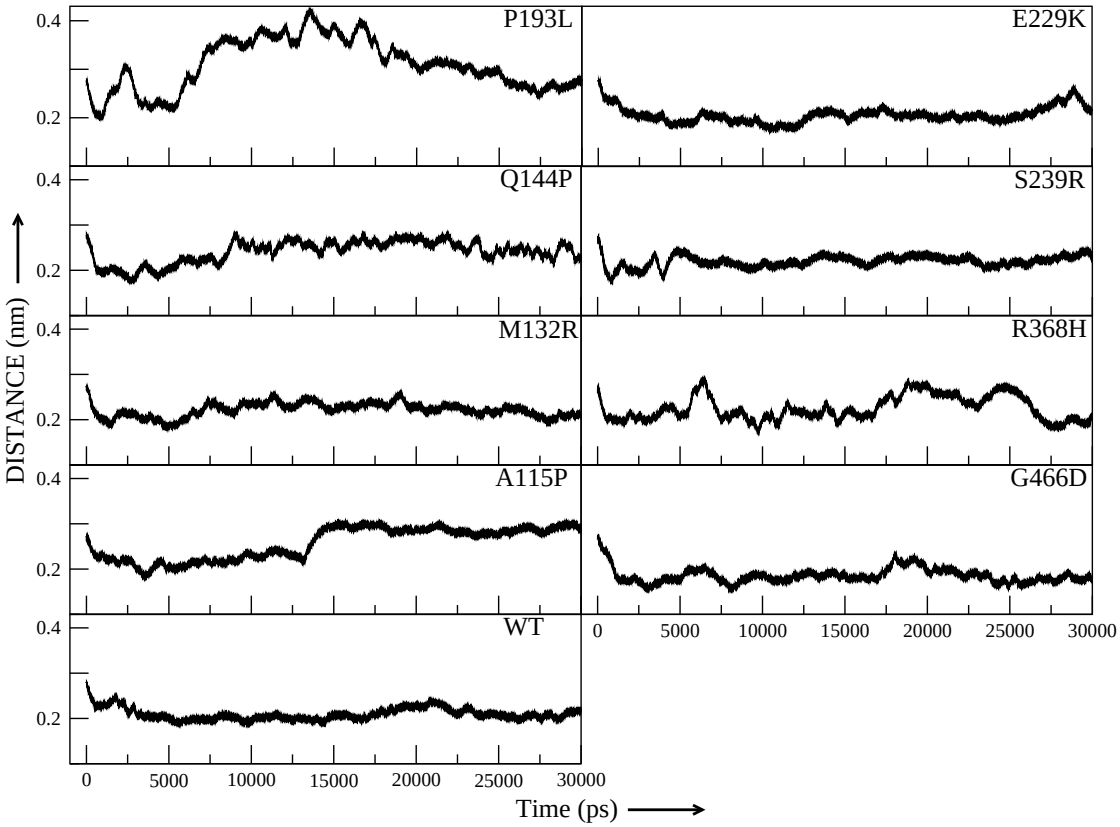


Figure 3.15: Trajectories of Coordination distance between CYS-SG and Heme-FE

3.3.4 Deleterious nature of the mutations

The detailed structural analysis by means of MD simulations in this study confirms some of the earlier speculations about the effects of some of these mutations, as mentioned in the previous Chapter. The average structural properties though indicate similarity of the overall fold of MTs and WT, the precise differences in the RMSD trajectories which indicate the differences in the path of transition of structures from the starting conformation to their final states, despite the initial structures being identical (except at the mutation sites), clearly speaks of the influence of residue substitutions on the dynamics of the protein. The RMSF data indicate that mutations are characterized by subtle but significant increase in the flexibility of the molecule. Further, the individual MTs have specific regions of increased or decreased fluctuations compared to WT, some of these located at spatially distant sites and forming part of FIRs.

It is also observed that mutations have an effect on the integrity of the active site pocket. From functional point of view, the active site regions of enzymes are fairly rigid structures, allowing for specific interactions of functional groups of the enzyme and substrate. Any change in the structure of the active site region can affect the substrate binding and functional interactions between the substrate and the enzyme. The volume of the SBR gives a measure of the disruptive effect due to mutations. As revealed in this study, the volume of the substrate-binding pocket is different in all MTs except A115P, as compared to WT. In P193L and G466D the volumes of SBRs are larger than the WT, whereas in the other five MTs (M132R, Q144P, E229K, S239R and R368H) the SBRs are smaller in volume as compared to the WT. A115P, which shows severe effect on HBR, seems to have no effect on SBR. In terms of the most affecting mutation on SBR, G466D stands out distinctly.

The MTs also reveal changes in the conformation and dynamics of the SAC region

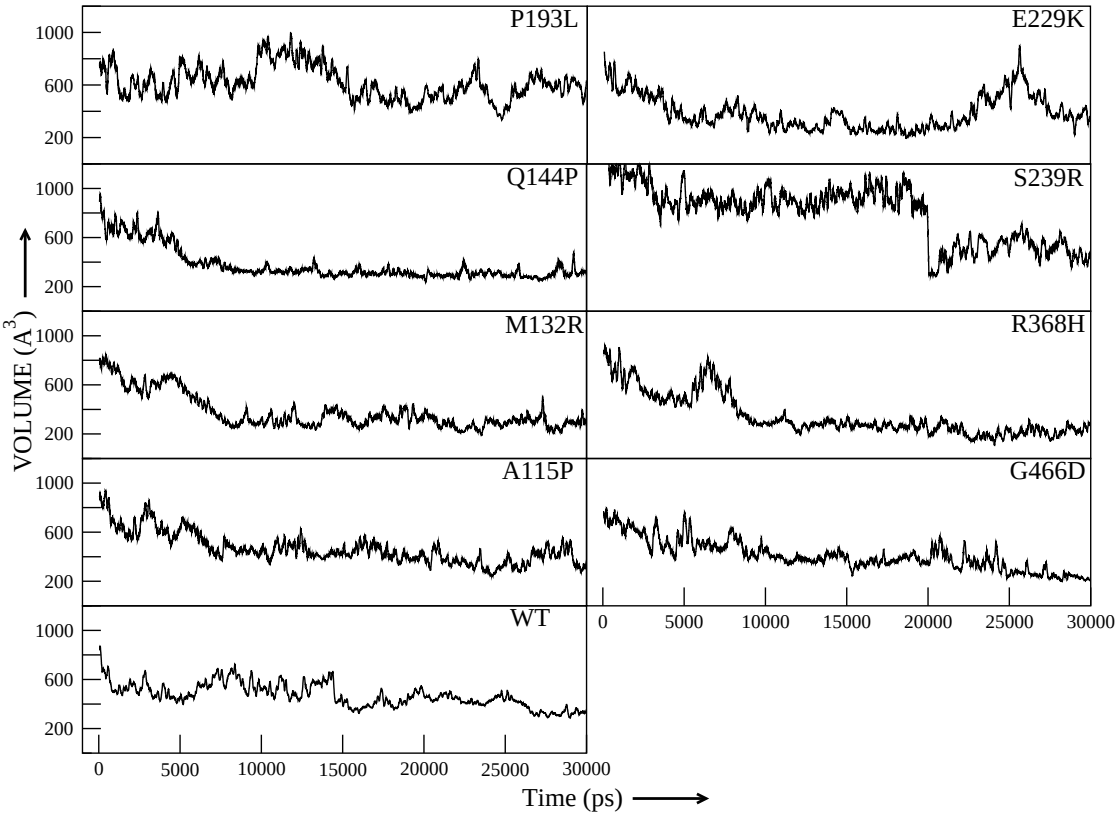


Figure 3.16: Trajectories of the Volume of SBR

wherein; in all the MTs the size of the channel opening is greater than WT. Moreover, in some of the MTs (M132R, Q144P, P193L and E229K) the SAC flutters irregularly which is most severe in P193L (Figure 3.14), during the course of the entire simulation probably indicating disruption of substrate recognition and also its accessibility. Further, the correlation between the MTs and the WT was examined, for residue-wise HB occupancies and the RMSFs where, both the properties are a measure of the nature of the 3D fold of the protein. It was found that with the WT, the MTs Q144P, S239R, R368H, E229K and A115P show poor correlation ($C_{ij} \leq 0.7$) with respect to HBs whereas MTs M132R, E229K, P193L, R368H, Q144P and G466D show poor correlation with respect to RMSFs (Table 3.7).

Furthermore, the extent of overall deviations of all the average properties of the MTs compared to the WT place the MTs Q144P, G466D, R368H, A115P, P193L, M132R, E229K and S239R in the descending order of deleteriousness (refer to the Table 3.1). But the deleteriousness observed in a specific FIR could be more severe in some MTs than others. For example, the Cys-heme co-ordination bond distance can be used to quantify the extent of structural deviation in the HBR in MTs as compared to the WT. In P193L, this distance is about 3.3Å which is more than the usual co-ordination bonding distance in CYPs (≈2Å). In MTs A115P, Q144P, M132R, R368H, G466D the distance is a slightly more than that of WT, ranging between 2.2Å to 2.9Å. E229K has similar distance as that of WT (2.1Å) while in S239R it is 1.8Å. Thus the HBR, from Cys-heme co-ordination bond point of view, is affected severely by P193L mutation and less severely by A115P than the other mutations. Other MTs having co-ordination distance between 1.8Å and 2.5Å may have proper heme binding. The Cys-heme distance roughly correlates with the RMSD of the heme co-factor from initial conformation. The other indicator of deleteriousness at the HBR, viz., RMSD of heme from its initial conformation indicates A115P as the most deleterious mutation.

Generally, loss of protein function can happen when key residue substitutions occur in catalytic/substrate binding sites. However, several disease-causing mutations have been reported which do not occur in catalytic/substrate binding sites. It is argued that such mutations manifest their deleterious effects by bringing in changes in structural characteristics such as a change in surface charge distribution, disruption of packing in protein core regions, etc. In general, any AA substitution, whether pathogenic or non-pathogenic, leave some impact on the protein structure. The extent of impact is dependent on the position and the nature of the AA that is replaced, as well as on the newly introduced AA. Some replacements are well accommodated in the protein structure either with or without any structural changes in the protein such as rearrangements in packing of secondary structural elements in the core (Nagarajaram *et al.*, 1999; Reddy & Blundell, 1993; Reddy *et al.*, 1999). In the case of the pathogenic mutations investigated in this study, it has been found that they can be accommodated into the protein structure but are associated with some structural changes that are functionally significant.

Panicker et al (Panicker *et al.*, 2004) have made a genotype-phenotype correlation studies and have indicated severity of disease manifestation for some of the PCG mutations found in Indian population. They have reported based on the clinically observed phenotype, the percentages of severe phenotypes in the Indian population associated with various mutations in at least one eye. They are: P193L (62.5%); E229K (80%) and R368H (72%). These percentages represent the cases in which severe disease phenotypes were observed, but they do not really compare the extent of disease severity associated with each mutation. Further, incomplete penetrance of the PCG mutations makes the assessment of the deleteriousness of each mutation difficult (Bejjani *et al.*, 2000). The MD simulation data presented here on the disruptive effect on specific FIRs could be validated with experimental evidences and

then effectively be used for disease prognosis.

3.4 Conclusion

Previous homology modeling study on CYP1b1 (Stoilov *et al.*, 1998) suggested that some of the mutations (W57C, G61E, G365W, P379L, R390H, E387K, P437L, and R469W) might disrupt either the hinge region or the conserved core of the protein. In another study, homology modeling of MT forms of CYP1b1 indicated that one (R444Q) out of the four (D192V, A330F, V364M, R444Q) mis-sense mutations in the protein caused significant structural changes while the remaining were structurally neutral (Mashima *et al.*, 2001). Although qualitative analysis and modeling studies give clues about the possible effects of disease causing mutations in proteins, it was unclear how precisely residue changes in sites other than functionally important regions, can bring about changes in protein structure, that in turn have deleterious effect on protein function. These aspects have been studied effectively by MD simulations, as seen in this Chapter.

From the observations made in this Chapter, it is evident that MD simulation has a vast potential to elucidate the structure-function relationships in greater detail, so as to precisely understand the function impeding mechanisms. The current MD simulation study on CYP1b1 clearly shows the underlying structural disruptions leading to loss of function as a consequence of disease mutations. It can be postulated in the present context that the PCG MT proteins carrying the pathogenic mutations fold into WT like structure, but with changed structural properties, which are detrimental to the WT like function. Changes in WT like structural properties may lead to impairment of one or more of the following: substrate access and binding abilities, interaction with the reductase, electron transfer ability and efficient binding of heme.

4

Essential Dynamics Studies

4.1 Introduction

Principal Component Analysis also popularly known as Essential Dynamics (ED), provides a way of dimensionality reduction to interpret high dimensional data in terms of only a few dimensions, thereby simplifying the analysis of complex data while still retaining its important features (Amadei *et al.*, 1993). ED has been used to get insights into the functional and mechanistic aspects of protein folding/unfolding (Chen *et al.*, 2005), catalysis (Bahar *et al.*, 1999; Nunez *et al.*, 2006; Yildirim & Doruker, 2004), etc., and many of the observations have been found to be correlating with the experimental findings (Delaunay-Bertoncini *et al.*, 2003; Gogos *et al.*, 2000).

Essential motions in proteins are functionally important (Berendsen & Hayward, 2000; Pan *et al.*, 2005) and correspond to the large-collective and anharmonic motions, in contrast to the near-constraint Gaussian motions that are harmonic in nature (Hayward *et al.*, 1994). ED has been used in the study of domain

movements (Snow *et al.*, 2007), opening and closure of substrate-binding pockets (Ota & Agard, 2001), loop movements in the catalytic sites (Nunez *et al.*, 2006) etc., of several proteins and protein-ligand complexes.

In the earlier chapter on MD simulation analysis of the WT and the PCG-MT forms of the Human CYP1b1 by means of MD simulations of 30ns duration, it was revealed that the WT and MTs show distinct differences in the time evolution as well as time averaged values of various structural properties, especially those of the functionally important regions. The reduction or complete absence of HBs holding the heme and variation in CYS-heme co-ordination distance in the MTs, indicated that MT structures may have impaired heme binding ability. The MTs (especially P193L and G466D) are also found to have a distorted substrate binding pocket, either smaller or larger than that found in the WT.

Further, it was observed that the MTs have changes in the conformation and dynamics of the SAC region, probably affecting substrate recognition and accessibility. These observations gave some insights into the possible structural characteristics of disease MTs. As discussed above, the functionally relevant motions in the proteins could be effectively studied using the Essential Dynamics analysis, and thus provide better insights into the functional aspects of effects of Disease causing Mutations. In this chapter the Essential Dynamics analysis of the WT and MT structures of CYP1b1 is discussed.

4.2 Materials and Methods

The details of the protocols used for Modeling and MD simulations has been discussed in Chapter 2 and Chapter 3 respectively. For the purpose of ED analysis, MD simulations that have been carried out for 30ns were extended till a total time of 50ns. The

trajectory data of 50 nano-seconds were used for ED analysis. ED analysis was done using the functions of GROMACS. Only the portions of the trajectories subsequent to the time of stabilization (see Table 3.1 in Chapter 3) of the C^{α} RMSD from initial structures were used for ED analysis. The structures were fitted using least squares fitting (LSF) procedure onto the average structure, using the C^{α} atoms that are least mobile, by choosing residues showing less than 1Å RMS fluctuation. The covariance matrices of the C^{α} atoms from their average positions were computed and diagonalized to get the Eigen Vectors (EV) and the corresponding Eigen Values (EL).

4.2.1 Principal components

The principal components p_i(t) were calculated by projecting the trajectory onto the EVs using the equation 4.1 (Spoel *et al.*, 2006), where, M is diagonal matrix containing the masses of the atoms and R is the orthonormal transformation matrix, the columns of which are the EVs, also called principal or essential modes.

$$p_i(t) = R^T M^{-\frac{1}{2}}(x(t) - \langle x \rangle) \tag{4.1}$$

4.2.2 Overlap analysis

The similarity between the EV sets of any two simulations is estimated by their subspace overlap (Hess, 2002). The overlap of the subspace spanned by m orthonormal vectors w_1,... ,w_m with a reference subspace spanned by n orthonormal vectors v_1,... , v_n is quantified using the equation 4.2 (Spoel *et al.*, 2006):

$$overlap(v,w) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}(v_i.w_j)^2 \tag{4.2}$$

In addition to subspace overlap, to check the reliability of the ED analysis results, the matrices of the squared inner products between the EV sets were also calculated.

4.2.3 Cosine content

The cosine content of the PCs, which is a negative index of the similarity to the random diffusion process, was also calculated. The PCs for random diffusion are cosines with the number of periods equal to half the principal component index (Hess, 2000). The ELs are proportional to the index to the power -2. The cosine content is defined using equation 4.3 (Spoel *et al.*, 2006).

$$\left(\frac{2}{T}\right)\left(\int_0^T \cos(i\pi t)p_i(t)dt\right)^2\left(\int_0^T p_i^2(t)dt\right)^{-1} \tag{4.3}$$

4.2.4 Hurst Exponent

The Hurst exponent (Saarela *et al.*, 2002) , which is used to estimate the degree of randomness in a time series data, is calculated as follows. The time series data ξ_t is divided into periods of length τ. In each window, the average $\langle\xi_\tau\rangle$, cumulative sums X(t,τ) (Equation 4.5), standard deviation S(τ) (Equation 4.4) and range, R(τ) (Equation 4.6) are calculated. Then, the average value of R(τ)/S(τ) over all the windows, is taken as an estimate of rescaled range for that τ value. This process is repeated for progressively larger τ-values. A regression line is then fitted over the data points of log[R(τ)/S(τ)] and the corresponding log(τ) values. The slope of the line gives an estimate of Hurst exponent.

$$X(t,\tau) = \sum_{i=1}^{t}(\xi_i - \langle\xi\rangle_\tau) \tag{4.4}$$

$$S(\tau) = \left[\frac{1}{\tau}\sum_{t=1}^{\tau}(\xi_i - \langle\xi\rangle_\tau)^2\right]^{\frac{1}{2}} \tag{4.5}$$

$$R(\tau) = [X(t,\tau)]_{\max} - [X(t,\tau)]_{\min}\,(1 \le t \le \tau) \tag{4.6}$$

The Hurst exponent H value near to 0.5 indicates random process. H value between 0.5 and 1 characterizes persistent behavior and H value between 0 and 0.5 characterizes anti-persistent behavior.

4.2.5 Computation of HBs

The HBs (Hb) in the structures were computed using HBPLUS (McDonald & Thornton, 1993) as explained in Chapter 3.

4.2.6 Combined ED analysis

The Combined ED analysis was done to find the common modes of motion present in all the molecules and to compare similarities and differences in the motions along each common mode. The individual trajectories taken after their equilibration times were concatenated and the covariance matrix of C^{α} atoms was constructed from the concatenated trajectories. The procedure of LSF similar to the one used for normal ED analysis was adopted, wherein the C^{α} atoms that had less than 1Å RMS fluctuation and that are common among all the structures were used for LSF in the covariance matrix calculation. This matrix was then diagonalized to yield the EVs and the corresponding ELs, for the combined trajectories. PCs for the individual trajectories were then calculated by projecting the individual trajectories onto the combined EVs, with the global average structure as the reference. The corresponding RMS fluctuations for each of the PCs were also computed.

4.3 Results and Discussion

As mentioned earlier, all the MD simulations were carried out for 50 ns, and portions of the trajectories after C^{α} RMSD stabilization were used for PCA analysis, resulting in simulation lengths of 35 ns or more in each case. Only C^{α} atoms were chosen for the analysis, as PCA using only C^{α} atoms has been shown to give a good description of the overall essential motions (Amadei *et al.*, 1993). Before commencing the PCA, the rigid body rotational and translational motions were removed. Furthermore, for the construction of covariance matrices, only the C^{α} atoms that have RMS fluctuation less than 1Å were considered as reference atoms for LSF, to get a better fit and description of the collective motions. The importance of fitting procedure has previously been studied and found to influence the sensitivity of identification of domain boundaries (Arnold & Ornstein, 1997) and the results of PCA analysis (Abseher & Nilges, 1998).

4.3.1 The Essential motions

The reliability of ED results depends on the conformational space that is sampled in the simulation. Though it was shown that reliable information could be obtained even from short simulations (de Groot *et al.*, 1996a), it has also been suggested that the simulation length should be as long as possible, to overcome the mis-interpretation of diffusive motions resulting from short simulations, as essential motions (Hess, 2000). Thus, it is necessary to ensure that the Essential motions are sufficiently converged. The subspace overlap of EV sets (usually the first 10), calculated between the first and second half portions of a simulation, gives a measure of the similarity of essential space spanned in the two halves (Hess, 2000). An overlap value close to 1 indicates similar subspaces or the directions of the collective modes, while a value near to zero indicates

completely orthogonal subspaces.

Table 4.1(A) gives the values of subspace overlap of first 10 EVs, between the first and second half portions of the WT and MT simulations. It can be seen that the subspace overlap ranges from 0.35 to 0.45. The overlap values are lower compared to some earlier studies (de Groot *et al.*, 1996a). The reason for this seems to be due to the larger size of protein (478aa) used in the current simulation. It was previously suggested that a few hundred picoseconds of simulation can give a rough approximation of the essential subspace for a small protein, and considerable amount of noise exists in the description of essential and near-constraint subspace (Amadei *et al.*, 1993; van Aalten *et al.*, 1995b; van Aalten *et al.*, 1995a). Thus, the descriptions of essential and near-constraint subspaces in larger proteins like p450s will be associated with more noise, which is reflected in the lower overlap values obtained in this study. CYP1b1 protein also has heme co-factor at its center and is not clearly distinguishable into multiple domains, suggesting a rigid molecule in terms of inter-domain collective motions. Thus, the percentage of noise existing in the description of essential and near-constraint subspaces can be expected to be high giving rise to lower subspace overlap values.

When the subspace overlap was calculated between successive time windows in the simulation , it is observed that the overlap among successive windows increase towards the end of simulation (Figure 4.1). It is also observed that the overlap values are higher when the successive windows used for calculation are larger. Thus, the essential modes become more consistent with increase in simulation time and longer simulation times lead to a more consistent description of the essential motions.

Similar conclusions on the subspace overlap can be drawn from the matrices of the squared inner products of EVs , calculated between the first and second half portions of simulations. The matrix pertaining to WT molecule is shown in Figure 4.2(a). Similar

Table 4.1: Subspace overlap[†] of the first 10 Eigen Vectors, between the first and second half simulation periods in WT and MTs.

	A (Overlap of first 10 EVs)	B (no of EVs representing 85% of TPF)	C (% of TPF represented by first 4 EVs)
WT	0.412	53	50
A115P	0.393	41	58
M132R	0.37	28	65
Q144P	0.422	39	62
P193L	0.335	42	56
E229K	0.448	64	51
S239R	0.395	46	54
R368H	0.388	48	53
G466D	0.374	59	48

[†]A value near to 1 indicates identical subspaces, and a value near to zero indicates completely orthogonal subspaces.

Number of Eigen Vectors needed to represent 85 % of the Total positional fluctuation (TPF) in WT and MT simulations.

Percentage of Total positional fluctuation (TPF) explained by the first 4 Eigen Vectors in WT and MTs.

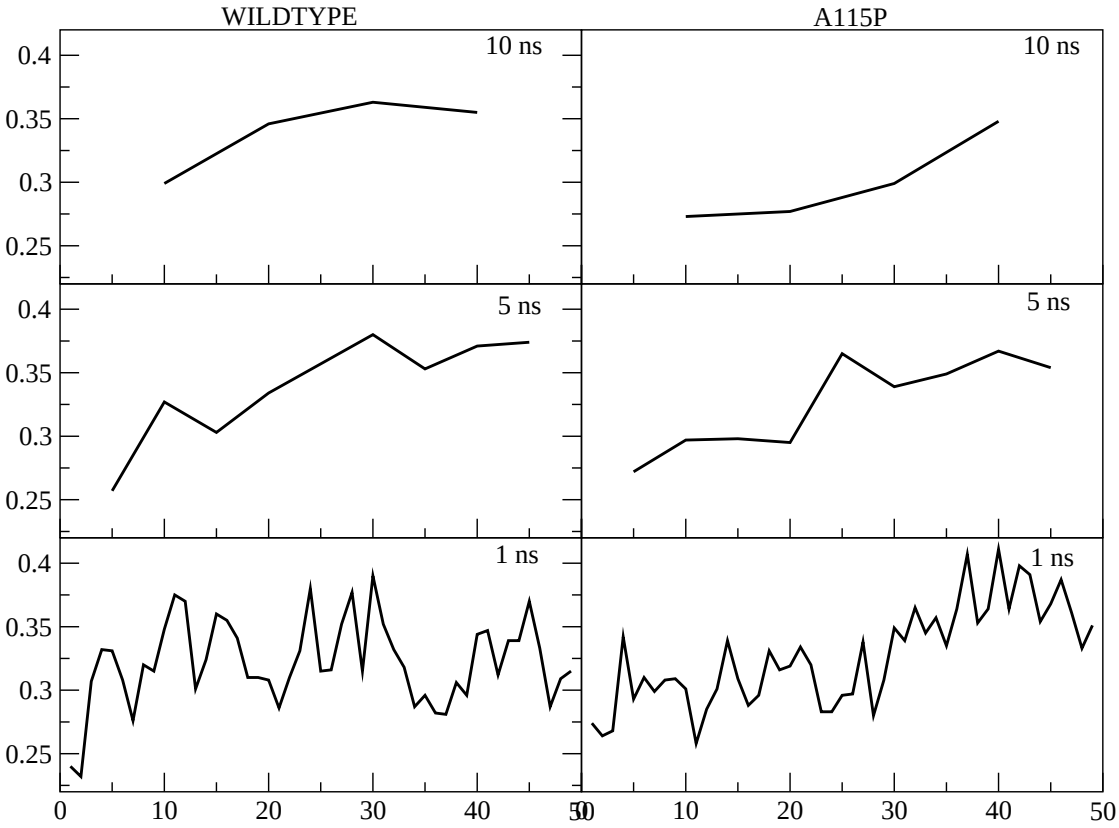


Figure 4.1: Variation of the Overlap values of the first 10 EVs calculated between successive windows of trajectory A. Window size = 1ns B. Window size = 5ns CA Window size=10ns

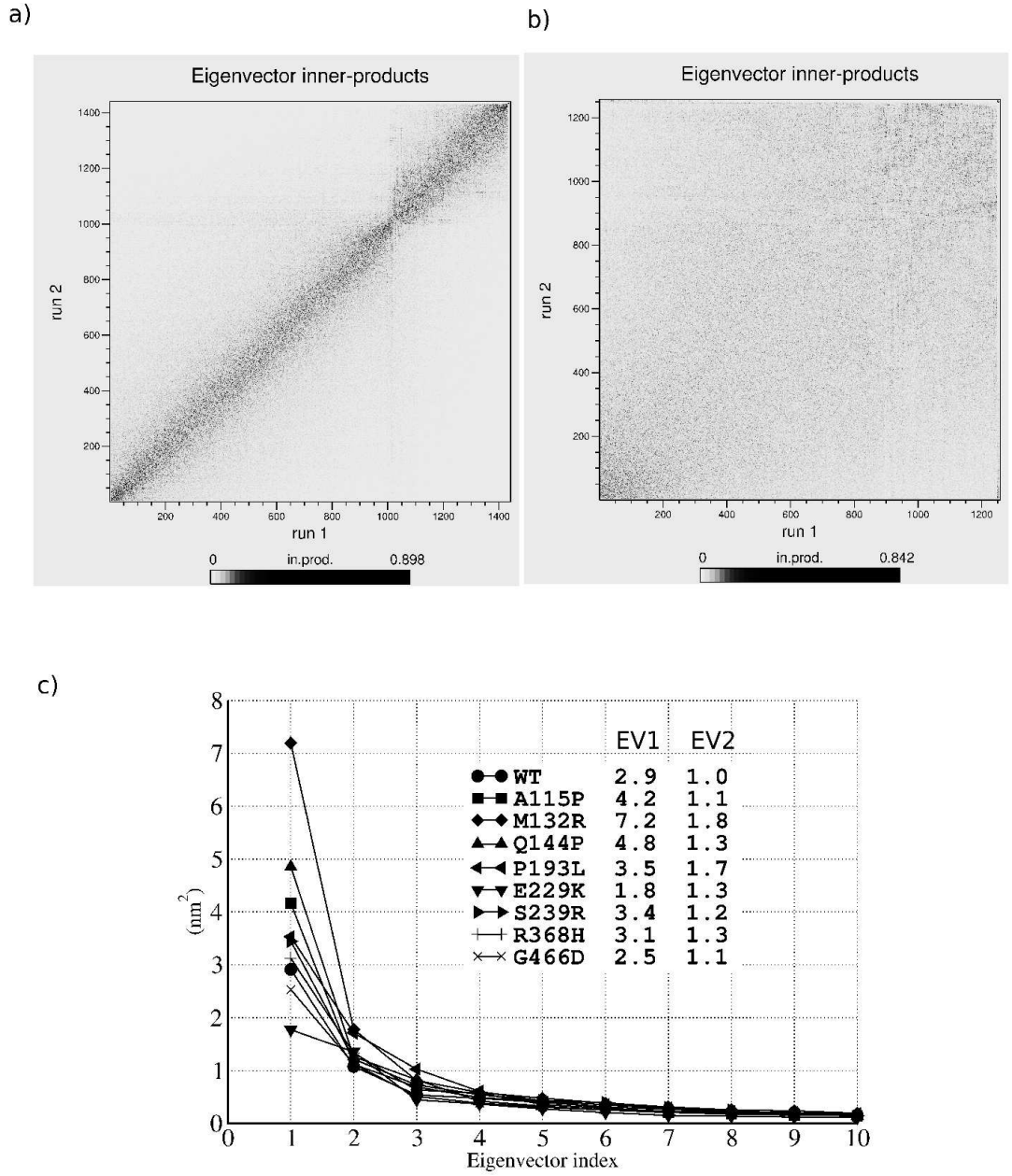


Figure 4.2: (a) Matrix of the squared inner products of Eigen Vectors between the first and second half of simulation periods in WT. (b) Matrix of the squared inner products of Eigen Vectors between the WT and an unrelated protein simulation. The brightness and contrast of (a) and (b) was adjusted for clarity. (c) Eigen Values of the covariance matrices of WT and MT simulations. The Eigen Values for the first and second Eigen Vector are indicated in the inset.

trend was seen in the case of MTs (Figure 4.3). In general, most of the higher inner products are concentrated along the diagonals, indicating that the EVs from first half portion of simulation have appreciable overlap with EVs of about the same rank from the second half portion of simulation. This in turn indicates that the Essential spaces spanned in the first and second half portions of the simulation are similar. The top ranking EVs show the concentration of higher inner products near to the diagonal, and their spread from the diagonal increases towards the higher or near-constraint EVs. This matrix can be compared with the inner-product matrix (Figure 4.2(b)) calculated between EVs of WT and an unrelated protein simulation, which shows a greater spread of higher inner products from the diagonal.

When the actual values of inner products of EVs between the two halves of the simulation are observed (Table 4.2), it was noticed that the first EVs in all the molecules except P193L and E229K, distinctly had high inner products ($>$0.7). In P193L and E229K, the first and second EVs have high inner products. The subsequent EVs however have lesser overlap with any other EVs. This, indicates that the first EVs mainly describe essential subspace of the protein. Thus, the essential modes, which are anharmonic, show considerable overlap between the two halves of the simulations and hence reasonably converged to be used for further analysis. Similar observation on the importance of only the first EVs in describing the collective motions, was made in earlier study (Peters *et al.*, 1997).

It has been earlier suggested that that the PC1 in an ED analysis can arise due to random diffusion (Hess, 2000). We therefore computed the Hurst exponent (Saarela *et al.*, 2002) (H) for the PC1, which is an indicator of randomness in a time series data. 'H' value near to 0.5 indicates random behavior while the value close to 1 indicates time correlations in the data. The 'H' exponents calculated for PCA were about 0.9 for all the structures indicating a non-random nature. We also calculated the cosine

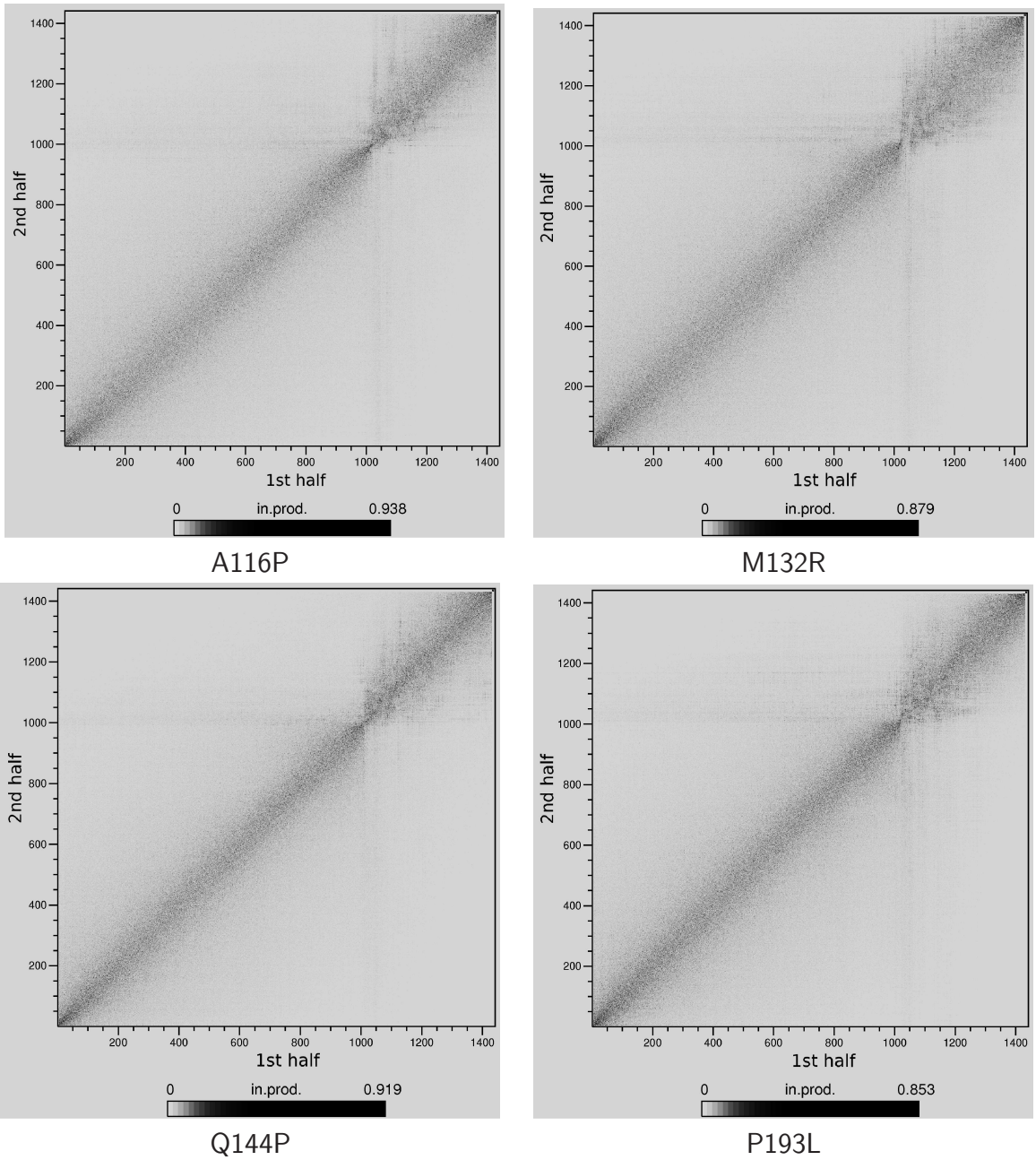
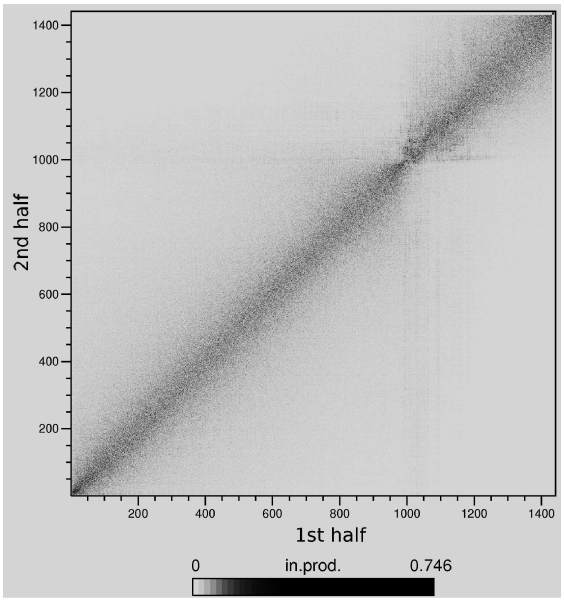
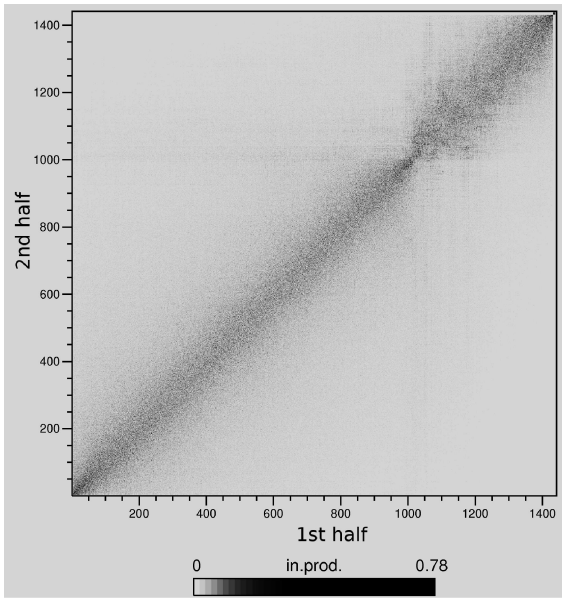


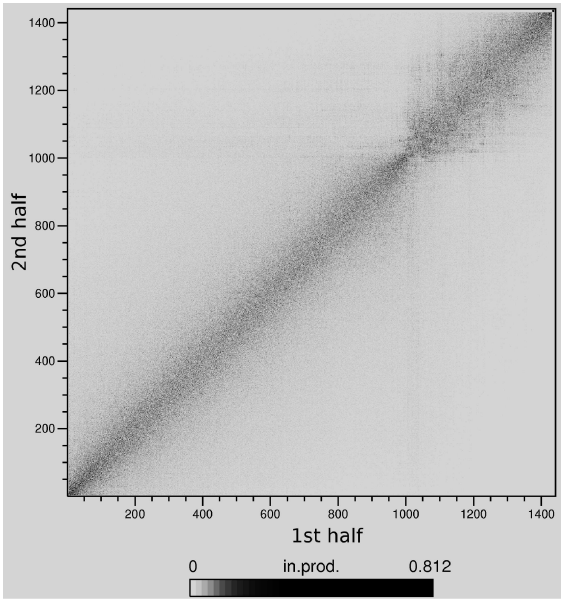
Figure 4.3: Matrix of the squared inner products of Eigen Vectors between the first and second half of simulation periods in the MTs. The brightness and contrast of the graphs was adjusted for clarity.



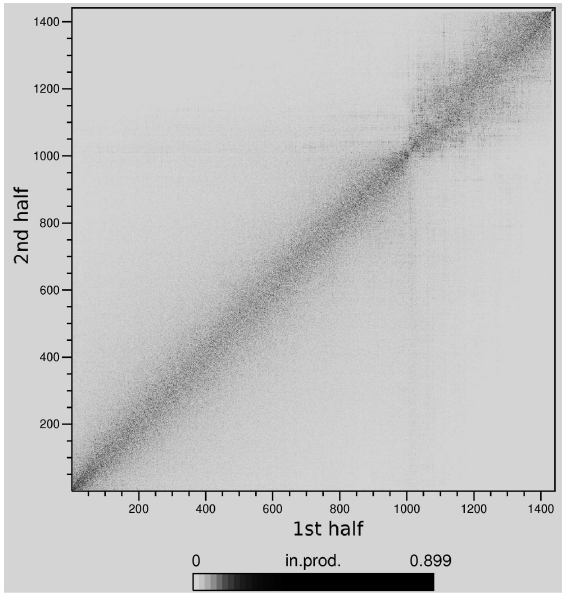
E229K



S239R



R368H



G466D

Figure 4.3 . . . continued

Table 4.2: Values of squared inner products calculated between two halves of simulations in WT and MT simulations. Only the inner products among the first 4EVs are shown.

WT	1	2	3	4
1	0.898	0.314	0.045	0.045
2	0.359	0.247	0.022	0.045
3	0.135	0.135	0.135	0.067
4	0.000	0.112	0.157	0.090

A115P	1	2	3	4
1	0.938	0.258	0.070	0.117
2	0.188	0.047	0.094	0.117
3	0.141	0.234	0.164	0.399
4	0.023	0.000	0.188	0.070

M132R	1	2	3	4
1	0.879	0.395	0.132	0.088
2	0.373	0.308	0.066	0.088
3	0.176	0.417	0.044	0.242
4	0.022	0.132	0.088	0.286

Q144P	1	2	3	4
1	0.919	0.344	0.092	0.000
2	0.230	0.023	0.322	0.069
3	0.161	0.390	0.367	0.161
4	0.023	0.069	0.253	0.046

P193L	1	2	3	4
1	0.576	0.597	0.320	0.171
2	0.384	0.213	0.085	0.000
3	0.299	0.085	0.107	0.213
4	0.235	0.085	0.043	0.043

E229K	1	2	3	4
1	0.168	0.746	0.019	0.298
2	0.727	0.336	0.019	0.168
3	0.205	0.037	0.000	0.019
4	0.242	0.168	0.075	0.037

S239R	1	2	3	4
1	0.702	0.488	0.312	0.117
2	0.410	0.058	0.117	0.039
3	0.117	0.078	0.078	0.097
4	0.312	0.117	0.273	0.273

R368H	1	2	3	4
1	0.751	0.507	0.264	0.000
2	0.406	0.203	0.020	0.162
3	0.304	0.041	0.203	0.000
4	0.101	0.406	0.264	0.183

G466D	1	2	3	4
1	0.742	0.584	0.180	0.000
2	0.337	0.112	0.067	0.270
3	0.270	0.225	0.135	0.135
4	0.157	0.000	0.180	0.337

content for PC1, since cosine content of >0.5 (usually the PC1), had been suggested to be a probable indicator of random-diffusion dynamics (Hess, 2000). The cosine content for the PC1 in all the structures were about 0.8. This result is in conflict with the result from the 'H' exponent analysis. However, as shall be noted from the later sections, it is observed that the loop regions are the major contributors to the collective motions along PC1 in all the simulations. This observation is in contrary to the expected uniform distribution of collective motions throughout the structure, had it been due to random diffusion. Thus a combination of methods discussed above can provide better means of checking the reliability of Essential motions.

4.3.2 Overall nature of Essential motions of WT and MTs

The ELs shown in Figure 4.2(c) show a steep fall from the 1^{st} EV to a value below 1 in the 3^{rd} EV, and then with a slow decreasing trend die down to zero, till the last EV. The purpose of ED is to reduce the dimensional complexity of the data, so that important features of the data can be studied using only few dimensions. On an average about 50 EVs are required to represent 85% of the total positional fluctuations (TPF) in all the structures (Table 4.1(B)). A comparison of results obtained in some earlier ED studies, for example; the ED analysis of 100ps simulation of Bovine Pancreatic Traipsing Inhibitor (57aa) (Hayward & Go, 1995), and of variable simulation lengths of Copper Plasticine(100aa) and assuring(130aa) (Arcangeli *et al.*, 2001), indicated that the number of EVs required to represent about 85% of TPF, is directly proportional to the size of the protein and inversely proportional to the simulation time. In the current study also, ED analysis of simulations of varying window lengths indicated a similar trend (data not shown).

The first four EVs were selected for further analysis according to the Cat ell's scree

test (Catell, 1996), which suggests that the subsequent components, after the point where the graph of ELs makes an elbow bend and becomes less steep, could be dropped off. Moreover, the probability distributions of the values of projections of trajectories, onto the EVs of smaller ELs are similar to Gaussian distribution calculated for the corresponding mean and standard deviation, indicating motions of random nature (Figure 4.4). Table 4.1(CA gives the percentage of the total positional fluctuation, explained by the first 4 EVs in the WT and MT simulations. It can be observed that except G466D and E229K, all the MTs are associated with higher mean square fluctuations than the WT. The values of subspace overlap of the first 10 EVs among the WT and MT simulations are low, ranging from 0.23 to 0.32, indicating considerable differences in collective motions, in each case (Table 4.4). As it was earlier mentioned that the essential motions are mainly described by the EV1, in the following sections unless specifically mentioned, the description pertaining to the collective motions will be referring mainly to the EV1.

The magnitude of fluctuation along the EV1, which is the direction along which the maximum motion of a molecule is observed, is significantly different among the structures, which in turn indicates the differences in the overall motions. The MTs (except E229K and G466D) are associated with collective motions of greater magnitude than the WT. These observations correlate with the number of Hb interactions present in the structures. Table 4.6 gives the number of Hbs with $\geq$50% occupancy and occurring between different secondary structures. These Hbs are further classified into A) Short range (between residues which are sequentially $<$50 residues apart) and B) Long range (between residues which are sequentially $\geq$50 residues apart). The reason for adopting this limit of sequence distance was that, the longest structure in the protein- the I-helix is about 30 AAs in length and together with the secondary structures adjacent to its two ends, spans about 50 AAs (Figure 7). Thus, this sequence

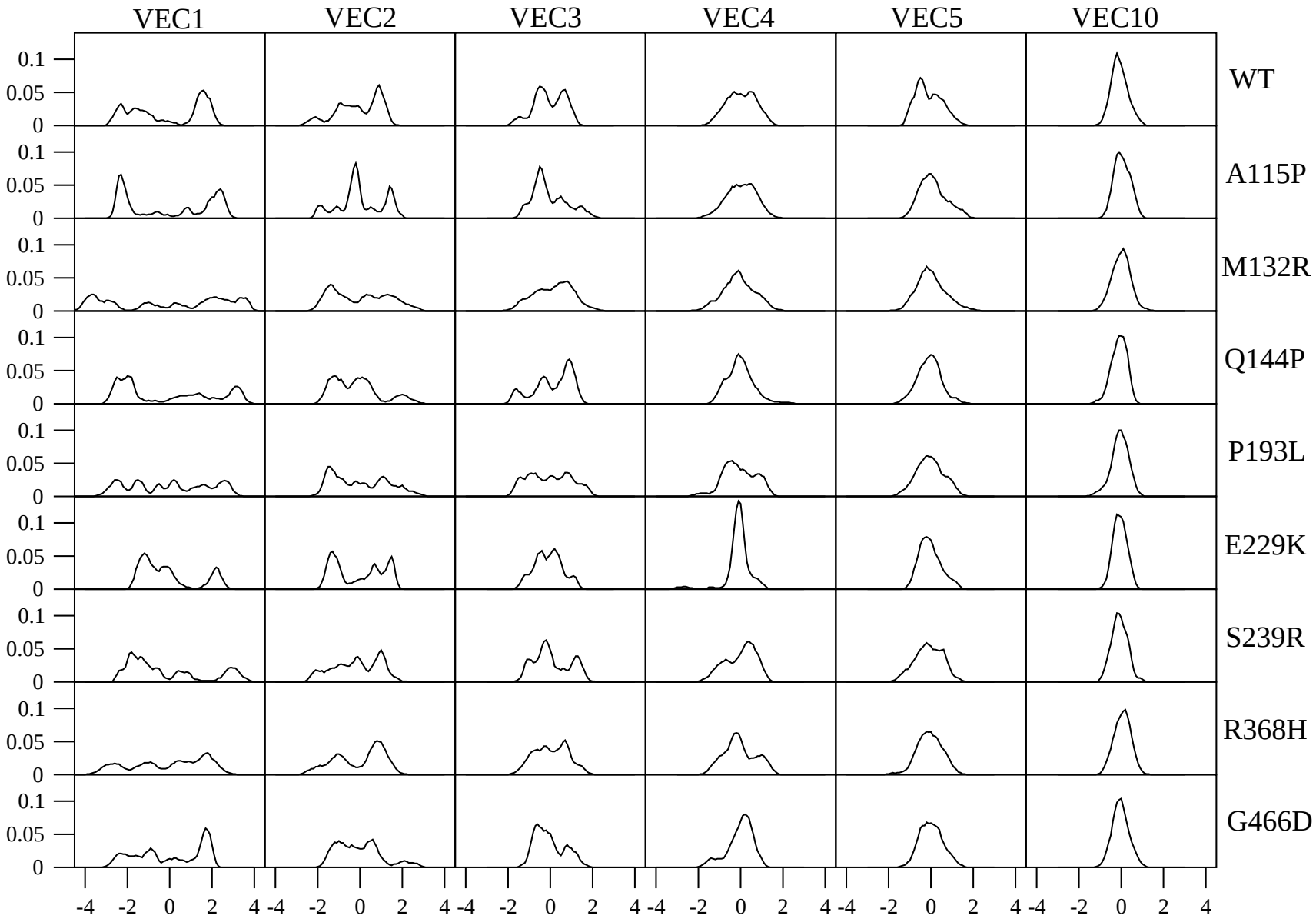


Figure 4.4: Probability distribution of the values of projections of trajectories onto the Eigen Vectors

Table 4.4: The subspace overlap values of the first 10 EVs calculated from WT and MT simulations.

	WT	**66**	**83**	**95**	**144**	**180**	**190**	**319**	**417**
WT	1.000	0.267	0.264	0.286	0.244	0.306	0.306	0.255	0.282
66	0.267	1.000	0.244	0.288	0.239	0.283	0.264	0.273	0.265
83	0.264	0.244	1.000	0.276	0.247	0.269	0.260	0.261	0.282
95	0.286	0.288	0.276	1.000	0.269	0.294	0.293	0.277	0.270
144	0.244	0.239	0.247	0.269	1.000	0.309	0.246	0.278	0.241
180	0.306	0.283	0.269	0.294	0.309	1.000	0.317	0.313	0.289
190	0.306	0.264	0.260	0.293	0.246	0.317	1.000	0.266	0.291
319	0.255	0.273	0.261	0.277	0.278	0.313	0.266	1.000	0.273
417	0.282	0.265	0.282	0.270	0.241	0.289	0.291	0.273	1.000

Table 4.6: Number of Inter-secondary structural Hydrogen bonds with $\geq$ 50% occupancy

	A(Total)	**B(Sequence separation >50 residues) Long range**	**C(Sequence separation $\leq$ 50 residues) Short range**
WT	109	37	72
A115P	113	44	69
M132R	96	22	74
Q144P	112	33	78
P193L	105	37	68
E229K	120	46	74
S239R	99	33	66
R368H	97	30	67
G466D	86	31	55

separation was considered the minimum limit to mean long-range interactions.

From the table it can be seen that the magnitudes of the collective motions along EV1 approximately correlate negatively with the total number of inter secondary structural Hbs. Further, the long range Hbs seem to be an important factor influencing the collective motions, since these are present between distant regions of the protein and their presence or absence would have effects on the large motions. Thus, M132R, which has the largest collective motions, has the least number of long range Hbs and E229K, which has the least collective motion, has the largest number of long-range Hbs. In contrast to the EV 1, the ELs for EVs 2, 3 and 4 are similar for all the structures and moreover these values are significantly lower than those for first EV.

The nature of collective motions can be better found, using projection of trajectory along any two EVs. The 2D projection of the trajectories along the 1^{st} and 2^{nd} EVs is shown in Figure 4.5. The data points in the plot are color coded according to the simulation time, with the initial $1/5$ portion of the trajectory colored in red, and the subsequent portions colored in pink, yellow, cyan and blue. This gives a vivid picture of the conformational sampling by the molecules during different time windows of the simulation. The spread of the data points indicates the extent of conformational sampling or the essential motions in the molecules, along the two most significant directions. The spread or extent of conformational sampling denoted by P_δ is calculated using the equation 4.7,

$$P_\delta = \frac{\sum_{t=1}^{N} \sqrt{(P_1(t)-\overline{P_1})^2+(P_2(t)-\overline{P_2})^2}}{N} \tag{4.7}$$

where, N is the number of time-steps, P_1 and P_2 are the projection values along EV1 and EV2 respectively at time-step 't' and $\overline{P_1}$ and $\overline{P_2}$ are the time averaged projection values along EV1 and EV2. From P_δ values shown in Table 4.8, E229K and G466D

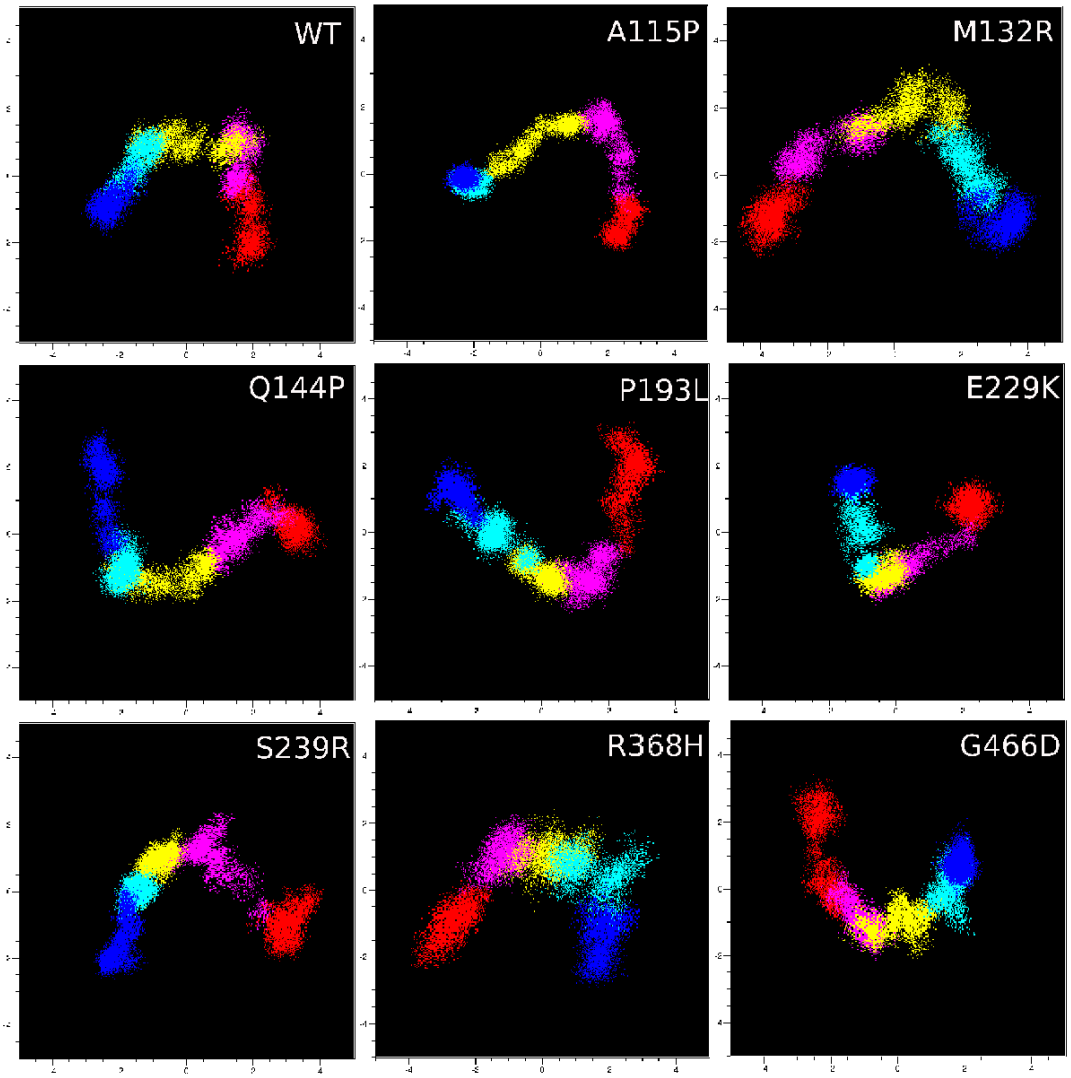


Figure 4.5: 2D projection of the WT and MT trajectories onto the first and second Eigen Vectors. All the sub figures are drawn to the same scale. In each sub figure the X-axis and Y-axis denote projection values (in nanometers) along Eigen Vectors 1 and 2 respectively. The colors red, pink, yellow, cyan and blue respectively indicate the five equal and consecutive time windows of the simulation.

have lesser conformational freedom than WT, while other MTs, especially M132R have larger conformational freedom than WT. In addition, it is observed that the extents of conformational sampling during different time windows vary among the structures. These conformational changes happen during different times among the structures. It can be observed that the projections contain variable number of clusters, separated by small or large gaps indicating conformational transitions, specific to each case (Figure 4.6).

The resident times in each cluster thus vary among the structures indicating that the periods of certain essential motions as well as their magnitudes are different among the molecules. The clusters are distinct in some (A115P, M132R, Q144P, E229K and S239R) while in others, there is less demarcation (WT, P193L, R368H and G466D). Thus, in the former case it may indicate that the molecule spends considerable time in one minimum before entering another minimum. In the latter cases, the molecules experience a smooth transition from their initial to final conformations. The 2D projections in WT and MTs along the third and fourth PCs show more compact and grossly similar clusters, centered on their equilibrium projections (data not shown). However the spread within these clusters and the time windows during which there is greater spread varies considerably between the structures.

4.3.3 The distribution of essential motions in the of WT and MT structures

The distribution of fluctuations along EV1 is shown in Figure 4.7. It is apparent from the figure that the distributions of collective motions in the WT and MTs are different. A portion of the native dynamic property is preserved in the mutations, since much of the collective motion is concentrated about the loop regions, followed by helical and

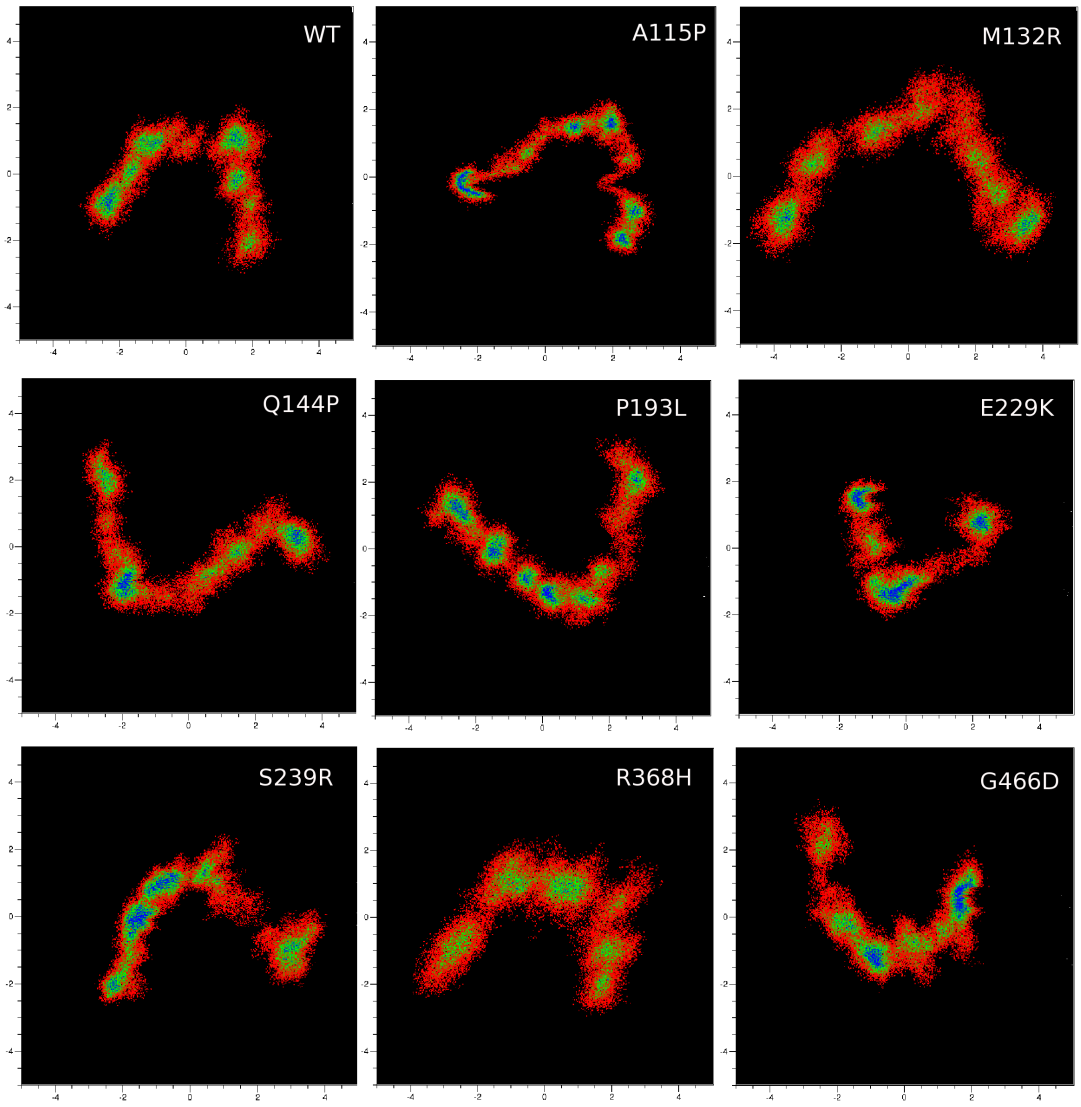


Figure 4.6: 2D projection of the WT and MT trajectories onto the first and second Eigen Vectors. All the sub figures are drawn to the same scale. In each sub figure the X-axis and Y-axis denote projection values (in nanometers) along Eigen Vectors 1 and 2 respectively. The colors from red through pink, yellow, cyan and blue indicate, in ascending order the intensity of visit to that conformation.

Table 4.8: The P_δ values of 2D projection plots (Figure 4.5 and Figure 4.6) to denote the spread of data points.

	P_δ Value
WT	191
A115P	205
M132R	276
Q144P	228
P193L	225
E229K	167
S239R	212
R368H	204
G466D	188

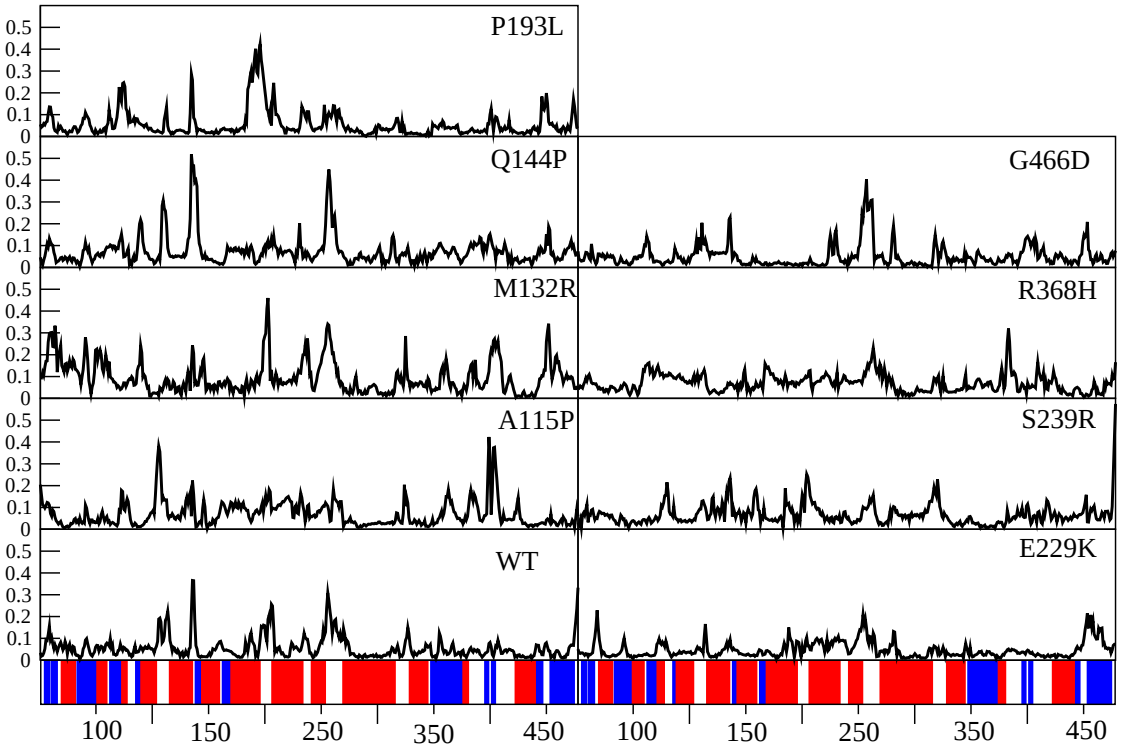


Figure 4.7: Residue wise distribution of RMSF along Eigen Vector 1 in WT and MT simulations. The regions of Helices and Strands are indicated by red and blue shades respectively, at the bottom of the figure.

strand regions. However, MTs differ from the WT with respect to the magnitudes of collective motions and absence of and/or presence of additional regions involved in collective motions.

In general, all the mutations have an impact on the structural core of the protein. Figure 4.8(a) gives the percentage of residues belonging to the Secondary structural regions involved in large collective motions along EV1. In all the MTs as compared to WT, relatively large percentages of residues, which are in α-helices or β-strands or both (see Figure 3.13 for definitions of secondary structures), contribute to large collective motions.

This indicates that in MTs some of the relatively intact core regions acquire large collective motions with other highly mobile regions of the protein. In addition to this, the effect of mutations is compounded, as collective motions involving the structural core are also found along EV 3 and EV 4 (Figure 4.9). Since the EVs are mutually orthogonal in their direction, existence of collective motions in the structural core, along multiple EVs denotes further complexity of motions.

The effects of mutations are also reflected in the dynamics of the three functionally important regions (Achary *et al.*, 2006) of the protein, i.e., the Heme-binding region (HBR), Substrate-binding region (SBR) and the Substrate access channel (SACA region (Figure 4.8(b)). The HBR and the SBR are required to be structurally rigid regions as they are involved in the binding of heme co factor and substrate respectively. Clearly in WT none of the HBR residues show collective motions along the EV1, but in all the MTs except P193L, parts of the HBR region show considerable collective motions along EV1. SBR is mostly intact in WT with few residues (5%) having collective motions along EV1. These residues are in the N-terminal of the F/G loop (a component of SACA that actually comprises the roof of the SBR. All the MTs except A115P and

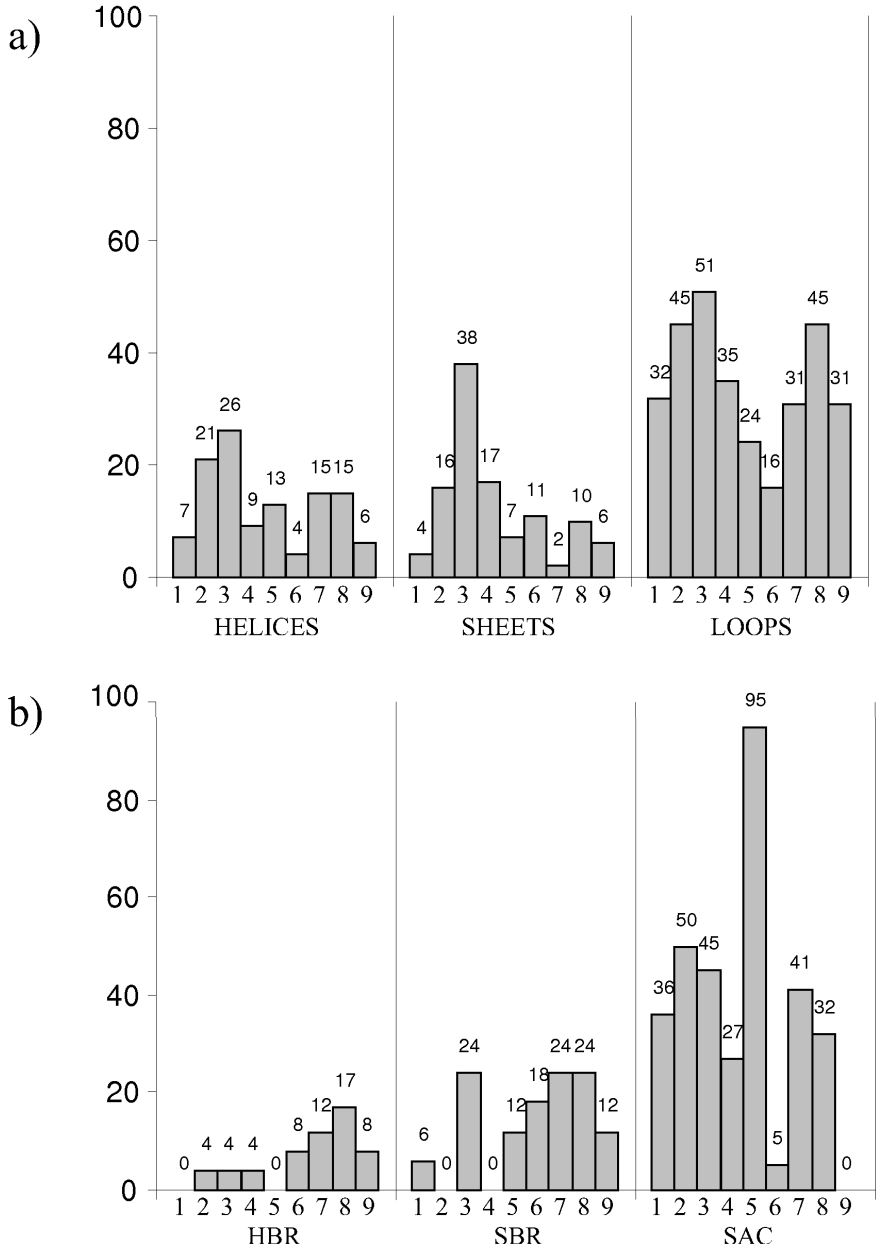


Figure 4.8: Percentage of a.Secondary Structures (Helices, Strands and Loops) and b.FIRs (HBR, SBR and SAC having RMSF >1A along Eigen Vector 1, in WT and MTs. The WT and MTs, A115P, M132R, Q144P, P193L, E229K, S239R, R368H and G466D (indicated by numbers 1 through 9). The values are annotated over the bars.

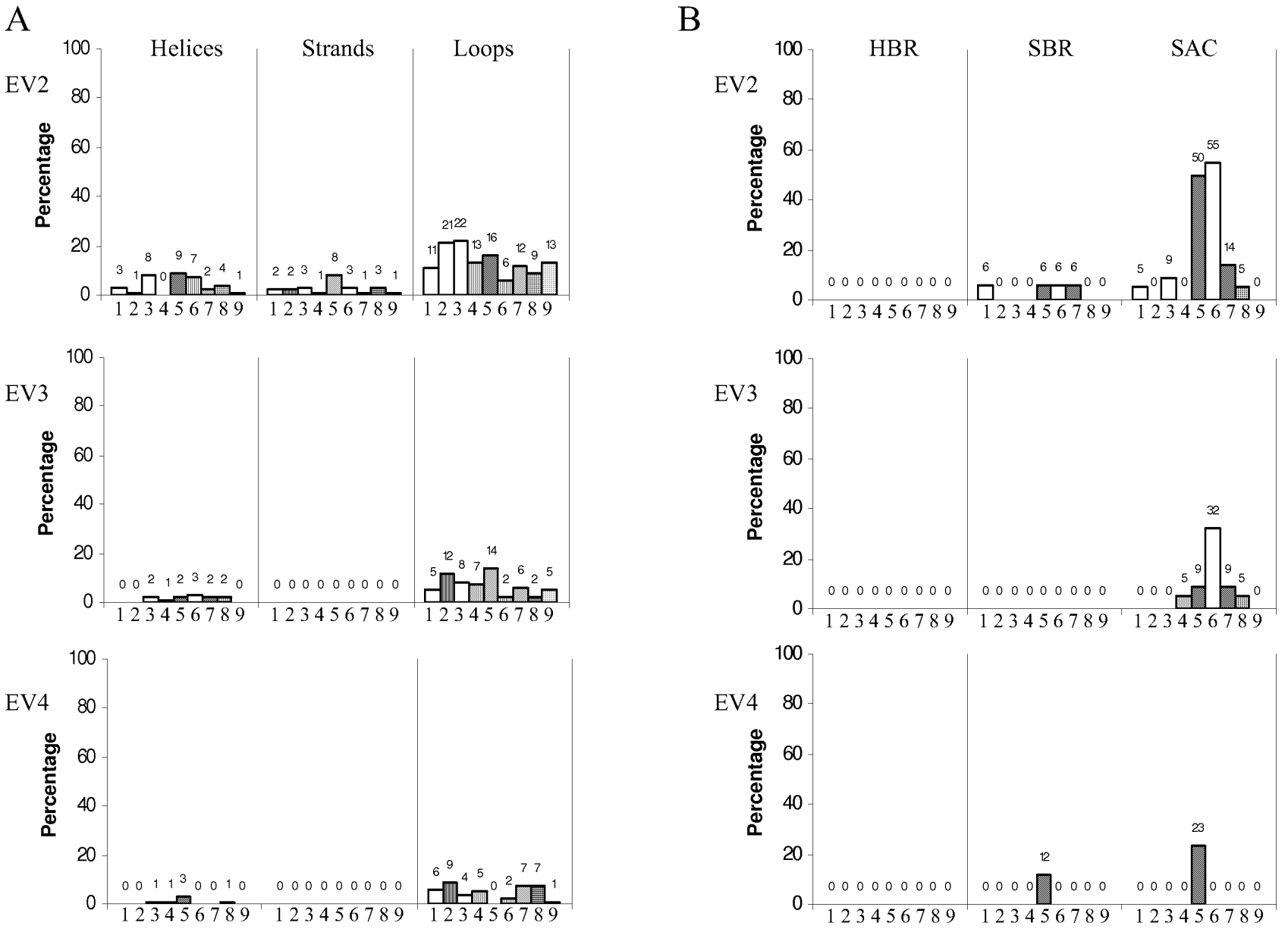


Figure 4.9: Percentage of A. Secondary Structures (Helices, Strands and Loops) and B. FIRs (HBR, SBR and SACA having RMSF($\geq$1Å) along Eigen vector 2, 3 and 4 in WT and MTs. The WT and MTs, A115P, M132R, Q144P, P193L, E229K, S239R, R368H and G466D (indicated by numbers 1 through 9). The values are annotated over the bars.

Q144P, have collective motions involving larger portions of SBR. A major portion of SAC region (defined as the region comprising F/G loop and B'-helix), which is involved in the recognition of substrate and its access, has collective motions in WT and many of the MTs. The MTs, G466D followed by E229K have negligible collective motion involving SAC along EV1. P193L is abnormal in that it has all of the SAC including the F/G loop and the B'-helix having collective motion. In the case of MTs, especially P193L and E229K, much of the SAC region has collective motions along EV2, 3 and 4 also (Figure 4.9). Thus, the SAC fluctuations as observed in the WT are not unidirectional in the MTs, but more complex. All these observations indicate that MTs deviate from the native dynamics in HBR, SBR and SAC regions.

A clearer picture is obtained from the representation of average structures of the WT and MTs, which are colored for each residue, according to decreasing magnitude of RMSF along EV1 (Figure 4.10 and Table 4.9). In WT, the region comprising the F/G loop, the β-rich domain and the B'-helix, which forms the opening of the substrate access channel (Box-1) has collective motions together with the loops C/D, D/E and H/I, that are on the other side of the protein where the electron donating partner, the p450-reductase interacts. This suggests that, within the constraint of simulation time, the principal functional motions in the native enzyme comprise the region of the SAC including the β-rich domain, and the loops present in the region of p450-reductase protein interaction.

Much of the core region is relatively static including the heme and substrate-binding regions (Box-2), which harbors heme co-factor and the structurally conserved I-helix. These observations agree with earlier studies which suggested that F/G loop and the adjoining β-rich domain are highly mobile regions in CYPs, and together with B'-helix form the opening of substrate access channel (SAC), aiding in substrate access and product exit (Ludemann *et al.*, 2000a; Winn *et al.*, 2002). In the case of MTs

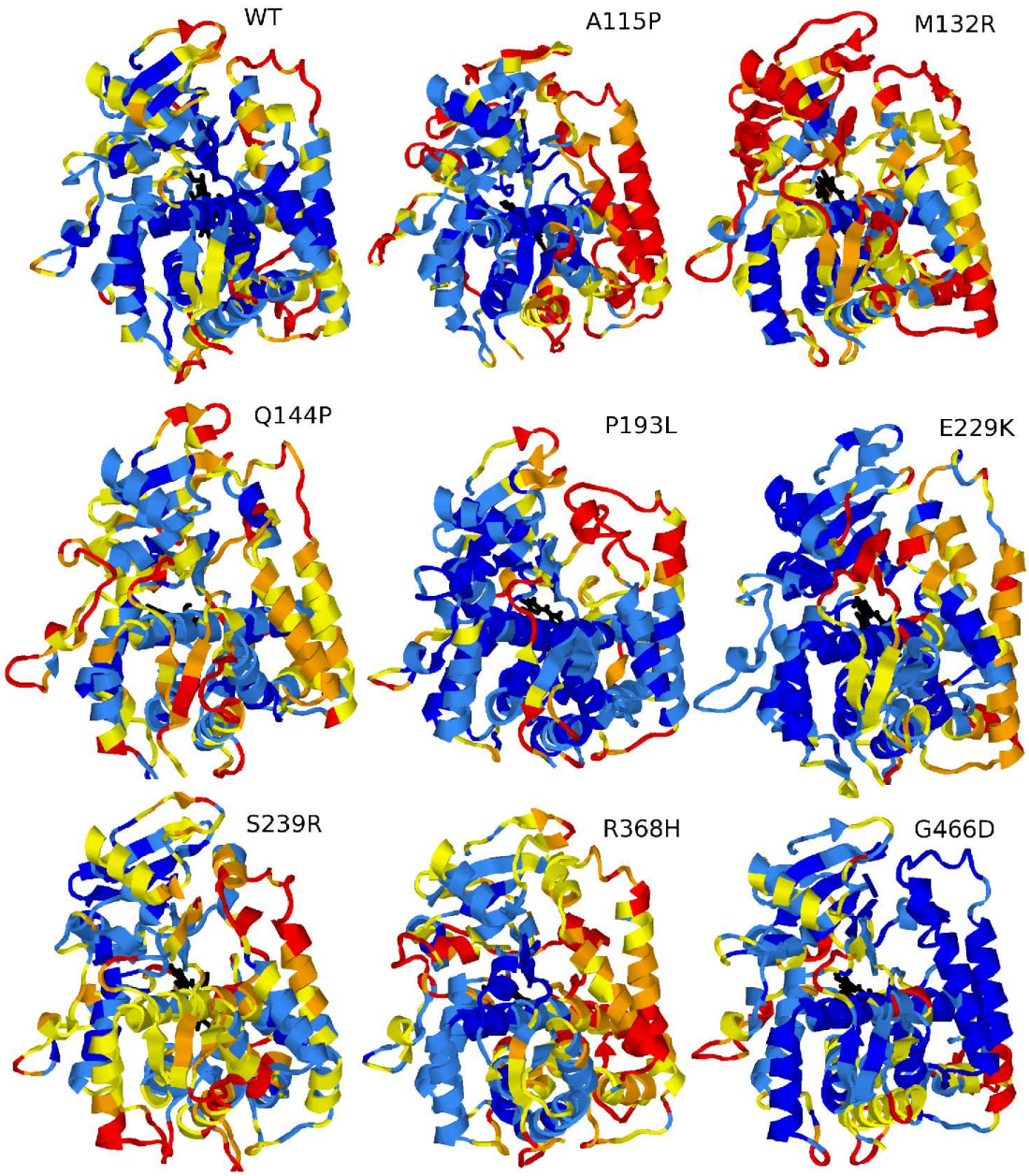


Figure 4.10: The distribution of RMSF in WT and MTs along Eigen Vector 1. The colors indicate the magnitude of fluctuations; Blue (0Å to 0.25Å), Sky-blue (0.25Å to 0.5Å), Yellow (0.5Å to 0.75Å), Orange (0.75Å to 1Å) and Red (=1Å). The average structures were used for representation, after superimposing onto the coordinate frame of the WT average structure.

A115P, Q144P, P193L, S239R and R368H the B'-helix of SAC is also involved in large collective motions along PC1. In the WT however, the B'-helix is relatively less mobile than the F/G loop and the β-rich domain, as it has stabilizing Hb interactions with β-sheet S5, B'/C loop of the β-rich domain and the G-helix; which is consistent with the reported literature (Negishi *et al.*, 1996). In all the MTs except E229K and G466D, one or all of these Hb interactions are absent. The MTs E229K and G466D do not involve either F/G loop, the B'-helix or the β-rich domain in the collective motions.

As previously noted, collective fluctuations in the relatively rigid regions of the structure, that is the HBR and the SBR can be observed in all the MTs. There is involvement of one or more of the secondary structures in SBR and HBR, including the α-helices CA I, F, G, L. M132R has the involvement of all of these conserved core structures in the collective fluctuation. A115P has the entire F and G helices in collective fluctuations. This is also found to a lesser extent in Q144P, S239R and E229K. P193L has the highest collective fluctuations in the SAC region, including the whole of B'-helix. From these results, it can be said that one of the primary effects of mutations may be impaired substrate recognition and/or access, as a result of altered dynamics of the SAC (especially in P193L). Many of the mutations are also associated with changed collective dynamics that involve the structural core of the HBR and SBR, probably resulting in compromised heme-binding or substrate-binding.

4.3.4 Essential motions along the common EVs of the WT and MTs

EVs calculated for combined trajectories will give a quantitative comparison of ED between the WT and MT simulations, since collective motions are studied along the common directions. The average projections (Figure 4.11) represent the equilibrium

conformations along the EVs. The differences in average projections indicate a 'static shift' in the equilibrium conformations. It can be observed that up to EV8 'static shift' is observed in all the cases, which is due to the differences in equilibrium conformations with respect to the other eight structures, indicating significant differences among the equilibrium structures of WT and MTs.

In addition to the static shift, there are also differences in the magnitude of collective dynamics among the structures along the common EVs, which are described mainly by the EV 9 and the then higher EVs (Figure 4.12). The WT and MT structures have the largest collective fluctuations along some of these EVs, visible as peaks in the graphs. From a comparison of the inner products, it was found that the EVs corresponding to the peaks have some overlap (inner product $= 0.3$) with the top ranking EVs (most predominantly EV1 and EV2), of the respective individual simulations. M132R has the highest mean square fluctuation along EV9, followed by Q144P and S239R. All other structures show negligible fluctuations in this direction. Thus, the fluctuations in M132R represented by EV9 are unique, which are found to a lesser extent only in Q144P and S239R. The collective fluctuations in P193L, S239R, R368H and Q144P are represented by EV10. The fluctuations described by EV11 are unique to A115P and Q144P. EV12 describes the collective fluctuation in Q144P and A115P. The collective fluctuations in G466D and to lesser extent in S239R are represented by EV13. The direction described by EV14 is represented by fluctuations in R368H and WT. P193L is again unique since it is the sole structure having collective fluctuations along EV15. EV16 is represented by only the WT, while not sharing with any other molecules. A structural representation of distribution of fluctuations along common EVs can be found in Figure 4.13.

It can be noted that none of the MTs have significant fluctuations along the EV16, the vector along which the WT was having maximum fluctuation, indicating that the

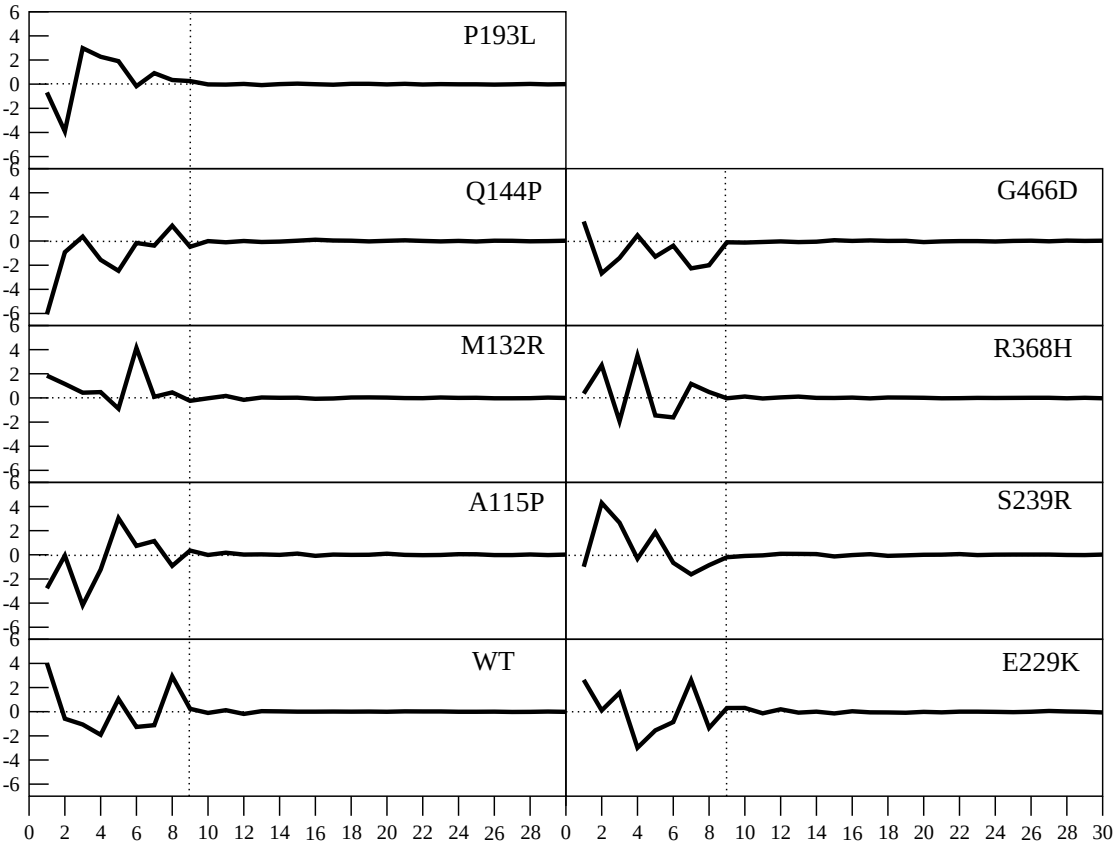


Figure 4.11: Average projections of individual WT and MT trajectories onto the common Eigen Vectors.

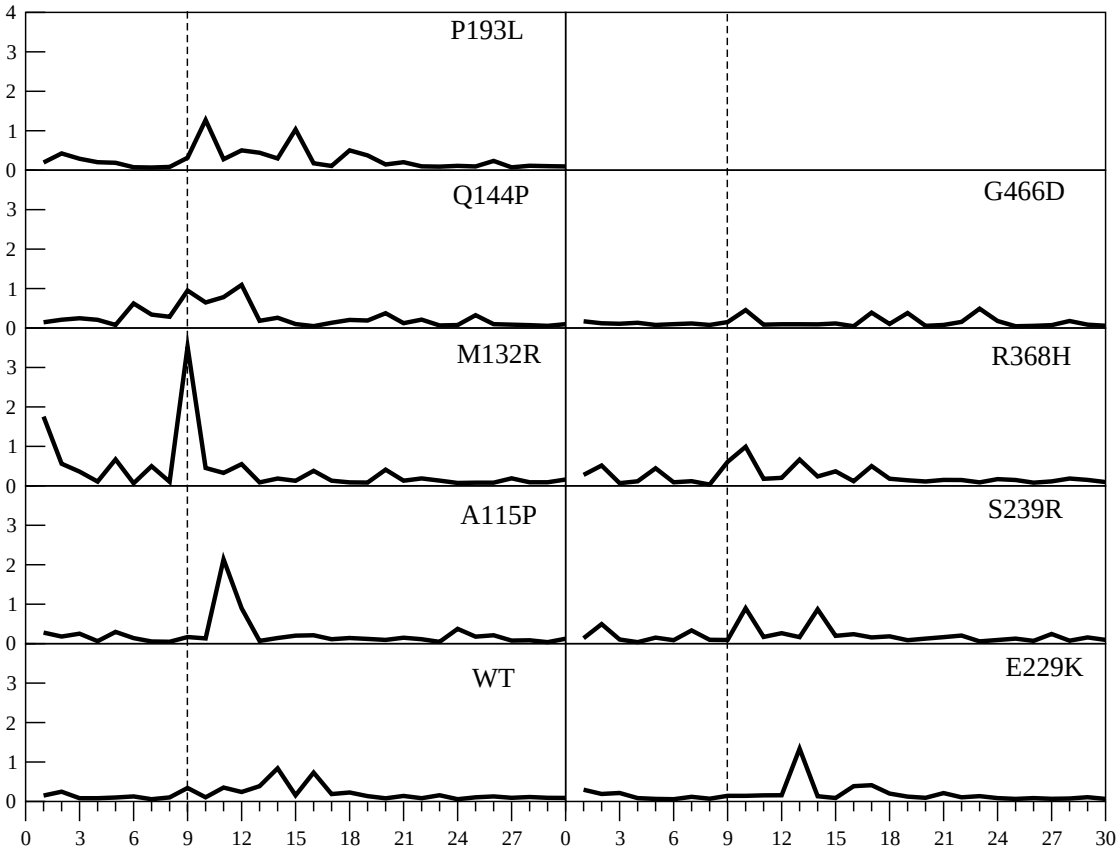


Figure 4.12: Mean square fluctuation of the WT and MT trajectories along the common Eigen Vectors.

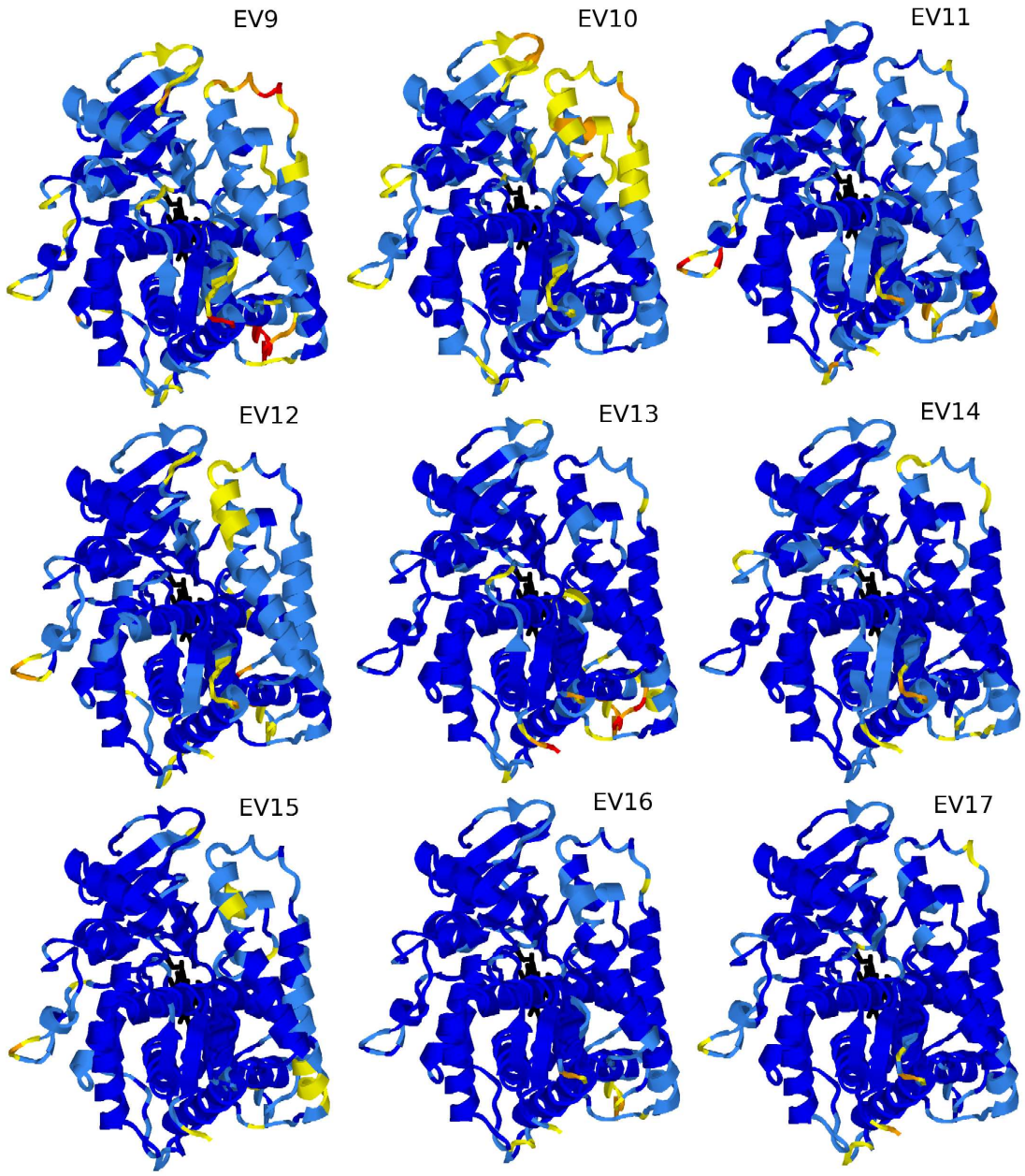


Figure 4.13: Representation of the Common collective fluctuations along the common EVs 9 to 17. The magnitudes of fluctuations in WT and MTs are given by Mean square fluctuation, along these vectors as shown in Figure 4.12. The colors (as indicated in Figure 4.13), indicate the distribution according to the magnitude of RMSF

native pattern of collective motions are lost in the MTs. From these observations it is evident that the essential motions observed in MTs are not similar to that of WT and MTs have unique collective motions some of these being common among them.

4.3.5 The use of homology models in ED studies

The current analysis of Essential motions in CYP1b1 and its MT forms is based on homology models. It is essential that the starting structures used for the simulations should be reasonably accurate, to avoid the possibility that differences in the observed dynamics are artifacts arising from using inaccurate starting structures. The 3D structure among Cytochrome p450s is highly conserved despite low sequence similarities and such high structural conservation among p450s is perhaps due to their common mechanism of electron and proton transfer and oxygen activation (Werck-Reichhart & Feyereisen, 2000). It is observed that the RMSD between CYP proteins doesn't substantially increase with decrease in their pair-wise sequence identities (Figure 4.14), indicating structural conservation even among CYPs with low sequence identities. This justifies the utility of CYP homology models in the structural studies.

Furthermore, to test the reliability of the simulations and verify that the differences in dynamics between the WT and MTs are the true reflection of their properties and not the possible artifacts of incorrect initial structures, a simple test was carried out, in which two of the MTs showed most deviation from the WT, were re-mutated back to WT and subjected to MD simulations. The MTs after reverting to WT acquired structural properties similar to the WT (Table 4.10), indicating that the results of the current study are not artifacts of incorrect initial structures.

Table 4.9: Structures showing correlated fluctuations greater than ≥1Å along PC1 in WT and MTs. These are quantitatively identified using the criteria, wherein a secondary structure was considered to be involved in collective fluctuation, only if ≥50% of its residues have fluctuation ≥1Å. See Figure 7 of the main manuscript for definitions of secondary structures.

Structure	Helices	Sheets	Loops
WT			N-term, C/D, D/E, F/G, H/I, C-term
A115P	B'	$S^{2.2}$, $S^{4.1}$, $S^{4.2}$	N-term, D/E, E/F, F/G, G, H/I, K'/L
M132R	A, B, G, H	$S^{1.1}$, $S^{2.1}$, $S^{4.1}$	N-term, A/B, B/B', B'/C, D/E, F/G, G/H, H/I, J/K, K/K', K'/L
Q144P	B'		N-term, B'/C, D/E, F/G, H/I, K'/L, C-term
P193L	B'		N-term, B/B', D/E, F/G, G/H, K'/L, C-term
E229K		$S^{3.1}$, $S^{3.2}$	N-term, D/E, C-term
S239R	B'		N-term, B'/C, D/E, E/F, F/G, H/I, J/K, K'/L, C-term
R368H	K'		N-term, B/B', B/B', E/F, H/I, K/K', K'/L
G466D		$S^{4.1}$, $S^{4.2}$	B/B', C/D, D/E, H/I, K'/L, C-term

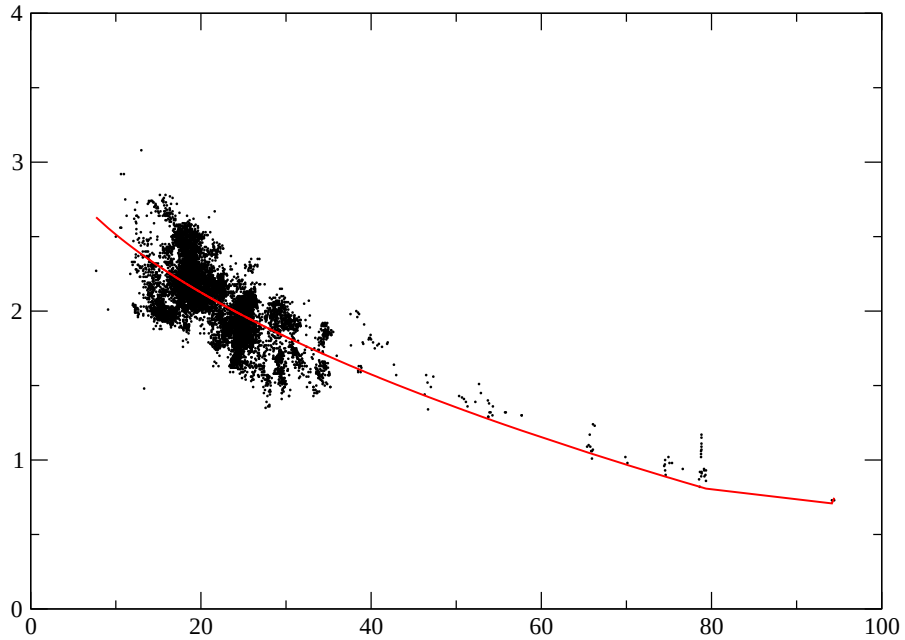


Figure 4.14: The behavior of RMSD vs % sequence identity among the known p450 structures. The Stamp structural alignment program was used for this purpose

Table 4.10: Time Average structural properties of the simulations of WT, MTs and MTs reverted to WT.

	WT	M132R	M132R to WT	P193L	P193L to WT
CYS470(SG)-Heme(FE) bond length (Å)	2.1	2.2	2.1	3.0	2.3
Volume of SBR ($Å^3$)	343	385	381	605	400

4.4 Conclusion

ED analysis or PCA was performed on the Human CYP1b1 and its PCG mutant forms. For this MD simulation data of minimum 35 nanoseconds (ns) in all cases were used. The ELs indicated that only the first few EVs (rather just the first EV) in both the WT and MTs were sufficient to describe majority of the collective motions, which implies that only one or two largest collective motions amply describe the functional motions of this enzyme. Analysis of WT revealed that the F/G loop, B'-helix and β-rich domain constituting the SAC region which were earlier shown to be involved in the substrate access in p450 enzymes, exhibit collective motions. The Collective motions observed in this region of high flexibility (Chapter 3), could be an important functional property in substrate recognition and access.

Further, it has been observed that the other FIRs, namely HBR and SBR regions, do not show large collective motions in the WT. MTs however, differed from the WT in the nature of these collective motions, at the HBR and SBR regions. On the whole, it was seen that the collective motions in MTs are associated with reduced integrity of dynamics at the functionally important regions of the enzyme. Thus, in this work some insights were gained about the functional motions present in the CYP1b1 enzyme, and the possible effects of disease causing mutations on them.

5

Ligand Binding Studies of WT and MT CYP1b1

5.1 Introduction

Structural analysis of the WT and PCG MT forms of Human CYP1b1, as described in Chapters 2 to 4, gave valuable insights into the nature of possible structural changes as a result of mutations. The modeling and simulation studies especially focusing on the FIRs suggested that altered structural properties of these regions could be detrimental to the enzyme functionality. Hence, it would be interesting to validate some of the observed structural phenomenon of the MTs.

From the inferences made in the previous chapters, it is evident that mutations could have disruptive effect on the function, by a destabilization of the native properties of the FIRs. It was observed that the HBR and SBR regions in the MTs had altered structural and dynamic properties. Since these regions constitute the site of the catalytic activity of the enzyme, a comparative study of the binding of the substrate

in the active site of WT and MTs, will be helpful in gaining insights into the possible differences in the substrate binding and catalysis.

In this regard, Molecular docking is useful to find the optimum binding orientations of the ligand in the active sites of receptors. Various methods of docking exist as discussed in Chapter 1, each having specific advantages and disadvantages. Judicious selection of these methods and the docking parameters is required for getting the results that are biologically relevant. In this chapter, the Molecular docking study of the MD structures of WT and MT CYP1b1 with Estradiol (E2), is described.

The ligand that CYP1b1 metabolizes *in vivo* in the context of PCG is not known. However, CYP1b1 is found to be involved in the metabolism of the endogenous steroid E2, the metabolites of which have been implicated in carcinogenicity (Liehr & Ricci, 1996; Spink *et al.*, 1998). CYP1b1 was shown to catalyze the conversion of E2 to 2 and 4 hydroxylated metabolites (Figure 5.1) (Hayes *et al.*, 1996). *In vitro* kinetic studies of CYP1b1 harboring common polymorphisms, indicated altered activities in the conversion of E2 to its 2 and 4 hydroxylated forms (Shimada *et al.*, 1999). Further, E2 hydroxylation is considered to be the characteristic reaction catalyzed by CYP1b1 (Murray *et al.*, 2001). Therefore, E2 was considered as the ligand for the current comparative docking analysis of the WT and MT CYP1b1.

5.2 Material and Methods

Dockings were performed using GOLD software (http://gold.ccdc.cam.ac.uk), which uses genetic algorithm for finding the binding modes of the ligand. The Docking protocol has four main steps; a) Ligand preparation b) Receptor preparation c) Docking using a search algorithm and d) Analysis of the binding modes using a scoring function. Docking calculations were done on Dell Poweredge 6800 server running on Linux.

5.2.1 Ligand preparation

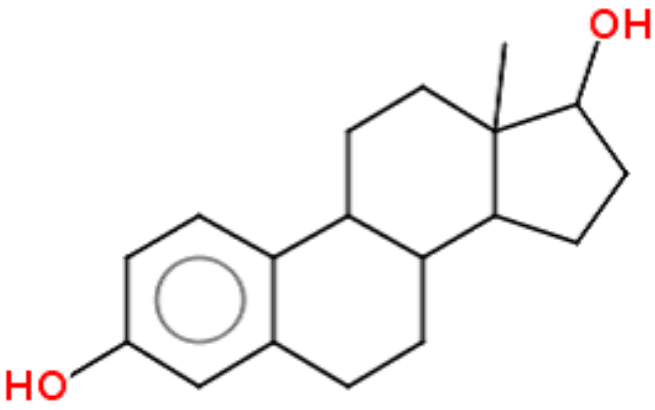
The common endogenous steroid E2 was used for docking. The molecular formula for the ligand was obtained from pubchem database (http://pubchem.ncb.nlm.nih.gov/) and the 3D structure was built using Java Molecular Editor software (http://www.molinspiration.com/jme/) (Figure 5.1)

5.2.2 Receptor preparation

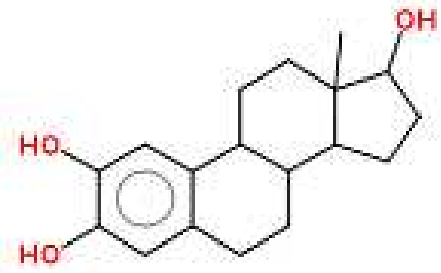
Multiple receptor structures in each case for WT and MT structures were used for docking experiment. The MD simulation snapshots were clustered using the Jarvis-Patrick algorithm (Jarvis & Patrick, 1973) implemented in the GROMACS program g_cluster. The Substrate Binding residues (SBR) were used for least squares fit. The number of nearest neighbors considered for clusters was 10 and a minimum of 3 identical nearest neighbors were required to form a cluster. The cluster centers of the top 3 highly populated clusters were selected for further analysis. The structures are energy minimized in GROMACS using the steepest descent with a termination criteria of $'$emtol=1 KJ/mol$'$, followed by conjugate gradient method with a termination criteria of $'$emtol = 0.01KJ/mol$'$. PROCHECK was used to check the conformation to be free from any bad contacts and the structures are converted to SYBYL-mol2 format using BABEL software, for use with the GOLD program.

5.2.3 Docking using a Search Algorithm

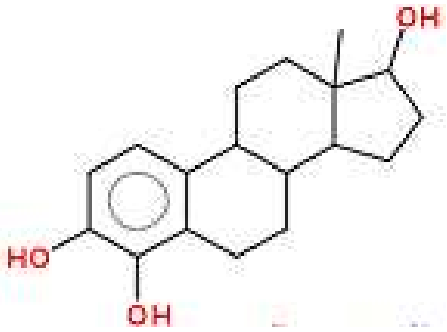
The residues of the SBR region (as explained in Chapter 2) were used to define the putative ligand binding sites and the atoms corresponding to SBR in each case were written to the file $'$cavity.atoms$'$. The Genetic Algorithm settings in GOLD were set to



2-OH-Estradiol



17b-Estradiol



4-OH-Estradiol

Figure 5.1: The structure of Estradiol (E2) and the two main catalytic products that show the site of oxidation.

′Automatic′ mode with an auto scaling of 1.0. By this method the program determines itself the optimum run parameters depending on the nature of the ligand and the receptor active site. Thus, the parameters like the crossover frequency, number of genetic algorithm runs, mutation rates etc are automatically adjusted by the program. The ligand atom types were reset in GOLD and a flood-fill center radius of 20Å was used to define the search space. The docking results were scored using the GOLD scoring function. In each case 50 Genetic Algorithm (GA) runs were performed.

5.2.4 Analysis of Ligand binding modes

The docking results of the 50 GA runs were clustered using an rmsd cutoff of 0.75Å and the best scoring conformation in each cluster was selected. The distance of the ligand atoms that are the sites of oxidation, that is closest to the heme FE and the average distance of the ligand were computed. The number of HBs and the number of non-bonded contacts between the ligand and the receptor were computed using LIGPLOT (Wallace *et al.*, 1995) program.

The volume of the SBR was found using the program POCKET (Levitt & Banaszak, 1992), which identifies cavities in the proteins. Since the program cannot handle hetero atoms like heme, the residue type for heme is changed to that of an AA. The program was used to find the volume of the SBR regions of the receptor structures of the current study. The SBR volumes of the known crystal structures of p450-ligand complexes from PDB were also computed for reference. While doing this, the ligand molecule from the complexes were removed before the computation of the volume. The figures of the best docked solutions were generated using the SPDBV (Kaplan & Littlejohn, 2001) software.

5.3 Results and Discussion

5.3.1 Modeling of Ligand

Docking studies were carried out to study the binding modes of E2, in the active site of the WT and MT structures. GOLD uses genetic algorithm to find the optimal ligand binding modes. Specifying the approximate location of binding site by the user significantly narrows down the search space and time. Thus, the Substrate binding Residues (SBR) as computed in Chapter 2 were specified to the program to define the search space.

5.3.2 Modeling of the Receptor structures

As mentioned earlier, several receptor structures generated from the MD simulations, were used in each case of WT and MT molecules, since this accounts for the flexibility at the binding site during the simulations. The local conformation at the substrate-binding site can vary during the course of the simulation. The receptor structures used for docking calculations were selected from the clustering of the trajectories. Since only the local conformation of the SBR is required for the docking calculations, a clustering based on the least squares fit of the SBR residues was done, to get the SBR conformations that have longer resident times. The conformations closest to the cluster centers were saved for further analysis. Since the structures thus obtained are from a dynamic simulation run, in which the atoms are in constant motion, these are energy minimized to correct any unrealistic bond lengths or bond angles.

The three cluster centers selected for docking are the three most populous conformations of the active site of the molecule during the simulation. The C^{α}-rmsd of the SBR residues between any of the three clusters in WT and the MT clusters ranges

from 2Å to 4Å indicating significant differences in the conformation at the active site between the WT and the MTs. Table 5.1 gives the volumes of the active sites of WT and MT molecules, corresponding to the structure of the three cluster centers. The active site volume in the case of WT for the first two conformations is around 270Å^3, while it is more in the third cluster. In some MTs, A115P, M132R, P193L, E229K and G466D the volume has more variation among the three clusters compared to WT, with an average volume higher than the WT. The active site volumes in WT and MTs are within the range found in the crystal structures of p450-ligand complexes. While in some of the MTs, the volumes are towards the higher extreme of the range observed among p450-ligand complexes.

5.3.3 The Protein Ligand Interactions

In each case, the solutions of 50 GA runs were ranked according to the GOLDSCORE fitness function. The scoring in the function is done based on the energy terms namely; protein-ligand HB energy, protein-ligand van der Waals energy, ligand internal van der Waals energy and ligand torsional strain energy. GOLD has an inbuilt clustering method based on the rmsd values, which is used to cluster the ligand poses having rmsd of $<$0.75Å . The top scoring ligand of the first cluster is the best pose according to the fitness function. The ligand orientation or the binding pose should be similar to that which is biologically meaningful, wherein the ligand binds in an orientation, in which the reactant atoms; the site of oxidation is proximal to the FE of heme.

We have examined the available crystal structures of CYP-ligand complexes to find the average distance of ligand atoms from the heme FE. In majority of the complexes, the distance is between 7Å to 13Å (Figure 5.2a). However, distances of less than 7Å and more than 13Å are also observed in some cases. The variation in the binding distance

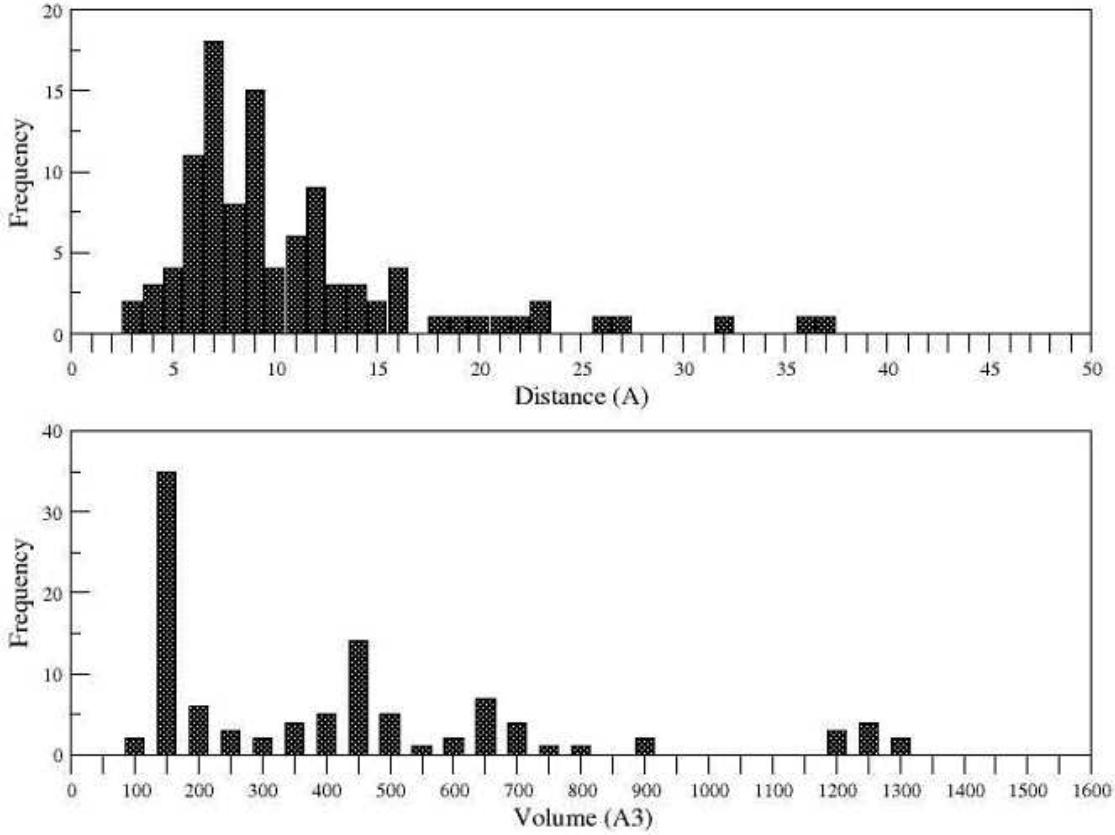


Figure 5.2: The frequency distribution of (a) Average distance (Å) of ligand from heme FE and (b) Volume ($\mathring{A}^3$) of the active site region, in 105 p450-ligand complexes obtained from the PDB

in the CYP-ligand complexes can be due to the differences in the nature of substrate binding region, and the bound ligands. The frequency distribution of volume of the active site region (Figure 5.2b) shows that, in most of the complexes the active site volume is about 150Å^3. The volume in many other complexes is about 450Å^3. In some complexes the SBR volumes are higher in the range of 500Å^3 to 1000Å^3. Thus, both the ligand binding distance and the SBR volume show considerable variation among the p450-ligand complexes. Moreover, the ligands in some of the complexes are inhibitors or enhancers and they may have binding sites quite different from the substrate-binding site. In the present study, the best ligand binding poses for the receptor structures representing the three clusters in the WT and MTs, are compared.

5.3.4 The Protein ligand Interactions

Table 5.1 gives the GOLD fitness scores for the best docking solutions for E2 binding. The scores are in the range of 40 to 50 in the WT and MT molecules. The table also gives the distances at which the ligand is positioned from heme FE. The average distance of the ligand atoms from heme FE, and the distances of the sites of 2-hydroxylation and 4-hydroxylation from FE are given. In WT E2 is bound at a distance of 10Å in the first cluster and about 12Å in the second and third clusters respectively (Table 5.1). The 2 and 4-hydroxylation sites are at 8Å and 11Å respectively in the 1st cluster and about 11Å and 12Å respectively in the second and third clusters.

The ligand makes single HB interaction with I350 of β-strand S9 or K463 of β-strand S17 respectively in the first and second clusters. The other interactions are those of Vander Waals contacts with the receptor, which is the largest in the case of cluster2. In the case of MTs the distance at which the ligand is bound varies considerably between the three clusters studied, indicating changes as noted earlier, in

conformation of the active site in the MTs. This is also seen in Table 5.1 where in the volume of the active site is more varying in the MTs than the WT. The GOLD fitness function has been optimized for a better prediction of ligand binding positions rather than ligand binding affinities (Jones *et al.*, 1997), and thus the best scoring solutions obtained in the various docking runs of WT and MTs are an indication of the best ligand binding positions rather than their relative strength of ligand binding. Hence, no attempt was made to compare the binding affinities in WT and MTs, but rather the orientations and binding positions, have been compared.

Table 5.3 gives the list of non-bonded contacts between the Protein residues and the ligand. In WT, the ligand has non-bonded contacts with the residues of F, G, I helices, β-strands S9 and S17 and loops B'/S6, K/S9. In the first cluster, the ligand is oriented perpendicular to the axis of the I-helix, having non-bonded contacts with loop K/S9 and β-strand S9 near the roof of the SBR. In the other two clusters, the orientation of the ligand is parallel to I-helix and makes additional contacts with the residues of loop B'/S6. Figure 5.3.4 gives the schematic representation of ligand binding in WT and MTs in the first three clusters. In the first cluster in WT, E2 is bound with its phenol group oriented towards the heme. The 2-hydroxylation site is close to the FE at a distance of 8Å . In the second cluster, the ligand is oriented parallel to I helix and the heme plane, while in the third cluster, the ligand is positioned more towards the B helix and parallel to I helix. In both second and third clusters, the reactive sites in the ligand are at distance of 10-12Å from FE. The ligand is also having stacking interactions with aromatic side chains F231, F261, F134 (Table 5.3) in the second and third clusters, which is not found in the first cluster. The main difference between the ligand poses in second and third cluster is that in the second cluster, the 2 and 4-hydroxylation sites are near to the heme, whereas in third cluster,

Table 5.1: The Volume of SBR of the three clusters of WT and MT structures. The Gold Fitness Scores of the docking calculations, the number of HB interactions, number of non-bonded contacts and the Distance of Ligand atoms from Heme FE.

	Cluster No.	Volume of SBR (Å^3)	GOLD Fitness Score	Ligand-FE Average Distance (Å)	Ligand 2OHsite-FE Distance (Å)	Ligand 4OHsite-FE Distance (Å)	Ligand-Receptor H-bonds	Ligand-Receptor NB-contacts
WT	1	275	38	10	8	11	1	53
	2	261	46	12	11	12	1	85
	3	432	46	12	11	13	0	66
A115P	1	279	44	11	7	8	0	65
	2	280	41	11	7	7	0	73
	3	649	46	9	13	12	0	84
M132R	1	695	40	7	3	5	0	69
	2	324	46	6	8	7	1	94
	3	165	49	7	11	10	0	86
Q144P	1	308	45	6	10	9	0	90
	2	260	44	7	10	10	1	84
	3	310	42	7	10	11	2	70
P193L	1	139	39	7	10	10	0	67
	2	683	42	14	15	14	0	81
	3	434	38	14	18	18	0	56
E229K	1	235	42	7	5	6	1	68
	2	356	39	13	17	17	1	57
	3	784	44	11	13	11	1	74
S239R	1	236	49	8	4	3	0	73
	2	274	51	6	10	9	0	94
	3	188	61	6	3	3	0	127
R368H	1	215	44	5	5	5	2	139
	2	216	40	13	15	14	0	81
	3	241	33	12	12	14	0	106
G466D	1	598	42	6	8	6	0	78
	2	306	39	9	9	6	2	69
	3	319	40	8	3	4	0	87

Table 5.3: Structures along with the residues involved in Non-bonded contacts with E2 in WT and MT structures

MOL	CLS	B'	E	FE	F	G	I	K'	K	K/S9	B'/S6	S12	S17	S5	S9
WT	1	-	-	-	F231, G232, V235, G236, A237	F261	G329, A330, D333	-	-	V395	F134	-	-	-	V397, T398, I399
	2	-	-	-	F231, G232, G236, A237	F261	T325, D326, G329, D333	-	-	-	S127, G129, R130, F134	-	K512	-	-
	3	-	-	-	F231, G232, G236	F261	D326, G329, D333	-	-	-	S127, G129, R130, S131, A133, F134	-	-	-	-
A115P	1	S122, F123, V126	-	-	V235	-	D326, A330, S331, T334	-	-	-	R130, F134	-	-	-	-
	2	F123, V126	-	528	-	-	D326, A330, S331, T334	-	-	-	S127, R130, F134	-	-	-	-
	3	S122, F123, V126	-	528	V235	-	A330, S331, T334	-	-	-	R130, F134	-	-	-	-
M132R	1	-	-	528	-	-	A330, T334	Q424	S392, S393	F394, V395	-	-	-	-	T398, I399
	2	-	-	528	-	-	T325, D326, G329, A330, T334	Q424	-	F394	S131, F134	-	-	-	I399

Continued. . .

Table 5.3 . . . Continued

MOL	CLS	B'	E	FE	F	G	I	K'	K	K/S9	B'/S6	S12	S17	S5	S9
	3	-	-	528	F231, G232, V235	-	G329, A330, D333, T334	Q424	-	F394	F134	-	-	R117	I399
Q144P	1	-	-	528	-	-	D326, A330, S331, T334	Q424	-	F394, V395	A133	-	-	-	V397, T398, I399
	2	-	-	528	-	-	D326, A330, S331, T334	Q424	-	F394	G129, R130, A133, F134	-	-	-	T398, I399
	3	-	-	528	-	-	D326	-	-	V395	S127, G128, R130, A133, F134	-	-	-	V397, I399, P400
P193L	1	-	-	528	N228	-	D326, G329, A330, D333, T334	-	-	V395	R130, A133	-	-	-	T398, I399
	2	-	R194	-	L225, N228, E229, G232, R233, T234, L240	-	G329, Q332, D333	-	-	-	-	-	K512	-	-
	3	R124, V125, V126	-	-	-	-	D326	-	-	-	G128, R130, A133, F134	-	-	F120	I399
E229K	1	-	-	528	F231	-	T325, D326, G329, A330	-	-	-	A133	-	-	R117	T398, I399

Continued. . .

Table 5.3 . . . Continued

MOL	CLS	B'	E	FE	F	G	I	K'	K	K/S9	B'/S6	S12	S17	S5	S9
	2	-	-	-	G232, G236, G238	-	D333, T334, T337	-	-	F394, V395	-	Y507	T510, I511, K512	-	-
	3	-	-	-	F231, G232, V235	-	T325, D326, G329, D333	-	-	-	S127, G129, R130, S131, A133, F134	-	-	-	-
S239R	1	S122, F123, V126	-	528	-	-	A330, S331, T334	-	-	-	S127, A133	-	-	R117	T398, P400
	2	F123, V126	-	528	-	-	A330, S331, T334	-	-	-	S127, M132, A133	-	-	-	T398, I399, P400
	3	V126	-	528	-	-	A330, S331, T334	-	-	V395	S127, M132, A133	-	-	-	T398, I399
R368H	1	-	-	528	-	-	I327, A330, T334	Q424	-	V395	A133	-	-	-	V397, T398, I399
	2	-	-	528	-	-	A330, D333, T334, T337	W425	S393	F394, P396	S127	L509	T510, I511	-	-
	3	F123	-	528	N228, F231, G232, V235	-	G329, A330, D333, T334	-	-	F394, V395	S127, G128, G129	L509	T510	-	-
G466D	1	-	-	528	-	-	D326, A330, S331	-	-	V395	M132, A133	-	-	-	V397, I399
	2	-	-	-	-	-	D333, T334, T337	-	-	F394, V395, P396	F134	L509	T510, K512	-	V397, I399

Continued. . .

Table 5.3 . . . Continued

MOL	CLS	B'	E	FE	F	G	I	K'	K	K/S9	B'/S6	S12	S17	S5	S9
	3	-	-	528	F231, V235	-	G329, A330, D333, T334	-	-	V395	A133	-	K512	-	I399

there is a 180°turn in the ligand making the reactive groups away from the heme. The orientation in the second cluster seems to be the functionally correct binding pose.

In MT A115P, in all the three clusters, the ligand has additional non-bonded contacts with B-helix and heme, apart from contacts with I-helix and loop B'/S6. The heme is displaced in position with respect to WT. The orientation of the ligand is perpendicular to the axis of I-helix similar to that in cluster 1 of WT. As seen from Figure 5.3.4, in this orientation there is no proper stacking interaction with the aromatic residues in the active site. The ligand protrudes into the space available between I-helix and heme, thus located at a shorter distance from FE of heme. In M132R, there is a similar change in the position of heme and change in the conformation near the heme-binding region as that of A115P, thus creating more volume. The ligand as a result binds close to the heme, making contacts predominantly with heme and I helix, and lesser contacts with K-helix, and loops K/S9, B'/S6. In the first two clusters the ligand orients parallel to the axis of heme and I-helix, but more closer to the heme as seen from Table 5.1. As seen from figure 5.2, the ligand at this position has no interactions with phenyl residues as found in WT. In the third cluster, the ligand orients vertically, similar to that in cluster 1 of WT, but the ligand is much closer to heme. In Q144P, in all the three clusters, the ligand is oriented in a horizontal position similar to that found in WT but at a distance of much closer about 7Å, to heme. The interactions with phenyl residues are again absent and the ligand has non-bonded interactions with the same structures as seen in M132R.

In P193L, the ligand orientation is similar to the cluster 1 of WT but is much closer to I-helix as seen from the interactions in Table 5.3. In the second cluster as seen from the figure 5.2, the ligand is located parallel to the heme plane and I-helix but much away from the heme making more contacts with the F-helix. The distance from the FE of heme as seen from Table 5.1 is 14Å, which is higher than in any other structures.

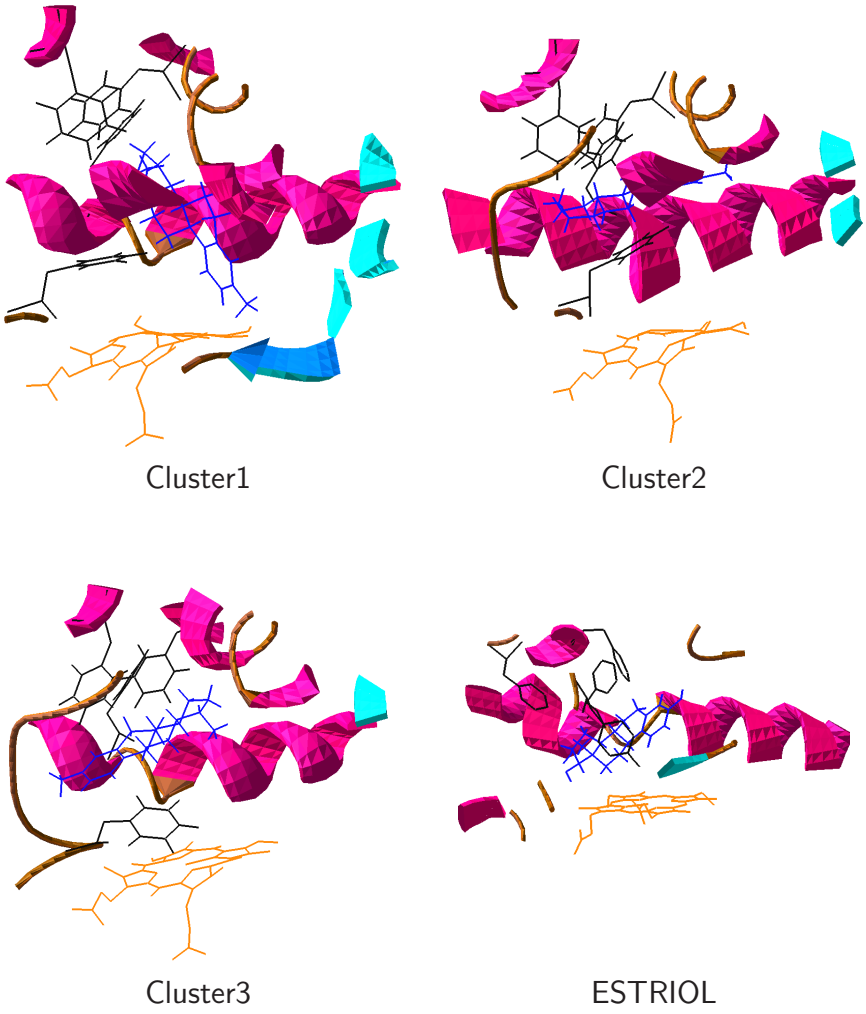
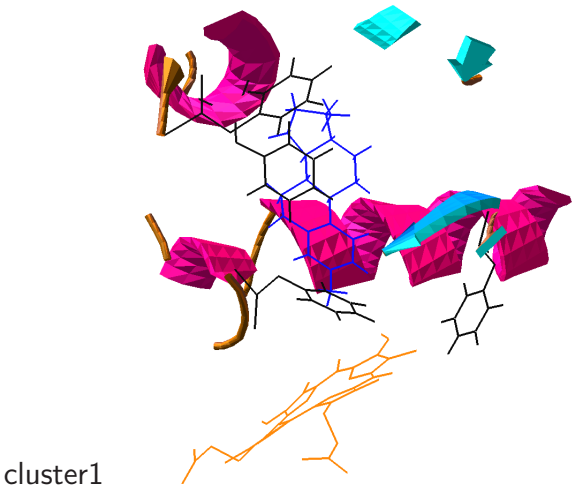
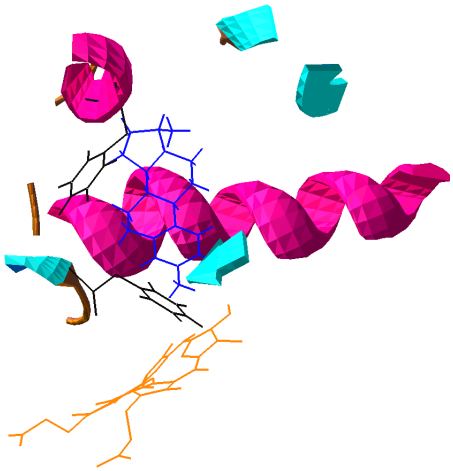


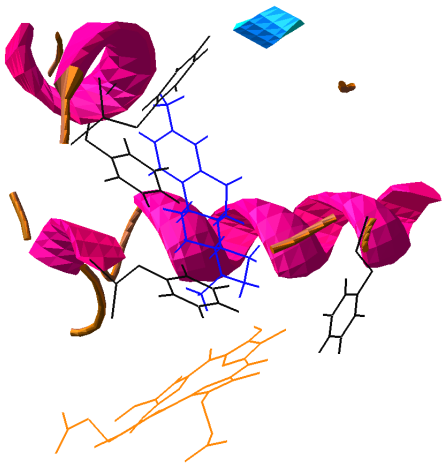
Figure 5.3: Projection diagrams of Estradiol docked into the three clusters of WT CYP1b1. The schematic figure of the reference complex- CYP51-ESTRIOL complex is also shown. Estradiol is colored in blue. Heme is colored in orange. The receptor's helices and Sheets and loops within 5Åfrom the ligand are represented in ribbon form and colored in red and blue and brown respectively. Aromatic sidechains within 5Åof the ligand are also shown in black color. For a better comparison part of I helix is also shown in the background in ribbon. The schematic representations of Estradiol docked into the three clusters in the Mutant structures are shown in subsequent figures; Figure 5.3.4 to Figure 5.3.4.



cluster1

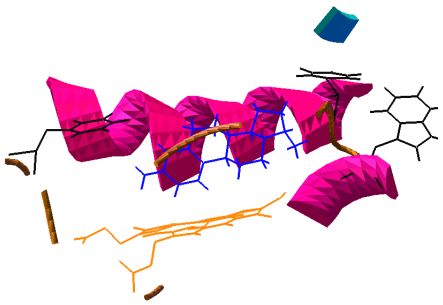


cluster2

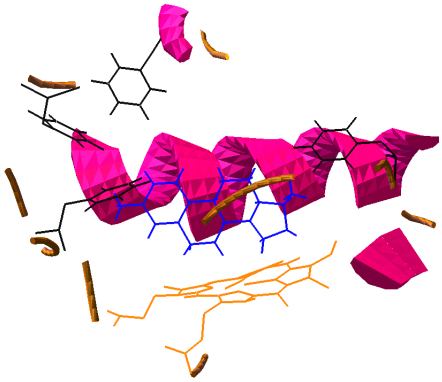


cluster3

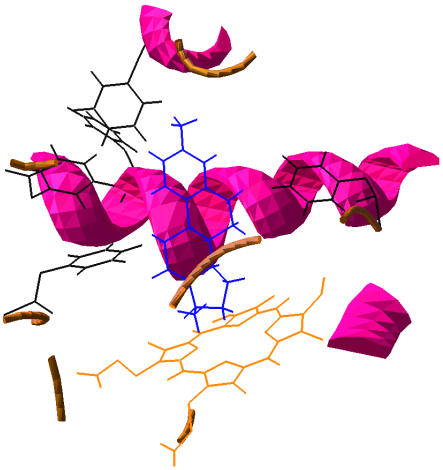
Figure 5.3.4 . . . continued. Projection diagram of Estradiol bound to the 3 clusters of A115P



cluster1

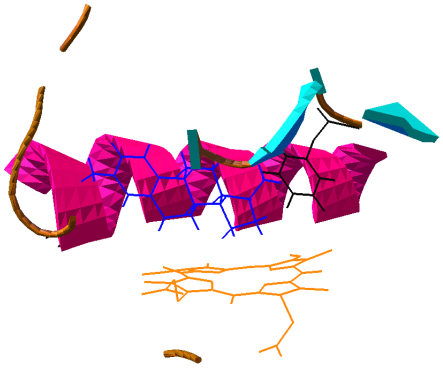


cluster2

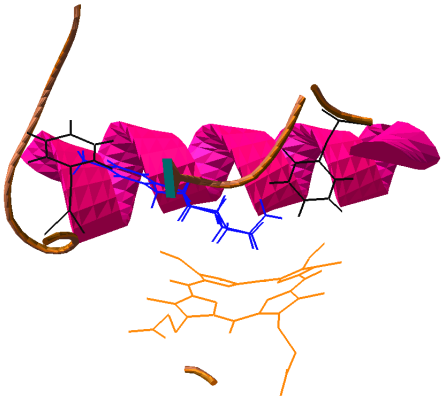


cluster3

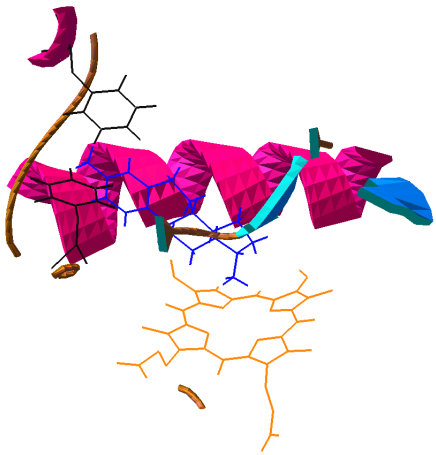
Figure 5.3.4 . . . continued. Projection diagram of Estradiol bound to the 3 clusters of M132R



cluster1

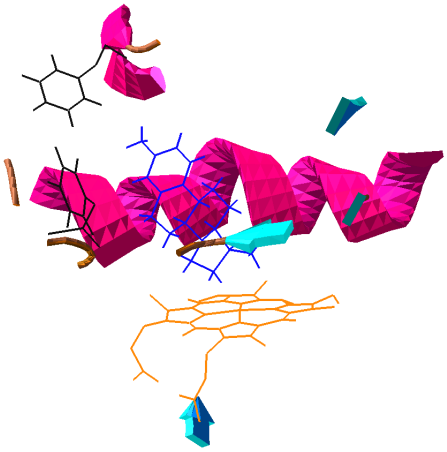


cluster3

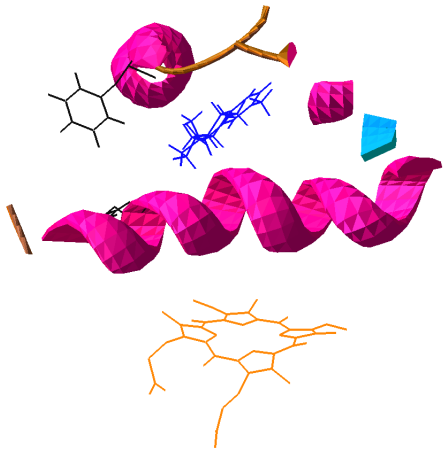


cluster3

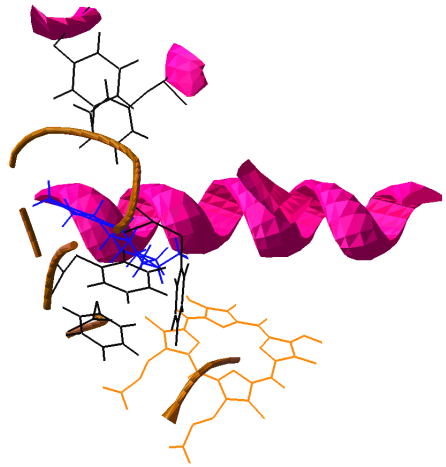
Figure 5.3.4 . . . continued. Projection diagram of Estradiol bound to the 3 clusters of Q144P



cluster1

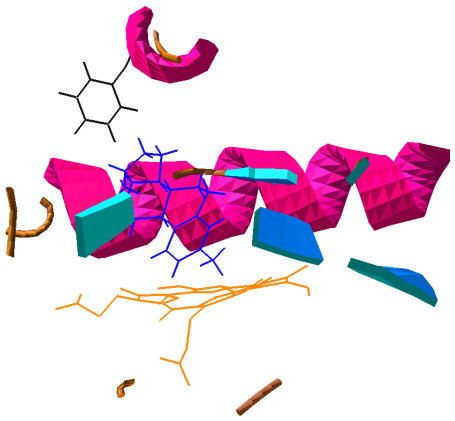


cluster2

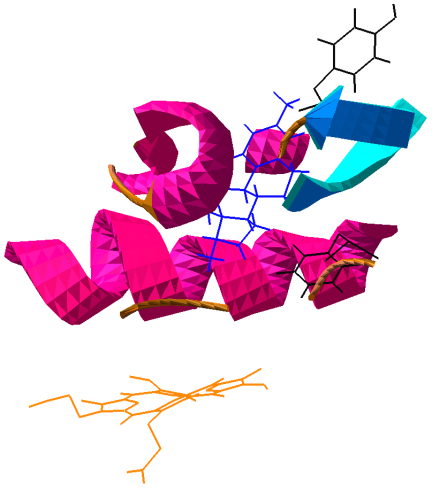


cluster3

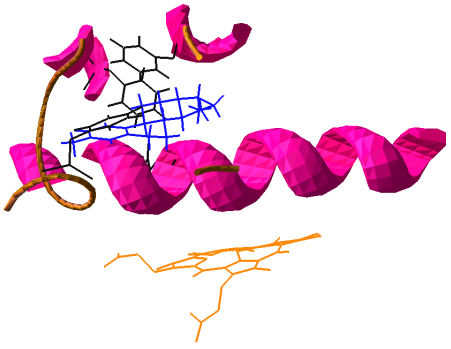
Figure 5.3.4 . . . continued. Projection diagram of Estradiol bound to the 3 clusters of P193L



cluster1

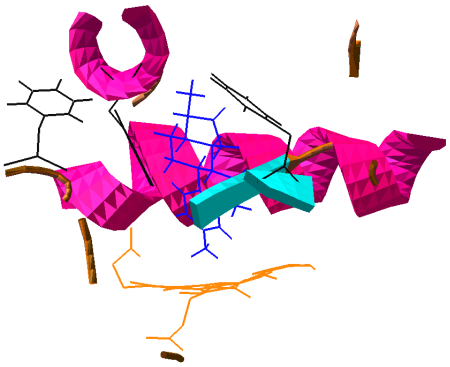


cluster2

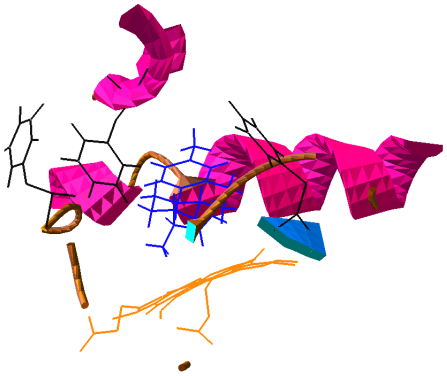


cluster3

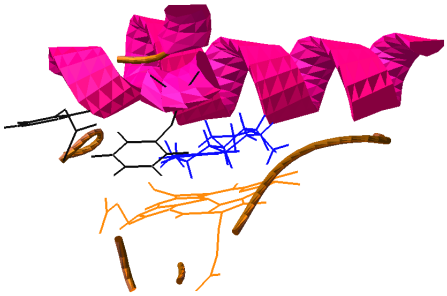
Figure 5.3.4 . . . continued. Projection diagram of Estradiol bound to the 3 clusters of E229K



cluster1



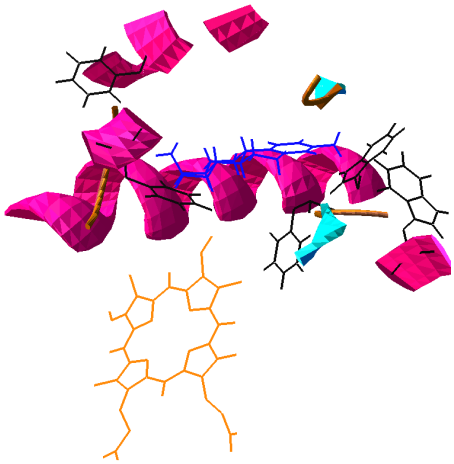
cluster2



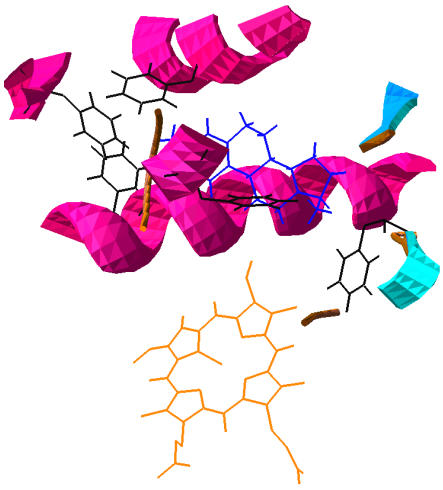
cluster3

Figure 5.3.4 . . . continued. Projection diagram of Estradiol bound to the 3 clusters of S239R

cluster1

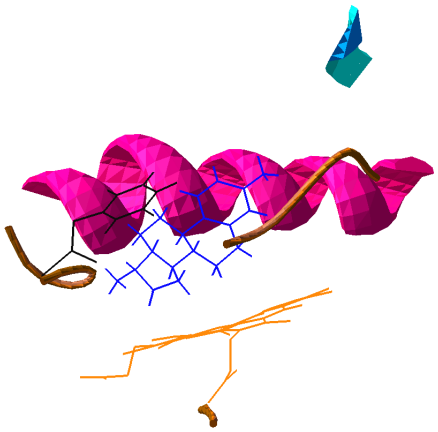


cluster2

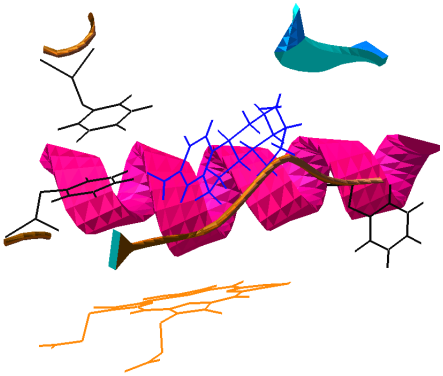


cluster3

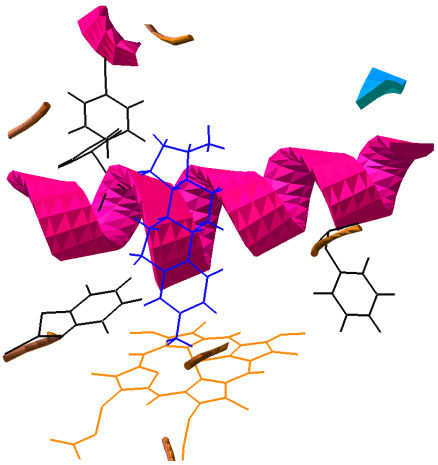
Figure 5.3.4 . . . continued. Projection diagram of Estradiol bound to the 3 clusters of R368H



cluster1



cluster1



cluster1

Figure 5.3.4 . . . continued. Projection diagram of Estradiol bound to the 3 clusters of G466D

The binding at a much farther place from heme indicates a larger volume of active site, which can be observed in Table 5.1. The orientation of ligand in cluster 3 is more towards the B-helix and loop B'/S6. As seen from the figure 5.2, the ligand makes interactions with the aromatic residues, but located much farther from the reaction center of heme.

In E229K, in the first cluster the ligand is oriented perpendicular and close to heme, with non-bonded contacts with heme and I-helix. In the second cluster too the ligand is perpendicular to heme but situated away from heme and having contacts with F-helix and loop K/S9 and β-strand S17 located towards the roof of the SBR. As in other cases, the perpendicular orientation of the ligand has no stacking interactions with the aromatic side chains in the active site. Ligand orientation in cluster 3 is similar to that observed in cluster 3 of P193L, in terms of stacking interactions with aromatic residues, but the ligand is located closer to I-helix than B helix as observed in P193L. In S239R, in ligand is in a similar orientation in the first 2 clusters as seen from the non-bonded contacts list in Table 5.3. The ligand is much closer to heme and I-helix and is pointing towards the Substrate access channel, as seen from the contacts with the B helix from Table 5.3. In the third orientation, the ligand is parallel to the plane of heme and located at about 6Å from heme FE.

In R368H, in the first cluster the ligand is oriented parallel to the heme plane at a close distance of 5Å. This results in contacts only with the heme and I-helix. In the second cluster, the ligand is located a distance of 13Å similar to the WT, but located more towards the I-helix making contacts with it, in contrast to the contacts observed in WT with the F-helix and loop K/S9 along with I-helix. This precludes from proper interaction of the ligand with the aromatic residues in the active site. In the third cluster, the orientation of the ligand is partly similar to the cluster 2 in WT having similar contacts. However from Table 5.1, it can be noted that the ligand orientation

differs from WT in that, its 2 and 4 hydroxylation sites are farther away from heme FE. As seen from the figure 5.2, in clusters 2 and 3, a change in the relative position of heme with respect to the active is observed.

In G466D, from Table 5.1 it can be seen that, the SBR volume in the first cluster is higher than the WT and the ligand is bound at a closer distance of about 6Å from heme FE, thus making contacts predominantly with heme and I-helix. The ligand orientation in the second cluster is similar to that found in the third cluster of WT but oriented more towards the I-helix with no contacts from the F-helix. The ligand conformation in the third cluster is perpendicular to the heme FE, accommodating the minor change in the position of the heme relative to the active site.

5.3.5 Comparative substrate binding analysis

From the above observations, the binding of the ligand in the active site of CYPs seem to predominantly dependent on the non-bonded or hydrophobic interactions, that the ligand makes with the protein. There were only occasional occurrences of 1 or 2 HBs between the ligand and the receptor. This is plausible as the p450 ligands are hydrophobic in nature. The GOLD Fitness scores and the NB contacts have some correlation, indicating the importance of the non-bonded contacts in ligand binding. CYP1b1 active site region has a cluster of side chains of aromatic AAs, which stack with the ligand molecules. Such hydrophobic nature of the active site was also observed in the crystal structures of CYP3A4 (Williams *et al.*, 2004) and CYP24A1 (Masuda *et al.*, 2007).

In the current study, the ligand pose corresponding to the second cluster is similar to that found for 'Estriol' in CYP51 (Podust *et al.*, 2004) (PDB-ID: 1x8v) which is structurally similar to E2, among the known p450-ligand complexes. In this complex

Estriol is bound at a distance of 8Å (average distance of all atoms from FE)(see Figure 3 (Estriol)). The equivalent sites corresponding to the 2 and 4-hydroxylation sites of E2 are also at a distance of 8Å from heme FE. The orientation of the ligand is parallel to the heme plane and along the axis of I helix. There are also stacking interactions with the aromatic side chains Y76, E78, F255, similar to those found in the current study. Thus, from these observations, the docking pose observed in the cluster 2 of WT is considered as reference to compare and contrast the ligand poses in the other MTs.

In earlier studies it was suggested that the SBR region of the p450s are relatively rigid and may not have large conformational changes upon ligand binding. The crystal structures of CYP3A4, ligand free and ligand complexed forms do not exhibit any significant conformational changes due to ligand binding (Williams *et al.*, 2004). However conformational changes upon ligand binding are observed in the case of CYP2B4 (Scott *et al.*, 2004). The conformational changes in the active site of the receptor seem to depend on the size of the bound ligand. In the current study of docking of E2, a moderate sized substrate, into CYP1b1 and its MT receptor structures, derived from homology modeling and MD simulation, the use of apo-form of enzyme for the docking analysis is justified based on the above observations.

Docking of E2 into a homology model of CYP1b1 which is based on the structure of CYP2C5, has been studied previously (Lewis *et al.*, 2003). In this study, the 4-hydroxylation site of E2 was found to be about 7Å from heme FE and the ligand binding to CYP1b1 had been proposed to be stabilized by a combination of $\pi - \pi$ stacking with aromatic residues and HBing. The ligand HB interactions in this model differ from the model in the current study. However, the $\pi - \pi$ stacking interactions are similar to those observed in current model. In the current study, the CYP1b1 model is based on CYP2c9 a human p450 template, solved at a better resolution. Further, a

rigorous method of MD simulation followed by clustering was employed to select the receptor structure.

Earlier studies on some of the CYP1b1 MTs and polymorphisms indicated altered activity towards E2. The MTs, G61E and R469W had compromised E2 hydroxylation activity (Jansson *et al.*, 2001) while two polymorphic variants, R48G and A119S did not show any change in activity (McLellan *et al.*, 2000). However, in another study the polymorphic variants, R48G, A119S, V432L exhibited altered kinetics while, N453S exhibited no change in E2 hydroxylation (Aklillu *et al.*, 2002). CYP1b1 activity assays using substrates other than E2 also indicated differences in the activity between WT and MTs. In a study involving the activity of CYP1b1 towards the substrate namely, (-)benzo[a]pyrene-7R-trans-7,8-dihyrodiol (B[a]P-7,8-diol), the protein mutations implicated in PCG significantly decreased the metabolism of the substrate, while minor differences in the activity were observed in case of the common polymorphisms (Mammen *et al.*, 2003). In an ethoxyresorufin O-deethylase assay, the CYP1b1 MT E229K showed decreased activity than the WT (Jeannot *et al.*, 2007). In these studies, polymorphic/Mutation sites do not form the part of SBR region, but have an affect on the catalytic activity of the enzyme. The mechanism of loss of activity thus could be indirectly through structural and dynamic changes in the enzyme, leading to altered substrate binding or access, as observed for the mutations studied in the current study.

The difference in the ligand interactions in MTs compared to the WT as seen in the current study is a result of change in the geometry of the SBR and change in the position of the heme relative to the active site. In WT, the ligand does not make non-bonded contacts with heme, but in most of the MTs, the ligand makes non-bonded contacts with heme. As seen earlier, the orientation of the ligand in WT is similar to that found in the complex of Estriol with CYP51. The interaction with residue F261

of G-helix, which is found in WT is absent in all the MTs. An important difference in ligand-protein interactions between the WT and MTs is the presence of stacking interaction with phenyl residues in WT and their absence or reduced interactions in the MTs.

5.4 Conclusion

In this chapter, docking studies conducted on WT and PCG MT structures of CYP1b1 using E2 as substrate was described. The docking was performed using GOLD, which is based on genetic algorithm and considers ligand conformational flexibility while searching for docking solutions. In order to bring in conformational flexibility for the receptor also, more than one receptor structures were used, which were obtained during MD simulations. In conclusion, the current study could explain the nature of differences in geometry at the active site region in WT and MTs which might preclude favorable protein-ligand interactions in the MTs, thereby resulting in compromised catalytic activity.

6

Conclusion

Human Genetic Diseases are conditions resulting from changes occurring in the genome among which, the Single Nucleotide Polymorphisms are the largest cause of genetic variability. The non-synonymous SNPs which are implicated in many diseases, manifest their effects by affecting the protein's function. Either the protein's catalytic function is directly affected, or its interactions with other proteins is affected in a multi-protein complex, leading to increase or decrease in the rate of the metabolic pathway. Methods using sequence based information, including evolutionary information, are useful to predict the deleterious nature of SNPs. Inclusion of structural features into these methods further improves the accuracy of these predictions. These statistical methods help in discriminating between the neutral and potentially harmful SNPs. However, in order to understand the nature of deleterious effects caused by SNPs in specific proteins, a focused and detailed investigation is required. The current work focuses on elucidating the effects of some known deleterious mutations in the CYP1b1 protein which are implicated in the disease, Primary Congenital Glaucoma. Though it was established that these mutations in CYP1b1 are involved in PCG phenotype, it was

not clear how they affect the protein's function. Thus, the investigation is focused on comparing qualitatively and quantitatively, the structural and dynamic properties of WT and MT proteins, so as to identify the nature of changes in the protein, which are not conducive for function.

The initial comparative sequence analysis using residue conservation pattern, and position wise entropy values, indicated that the disease mutations are least represented indicating their incompatibility with the structure. The functionally important regions of the protein, especially the Heme binding region were found to be sensitive for residue changes than the rest of the protein. The mutations were mapped onto the homology model built using Human CYP2c9 as template. It was noted that though some of the mutations formed part of FIRs, other mutations are spatially far from the FIRs, raising the question of how these could exert an impact on the protein's function. This necessitated the detailed investigations by MD simulations, to study how the effects of mutations are propagated into the structure, that may cause deleterious changes affecting the protein's function.

The MD simulations, which were performed for 30 nanoseconds, were found to be sufficiently stable, from the analysis of the trajectories of several structural properties. It was observed that the MTs exhibit an overall similar fold as that of WT, but with changed structural properties, which are detrimental to the WT like function. The properties calculated from the stabilized portions of the trajectories indicated specific differences between the WT and MTs. The mutant structures showed an increase in the overall flexibility of the protein. It is also observed that mutations have an effect on the integrity of the active site pocket, found from the variation in the volume of the SBR. In P193L and G466D the volumes of SBRs are larger than the WT, whereas in the other five MTs (M132R, Q144P, E229K, S239R and R368H) the SBRs are smaller in volume as compared to the WT. A115P, which shows severe effect on HBR, seems

to have no effect on SBR. The MTs also revealed changes in the conformation and dynamics of the SAC region wherein; in all the MTs the size of the channel opening is greater than WT. Moreover, in some of the MTs (M132R, Q144P, P193L and E229K) the SAC flutters irregularly which is the most severe in P193L, during the course of the entire simulation, probably indicating disruption of substrate recognition and also its accessibility. The Cys-heme co-ordination bond distance quantifies the extent of structural deviation at the HBR. In P193L, this distance is about 3.3Å, which is more than the usual coordination bonding distance in CYPs (2Å). In MTs A115P, Q144P, M132R, R368H, G466D the distance is a little over the WT, ranging between 2.2Å to 2.9Å. The distance in E229K is similar to that of WT (2.1Å) while in S239R it is 1.8Å. Thus the HBR, from Cys-heme co-ordination bond point of view, is affected severely by P193L mutation and less severely by A115P, than the other mutations. Other MTs having co-ordination distance between 1.8Å and 2.5Å may have proper heme binding. Thus, each of the mutations seemed to affect the protein's function in a specific way. The mutations are accommodated into the protein structure but are associated with some structural changes that are functionally significant.

To further probe into the structural differences between the WT and the MTs, Essential dynamics analysis was performed using extended MD simulations. This technique is used to extract functionally significant collective motions in proteins. The Eigen values of the covariance matrix indicated that in each case, the first Eigen vector itself could adequately describe the functional motions in the enzyme. In all the structures, collective motions with varying magnitudes, were observed in the F/G loop and the B'-helix, the structures constituting the SAC region of the protein. In addition to this, the MTs differed from the WT in the collective motions in the HBR and SBR regions. Compared to the WT, the MTs showed a reduced integrity of dynamics of FIRs.

The structural analysis studies indicated that the MTs could have disruptive effect on the protein's function, by a destabilization of the native properties of the FIRs, especially that of HBR and SBR that comprise the active site. To validate these observations, a molecular docking study using Estradiol as the ligand, was conducted. The MTs, with changed geometry of the SBR and changed position of the heme relative to the active site, showed differences in the ligand orientation and its interactions with the protein. The ligand-protein interactions found in the WT were not found in the MTs, which may result in compromised substrate binding.

An important feature of this study is to elucidate the mechanism of how certain polymorphic/Mutation sites that were not part of FIRs, can have deleterious effects on the protein. The comparative analysis of various structural properties presented in this study, could be effectively incorporated into the large scale prediction methods. However, as indicated earlier the choice of properties analyzed will be specific to the protein that is studied. Incorporation of data from Molecular dynamics simulations, including some functional data on collective or functional motions of the molecules, can increase the sensitivity of the existing predictive methods in discriminating between neutral and disease causing SNPs. With such an effort a standardized protocol could be developed that can give an index of deleterious for specific mutations, which can be used in disease prognosis. In the current investigation, which is based on sequence analysis and extensive structure based analysis of some deleterious point mutations in CYP1b1 protein, the MTs were compared with the WT, using several local and global structural features, with a focus on the structural changes in the functionally important regions of the protein. The results presented in this work, in the form of comparative analysis of various static and dynamics properties, could be implemented into the statistical prediction methods, to discriminate the deleterious SNPs from neutral ones. In addition, the current investigation shows that, in the known cases of mutations involved in

disease, structural analysis using MD simulations helps in detailed understanding of the mutant enzyme properties, that can aid in the process of drug development.

Appendix

A

Modeling of DNA-PBD Interactions

A.1 Introduction

PBD or pyrrolo-benzo-diazepine class of molecules, derived from *Streptomyces sp.* are well known for their DNA binding property(Thurston, 1993). They selectively bind to the minor groove of DNA at the Purine-Guanine-Purine sequences forming covalent interaction(Boyd *et al.*, 1990). They have been shown to inhibit endonuclease enzyme cleavage of DNA and block transcription by inhibiting DNA polymerase in a sequence specific manner(Puvvada *et al.*, 1997). They exert their biological activity through covalent binding via their N10-C11 imine/carbinolamine moiety to the C2-amino position of a guanine residue within the minor groove of DNA (Petrusek *et al.*, 1982). Molecular modeling, solution NMR, fluorimetry, and DNA foot printing experiments have shown that these molecules have a preferred selectivity for Pu-G-Pu sequences (Hurley *et al.*, 1988) and can be oriented with their A-rings pointed either toward the 3' or 5' end of the covalently bonded DNA strand. Thus, they are potential candidates for rational drug design for therapy of gene related

diseases including cancers (Neidle *et al.*, 1994).

There have been attempts to develop PBD molecules using different functional groups, to increase their binding and selectivity for DNA sequences(Cooper & Hagan, 2002). It has been observed that there has been extensive improvement in the biological activity due to the cross-linking property, by the presence of two imine functionalities. In this context, mixed dimers of PBD that contain an imino functionality in one of the PBD rings and a secondary amine group in the other, and linked at the C8 position by a suitable alkane spacer, have been synthesized to study thier DNA binding properties (Kamal *et al.*, 2004). In the current study, which forms part of a collaborative project between our group and the Bio-transformation group at IICT, Hyderabad, molecular modeling and simulation studies of some of the newly synthesized mixed dimers of PBD were carried out to compare their DNA binding ability.

A.2 Material and Methods

The PBD dimers contain an imino functionality in one of the A-rings and a secondary amine (N10) group in another, and which are linked at the C8 position by a suitable alkane spacer. All the molecular modeling and simulations were performed using INSIGHT-II suite of software (http://www.accelrys.com/products/insight/), running on Silicon Graphics OCTANE system.

Modeling of DNA duplex and PBD dimer structures

The 15-mer DNA sequence GGGGAGAGAGAGGGG a symmetric sequence about the central triplet AGA, which is the most preferred site for PBD binding was considered. BIOPOLYMER module of INSIGHT-II was used to build the B-DNA duplex structure. The four PBD dimers; 1, 4, 5a, and 5b used in this study were separately constructed

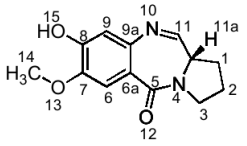
using the BUILDER module of INSIGHT-II (Figure A.1). Initially PBDs were sketched in 2D (two dimensions) and then converted into 3D (three dimensions) using 2D-3D converter tool of the BUILDER module. Care was taken to see that there is C11(S)-geometry for all PBD dimers constructed, as this stereochemistry is known to lead to energetically favored adduct with that of Guanine of B-DNA duplex structure (Thurston, 1993).

Docking studies

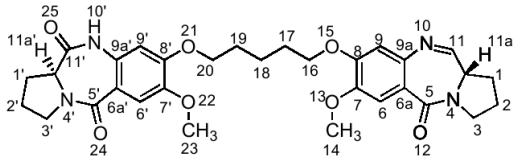
PBDs were manually docked into the minor groove of DNA such that the N10-C11 imine functionality and the exocyclic C2-amino group of Guanine (G8) are nearly at bonding distance and a covalent bond was then formed using the 'create-bond' tool. After the bond was created, the PBD in the minor groove was manually oriented such that the A-ring is oriented toward the 3' end of DNA to which it is covalently linked. Dihedral angles about C-C bonds were manually adjusted such that PBD dimers form an isohelical fit within the minor groove of the B-DNA duplex structure. CVFF force field was used to assign the charges and the potentials required for energy calculations. The complexes formed were subjected to energy minimization (EM) using conjugate gradient method till they were fully converged that is, till the energy gradient was nearly equal to 0.001 kJ.mol^{-1}.nm^{-1}. Constraints were applied to fix the DNA during EM.

Molecular dynamics

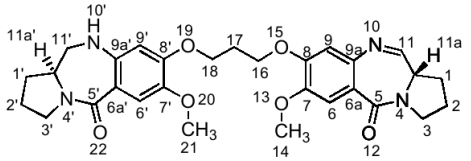
Molecular dynamic studies were carried using the following protocol: heating phase (equilibration) = 30 ps and sampling phase = 100 ps. During simulations constraints were applied to fix the DNA structure. Intermittent structures of the complex formed at every 10 ps of simulation were collected and subjected to EM. The minima of all



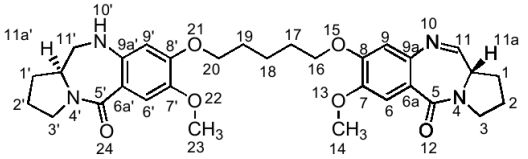
1



4



5a



5b

Figure A.1: Structures of the PBD molecules. DC-81 (1); Mixed imine-amide PBD dimers (4, 5a and 5b)

the snap shots were examined in order to select the lowest energy conformation as a representative of DNA-PBD complex for further studies.

Energy of interaction

The energy of the PBD-DNA complex ($E_{complex}$) and the energies of DNA (E_{DNA}) and PBD (E_{PBD}) individually after separating from the complex were calculated. Energy of interaction (E_{int}) between DNA and PBD complex were calculated using the equation A.1; where, E_{int} = energy of interaction of the complex, $E_{complex}$ = total energy of the complex, E_{DNA} and E_{PBD} are the individual total energies of the DNA and the PBD molecules calculated after they are separated from each other.

$$E_{int} = E_{complex} - (E_{DNA} + E_{PBD}) \tag{A.1}$$

A.3 Results and Discussion

Molecular Modeling

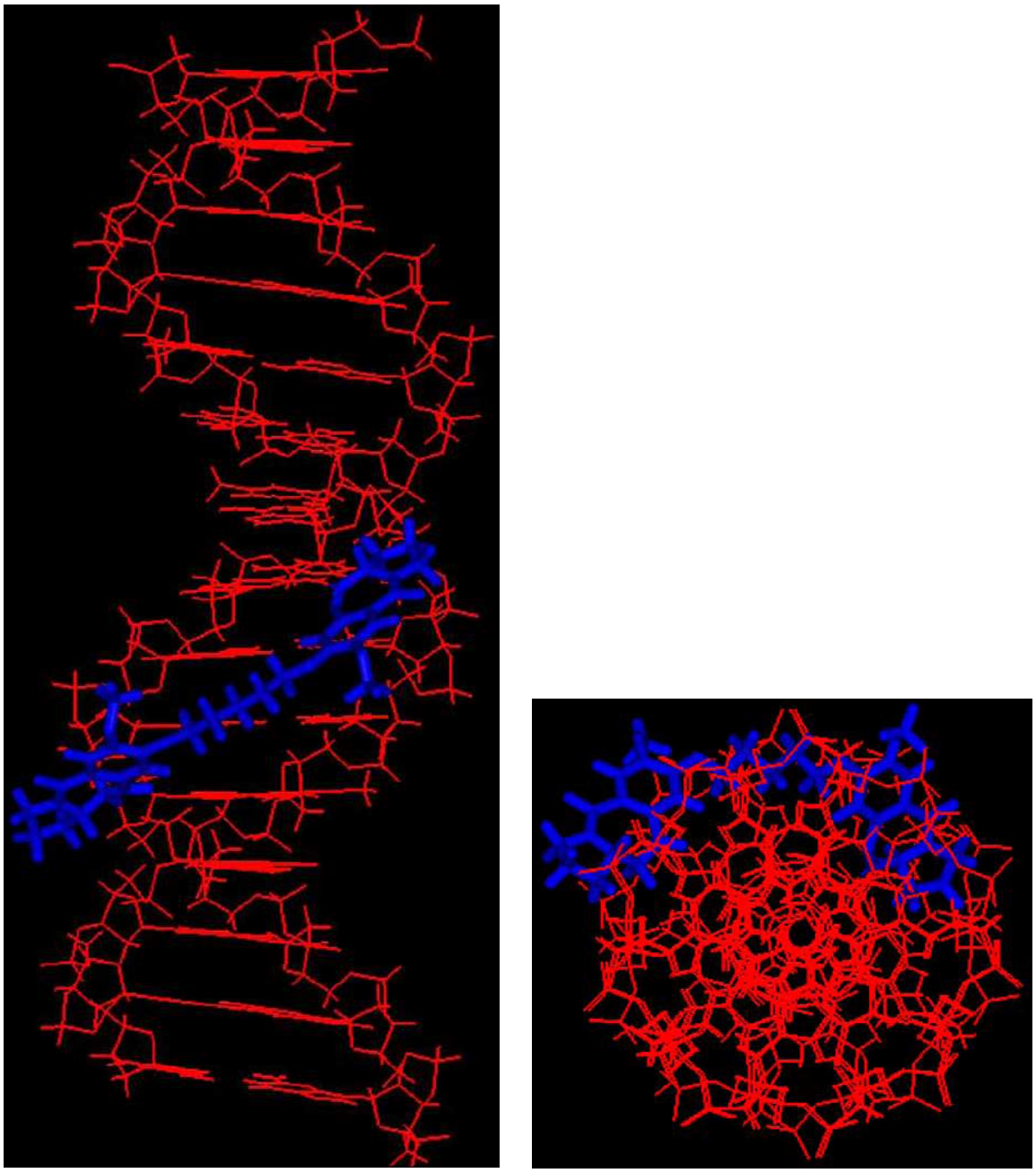
Modeling of the complexes of PBD dimers; 1, 4, 5a, and 5b with the DNA has been carried out as described in the methods section (Figure A.1). A B-DNA duplex structure has been considered with a sequence 5′-GGGGAGAGAGAGGGG-3′- a symmetric sequence about the central triplet AGA, which is the most preferred site for PBD binding(Thurston, 1993). Each of the DNA-PBD complexes has been subjected to molecular dynamics (MD) simulations followed by energy minimization of snap shots collected at regular time intervals during the MD simulations. After energy minimization all the energy minima obtained have been compared with each other and the one with the lowest energy has been picked as the representative of the DNA-PBD

complex (Figure A.2).

The least energy complex thus picked up has been selected to calculate the interaction energy as a measure of stability of the complex. The interaction energies between DNA and PBDs are given in Table A.1. It can be seen from this table that the imine-amide PBD dimer 4 renders more stability to the complex compared to the imine-amine PBD dimers (5a and 5b), and the PBD monomer 1. This property correlates with the experimentally determined values of the DNA melting temperature in these complexes (Table A.2) (Kamal *et al.*, 2004). The complexes are characterized by the presence of a number of nonbonded interactions (Table A.3) formed between DNA and PBD molecules in addition to the covalent linkage formed between the imine PBD subunit and exocyclic C2-amino group of the Guanine (G8).

The dimeric PBDs (4, 5a, and 5b) offer more favorable non-bonded interactions as compared to the monomeric DC-81 (1), evident from their respective non-bonded energy terms (Table A.1), which are about 50 Kcal.mol^{-1} less as compared to DC-81. Among the dimeric PBDs the imine-amide PBD dimer 4 is associated with the lowest nonbonded energy term and is stabilized by the hydrogen bonding interaction between the carbonyl functionality and the amino group of G12. This is precluded in 5a and 5b because of the absence of the required carbonyl functionality (see Table A.3). Between 5a and 5b, the latter has better nonbonded and bonded energy components and these together render higher stability for 5b by about 11 Kcal mol^{-1} as compared to 5a.

The modeling of the complex of DNA with the cross-linking PBD dimer 3 has not been considered in the present study. In this case, the stability of the complex is predominantly due to the presence of additional covalent interactions whereas in the noncross-linking molecules, nonbonded interactions play an important role in stabilization of their respective complexes.



A. Side-on view B. Down the helix axis

Figure A.2: Projection diagram showing the DNA-5b comples;(a)side-on view; (b) down the helix axis

Table A.1: The values of energy of interactions calculated for the DNA-PBD dimer complexes

PBD molecule	Total	Bonded	Non-bonded	
			Van der Waals	Coulombic
1	-85.2	-14.2	-33.7	-37.3
4	-145.9	-12.2	-84.3	-49.4
5a	-132.1	-12.6	-71.8	-47.7
5b	-143.6	-14.8	-68.7	-60.1

Table A.2: Thermal denaturation data for DC-81 (1), DSB-120 (3), mixed imine-amide PBD dimer (4) and compounds 5a and b with calf thymus DNA

PBD compound	[PBD]:[DNA] molar ratio	Induced ΔTm (^{o}C) after incubation at 37^{o}C for	
		0h	18 h
5a	1:5	8.7	9.7
5b	1:5	10.6	11.0
4	1:5	14.0	17.0
3	1:5	10.2	15.4
1	1:5	0.3	0.7

†For CT-DNA alone at pH 7.00 +/- 0.01, Tm = 69.2 ^{o}C +/- 0.01 (mean value from 30 separate determinations), all ΔTm values are +/- 0.1 - 0.2 ^{o}C.

†For a 1:5 molar ratio of [ligand]/[DNA], where CT-DNA concentration = 100 microM and ligand concentration = 20 microM in aqueous sodium phosphate buffer [10 mM sodium phosphate + 1 mM EDTA, pH 7.00 +/- 0.01].

Table A.3: List of hydrophobic contacts and hydrogen bond interactions

	DC-81	4	5a	5b
Hydrophobic interactions				
	G8.C2'—C1	G8.C2'—C1	G8.C2'—C1	G8.C2'—C1
	A9.C2'—C1	A9.C2'—C1	A9.C2'—C1	A9.C2'—C1
	G10.C2'—C6	G10.C2'—C6	G10.C2'—C9	G10.C2'—C9
	C8.C2'—C9	A11.C2'—C19	A11.C2'—C9'	A11.C2'—C18
	T9.C2'—C6a	G12.C2'—C9'	G12.C2'—C6'a	G12.C2'—C9'
	C10.C2'—C2	G13.C2'—C1'	G13.C2'—C1'	G13.C2'—C1'
		C6.C2'—C6'a	C6.C2'—C1'	C6.C2'—C1'
		T7.C2'—C9'	T7.C2'—C6'a	T7.C2'—C9'
		C8.C2'—C9	C8.C2'—C9	C8.C2'—C9
		T9.C2'—C9	T9.C2'—C6a	T9.C2'—C6a
		C10.C2'—C2	C10.C2'—C2	C10.C2'—C2
Hydrogen bonded interactions[a]				
	C8.O2—N10	G12.N2—O25	C8.O2—N10	C8.O2—N10
	C8.O2—N10			

[a]Imine N of PBD in all the complexes (1, 4, 5a, and 5b) is at a favorable distance ($\tilde{3}$.3Å) to imino group of A9 of DNA however the angle is not within hydrogen bonding limits (NH—N $\tilde{7}7^{o}$).

A.4 Conclusion

The DNA binding ability for the noncross-linking PBD dimers is noticeable inspite of the introduction of amine functionality in one of the A-rings in comparison to DC-81. Modeling studies suggest that apart from the covalent linkage formed between the imine PBD subunit and G8, there are a number of favorable van der Waals and coulombic interactions that are formed between the DNA and the mixed imine-amine PBD dimer. The complex formed by compound 5b (with a five carbon chain linker) is energetically more stable than the complex formed by 5a as the extra alkane spacer units offer the molecule to make extra favorable van der Waals and coulombic interactions with DNA. However, 5b exhibits lower ΔT_m value in comparison to 4 because of the absence of carbonyl functionality in the complex. Of all the molecules imine-amide PBD dimer 4 forms the most stable complex. Nevertheless, the imine-amine PBD dimers also exhibit interesting profile of DNA binding ability in contrast to PBD monomer (DC-81). This investigation exhibits the role played by noncovalent interaction of PBD secondary amine subunit. In addition, these compounds show promising *in vitro* cytotoxicity in different cancer cell lines.

B

Modeling Studies on Interaction of Lambdoid N With Transcription Elongation Complex

B.1 Introduction

Transcription is the key process by which the information stored in the DNA is decoded and expressed through proteins. This is catalyzed by the enzyme RNA polymerase with the help of other accessory factors. The core enzyme of bacterial RNA polymerase is a multisubunit protein, containing five subunits; $2\alpha, \beta, \beta'$ and ω. Two α subunits help in the assembly of the enzyme and regulatory factors. The β subunit has the polymerase activity. The β' subunit helps to bind to DNA non specifically. The ω subunit is known to function like a chaperone to restore the β' subunit. In addition to these, the polymerase requires the σ subunit to selectively bind to the promoter specific regions.

Transcription can be viewed as a multistep process, which includes the stages of; binding of the polymerase to the DNA (transcription initiation), elongation of the RNA chain (transcription elongation), and finally dissociation of the RNA polymerase enzyme complex from the DNA (transcription termination), upon receiving the termination signals. At each stage of transcription, the RNA polymerase is associated with specific transacting factors.

At the stage of transcription elongation, the RNA polymerase elongation complex (E) consists of RNA polymerase (RNAP), DNA and nascent RNA. It is extremely stable, yet dynamic in nature (Arndt & Chamberlin, 1990). This stability of the E arises from RNAP-DNA, RNAP-RNA and RNA-DNA interactions (Korzheva *et al.*, 2000; Darst, 2001). The E becomes unstable and dissociates when it transcribes specific DNA sequences called intrinsic terminators (Uptain *et al.*, 1997; von Hippel, 1998; Nudler & Gottesman, 2002), or in response to a factor called Rho (Richardson, 2002). RNA made from the factor-independent terminator sequences forms a stable hairpin followed by U-rich sequences. The E at a terminator sequence consists of a very weak RNA:DNA hybrid with A:U base-pairing and an RNA hairpin in the RNA-exit channel.

Specific modifications of the E can help to overcome both types of termination signals, and this process is called transcription antitermination. Lambdoid phages code for transacting proteins such as N and Q, or cis-acting RNA, HK022 PUT RNAs, which interact with the E to make it termination-resistant (Friedman & Court, 1995; Weisberg & Gottesman, 1999). These are called transcription antiterminators. N is a small basic protein belonging to the arginine-rich motif (ARM) family of RNA-binding proteins. N binds to an RNA stem-loop structure, encoded by the 'nut' site, present in the early transcripts of lambdoid phages (Lazinski *et al.*, 1989; Chattopadhyay *et al.*, 1995). The 'nut'-bound N, together with different Nus factors of Escherichia coli, bind to the E and convert it into a termination-resistant state (Greenblatt *et al.*, 1993).

This allows full expression of downstream genes of the phage. The modified E can overcome Rho-independent and dependent terminators in a processive manner.

N interacts with 'nut' RNA through its N-terminal ARM sequences and with the E through its C-terminal domain (Mogridge *et al.*, 1998). Although RNAP mutants defective in λ phage growth have been reported (Georgopoulos, 1971; Ghysen & Pironio, 1972; Jin *et al.*, 1988; Obuchowski *et al.*, 1997; Schauer *et al.*, 1996; Sternberg, 1976; Szalewska-Palasz *et al.*, 2003), the interacting surface on the E for N has remained elusive, as has the mechanism by which N imparts the antitermination property to RNAP. In the current work, the interacting regions of E with N protein is investigated using modeling and mapping studies, to better understand the anti-termination mechanism. This study supports the results obtained from *in vitro* mutagenesis studies (Cheeran *et al.*, 2005).

B.2 Material and Methods

The individual structures of the five subunis of *E. coli* RNA polymerase are available. However, the three-dimensional structure of the complete enzmyme core was not available. So, in this study, a model of the Thermus acqaticus holo enzyme in complex with a RNA:DNA hybrid was used, which was kindly provided by Seth Darst. Through *in vitro* studies (Cheeran *et al.*, 2005), five mutations were identified in the β (G1045D) and β' (P251S, P254L, R270C and G336S) subunits of RNAP that are specifically defective for antitermination by N protein of the lambdoid phage, H-19B. The positions of these mutations were putatively identified in E.coli RNAP enzyme, using the Thermus acqaticus model.

For this, structure based sequence alignments between *E. coli* RNA polymerase and *Thermus acquaticus* RNAP subunit sequences were obtained from the HOM-

STRAD database. The corresponding β and β' subunit sequences of the E model were aligned to these profiles using CLUSTALX (Thompson *et al.*, 1994). The alignments were corrected manually to properly represent the gaps occurring in the model structure. E.coli mutations were then mapped onto the structurally equivalent positions in the E model. The distances between these mutated amino acid residues and the different structural elements (RNA, DNA and the β'-rudder) were obtained by computing the distances between these structures. For this, the distances between the C^{α} atom of the mutant amino acid residues and the phosphorus atoms in each of the nucleotides of the DNA/RNA, were considered.

B.3 Results and Discussion

In vitro studies (Cheeran *et al.*, 2005), identified five mutations in the β (G1045D) and β' (P251S, P254L, R270C and G336S) subunits of RNAP that are specifically defective for antitermination by N protein of the lambdoid phage, H-19B. In addition, a mutation in the C-terminal domain of N- L108F, suppresses the defect of β'-P254L. Purified mutant holoenzymes showed less processive antitermination. These together suggest that H-19B N exerts its effect on E through the region defined by these mutation positions. These mutant RNAPs were severely defective for phage H-19B growth but were either partially (for β'-mutants) or not (for G1045D) defective for phage λ growth. The partial defect of β'-mutants for the growth of λ phage may suggest that the site of action of λ N is also near this region. Any transcription antitermination mechanism should involve one or all of the following.

1. An antiterminator may prevent or delay the terminator hairpin folding in the RNA- exit channel.
2. It may stabilize weak interactions in the RNA:DNA hybrid.

3. It may alter the interactions of the clamp domain of RNAP with the downstream duplex DNA.

The model of E complex containing the DNA:RNA hybrid, was based on the crystal structures of *Thermus aquaticus* RNAP (Zhang *et al.*, 1999), the E of yeast RNAPII (Gnatt *et al.*, 2001) and cross-linking data on protein-nucleic acids interactions from the E of *E. coli* RNAP (Korzheva *et al.*, 2000). The equivalent amino acids of these mutations in the model structure were determined by structural alignment (Figure B.1) using CLUSTALX. All these changed amino acids, except β'-P254, are highly conserved among the bacterial RNAPs.

Figure B.2 shows the locations of the mutations. In the model structure, these mutations map very close to the RNA:DNA hybrid, the DNA template strand, the RNA-exit channel and important structural elements like the rudder and the lid. A schematic diagram describing the numbering scheme of the RNA:DNA hybrid used in the model is shown in Figure B.3. RNA positions K1 to K8 are in the RNA:DNA hybrid and positions K9 to K15 are in the RNA-exit channel. Numbers of DNA residues are indicated, and template (T) and non-template (NT) DNAs are labeled. The distances of each of these mutations from DNA, RNA and rudder elements are shown in Table B.1. Nearest residue positions are shown in parentheses. For example, P251S is 6.2Å away from the K10 position of nascent RNA, 15.3Å away from the 31T of template DNA and 19.2Å away from the 325K amino acid in the rudder element. The rudder and lid are the two loop structures, and their role was originally proposed to guide nascent RNA into the RNA-exit channel (Gnatt *et al.*, 2001).

Recent deletion analysis of the rudder and lid regions was shown to destabilize E (Kuznedelov *et al.*, 2002). β'-P251S and β-P254L are located in the lid structure and close to the -10 position of RNA in the RNA-exit channel. β'-G336S comes within

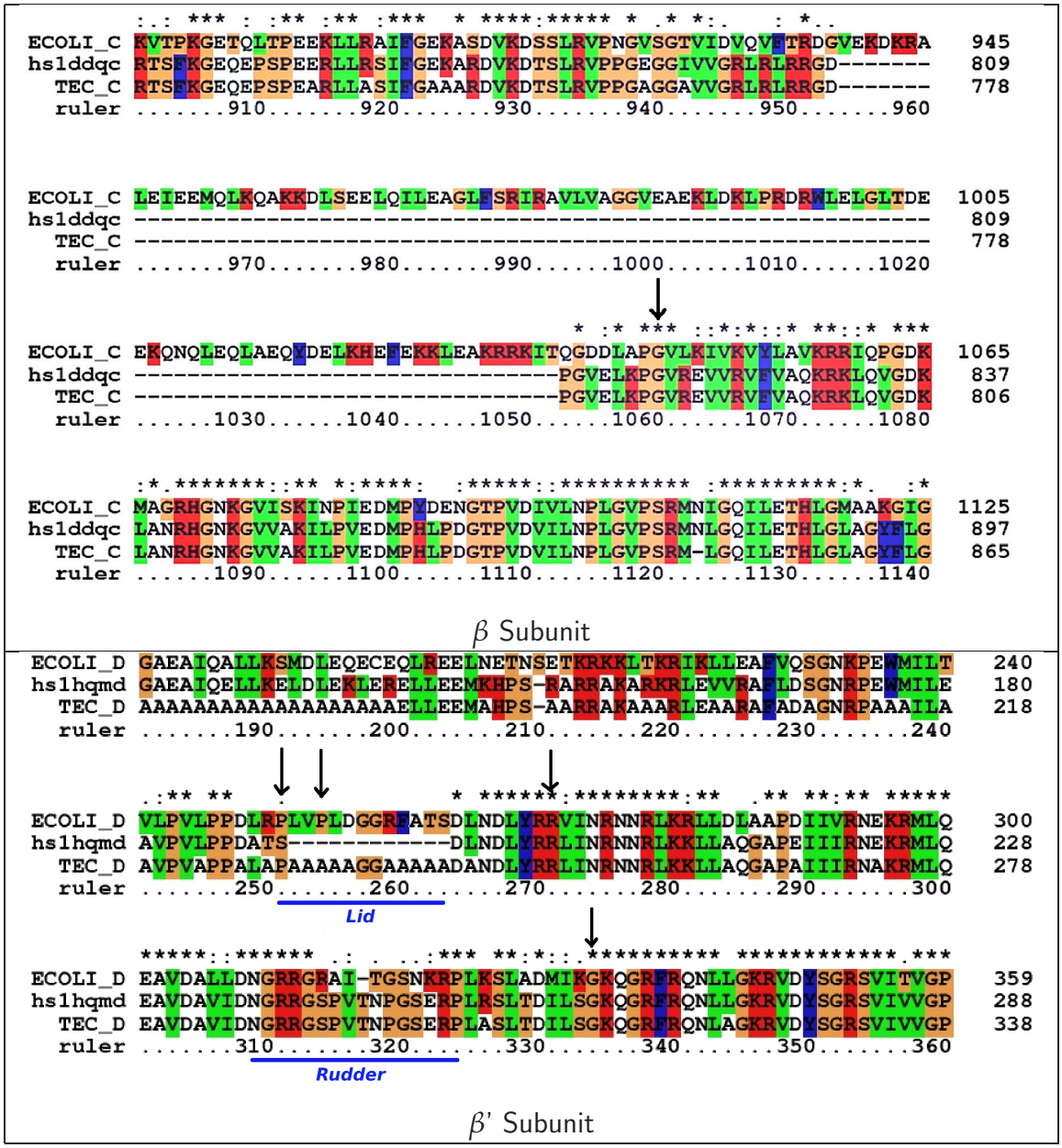


Figure B.1: Structural alignment of the relevant regions of the β and β' subunits of E. coli RNAP with the equivalent regions of *T. acquaticus* RNAP using CLUSTALX (Thompson *et al.*, 1994). This was done to identify the equivalent position of the mutations in the model structure of E. The position of the mutations is shown by arrows, and structurally important elements such as the β' rudder and the β' lid regions are underlined. hs1ddqc and hs1hqmd are the PDB identifier numbers of *T. acquaticus* RNAP C and D chains, respectively, obtained from the HOMSTRAD database. TEC_C and TEC_D stand for ternary elongation complex C and D chains, respectively.

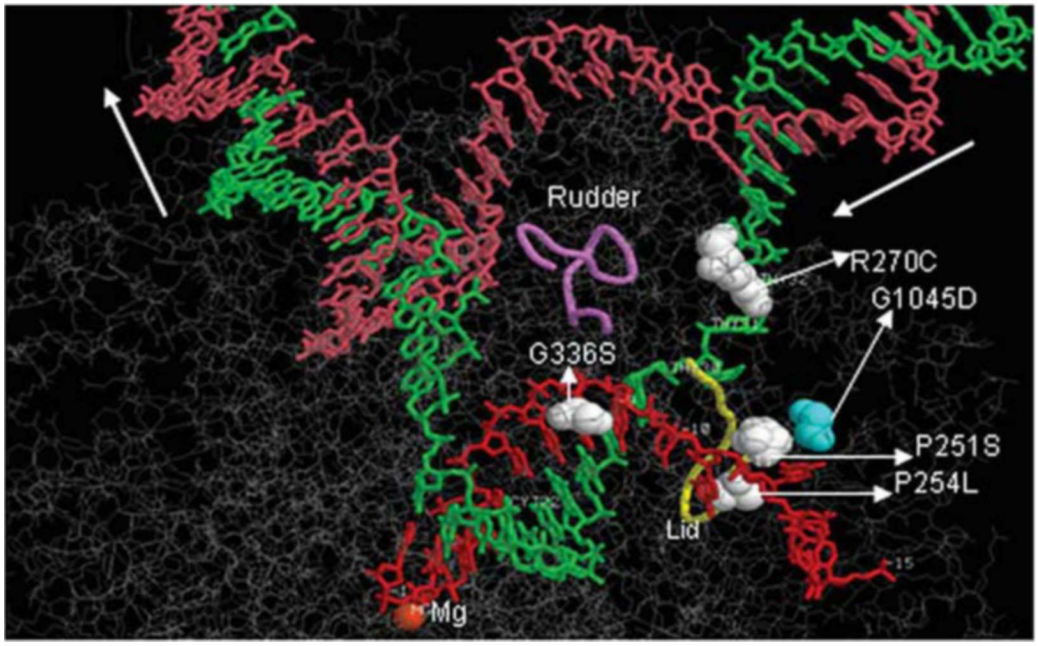


Figure B.2: An expanded view of the nucleic acid framework of the model structure of E using RASMOL. Locations of different mutations to the equivalent positions of this model are shown as space-filled in the wire-framed grey background of other parts of RNAP. The color-coding of template DNA, non-template DNA and RNA are green, pink and red, respectively. Active-site Mg2C is shown in orange. Flexible loop structures of rudder and lid are indicated.

Figure B.3: The numbering scheme of the RNA: DNA hybrid used in the model

Table B.1: The distances of each mutation from the nearest residues in the DNA, RNA or rudder element

Mutation	Distance from RNA Å	Distance from DNA Å	Distance from rudder
rpoC			
P251S	6.2 (-10)	15.3 (31T)	19.2 (325K)
P254L	11.9 (-10)	14.7 (30T)	23.8 (322R)
R270C	19.3 (-9)	15.0 (32T)	13.1 (314R)
G336S	9.3 (-8)	12.2 (22C)	10.3 (325K)
rpoB			
G1045D	20.4 (-11)	14.8 (31T)	24.3 (316I)

10Å of the upstream edge of the RNA:DNA hybrid and the rudder element. β'-R270C is located close to the rudder and the template DNA upstream of the RNA:DNA hybrid, whereas β-G1045D is situated a small distance away from the nucleic acid framework but comes within 12Å of P254L.

The mutations were then mapped onto the space-filling model of the E to determine the surface-accessibility of these amino acid residues (Figure B.4). β-G1045D and β'-R270C are on the surface, β'-P251S is partially exposed, whereas β'-P254L and β'-G336S are buried inside the structure. If these amino acid residues were to take part in direct interactions with N, a major conformational change in E would be required to get access to the β'-P254L and β'-G336S amino acid residues. RNA polymerase mutations in β (Georgopoulos, 1971; Ghysen & Pironio, 1972; Sternberg, 1976; Jin *et al.*, 1988) and α subunits (Schauer *et al.*, 1996; Obuchowski *et al.*, 1997; Szalewska-Palasz *et al.*, 2003), defective in supporting the growth of λ phage as well as impaired for N-mediated antitermination, have been reported. The amino acid substitutions in the mutant RNAPs cluster very close to the RNA:DNA hybrid at the beginning of the RNA-exit channel of the E. Thus, it is suggested that the action of H-19B N is exerted through the region defined by these amino acids.

Wild-type N stabilizes the E at terminator sites and in this modified E, a part of the terminator hairpin may form but appears to be unstable. Due to the spatial location of the mutations close to the RNA:DNA hybrid and the upstream part of the RNA-exit channel, these altered amino acid residues can affect the inter- actions of RNAP with the RNA:DNA hybrid and with RNA in the exit channel, which in turn can impair the process of antitermination. Thus, it is proposed that the action of N close to the active center alters the RNAPnucleic acid interactions around the RNA:DNA hybrid, which impairs proper folding of the terminator hairpin or stabilizes the weak RNA:DNA hybrid, or both.

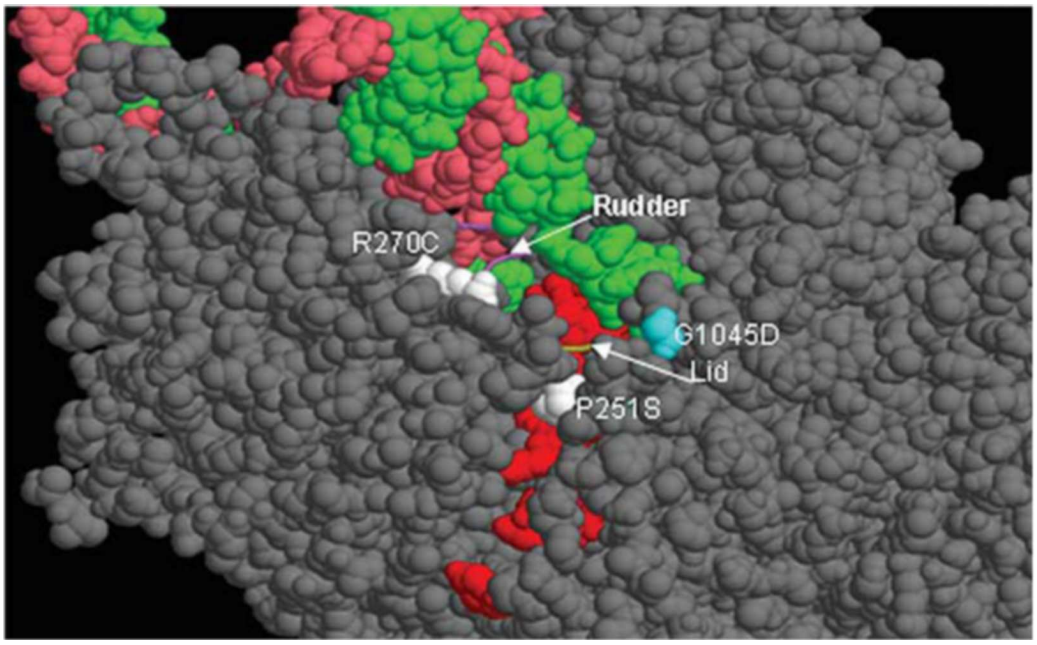


Figure B.4: Space-filled model of the E surrounding the nucleic acid framework, showing the surface accessibility of the mutated amino acid residues. G1045D, R270C and P251S are visible on the surface, while other mutants are buried inside the structure. The color coding is the same as in Figure B.2.

B.4 Conclusion

The results do not show unequivocally that the identified positions of amino acids in RNAP are involved in direct interaction with H-19B N and it is possible that the effects of these mutations are indirect. N may bind to a different part of RNAP, and the effect might be exerted allosterically through this region. Alternatively, it is possible that these changes affect NusA and NusG functions in the antitermination complex. On the basis of this result and the location of the amino acid substitutions in the E, it is proposed that the action of H-19B N close to the active center of RNAP alters the RNAP interactions around the RNA:DNA hybrid which, in turn, can impair proper folding of the terminator hairpin or stabilize the weak RNA:DNA hybrid or both. If the position of the mutations defines the site of action of N, altered interactions at the beginning of the RNA-exit channel and the upstream part of the RNA:DNA hybrid can prevent the formation of the base of the hairpin.

Alternatively, this altered interaction in modified E can stabilize the RNA:DNA hybrid, which will also prevent the completion of the hairpin folding because melting of the hybrid is essential for the hairpin to form (Komissarova *et al.*, 2002). Affecting the proper formation of the terminator hairpin by H-19B N is consistent with an earlier proposal of prevention or delayed folding of the hairpin in the presence of λ N (Gusarov & Nudler, 2001). In addition to the effects on hairpin formation, it is possible that altered interactions surrounding the RNA:DNA hybrid could have an allosteric effect on the clamp domain that holds the downstream duplex DNA (Gnatt *et al.*, 2001). Further biochemical and structural probing of the N-modified E would validate these propositions.

Bibliography

Abseher, R., & Nilges, M. 1998. Are there non-trivial dynamic cross-correlations in proteins? *J mol biol*, **279**(4), 911–20.

Achary, M. S., Reddy, A. B., Chakrabarti, S., Panicker, S. G., Mandal, A. K., Ahmed, N., Balasubramanian, D., Hasnain, S. E., & Nagarajaram, H. A. 2006. Disease-causing mutations in proteins: structural analysis of the cyp1b1 mutations causing primary congenital glaucoma in humans. *Biophys j*, **91**(12), 4329–39.

Akarsu, A. N., Turacli, M. E., Aktan, S. G., Barsoum-Homsy, M., Chevrette, L., Sayli, B. S., & Sarfarazi, M. 1996. A second locus (glc3b) for primary congenital glaucoma (buphthalmos) maps to the 1p36 region. *Hum mol genet*, **5**(8), 1199–203.

Aklillu, E., Oscarson, M., Hidestrand, M., Leidvik, B., Otter, C., & Ingelman-Sundberg, M. 2002. Functional analysis of six different polymorphic cyp1b1 enzyme variants found in an ethiopian population. *Mol pharmacol*, **61**(3), 586–94.

Alakent, B., & Doruker, P. 2004. Application of time series analysis on molecular dynamics simulations of proteins: a study of different conformational spaces by principal component analysis. *J chem phy*, **121**(10), 4759–69.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. 1990. Basic local alignment search tool. *J mol biol*, **215**(3), 403–10.

Altschul, S. F., Madden, T. L., & Schaffer, A. A. 1997. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids res*, **25**(17), 3389–402.

Amadei, A., Linssen, A. B., & Berendsen, H. J. 1993. Essential dynamics of proteins. *Proteins*, **17**(4), 412–25.

Anfinsen, C. B. 1972. The formation and stabilization of protein structure. *Biochem j*, **128**(4), 737–49.

Anfinsen, C. B. 1973. Principles that govern the folding of protein chains. *Science*, **181**(96), 223–30.

Arcangeli, C., Bizzarri, A. R., & Cannistraro, S. 2001. Concerted motions in copper plastocyanin and azurin: an essential dynamics study. *Biophys chem*, **90**(1), 45–56.

Arndt, K. M., & Chamberlin, M. J. 1990. Rna chain elongation by *Escherichia coli* rna polymerase. factors effecting the stability of elongating ternary complexes. *J mol biol*, **213**, 79–108.

Arnold, G. E., & Ornstein, R. L. 1997. Molecular dynamics study of time-correlated protein domain motions and molecular flexibility: cytochrome p450bm-3. *Biophys j*, **73**(3), 1147–59.

Bahar, I., Erman, B., Jernigan, R. L., Atilgan, A. R., & Covell, D. G. 1999. Collective motions in hiv-1 reverse transcriptase: examination of flexibility and enzyme function. *J mol biol*, **285**(3), 1023–37.

Baker, D., & Sali, A. 2001. Protein structure prediction and structural genomics. *Science*, **294**(5540), 93–6.

Bejjani, B. A., Lewis, R. A., Tomey, K. F., Anderson, K. L., Dueker, D. K., Jabak, M., Astle, W. F., Otterud, B., Leppert, M., & Lupski, J. R. 1998. Mutations in cyp1b1, the gene for cytochrome p4501b1, are the predominant cause of primary congenital glaucoma in saudi arabia. *Am j hum genet*, **62**(2), 325–33.

Bejjani, B. A, Stockton, D. W., & Lewis, R. A. 2000. Multiple cyp1b1 mutations and incomplete penetrance in an inbred population segregating primary congenital glaucoma suggest frequent de novo events and a dominant modifier locus. *Hum mol genet*, **9**, 367–74.

Belkina, N. V., Skvortsov, V. S., Ivanov, A. S., & Archakov, A. I. 1998. [modeling of a three-dimensional structure of cytochrome p-450 1a2 and search for its new ligands]. *Vopr med khim*, **44**(5), 464–73.

Berendsen, H. J., & Hayward, S. 2000. Collective protein dynamics in relation to function. *Curr opin struct biol*, **10**(2), 165–9.

Berendsen, H. J. C., Postma, J. P. M., DiNola, A., & Haak, J. R. 1984. Molecular dynamics with coupling to an external bath. *J chem phy*, **81**, 3684–90.

Betz, S. F. 1993. Disulfide bonds and the stability of globular proteins. *Protein sci*, **2**(10), 1551–8.

Bohm, H. J. 1998. Prediction of binding constants of protein ligands:a fast methods for the prioritization of hits obtained from de novo design or 3d database search programs. *J comput aided mol des*, **12**(4), 309–23.

Bonneau, R., & Baker, D. 2001. Ab initio protein structure prediction: progress and prospects. *Annu rev biophys biomol struct*, **30**, 173–89.

Bork, P. 1991. Shuffled domains in extracellular proteins. *Febs lett*, **286**, 47–54.

Bowie, J. U., Luthy, R., & Eisenberg, D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**(5016), 164–70.

Boyd, F. L., Stewart, D., Remers, W. A., Barkley, M. D., & Hurley, L. H. 1990. Characterization of unique tomaymycin-d(cicgaattcicg)2 adduct containing two drug molecules per duplex by nmr, fluorescence, and molecular modeling studies. *Biochemistry*, **29**(9), 2387–403.

Braun, W., & Go, N. 1985. Calculation of protein conformations by proton-proton distance constraints. a new efficient algorithm. *J mol biol*, **186**(3), 611–26.

Brenner, S. E., Chothia, C., & Hubbard, T. J. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc natl acad sci u s a*, **95**(11), 6073–8.

Buchanan, S. G., & Sauder, J. M. 2002. The promise of structural genomics in the discovery of new antimicrobial agents. *Curr pharm des*, **9**(13), 1173–88.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G. Q., & Lander, E. S. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat genet*, **22**(3), 231–8.

Castrignano, T., De Meo, P.D., Cozzetto, D., Talamo, I. G., & Tramontano, A. 2006. The pmdb protein model database. *Nucleic acids res*, **34**, D306–D309.

Catell, R. B. 1996. The scree test for the number of factors. *Multivariate behav. res.*, **1**, 245–276.

Chasman, D., & Adams, R. M. 2001. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J mol biol*, **307**(2), 683–706.

Chattopadhyay, S., Garcia-Mena, J., DeVito, J., Wolska, K., & Das, A. 1995. Bipartite function of a small rna hairin in transcription antitermination in bacteriophage lambda. *Proc natl acad sci usa*, **92**, 4061–4065.

Cheeran, A., Babu Suganthan, R., Swapna, G., Bandey, I., Achary, M. S., Nagarajaram, H. A., & Sen, R. 2005. Escherichia coli rna polymerase mutations located near the upstream edge of an rna:dna hybrid and the beginning of the rna-exit channel are defective for transcription antitermination by the n protein from lambdoid phage h-19b. *J mol biol*, **352**(1), 28–43.

Chen, C., Xiao, Y., & Zhang, L. 2005. A directed essential dynamics simulation of peptide folding. *Biophys j*, **88**(5), 3276–85.

Cheung, Y. L., Kerr, A. C., McFadyen, M. C., Melvin, W. T., & Murray, G. I. 1999. Differential expression of cyp1a1, cyp1a2, cyp1b1 in human kidney tumours. *Cancer lett*, **139**(2), 199–205.

Chothia, C. 1992. Proteins. one thousand families for the molecular biologist. *Nature*, **357**(6379), 543–4.

Chothia, C., & Lesk, A. M. 1986. The relation between the divergence of sequence and structure in proteins. *Embo j*, **5**(4), 823–6.

Chou, P. Y., & Fasman, G. D. 1974. Prediction of protein conformation. *Biochemistry*, **13**(2), 222–45.

Chou, P. Y., & Fasman, G. D. 1977. Beta-turns in proteins. *J mol biol*, **115**(2), 135–75.

Clark, M., Cramer, R. D., & Van Opdenbosch, N. 1989. Validation of the general-purpose tripos 5.2 force field. *J comput chem*, **10**, 982–1012.

Collins, F. S., Brooks, L. D., & Chakravarti, A. 1998. A dna polymorphism discovery resource for research on human genetic variation. *Genome res*, **8**(12), 1229–31.

Coon, M. J. 2004. Cytochrome p450: Nature's most versatile biological catalyst. *Ann rev pharm toxi*, **45**, 1–25.

Cooper, N., & Hagan, D. R. 2002. Synthesis of novel c2-aryl pyrrolobenzodiazepines(pbds) as potential antitumor agents. *Chem commun (camb)*, **16**, 1764–5.

Creighton, T. E. 1992. *Proteins:structures and molecular properties*.

Cupp-Vickery, J. R., & Poulos, T. L. 1995. Structure of cytochrome p450eryf involved in erythromycin biosynthesis. *Nat struct biol*, **2**(2), 144–53.

Dai, R., Zhai, S., Wei, X., Pincus, M. R., Vestal, R. E., & Friedman, F. K. 1998. Inhibition of human cytochrome p450 1a2 by flavones: a molecular modeling study. *J protein chem*, **17**(7), 643–50.

Darden, T., York D. Pedersen L. 1993. Particle mesh ewald: An n-log(n) method for ewald sums in large systems. *J. chem. phys*, **98**, 10089–10092.

Darst, S. A. 2001. Bacterial rna polymerase. *Curr opin struct biol*, **11**, 155–162.

de Groot, B. L., van Aalten, D. M., Amadei, A., & Berendsen, H. J. 1996a. The consistency of large concerted motions in proteins in molecular dynamics simulations. *Biophys j*, **71**(4), 1707–13.

de Groot, M. J., Vermeulen, N. P., Kramer, J. D., van Acker, F. A., & Donne-Op den Kelder, G. M. 1996b. A three-dimensional protein model for human cytochrome p450 2d6 based on the crystal structures of p450 101, p450 102, and p450 108. *Chem res toxicol*, **9**(7), 1079–91.

Delaney1992. 1992. Finding and filling protein cavities using cellular logic operations. *J mol graph*, **10**, 174–177.

Delaunay-Bertoncini, N., Pichon, V., & Hennion, M. C. 2003. Experimental comparison of three monoclonal antibodies for the class-selective immunoextraction of triazines. correlation with molecular modeling and principal component analysis studies. *J chromatogr a*, **999**(1-2), 3–15.

DeLucas, L. J., & Bray, T. L. 2003. Efficient protein crystallization. *J struct biol*, **142**(1), 188–206.

deLuise, V. P., & Anderson, D. R. 1983. Primary infantile glaucoma (congenital glaucoma). *Surv ophthalmol*, **28**(1), 1–19.

Dill, K. A. 1990. Dominant forces in protein folding. *Biochemistry*, **29**(31), 7133–55.

Dunbrack, R. L., & Karplus, M. 1993. Backbone-dependent rotamer library for proteins:applications to side-chain prediction. *J mol biol*, **230**, 543–574.

Dunbrack, R. L., & Karplus, M. 1994. Conformational analysis of the backbone dependent rotamer preferences of protein side chains. *Nat struct biol*, **5**, 334–40.

Eriksson, A. E., Baase, W. A., Zhang, X. J., Heinz, D. W., Blaber, M., Baldwin, E. P., & Matthews, B. W. 1992. Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*, **255**(5041), 178–83.

Fan, H., & Mark, A. E. 2004. Refinement of homology-based protein structures by molecular dynamics simulation techniques. *Protein sci*, **13**, 211–20.

Fieg, M., Onufriev, A., Lee, M. S., Im, W., Case, D. A., & Brooks, C. L. 2004. Performance comparison of genarilized born and poisson methods in the calculation of electrostatic solvation energies for protein structures. *J comp chem*, **25**(2), 265–84.

Finn, R. D., & Tate, J. 2008. The pfam protein families database. *Nucleic acids res*, **36**, D281–8.

Forrest, L. R., Tang, C. L., & Honig, B. 2006. On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. *Biophys j*, **91**(2), 508–17.

Friedman, D. I., & Court, D. L. 1995. Transcription antitermination: the lambda paradigm updated. *Mol microbiol*, **18**, 191–200.

Gabb, H. A., Jackson, R, M., & Sternberg, M. J. E. 1997. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J mol biol*, **272**, 106–120.

Garnier, J., Osguthorpe, D. J., & Robson, B. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J mol biol*, **120**(1), 97–120.

Gencik, A., Gencikova, A., & Ferak, V. 1982. Population genetical aspects of primary congenital glaucoma. i. incidence, prevalence, gene frequency, and age of onset. *Hum genet*, **61**(3), 193–7.

George, R. A., & Heringa, J. 2002. An analysis of protein domain linkers: their classification and role in protein folding. *Protein eng*, **15**, 871–879.

Georgopoulos, C. P. 1971. Bacterial mutants in which the gene n function of bacteriophage lambda is blocked have an altered rna polymerase. *Proc natl acad sci u s a*, **68**(12), 2977–81.

Gerstein, M., & Levitt, M. 1998. Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. *Protein sci*, **7**(2), 445–56.

Ghysen, A., & Pironio, M. 1972. Relationship between the n function of bacteriophage lambda and host rna polymerase. *J mol biol*, **65**(2), 259–72.

Gnatt, A. L., Cramer, P., Fu, J., Bushnell, D. A., & Kornberg, R. D. 2001. Structural basis of transcription: an rna polymerase ii elongation complex at 3.3 a resolution. *Science*, **292**(5523), 1876–82.

Gogos, A., Jantz, D., Senturker, S., Richardson, D., Dizdaroglu, M., & Clarke, N. D. 2000. Assignment of enzyme substrate specificity by principal component analysis of aligned protein sequences: an experimental test using dna glycosylase homologs. *Proteins*, **40**(1), 98–105.

Goodsell, D., & OLson, A. 1990. Automated docking of substrates to proteins by simulated annealing. *Proteins*, **8**, 195–202.

Graham-Lorence, S., & Peterson, J. A. 1996. P450s: structural similarities and functional differences. *Faseb j*, **10**(2), 206–14.

Graham-Lorence, S., Amarneh, B., White, R. E., Peterson, J. A., & Simpson, E. R. 1995. A three-dimensional model of aromatase cytochrome p450. *Protein sci*, **4**(6), 1065–80.

Greenblatt, J., Nodell, J. R., & Mason, S. W. 1993. Transcriptional antitermination. *Nature*, **364**, 401–406.

Guex, N., & Peitsch, M. C. 1997. Swiss-model and the swiss-pdbviewer: an environment for comparative protein modeling. *Electrophoresis*, **18**(15), 2714–23.

Gusarov, I., & Nudler, E. 2001. Control of intrinsic transcription termination by n and nusa: the basic mechanisms. *Cell*, **107**(4), 437–49.

Hakkola, J., Pasanen, M., Pelkonen, O., Hukkanen, J., Evisalmi, S., Anttila, S., Rane, A., Mantyla, M., Purkunen, R., Saarikoski, S., Tooming, M., & Raunio, H. 1997. Expression of cyp1b1 in human adult and fetal tissues and differential inducibility of cyp1b1 and cyp1a1 by ah receptor ligands in human placenta and cultured cells. *Carcinogenesis*, **18**(2), 391–7.

Halushka, M. K., Fan, J. B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R., & Chakravarti, A. 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat genet*, **22**(3), 239–47.

Hardin, C., Pogorelov, T. V., & Luthey-Schulten, Z. 2002. Ab initio protein structure prediction. *Curr opin struct biol*, **12**(2), 176–81.

Hasemann, C. A., Kurumbail, R. G., Bodupalli, S. S., Peterson, J. A., & Deisenhofer, J. 1995. Structure and function of cytochrome p450:a comparative analysis of three crystal structures. *Structure*, **2**, 41–62.

Hasler, J. A., Estabrook, R., Murray, M., Pikuleva, I., Waterman, M., Capdevila, J., Vjakumar, H., Helvig, C., Falck, J. R., Farrel, G., Kaminsky, L. S., Spivack, S. D., Boitier, E., & Beaune, P. 1999. Human cytochromes p450. *Mol asp med*, **20**, 1–137.

Hayes, C. L., Spink, D. C., Spink, B. C., Cao, J. Q., Walker, N. J., & Sutter, T. R. 1996. 17 beta-estradiol hydroxylation catalyzed by human cytochrome p450 1b1. *Proc natl acad sci u s a*, **93**(18), 9776–81.

Hayward, S., & Go, N. 1995. Collective variable description of native protein dynamics. *Ann. rev. phys. chem*, **46**, 223–250.

Hayward, S., Kitao, A., & Go, N. 1994. Harmonic and anharmonic aspects in the dynamics of bpti: a normal mode analysis and principal component analysis. *Protein sci*, **3**(6), 936–43.

Hellmold, H., Rylander, T., Magnusson, M., Reihner, E., Warner, M., & Gustafsson, J. A. 1998. Characterization of cytochrome p450 enzymes in human breast tissue from reduction mammaplasties. *J clin endocrinol metab*, **83**(3), 886–95.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., & Sippl, M. J. 1990. Identification of native protein folds amongst a large number of incorrect models. *J mol biol*, **216**, 167–180.

Hess, B. 2000. Similarities between principal components of protein dynamics and random diffusion. *Phys rev e stat phys plasmas fluids relat interdiscip topics*, **62**(6 Pt B), 8438–48.

Hess, B. 2002. Convergence of sampling in protein simulations. *Phys rev e stat nonlin soft matter phys*, **65**(3 Pt 1), 031910.

Hess, B., Bekker, H., Berendsen, H. J. C., & Fraaije, J. G. E. M. 1997. Lincs: A linear constraint solver for molecular simulations. *J comp chem*, **18**, 1463–72.

Hilbert, M., Bohm, G., & Jaenicke, R. 1993. Structural relationships of homologous proteins as a fundamental principle in homology modeling. *Proteins*, **17**(2), 138–51.

Ho, C. M. W., & Marshall, G. R. 1990. Cavity search: an algorithm for the isolation and display of cavity like binding regions. *J comput aid mol des*, **4**, 337–354.

Horovitz, A., Serrano, L., Avron, B., Bycroft, M., & Fersht, A. R. 1990. Strength and co-operativity of contributions of surface salt bridges to protein stability. *J mol biol*, **216**(4), 1031–44.

Hung, R. J., Boffetta, P., Brennan, P., Malaveille, C., Hautefeuille, A., Donato, F., Gelatti, U., Spaliviero, M., Placidi, D., Carta, A., Scotto di Carlo, A., & Porru, S. 2004. Gst, nat, sult1a1, cyp1b1 genetic polymorphisms, interactions with environmental exposures and bladder cancer risk in a high-risk population. *Int j cancer*, **110**(4), 598–604.

Hurley, L. H., Reck, T., Thurston, D. E., Langley, D. R., Holden, K. G., Hertzberg, R. P., Hoover, J. R., Gallagher, G. Jr., Faucette, L. F., & Mong, S. M. 1988. Pyrrolo[1,4]benzodiazepine antitumor antibiotics: relationship of dna alkylation and sequence specificity to the biological activity of natural and synthetic compounds. *Chem res toxicol*, **1**(5), 258–68.

Hutchinson, E. G., & Thornton, J. M. 1994. A revised set of potentials for beta-turn formation in proteins. *Protein sci*, **3**(12), 2207–16.

Inoue, K., Asao, T., & Shimada, T. 2000. Ethnic-related differences in the frequency distribution of genetic polymorphisms in the cyp1a1 and cyp1b1 genes in japanese and caucasian populations. *Xenobiotica*, **30**(3), 285–95.

Jansson, I., Stoilov, I., Sarfarazi, M., & Schenkman, J. B. 2001. Effect of two mutations of human cyp1b1, g61e and r469w, on stability and endogenous steroid substrate metabolism. *Pharmacogenetics*, **11**(9), 793–801.

Jarvis, R.A, & Patrick, E.A. 1973. Clustering using a similarity measure based on shared nearest neighbors. *Ieee trans. comput*, **22**, 1025–1034.

Jeannot, E., Poussin, K., Chiche, L., Bacq, Y., Sturm, N., Scoazec, J. Y., Buffet, C., Van Nhieu, J. T., Bellanne-Chantelot, C., de Toma, C., Laurent-Puig, P., Bioulac-Sage, P., & Zucman-Rossi, J. 2007. Association of cyp1b1 germ line mutations with hepatocyte nuclear factor 1alpha-mutated hepatocellular adenoma. *Cancer res*, **67**(6), 2611–6.

Jin, D. J., Cashel, M., Friedman, D. I., Nakamura, Y., Walter, W. A., & Gross, C. A. 1988. Effects of rifampicin resistant rpob mutations on antitermination and interaction with nusa in escherichia coli. *J mol biol*, **204**(2), 247–61.

Johnson, G. C., & Todd, J. A. 2000. Strategies in complex disease mapping. *Curr opin genet dev*, **10**, 330–334.

Johnson, M. S., Srinivasan, N., Sowdhamini, R., & Blundell, T. L. 1994. Knowledge-based protein modeling. *Crit rev biochem mol biol*, **29**, 1.

Jones, D. T., Taylor, W. R., & Thornton, J. M. 1992. A new approach to protein fold recognition. *Nature*, **358**(6381), 86–9.

Jones, G., Willett, P., Glen, R. C., Leach, A. R., & Taylor, R. 1997. Development and validation of a genetic algorithm for flexible docking. *J mol biol*, **267**(3), 727–48.

Jones, S., & Thornton, J. M. 2001. *Protein quaternary structure: Subunit-subunit interactions*. University college, london, england edn.

Jones, S., Stewart, M., Michie, A., Swindells, M. B., Orengo, C., & Thornton, J. M. 1998. Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein sci*, **7**, 233–242.

Jorgensen, W. L. 1991. Rusting of the lock and key model for protein-ligand binding. *Science*, **254**(5034), 954–5.

Kabsch, W., & Sander, C. 2004. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**(12), 2577–2637.

Kairys, V., & Gilson, M. K. 2006. Using protein homology models for structure-based studies: approaches to model refinement. *Scientificworldjournal*, **6**, 1542–54.

Kakiuchi, T., Isashiki, Y., Nakao, K., Sonoda, S., Kimura, K., & Ohba, N. 1999. A novel truncating mutation of cytochrome p4501b1 (cyp1b1) gene in primary infantile glaucoma. *Am j ophthalmol*, **128**(3), 370–2.

Kamal, A., Ramesh, G., Srinivas, O., Ramulu, P., Laxman, N., Rehana, T., Deepak, M., Achary, M. S., & Nagarajaram, H. A. 2004. Design, synthesis, and evaluation of mixed imine-amine pyrrolobenzodiazepine dimers with efficient dna binding affinity and potent cytotoxicity. *Bioorganic med chem*, **12**, 5427–5436.

Kaplan, W., & Littlejohn, T. G. 2001. Swiss-pdb viewer (deep view). *Brief bioinform*, **2**(2), 195–7.

Karplus, M., & Kuriyan, J. 2005. Molecular dynamics and protein function. *Proc nat acad sci*, **102**(19), 6679–6685.

Kelley, L. A., MacCallum, R. M., & Sternberg, M. J. E. 2000. Structural profiles in the program 3d-pssm. *J mol biol*, **299**(2), 501–22.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., & Phillips, D. C. 1958. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, **181**(4610), 662–6.

Kitchen, D. B., Decornez, H., Furr, J. R., & Bajorath, J. 2004. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews. drug discovery*, **3**(11), 935–49.

Komissarova, N., Becker, J., Solter, S., Kireeva, M., & Kashlev, M. 2002. Shortening of rna:dna hybrid in the elongation complex of rna polymerase is a prerequisite for transcription termination. *Mol cell*, **10**(5), 1151–62.

Korzheva, N., Mustaev, A. Kozlov, M., Malhotra, A., Nikiforov, V., Goldfarb, A., & Darst, S. A. 2000. A structural model of transcription elongation. *Science*, **289**, 619–625.

Krawczak, M., & Cooper, D. N. 1997. The human gene mutation database. *Trends genet*, **13**(3), 121–2.

Kuznedelov, K., Korzheva, N., Mustaev, A., & Severinov, K. 2002. Structure-based analysis of rna polymerase function: the largest subunit's rudder contributes critically to elongation complex stability and is not involved in the maintenance of rna-dna hybrid length. *Embo j*, **21**(6), 1369–78.

Laskowski, R. A., Moss, D. S., & Thornton, J. M. 1993. Main-chain bond lengths and bond angles in protein structures. *J mol biol*, **231**(4), 1049–67.

Lazinski, D., Grzadzielska, E., & Das, A. 1989. Sequence-specific recognition of rna hairpins by bacteriophage antiterminators requires a conserved arginine rich motif. *Cell*, **59**, 207–218.

Leach, A. R. 2001. *Molecular modelling: Principles and applications*. 2 edn.

Lengauer, T., & Rarey, M. 1996. Computational methods for biomolecular docking. *Curr opin struct biol*, **6**(3), 402–6.

Levitt, D. G., & Banaszak, L. J. 1992. Pocket: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J mol graph*, **10**(4), 229–34.

Lewis, D. F., Gillam, E. M., Everett, S. A., & Shimada, T. 2003. Molecular modelling of human cyp1b1 substrate interactions and investigation of allelic variant effects on metabolism. *Chem biol interact*, **145**(3), 281–95.

Lewis, P. N., & Momany, F. A. 1973. Chain reversals in proteins. *Biochem biophys acta*, **303**(2), 211–29.

Li, D. N., Seidel, A., Pritchard, M. P., Wolf, C. R., & Friedberg, T. 2000. Polymorphisms in p450 cyp1b1 affect the conversion of estradiol to the potentially carcinogenic metabolite 4-hydroxyestradiol. *Pharmacogenetics*, **10**(4), 343–53.

Li, H., & Poulos, T. L. 2004. Crystallization of cytochromes p450 and substrate-enzyme interactions. *Curr top med chem*, **4**(16), 1789–802.

Liehr, J. G., & Ricci, M. J. 1996. 4-hydroxylation of estrogens as marker of human mammary tumors. *Proc natl acad sci u s a*, **93**(8), 3294–6.

Lindahl, E, Hess, B, & van der Spoel, D. 2001. Gromacs 3.0: A package for molecular simulation and trajectory analysis. *J. mol. mod*, **7**, 306–317.

Ludemann, S. K., Carugo O., & Wade, R. C. 1997. Substrate access to cytochrome p450cam: A comparison of a thermal motion pathway analysis with molecular dynamics simulation data. *J. mol. model*, **3**, 369–374.

Ludemann, S. K., Lounnas, V., & Wade, R. C. 2000a. How do substrates enter and products exit the buried active site of cytochrome p450cam 1. random expulsion molecular dynamics investigation of ligand access channels and mechanisms. *J mol biol*, **303**(5), 797–811.

Ludemann, S. K., Lounnas, V., & Wade, R. C. 2000b. How do substrates enter and products exit the buried active site of cytochrome p450cam 2. steered molecular dynamics and adiabatic mapping of substrate pathways. *J mol biol*, **303**(5), 813–30.

Luthy, R., Bowie, J. U., & Eisenberg, D. 1992. Assessment of protein models with three-dimensional profiles. *Nature*, **356**(6364), 83–5.

Mammen, J. S., Pittman, G. S., Li, Y., Abou-Zahr, F., Bejjani, B. A., Bell, D. A., Strickland, P. T., & Sutter, T. R. 2003. Single amino acid mutations, but not common polymorphisms, decrease the activity of cyp1b1 against (-)benzo[a]pyrene-7r-trans-7,8-dihydrodiol. *Carcinogenesis*, **24**(7), 1247–55.

Manning, M. C. 2005. Use of infrared spectroscopy to monitor protein structure and stability. *Expert rev proteomics*, **2**(5), 731–43.

Mansuy, D. 1998. The great divesity of reactions catalyzed by cytochrome p450. *Comp biochem physiol part c*, **121**, 5–14.

Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F., & Sali, A. 2000. Comparative protein structure modeling of genes and genomes. *Annu rev biophys biomol struct*, **29**, 291–325.

Martin, J., Gibrat, J. F., & Rodolphe, F. 2006. Analysis of an optimal hidden markov model for secondary structure prediction. *Bmc struct biol*, **6**, 25.

Martin, S. N., Sutherland, J., Levin, A. V., Klose, R., Priston, M., & Heon, E. 2000. Molecular characterisation of congenital glaucoma in a consanguineous canadian community: a step towards preventing glaucoma related blindness. *J med genet*, **37**(6), 422–7.

Mashima, Y., Suzuki, Y., Sergeev, Y., Ohtake, Y., Tanino, T., Kimura, I., Miyata, H., Aihara, M., Tanihara, H., Inatani, M., Azuma, N., Iwata, T., & Araie, M. 2001. Novel cytochrome p4501b1 (cyp1b1) gene mutations in japanese patients with primary congenital glaucoma. *Invest ophthalmol vis sci*, **42**(10), 2211–6.

Masood, E. 1999. As consortium plans free snp map of human genome. *Nature*, **398**(6728), 545–6.

Masuda, S., Prosser, D. E., Guo, Y. D., Kaufmann, M., & Jones, G. 2007. Generation of a homology model for the human cytochrome p450, cyp24a1, and the testing of putative substrate binding residues by site-directed mutagenesis and enzyme activity studies. *Arch biochem biophys*, **460**(2), 177–91.

McDonald, I.K, & Thornton, J.M. 1994. Satisfying hydrogen bonding potential in proteins. *J.mol.biol*, **238**, 777–793.

McDonald, I., Naylor D. Jones D., & Thornton, J. 1993. *Hbplus: Hydrogen bond calculator version 2.25*. London: University College London.

McFadyen, M. C., Breeman, S., Payne, S., Stirk, C., Miller, I. D., Melvin, W. T., & Murray, G. I. 1999. Immunohistochemical localization of cytochrome p450 cyp1b1 in breast cancer with monoclonal antibodies specific for cyp1b1. *J histochem cytochem*, **47**(11), 1457–64.

McGuffin, L. J., Bryson, K., & Jones, D. T. 2000. The psipred protein structure prediction server. *Bioinformatics*, **16**(4), 404–5.

McLellan, R. A., Oscarson, M., Hidestrand, M., Leidvik, B., Jonsson, E., Otter, C., & Ingelman-Sundberg, M. 2000. Characterization and functional analysis of two common human cytochrome p450 1b1 variants. *Arch biochem biophys*, **378**(1), 175–81.

Mizuguchi, K., Deane, C. M., Blundell, T. L., Johnson, M. S., & Overington, J. P. 1998. Joy: protein sequence-structure representation and analysis. *Bioinformatics*, **14**(7), 617–23.

Mogridge, J., Legault, P., Li, J., van Oene, M. D., E., Kay. L., & Greenblatt, J. 1998. Independent ligand-induced folding of the rna-binding domain and two functionally distinct antitermination regions in the phage λ n protein. *Mol cell*, **1**, 265–275.

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., & Olson, A. J. 1999. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *J comp chem*, **19**(14), 1639–1662.

Muegge, I. 2006. Pmf scoring revisited. *J med chem*, **49**(20), 5895–902.

Murray, G. I., Taylor, M. C., McFadyen, M. C., McKay, J. A., Greenlee, W. F., Burke, M. D., & Melvin, W. T. 1997. Tumor-specific expression of cytochrome p450 cyp1b1. *Cancer res*, **57**(14), 3026–31.

Murray, G. I., Melvin, W. T., Greenlee, W. F., & Burke, M. D. 2001. Regulation, function, and tissue-specific expression of cytochrome p450 cyp1b1. *Annu rev pharmacol toxicol*, **41**, 297–316.

Murzin, A. G., & Brenner, S. E. 1995. Scop: ac structural classification of proteins database for the investigation of sequences and structures. *J mol biol*, **247**(4), 536–40.

Nagarajaram, H. A., Reddy, B. V., & Blundell, T. L. 1999. Analysis and prediction of inter-strand packing distances between beta-sheets of globular proteins. *Protein eng*, **12**(12), 1055–62.

Needleman, S., & Wunsch, C. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J mol biol*, **48**(3), 443–53.

Negishi, M., Iwasaki, M., Juvonen, R. O., Sueyoshi, T., Darden, T. A., & Pedersen, L. G. 1996. Structural flexibility and functional versatility of cytochrome p450 and rapid evolution. *Mutat res*, **350**(1), 43–50.

Neidle, S., Puvvada, M. S., & Thurston, D. E. 1994. The relevance of drug dna sequence specificity to anti-tumor activity. *Eur j cancer*, **30A**(4), 567–8.

Nelson, D. R., Koymans, L., Kamataki, T., Stegeman, J. J., Feyereisen, R., Waxman, D. J., Waterman, M. R., Gotoh, O., Coon, M. J., & Estabrook, R. W. 1996. P450 superfamily:update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics*, **6**, 1–42.

Ng, P. C., & Henikoff, S. 2001. Predicting deleterious amino acid substitutions. *Genome res*, **11**(5), 863–74.

Nishikawa, K. 1983. Assessment of secondary-structure prediction of proteins. comparison of computerized chou-fasman method with others. *Biochim biophys acta*, **748**(2), 285–99.

Nudler, E., & Gottesman, M. E. 2002. Transcription termination and anti-termination in *E. coli*. *Genes cells*, **7**, 755–768.

Nunez, S., Wing, C., Antoniou, D., Schramm, V. L., & Schwartz, S. D. 2006. Insight into catalytically relevant correlated motions in human purine nucleoside phosphorylase. *J phys chem a mol spectrosc kinet environ gen theory*, **110**(2), 463–72.

Obuchowski, M., Wegrzyn, A., Szalewska-Palasz, A., Thomas, M. S., & Wegrzyn, G. 1997. An rna polymerase alpha subunit mutant impairs n-dependent transcriptional antitermination in escherichia coli. *Mol microbiol*, **23**(2), 211–22.

Oesterheld, J. R. 1998. A review of developmental aspects of cytochrome p450. *J child adolesc psychopharmacol*, **8**(3), 161–74.

Ohtake, Y., Kubota, R., Tanino, T., Miyata, H., & Mashima, Y. 2000. Novel compound heterozygous mutations in the cytochrome p4501b1 gene (cyp1b1) in a japanese patient with primary congenital glaucoma. *Ophthalmic genet*, **21**(3), 191–3.

Oprea, T. I., Hummer, G., & Garcia, A. E. 1997. Identification of a functional water channel in cytochrome p450 enzymes. *Proc natl acad sci u s a*, **94**(6), 2133–8.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. 1997. Cath-a hierarchical classification of protein domain structures. *Structure*, **5**, 1093–1108.

Ota, N., & Agard, D. A. 2001. Enzyme specificity under dynamic control ii: Principal component analysis of alpha-lytic protease using global and local solvent boundary conditions. *Protein sci*, **10**(7), 1403–14.

Overington, J., & Johnson, M. S. 1990. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc biol sci*, **241**(1301), 132–45.

Pace, C. N., & Scholtz, J. M. 1998. A helix propensity scale based on experimental studies of peptides and proteins. *Biophysical j*, **75**, 422–427.

Pan, P. W., Dickson, R. J., Gordon, H. L., Rothstein, S. M., & Tanaka, S. 2005. Functionally relevant protein motions: extracting basin-specific collective coordinates from molecular dynamics trajectories. *J chem phys*, **122**(3), 34904.

Panicker, S. G., Reddy, A. B., Mandal, A. K., Ahmed, N., Nagarajaram, H. A., Hasnain, S. E., & Balasubramanian, D. 2002. Identification of novel mutations causing familial primary congenital glaucoma in indian pedigrees. *Invest ophthalmol vis sci*, **43**(5), 1358–66.

Panicker, S. G., Mandal, A. K., Reddy, A. B., Gothwal, V. K, & Hasnain, S. E. 2004. Correlation of genotype with phenotype in indian patients with primary congenital glaucoma. *Invest ophthalmol vis sci*, **45**(4), 1149–56.

Pearson, W. R. 1990. Rapid and sensitive sequence comparison with fastp and fasta. *Methods enzymol*, **183**, 63–98.

Peters, G. H., van Aalten, D. M., Svendsen, A., & Bywater, R. 1997. Essential dynamics of lipase binding sites: the effect of inhibitors of different chain length. *Protein eng*, **10**(2), 149–58.

Petrusek, R. L., Uhlenhopp, E. L., Duteau, N., & Hurley, L. H. 1982. Reaction of anthramycin with dna. biological consequences of dna damage in normal and xeroderma pigmentosum cell. *J biol chem*, **257**(11), 6207–16.

Pieper, U., Eswar, N., Davis, F., Madhusudhan, M. S., Rossi, A., Marti-Renom, M. A., Karchin, R, Webb, B., Eramian, D., Shen, M., Kelly, L., Melo, F., & Sali, M. 2006. Modbase, a database of annotated comparative protein structure models, and associated resources. *Nucleic acids res*, **34**, D291–D295.

Plasilova, M., Ferakova, E., Kadasi, L., Polakova, H., Gerinec, A., Ott, J., & Ferak, V. 1998a. Linkage of autosomal recessive primary congenital glaucoma to the glc3a locus in roms (gypsies) from slovakia. *Hum hered*, **48**(1), 30–3.

Plasilova, M., Gerinec, A., & Ferak, V. 1998b. [molecular diagnosis of mutations responsible for recurrent and severe forms of primary congenital glaucoma]. *Cesk slov oftalmol*, **54**(5), 281–8.

Podust, L. M., Yermalitskaya, L. V., Lepesheva, G. I., Podust, V. N., Dalmasso, E. A., & Waterman, M. R. 2004. Estriol bound and ligand-free structures of sterol 14alpha-demethylase. *Structure*, **12**(11), 1937–45.

Ponder, J. W., & Richards, F. M. 1987. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J mol biol*, **193**, 775–792.

Poulos, T. L. 2003. Cytochrome p450 flexibility. *Proc natl acad sci u s a*, **100**(23), 13121–2.

Pratt, R. E., & Dzau, V. J. 1999. Genomics and hypertension: concepts, potentials, and opportunities. *Hypertension*, **33**(1 Pt 2), 238–47.

Pugalenthi, G., & Shameer, K. 2006. Harmony:a server for the assessment of protein structures. *Nucleic acids res*, **34**, W231–4.

Puvvada, M. S., Forrow, S. A., Hartley, J. A., Stephenson, P., Gibson, I, Jenkins, T. C., & Thurston, D. E. 1997. Inhibition of bacteriophage t7 rna polymerase *in vitro* transcription by dna-binding pyrrolo[2,1,c][1,4]benzodiazepines. *Biochemistry*, **36**, 2478–2484.

Qian, N., & Sejnowski, T. J. 1988. Predicting the secondary structure of globular proteins using neural network models. *J mol biol*, **202**(4), 865–84.

Raghava, G. P. S. 2002. Apssp2: A combination method for protein secondary structure prediction based on neural network and example based larning. *Casp5*, A–132.

Rajamani, R., & Good, A. C. 2007. Rankig poses in structure-based lead discovery and optimization:current trands in scoring function development. *Curr opinion drug dis dev*, **10**(3), 308–15.

Ramachandran, G. N., & Kolaskar, A. S. 1974. The mean geometry of the peptide unit from crystal structure data. *Biochem biophys acta*, **359**, 298–302.

Ramachandran, G. N., & Mitra, A. K. 1976. An explanation for the rare occurrence of *cis* peptide units in proteins and polypeptides. *J mol biol*, **107**, 85–92.

Ramachandran, G. N., & Ramakrishnan, C. 1963. Stereochemistry of polypeptide chain configurations. *J mol biol*, **7**, 95–9.

Ramazzotti, M, Degl'lnnocenti, D, G, Manao, & Giampietro, Ramponi. 2004. Entropy calculator: getting the best from your multiple protein alignemnts. *Ital j biochem*, **53**(1), 16–22.

Ramensky, V., Bork, P., & Sunyaev, S. 2002. Human non-synonymous snps: server and survey. *Nucleic acids res*, **30**(17), 3894–900.

Raucy, J. L., & Allen, S. W. 2001. Recent advances in p450 research. *Pharmacogenomics j*, **1**(3), 178–186.

Reddy, A. B., Kaur, K., Mandal, A. K., Panicker, S. G., Thomas, R., Hasnain, S. E., Balasubramanian, D., & Chakrabarti, S. 2004. Mutation spectrum of the cyp1b1 gene in indian primary congenital glaucoma patients. *Mol vis*, **10**, 696–702.

Reddy, B. V., & Blundell, T. L. 1993. Packing of secondary structural elements in proteins. analysis and prediction of inter-helix distances. *J mol biol*, **233**(3), 464–79.

Reddy, B. V., Nagarajaram, H. A., & Blundell, T. L. 1999. Analysis of interactive packing of secondary structural elements in alpha/beta units in proteins. *Protein sci*, **8**(3), 573–86.

Richardson, J. P. 2002. Rho-dependent termination and atpases in transcript termination. *Biochem biophy acta*, **1577**, 251–260.

Richardson, J. S. 1981. The anatomy and taxonomy of protein structure. *Adv protein chem*, **34**, 168–340.

Rose, G. D., Gierasch, L. M., & Smith, J. A. 1985. Turns in peptides and proteins. *Adv protein chem*, **37**, 1–109.

Rost, B. 1997. Protein structures sustain evolutionary drift. *Fold des*, **2**(3), S19–24.

Rost, B. 1999. Twilight zone of protein sequence alignments. *Prot eng des sel*, **12**(2), 85–94.

Rost, B., & Sander, C. 1993. Prediction of protein secondary structure at better than 70accuracy. *J mol biol*, **232**(2), 584–99.

Rule, G. S., & Hitchens, T. K. 2006. *Fundamentals of protein nmr spectroscopy*. Springer, isbn 1-4020-3499-7 edn.

Ryckaert, J. P., Ciccotti, G., & Berendsen, H. J. C. 1977. Numerical integration of the cartesian equations of motion of a system with constraints; molecular dynamics of n-alkanes. *J comp phys*, **23**, 327–341.

Saarela, J. T., Tuppurainen, K., Perakyla, M., Santa, H., & Laatikainen, R. 2002. Correlative motions and memory effects in molecular dynamics simulations of molecules: principal components and rescaled range analysis suggest that the motions of native bpti are more correlated than those of its mutants. *Biophys chem*, **95**(1), 49–57.

Sali, A., & Blundell, T. L. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J mol biol*, **234**(3), 779–815.

Sarfarazi, M., Akarsu, A. N., Hossain, A., Turacli, M. E., Aktan, S. G., Barsoum-Homsy, M., Chevrette, L., & Sayli, B. S. 1995. Assignment of a locus (glc3a) for primary congenital glaucoma (buphthalmos) to 2p21 and evidence for genetic heterogeneity. *Genomics*, **30**(2), 171–7.

Schauer, A. T., Cheng, S. W., Zheng, C., St Pierre, L., Alessi, D., Hidayetoglu, D. L., Costantino, N., Court, D. L., & Friedman, D. I. 1996. The alpha subunit of rna polymerase and transcription antitermination. *Mol microbiol*, **21**(4), 839–51.

Scott, E. E., He, Y. A., Wester, M. R., White, M. A., Chin, C. C., Halpert, J. R., Johnson, E. F., & Stout, C. D. 2003. An open conformation of mammalian cytochrome p450 2b4 at 1.6-a resolution. *Proc natl acad sci u s a*, **100**(23), 13196–201.

Scott, E. E., White, M. A., He, Y. A., Johnson, E. F., Stout, C. D., & Halpert, J. R. 2004. Structure of mammalian cytochrome p450 2b4 complexed with 4-(4-chlorophenyl)imidazole at 1.9-a resolution: insight into the range of p450 conformations and the coordination of redox partner binding. *J biol chem*, **279**(26), 27294–301.

Shamim, M. T. A., Anwaruddin, M., & Nagarajaram, H. A. 2007. Support vector machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs. *Bioinformatics*, **23**(24), 3320–3327.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell system technicaljournal*, **27**, 623–656.

Shi, J., Blundell, T. L., & Mizuguchi, K. 2001. Fugue: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J mol biol*, **310**(1), 243–57.

Shimada, T., Watanabe, J., Kawajiri, K., Sutter, T. R., Guengerich, F. P., Gillam, E. M., & Inoue, K. 1999. Catalytic properties of polymorphic human cytochrome p450 1b1 variants. *Carcinogenesis*, **20**(8), 1607–13.

Shirley, B. A., Stanssens, P., Hahn, U., & Pace, C. N. 1992. Contribution of hydrogen bonding to the conformational stability of ribonuclease t1. *Biochemistry*, **31**(3), 725–32.

Shoichet, B. K., Kuntz, I. D., & Bodian, D. L. 2004. Molecular docking using shape descriptors. *J comp chem*, **13**(3), 380–397.

Simons, K. T., Kooperberg, C., Huang, E., & Baker, D. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J mol biol*, **268**(1), 209–25.

Sippl, M. J. 1993. Recognition of errors in three-dimensional structures of proteins. *Proteins*, **17**(4), 355–62.

Sippl, M. J., & Weitckus, S. 1992. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformation. *Proteins*, **13**, 258–271.

Smyth, M. S., & Martin, J. H. 2000. X ray crystallography. *Mol pathol*, **53**(1), 8–14.

Snow, C., Qi, G., & Hayward, S. 2007. Essential dynamics sampling study of adenylate kinase: comparison to citrate synthase and implication for the hinge and shear mechanisms of domain motions. *Proteins*, **67**(2), 325–37.

Spink, D. C., Spink, B. C., Cao, J. Q., DePasquale, J. A., Pentecost, B. T., Fasco, M. J., Li, Y., & Sutter, T. R. 1998. Differential expression of cyp1a1 and cyp1b1 in human breast epithelial cells and breast tumor cells. *Carcinogenesis*, **19**(2), 291–8.

Spoel, D. V., Lindahl, E., & Hess, B. 2006. *Gromacs user manual*. 3.3 edn.

Sternberg, N. 1976. A class of rifr rna polymerase mutations that interferes with the expression of coliphage lambda late gene. *Virology*, **73**(1), 139–54.

Stoilov, I., Akarsu, A. N., Alozie, I., Child, A., Barsoum-Homsy, M., Turacli, M. E., Or, M., Lewis, R. A., Ozdemir, N., Brice, G., Aktan, S. G., Chevrette, L., Coca-Prados, M., & Sarfarazi, M. 1998. Sequence analysis and homology modeling suggest that primary congenital glaucoma on 2p21 results from mutations disrupting either the hinge region or the conserved core structures of cytochrome p4501b1. *Am j hum genet*, **62**(3), 573–84.

Stoilov, I., Jansson, I., Sarfarazi, M., & Schenkman, J. B. 2001. Roles of cytochrome p450 in development. *Drug metabol drug interact*, **18**(1), 33–55.

Stoilov, I. R., Costa, V. P., Vasconcellos, J. P., Melo, M. B., Betinjane, A. J., Carani, J. C., Oltrogge, E. V., & Sarfarazi, M. 2002. Molecular genetics of primary congenital glaucoma in brazil. *Invest ophthalmol vis sci*, **43**(6), 1820–7.

Stoilov, I.R., & Sarfaraze, M. 2002. The third genetic locus (glc3c) for primary congenital glaucoma (pcg) maps to chromosome 14q24.3. *Opthalmol. vis. sci. suppl*, **43**, Abstract no. 3015.

Storbeck, K., Swart, P., & Swart, A. C. 2007. Cytochrome p450 side-chain cleavage: Insights gained from homology modeling. *Mol cell endocrinology*, **265-266**, 65–70.

Street, A. G., & Mayo, S. L. 1999. Intrinsic β-sheet propensities result from van der waals interactions between side chains and the local backbone. *Proc nat acad sci*, **96**(16), 9074–9076.

Sunyaev, S., Ramensky, V., & Bork, P. 2000. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends genet*, **16**, 198–200.

Sunyaev, S., Ramensky, V., Koch, I., Lathe, W, Kondrashov, A. S., & Bork, P. 2001. Prediction of deleterious human alleles. *Hum mol genet*, **10**(6), 591–7.

Sussman, J. L., & Lin, D. 1998. Protein data bank(pdb):database of three-dimensional structural information of biological macromolecules. *Acta crystallogr d biol crystallogr*, **54**, 1078–84.

Sutcliffe, M. J., Haneef, I., Carney, D., & Blundell, T. L. 1987. Knowledge-based modeling of homologous proteins, part 1: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Prot engg*, **1**, 377.

Sutter, T. R., Tang, Y. M., Hayes, C. L., Wo, Y. Y., Jabs, E. W., Li, X., Yin, H., Cody, C. W., & Greenlee, W. F. 1994. Complete cdna sequence of a human dioxin-inducible mrna identifies a new gene subfamily of cytochrome p450 that maps to chromosome 2. *J biol chem*, **269**(18), 13092–13099.

Szalewska-Palasz, A., Strzelczyk, B., Herman-Antosiewicz, A., Wegrzyn, G., & Thomas, M. S. 2003. Genetic analysis of bacteriophage lambdan-dependent antitermination suggests a possible role for the rna polymerase alpha subunit in facilitating specific functions of nusa and nuse. *Arch microbiol*, **180**(3), 161–8.

Tavernelli, L., Costesta, S., & DiLorio, E. E. 2003. Protein dynamics, thermal stability, and free energy landscapes: a molecular dynamics investigation. *Biophys j*, **85**(4), 2641–9.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. 1994. Clustalw: improving the sensitivity of progressive multiple sequence alignments through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids res*, **22**, 4673–80.

Thurston, D. E. 1993. *Molecular aspects of anticancer drug-dna interactions;macmillan:london*.

Tramontano, A., Leplae, R., & Morea, V. 2001. Analysis and assessment of comparative modeling predictions in casp4. *Proteins*, **45**(5), 22–38.

Tsigelny, I. F., Kotlovyi, V., & Wasserman, L. 2004. Snp analysis combined with protein structure prediction defines structure-functional relationships in cancer related cytochrome p450 estrogen metabolism. *Curr med chem*, **11**(5), 525–38.

Turner, P. J. 2005. *Xmgrace*. Center for Coastal and Land-Margin Research Oregon Graduate Institute of Science and Technology Beaverton, Oregon.

Uptain, S. M., Kane, C. M., & Chamberlin, M. J. 1997. Basic mechanisms of transcript eongation and its regulation. *Ann rev biochem*, **66**, 117–172.

Vadlamuri, S. V., Glover, D. D., Turner, T., & Sarkar, M. A. 1998. Regiospecific expression of cytochrome p4501a1 and 1b1 in human uterine tissue. *Cancer lett*, **122**(1-2), 143–50.

Valdar, W. S. 2002. Scoring residue conservation. *Proteins*, **48**(2), 227–41.

van Aalten, D. M., & Amadei, A. 1995. The essential dynamics of thermolysin: confirmation of the hige-bending motion and comparison of simulations in vacuum and water. *Proteins*, **22**, 45–54.

van Aalten, D. M., & Findlay, J. B. 1995. Essential dynamics of the cellular retinol-binding protein-evidence for ligand-induced conformational changes. *Protein eng*, **8**(11), 1129–35.

van Aalten, D. M., Findlay, J. B., Amadei, A., & Berendsen, H. J. 1995a. Essential dynamics of the cellular retinol-binding protein–evidence for ligand-induced conformational changes. *Protein eng*, **8**(11), 1129–35.

van Aalten, D. M., Amadei, A., Linssen, A. B., Eijsink, V. G., Vriend, G., & Berendsen, H. J. 1995b. The essential dynamics of thermolysin: confirmation of the hinge-bending motion and comparison of simulations in vacuum and water. *Proteins*, **22**(1), 45–54.

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., & Berendsen, H. J. 2005. Gromacs: fast, flexible, and free. *J comput chem*, **26**(16), 1701–18.

Venkatachalam, C. M., & Ramachandran, G. N. 1969. Conformation of polypeptide chains. *Annu rev biochem*, **38**, 45–82.

von Hippel, P. H. 1998. An integrated model of the transcription complex in elongation, termination, and editing. *Science*, **281**, 660–665.

Wallace, A. C., Laskowski, R. A., & Thornton, J. M. 1995. Ligplot: a program to generate schematic diagrams of protein-ligand interactions. *Protein eng*, **8**(2), 127–34.

Ward, J. J., McGuffin, L. J., Buxton, B. F., & Jones, D. T. 2003. Secondary structure prediction with support vector machines. *Bioinformatics*, **19**(13), 1650–5.

Wei, B. Q., Weaver, L. H., Ferrari, A. M., Matthews, B. W., & Shoichet, B. K. 2004. Testing a flexible-receptor docking algorithm in a model binding site. *J mol biol*, **337**(5), 1161–82.

Weisberg, R. A., & Gottesman, M. E. 1999. Processive antitermination. *J bacteriol*, **181**, 359–367.

Werck-Reichhart, D., & Feyereisen, R. 2000. Cytochromes p450: a success story. *Genome biol*, **1**(6), REVIEWS3003.

Wetlaufer, D. B. 1973. Nucleation, rapid folding, and globular intrachain regiosn in proteins. *Proc natl acad sci, usa*, **70**, 697–701.

Whitmore, L., & Wallace, B. A. 2007. Protein secondary structure analyses from circular dichroism spectroscopy: Methods and reference databases. *Biopolymers*.

Williams, P. A., Cosme, J., Sridhar, V., Johnson, E. F., & McRee, D. E. 2000. Mammalian microsomal cytochrome p450 monooxygenase: structural adaptations for membrane binding and functional diversity. *Mol cell*, **5**(1), 121–31.

Williams, P. A., Cosme, J., Ward, A., Angove, H. C., Vinkovic, M. D., & Jhoti, H. 2003. Crystal structure of human cytochrome p450 2c9 with bound warfarin. *Nature*, **424**(6947), 464–8.

Williams, P. A., Cosme, J., Vinkovic, D. M., Ward, A., Angove, H. C., Day, P. J., Vonrhein, C., Tickle, I. J., & Jhoti, H. 2004. Crystal structures of human cytochrome p450 3a4 bound to metyrapone and progesterone. *Science*, **305**(5684), 683–6.

Winn, P. J., Ludemann, S. K., Gauges, R., Lounnas, V., & Wade, R. C. 2002. Comparison of the dynamics of substrate access channels in three cytochrome p450s reveals different opening mechanisms and a novel functional role for a buried arginine. *Proc natl acad sci u s a*, **99**(8), 5361–6.

Wishart, D. 2005. Nmr spectroscopy and protein structure determination: application to drug discovery and development. *Curr pharm biotechnol*, **6**(2), 105–20.

Wuthrich, K. 1990. Protein structure determination in solution by nmr spectroscopy. *J biol chem*, **25**(36), 22059–62.

Wuthrich, K., Spitzfaden, C., Memmert, K., Widmer, H., & Wider, G. 1991. Protein secondary structure determination by nmr. application with recombinant human cyclophilin. *Febs lett*, **285**(2), 237–47.

Xiong, B., & Huang, X. Q. 2004. Conformational flexibility of beta-secretase:molecular dynamics simulation and essential dynamics analysis. *Acta pharmacol sin*, **25**(6), 705–13.

Xu, J., Baase, W. A., Baldwin, E., & Matthews, B. W. 1998. The response of t4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect. *Protein sci*, **7**(1), 158–77.

Yildirim, Y., & Doruker, P. 2004. Collective motions of rna polymerases. analysis of core enzyme, elongation complex and holoenzyme. *J biomol struct dyn*, **22**(3), 267–80.

Zhang, G., Campbell, E.A., Minakhin, L., Richter, C., Severinov, K., & Darst, S. A. 1999. Crystal structure of thermus aquaticus core rna polymerase at 3.3å resolution. *Cell*, **98**, 811–824.