Structure Prediction and Docking Interaction Analysis Of Human Autosomal Recessive Deafness Genes

Submitted By

Ambreen Qadir

31-FBAS/MSBI/F09

Supervised By

Dr. Asif Mir

Assistant Professor

Department of Environmental Science
Faculty of Basic & Applied Sciences
International Islamic University Islamabad

(2012)

MS
611.01816
AMS

1- Gene ; human

2. Biochemical genetics

In The Name Of
ALLAH
The Most Beneficent, the Most Merciful

This
Piece of work is
Dedicated to the People
Who
Contributed in My
Education

A

Dissertation submitted

to the

Department of Environmental Science,

International Islamic University, Islamabad

As

a partial fulfillment of the requirements

for the award of the degree of

Masters in Bio Informatics

MS (BI)

Declaration

I hereby solemnly declare that this research work "***Structure Prediction and Docking Interaction analysis of human autosomal recessive deafness genes (DFNB)***" neither as a whole nor as a part has been copied out from any source. It is further declared that I have done this research, with the accompanied report, entirely on the basis of my personal efforts, under the proficient guidance of my research supervisor Dr. Asif Mir.

Dated: 09-Mar-2012

Ambreen Qadir
MS (Bio Informatics)
31-FBAS/MSBI/F09

International Islamic University Islamabad

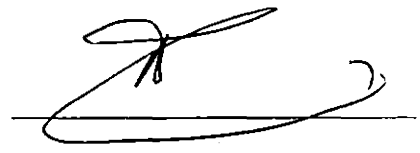
Dated: _____

Final Approval

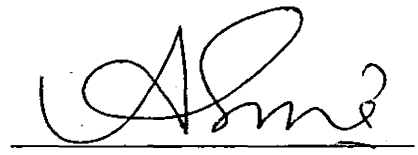
It is certificate that we have read the thesis submitted by Ms. Ambreen Qadir. To our Judgment that this project is of sufficient standard to warrant its acceptance by the International Islamic University, Islamabad for the M.S Degree in Bioinformatics.

Committee

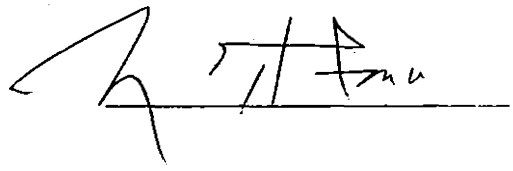
Exernal Examiner
Dr. Amir Ali Abbasi
Assistant Professor,
Quaid Azam University
Islamabad.



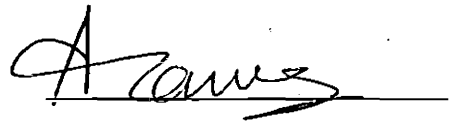
Internal Examiner
Dr. Asma Gul
Assistant Professor,
Department of Environmental Science,
International Islamic University,
Islamabad.



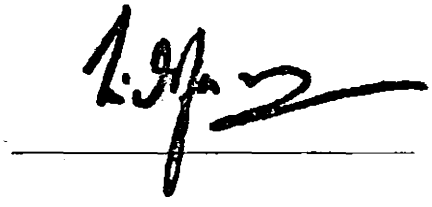
Supervisor
Dr. Asif Mir
Assistant Professor,
Department of Environmental Science,
International Islamic University,
Islamabad.



Co-Supervisor
Jabar Zaman Khan Khattak
Assistant Professor,
Department of Environmental Science,
International Islamic University,
Islamabad.



Dean FBAS
Dr. Muhammad Irfan Khan
International Islamic University
Islamabad, Pakistan



Contents

Acknowledgement

I simply bow my head before Allah Almighty for giving me faith in my abilities and enabling me to accomplish this research work. He showered on me His unremitting blessings throughout my life.

Parent's prayers remained as an asset for me throughout my career. Their relentless self-sacrifices always inspired me to put my best efforts in every challenge of my life. The humble accomplishment of this thesis would not have been possible without the contribution of many individuals, to whom I must express my appreciation and gratitude.

I am also grateful to Dr. Asif Mir my research supervisor for his guidance, encouragement and support in all phases of the research activity.

In the nutshell, it is the blessings of Almighty Allah, parent's prayers, the guidance of my supervisor Dr. Asif Mir, help of madam Hina and encouragement of my brothers-Muhammad Sadiq Qureshi and Muhammad Sajid Qureshi, which enabled me to complete this piece of work successfully.

Ambreen Qadir
MS (Bio Informatics)

List of Figures:

List of Tables:

Abstract

Although the genetic information needed for organisms development and life maintenance are encoded by the DNA molecule, the dynamic processes of life maintenance, replication, defense and reproduction are carried out by proteins. The functional ability of these macromolecules is result of their ability to build complex protein to accomplish a biological process.

Initially the ability to form protein-protein complexes was associated to only small number of proteins. However with the advancement in high throughput proteomics it has become clear that protein-protein interactions are norm and not exception.

Almost every essential processes occurring in living organisms, such as cell signaling, immune response, protein targeting and gene expression involves Protein–protein interactions (PPI) (Grosdidier *et al.*, 2009). The study of protein-protein interaction is important for in-depth understanding of disease at organism level (phizicky EM and fields S. 1995).

To date no in-silico study is conducted for protein-protein interaction analyzing autosomal recessive deafness genes. Our study is aimed to model the protein-protein docking interactions of our modeled proteins. The interacting Residues found by docking studies could serve as hot points for mutational studies of Autosomal recessive deafness. Protein–protein interactions knowledge of studied protein could be used to understand the complex metabolic interaction networks that occur in hearing process. The Predicted structure of autosomal recessive Genes could be used for generating more reliable Experimental models.

Chapter 1

Introduction

1. INTRODUCTION

Although the genetic information needed for organisms development and life maintenance are encoded by the DNA molecule, the dynamic processes of life maintenance, replication, defense and reproduction are carried out by proteins. Proteins are macromolecules they are the building blocks of all cells in our bodies and in all living creatures of all kingdoms. The functional ability of these macromolecules is result of their ability to build complex protein structures from amino acid sequences and to form protein complexes to accomplish a biological process.

Initially the ability to form protein-protein complexes was associated to only small number of proteins. However with the advancement in high throughput proteomics it has become clear that protein-protein interactions are norm and not exception.

Thus, protein function could be accurately understood by considering the larger context of the various binding complexes that each protein forms with interacting counterpart. Proteins are now considered as forming complex interaction networks controlled by highly efficient regulation.

Almost every essential processes occurring in living organisms, such as cell signaling, immune response, protein targeting and gene expression involves Protein–protein interactions (PPI) (Grosdidier *et al.*, 2009). Protein–protein interactions (PPI) knowledge is necessary to understand the complex metabolic interaction networks that take place in living organisms with the ultimate target of designing drugs for blocking or enhancing interactions of therapeutic interest. The study of protein-protein interaction is important for in-depth understanding of disease at organism level (phizicky EM and fields S, 1995).

Now-a-day targeting PPI of therapeutic interest has become hot research area. Although the number of three-dimensional (3-D) protein structures present in the Protein Data Bank (PDB) is rapidly growing, only a small fraction of numerous protein–protein complexes has been experimentally analyzed.

To date no *in-silico* study is conducted for protein-protein interaction analyzing autosomal recessive deafness genes. Our study is aimed to model the protein-protein docking interactions of our modeled proteins. For this we need the pdb structures of our selected genes.

Proteins structures determined by experimental methods such as X-ray crystallography, high-resolution Electron Microscopy and Nuclear Magnetic Resonance (NMR) methods of protein structure prediction are very expensive in terms of time and memory and each method has its own limitation. This problem can be overcome by using homology modeling techniques for structure prediction.

1.1. Protein Structure prediction Methods

Protein tertiary structure prediction can be broadly categories into three main classes

1.2. Ab initio Modeling

Ab initio protein structure prediction methods build protein 3D structures from sequence based on physical principles. The ab initio methods are important even though they are computationally demanding. Ab initio protein structure prediction would fail if Structure homologues are not available or possible undiscovered new fold exist in protein structure.

1.3. Protein Threading

This method takes benefit of the knowledge of existing structures and the principles by which they are stabilized. This method searches the amino acid sequence of an unknown structure against a database of solved structures. During the search a scoring function is used to map the sequence to the structures in the database. Protein threading for structure prediction is used when sequence identity is less than 30%.

1.3.1. Comparative Homology Modeling

Comparative homology modeling is based on the assumption that two evolutionary related proteins will share very similar structures. This assumption works well as it has been estimated that there are only around 2,000 distinct protein folds in nature, while the number of proteins goes up to millions (Zhang Y, 2008). It is fact that protein's fold is more evolutionarily preserved than its amino acid sequence , so target sequence can be

modeled with much accuracy on basis of distantly related template, provided that sequence similarity give clue to their relatedness. The sequence similarity greater than 30% sequence identity can be used for generating satisfactory homologous model.

The homology modeling process can be divided into four steps, template selection, target-template alignment, model building, and model evaluation.

1. Selection of Template

The most sensitive step in homology modeling is selection of suitable template. FASTA and BLAST employ the simplest method for template identification based on pair wise sequence alignments. More sensitive methods such as PSI-BLAST -iteratively update their position-specific scoring matrix to identify homologous sequences. Protein threading, fold recognition or 3D-1D alignment, can also be used for identifying homologous templates used in homology modeling (Marti-Rénom MA *et al.*, 2000).

2. Target-template alignment

The second step in the homology modeling is the alignment of the unknown sequence with the homologues structure sequences. One should bear in mind the correct options for the following Factors (1) Algorithm to be use for sequence alignment (2) Scoring matrix to be applied (3) Assigning the gap penalties.

3. Model building:

In this step the alignment generated in previous step is used for model building. Model building methods can be classified in three main classes.

3.1 Fragment assembly:

In fragment assembly method of homology modeling the final model is based upon conserved structural fragments identified in homologous structures. The variable regions are often built by using fragment libraries (Greer J, 1981).

3.2 Segment matching

This method splits the query sequence into a series of short segments, matched independently to its structure in Protein Data Bank. So the sequence alignment is done

for each segment rather than over the entire sequence. Template selection for each segment is done on basis of sequence similarity, comparisons of alpha carbon positions, and staric hindrance arising from the van-der-Waals interaction between target and template atoms (Levitt M,. 1992).

3.3 Satisfaction of spatial restraints

This method uses the same technique to generate the 3D structure as used in NMR spectroscopy. The probability density function for each restraint is constructed from alignment of target and template. These restraints are used for global optimization of the position for atoms in the protein structure to get the minimum energy state (Sali A & T.L. Blundell, 1993, Marc *et al*, 2000). The model is then derived by avoiding the violations of all the restraints. Due to flexibility of loops in aqueous solution this method is used mostly for loops modeling (Fiser et al., 2003).The most widely used software that satisfy spatial restraints while modeling structure is MODELLER (Sali *et al.*, 1993).

4. Model Evaluation

The methods employed for model evaluation are based upon statistical or physics-based energy calculation of residue by residue interaction frequencies among known protein structures. Evaluation programs include PROSA (Sippl, 1993), PROCHECK (Laskowski *et al.*, 1998), WHAT IF server (**http://swift.cmbi.ru.nl/WHAT IF/**) and RAMPAGE server **(http://swift.cmbi.kun.nl/swift/servers/modrama-submit.html).**

1.4. Mutation Modeling and Analysis of Proteins:

In the present study we also modeled the reported mutation of our modeled proteins. We used pymol (Delano W L, 2002) for normal and mutated structures analysis.

1.5. Protein Docking Analysis:

Docking studies have become nearly vital for study of macromolecular structures and interactions. Macromolecular modeling by Docking studies provides most detailed idea of drug-receptor interaction and has created a new rational approach to drug design where drug is designed based on its fit to three dimensional structures of receptor site, rather than by analogy to other active structures of random leads.

Protein docking is the task of calculating the 3D structure of a protein complex starting from unbound or model-built protein structures (Halperin *et al.*, 2002). The modeled proteins are subjected to protein-protein docking by using GRAMM-X (Tovchigrechko A and Vakser AI, 2006) and Hex software (Macindoe *et al.*, 2010). The docked complexes are analyzed using pymol (Delano W L, 2002).

1.6. Deafness Genetics:

Deafness is a condition in which the ability to perceive certain frequencies of sound is completely or partially impaired. Deafness is genetically diverse disorder and can result from environmental as well as genetic factors. In Pakistan the ratio of deafness is 1.6 percent of 1000. It is estimated that 70% of deafness results from inter cousin marriages. The genetically determined deafness can be broadly categories into two types; syndromic and non-syndromic forms. The syndromic forms of deafness include several hundred deafness syndromes.

In non-syndromic genetic deafness, autosomal recessive type is most prevalent (80%), while autosomal dominant accounts (20%), X-linked (1%), and mitochondrial (<1%) forms have been reported. The autosomal recessive deafness is usually more severe than the other forms and is attributed to cochlear defects. More than 400 genetic hear loss syndromes have been reported (Peterson MB and Willems PJ, 2006).

Non-syndromic deafness is an example of genetic heterogeneity. It is estimated that more than 70% of hereditary hearing loss is of non-syndromic nature (Peterson MB and Willems PJ, 2006). The gene loci for non-syndromic deafness are called DFN. Loci for genes inherited in autosomal dominant forms are termed as DFNA. Genes inherited in

an autosomal recessive forms are known as DFNB and genes inherited in an X-linked forms are termed as DFN.

Till now the contributions of several other DFNB genes to recessive deafness in Pakistani population is found (Khan *et al.*, 2007). Mutations of RDX is found (0.3%). MYO6 (1.2%), TRIOBP (1.6%), OTOF (2.3%) MYO15A (3.3%), .TMC1 (3.4%), SLC26A4 (4.7%) and GJB2 (6.1%) each account for 0.3-6.1% or recessive deafness (DFNB1, DFNB4, DFNB7/11, DFNB3, DFNB9, DFNB37 and DFNB24) respectively in Pakistani population (Khan *et al.*, 2007). These results show the genetic heterogeneity and large genetic load of deafness that is still unaccounted for in Pakistani population.

Approximately 129 different gene loci linked with non-syndromic hearing loss have been found (Van Camp G and Smith RJH, 2009). Currently 57 gene loci are found to be linked with autosomal dominant mode of inheritance, 62 loci are associated with non syndromic autosomal recessive mode of inheritance, 7 loci are characterized to X-chromosome linked and 4 loci are characterized as mitochondrial. Over all 21 genes have been characterized for autosomal dominant (DFNA), 27 genes as autosomal recessive (DFNB), and 2 genes for X-linked (DFN) disorders (see the Hereditary Hearing Loss Homepage: http://www.dnalab-www.uia.ac.be /dnalab/hhh). These genes control diverse functions by encoding transcription factors, ion channels and extracellular matrix components. The complete list of autosomal recessive deafness is listed in table 1.1. Selected genes for analysis are listed in Table 1.2. Brief introduction of our selected genes is as follow.

i) RDX

RDX (OMIM #179410) is a cytoskeletal protein that may be important in anchoring actin to the plasma membrane. It has sequence similarity to both ezrin and moesin proteins (Safran *et al.*, 2010).

Another study showed that mutation of the RDX gene cause non syndromic hearing loss at the DFNB24 locus. The mutations were supposed to disturb the actin-binding domain of the gene (Khan *et al.*, 2007). One study suggest the role of Radixin in cell polarity and distribution of Resistance related protein 2 (Mrp-2) in liver hepatocytes cells (suda *et al.*, 2010).

Member proteins of ezrin-radixin-moesin family are constituents of the submembrane cortex, especially in epithelial cells. Radixin is thus important constituent of sensory receptors. It is assumed that it is involved in anchoring of actin protein to membrane (Pataky et al., 2004).

ii) TMPRSS3

TMPRSS3 (OMIM #605316) encodes a protein that belongs to the serine protease family. Serine proteases are said to be involved in a variety of processes, thus they are known to be involve in a no of diseases (Masmoudi *et al.*, 2001). This gene was identified by linkage analysis as one of the cause of childhood onset autosomal recessive deafness (Masmoudi *et al.*, 2001). This gene is expressed in cochlea of ear and many other tissues, and is known to be involved in inner ear structure formation. This gene was also identified as a tumor associated gene that is over expressed in ovarian tumors (Wattenhofer *et al.*, 2002).

iii) MYO6

MYO6 (OMIM #607821) gene encodes Myosins protein with ATPase activity. MYO6 function in a various intracellular processes such as cell membrane transport and cell migration (Safran *et al.*, 2010). They are integral part of structure of Golgi apparatus via the p53-dependent pro-survival pathway (Jung *et al.*, 2006).

MYO6 is also known to involve in clathrin-mediated endocytosis in epithelial cells and in the target specific transport of DAB2 (Buss *et al.*, 2001). MYO6 is also known to be important for development and maintanence of stereocilia (Hertzano et al., 2008). MYO6 mutation is involved with autosomal recessive deafness (Ahmed *et al.*, 2003).

iv) ESRRB

ESRRB (OMIM #608565) gene encodes a protein similar to the estrogen receptor. The information about function is unknown however a homologue protein in mouse plays role in placental development (Ansar *et al.*, 2003).Another study has described the ESRRB gene role in Autosomal recessive deafness (Collin *et al.*, 2008).

v) GIPC3

GIPC3 (OMIM # 601869) gene encodes PDZ domain-containing protein GIPC3 (Saitoh *et al.*, 2002). GIPC3 is a member of the GIPC gene family which also includes GIPC1 and GIPC2 gene. GIPC3 is known to be important for sound signal acquisition and are vital for hair cells of the cochlea. GIPC3 mutation is also associated with sudden hearing loss (Nikoletta *et al.*, 2011).

vi) LRTOMT/COMT2

LRTOMT/COMT2 (OMIM #612414) encodes two different proteins. One of them called leucine-rich transmembrane protein with unknown function while the other is an O-methyltransferase. Defects in the O-methyltransferase protein is attributed to nonsyndromic deafness (Vanwesemael *et al.*, 2011; Ahmed *et al.*. 2008; Du *et al.*, 2008).

vii) HGF

HGF (OMIM #608265)gene encode Hepatocyte growth factor which by binding to the proto-oncogenic c-Met receptor regulates cell growth and cell motility, by activating a tyrosine kinase signaling pathway giving it central role in angiogenesis, tumorogenesis, and tissue regeneration. The protein is member of plasminogen subfamily of S1 peptidases but has no reported protease activity. Different isoforms of this gene are known resulted from alternate splicing (Safran *et al.*, 2010).

SN	DFNB Names Of Genes	Chromosome Location	Common Names	SN	DFNB Names Of Genes	Chromosome Location	Common Names
1	DFNB1	13q12	GJB2	20	DFNB42	3q13.31-q22.3	ILDR1
2	DFNB2	11q13.5	MYO7A	21	DFNB44	7p14.1-q11.22	unknown
3	DFNB3	17p11.2	MYO15A	22	DFNB45	1q43-q44	unknown
4	DFNB4	7q31	SLC26A4	23	DFNB46	18p11.32	unknown
5	DFNB5	14q12	unknown	24	DFNB47	2p25.1-p24.3	unknown
6	DFNB6	3p14-p21	TMIE	25	DFNB48	15q23-q25.1	unknown
7	DFNB7/11	9q13-q21	TMC1	26	DFNB49	5q12.3-q14.1.	MARVELD2
8	DFNB8	21q22	TMPRSS3	27	DFNB51	11p13-p12	unknown
9	DFNB9	2p22-p23	OTOF	28	DFNB53	6p21.3	COL11A2
10	DFNB12	10q21-q22	CDH23	29	DFNB59	2q31.1-q31.3	PJVK
11	DFNB13	7q34-36	unknown	30	DFNB55	4q12-q13.2	unknown
12	DFNB14	7q31	unknown	31	DFNB45	7q34-36	unknown
13	DFNB15	3q21-q25 19p13	GIPC3	32	DFNB61	7q22.1	SLC26A5
14	DFNB16	15q21-q22	STRC	33	DFNB63	11q13.2-q13.4	COMT2
15	DFNB17	7q31	unknown	34	DFNB62	12p13.2	unknown
16	DFNB18	11p14-15.1	USH1C	35	DFNB66/67	6p21.2-22.3	LHFPL5
17	DFNB21	11q	TECTA	36	DFNB73	1p32.3	BSND
18	DFNB19	18p11	unknown	37	DFNB65	20q13.2	unknown
19	DFNB20	11q25-qter	unknown	38	DFNB68	19p13.2	unknown

SN	DFNB Names Of Genes	Chromosome Location	Common Names	SN	DFNB Names Of Genes	Chromosome Location	Common Names
39	DFNB22	16p12.2	OTOA	57	DFNB74	12q14.2-q15	MSRB3
40	DFNB23	10p11.2-q21	PCDH15	58	DFNB77	18q12-q21	LOXHD1
41	DFNB24	11q23	RDX	59	DFNB79	9q34.3	TPRN
42	DFNB25	4p13	GRXCR1	60	DFNB84	12q21.2	PTPRQ
43	DFNB26	4q31	unknown	61	DFNB71	8p22-21.3	unknown
44	DFNB27	2q23-q31	unknown	62	DFNB81	19p	unknown
44	DFNB28	22q13	TRIOBP	63	DFN91	6p25	SERPINB6
46	DFNB29	21q22	CLDN14	64	DFNB95	19p13	GIPC3
47	DFNB30	10p11.1	MYO3A	65	DFNB85 17p12	q11.2 unknown	DFNB85 17p12
48	DFNB31	9q32-q34	WHRN	66	DFNB93	11q12.3	unknown
49	DFNB32	1p13.3-22.1	GPSM2				
50	DFNB33	9q34.3	unknown				
51	DFNB35	14q24.1-24.3	ESRRB				
52	DFNB36	1p36.3	ESPN				
53	DFNB37	6q13	MYO6				
54	DFNB38	6q26-q27	unknown				
55	DFNB39	7q21.1	HGF				
56	DFNB40	22q	unknown				

Table 1.1: Genes implicated in Autosomal Recessive deafness

DFNB Names Of Genes	Chromosome Location	Common Names
DFNB24	11q23	RDX
DFNB63	11q13.2-q13.4	LRTOMT/COMT2
DFNB8/DFNB10	21q22	TMPRSS3
DFNB35	14q24.1-24.3	ESRRB
DFNB37	6q13	MYO6
DFNB95	19p13	GIPC3

Table 1.2: Non-Syndromic Autosomal Recessive Genes selected for Analysis.

Chapter 2

Material and Methods

2. MATERIALS and METHODS

The chapter describes the method followed during the research along with the introduction to tools and software used.

Protein Modeling of DFNB genes

In the present research we have selected seven genes which are implicated in autosomal non-syndromic deafness. The X-ray crystallographic structures of these genes are not known but the protein sequences are known and well characterized. The overall aim of present study is to find the structure of selected genes and also to predict their protein-protein docking interaction. Protein-protein interactions are vital to every cellular process. Almost every major cellular process such as DNA replication, transcription, translation, post translational modification, cell cycle control, signal transduction involve protein-protein interaction. Protein–protein interactions (PPI) knowledge is necessary to understand the complex metabolic interaction networks that occur in living organisms, with the ultimate target of designing drugs of therapeutic interest. We used literature review and String data base for finding interactions of our modeled proteins. Modeled proteins are docked with our complementary interacting proteins using GRAMM-X (Tovchigrechko A and Vakser AI, 2006) and Hex software (Macindoe *et al.*, 2010).The docked complexes are analyzed using pymol (Delano W L, 2002).

During the study we used various automated tools for structure prediction of our selected genes. The tools used include SWISS-MODEL (Schwede *et al.*, 2003), 3Djigsaw Model (Bates *et al.*, 2001) and SAM-T08 (K Karplus ,2009). We also used MODELLAR (Martí-Renom *et al.*, 2002 ; Sali A & T.L. Blundell., 1993), which is stand alone program for protein structure prediction. The Tools and Databases used during the study are listed in Table 2.1.

2.1. Primary Structure analysis through Protparam

ProtParam Tool computes various physico-chemical properties of a protein from a protein sequence. The query can be submitted either by using Swiss-Prot/TrEMBL accession number or by submitting raw protein sequence. ProtParam Tool does not

predict post-translational modification for query protein ProtParam calculations are based on sequence compositional data (Gasteiger *et al.*, 2005).

2.2. Secondary Structure analysis through Scratch protein predictor

SCRATCH is a web server for predicting protein secondary, tertiary structure and other structural features of proteins. The SCRATCH finds domains, disulfide bridges, single mutation stability, molecular weight and Theoretical pI for proteins (Cheng J *et al.*, 2005). The input of server is in the foam of amino acid sequence of query protein. User can optimize desired options, than submits to the server. Results are emailed to the user. The server is available at (**http://www.igb.uci.edu/servers/psss.html**).

2.3. Tertiary Structure Prediction

i) SWISS-MODEL

The SWISS-MODEL workspace incorporates expert knowledge into easy-to-use web-based modeling server. It enables the user in building protein homology models at different levels of complexity. It eliminates the need of downloading and installing large program packages and databases (Schwede *et al.*, 2003). The results of studied genes are given in Table 3.3.

ii) 3D Jigsaw

The 3D-JIGSAW is automatic comparative modeling server for predicting the structure of query protein sequence (Bates *et al.*, 2001). The results of studied genes are given in Table 3.4.

iii) SAM-T08

The SAM-T08 is web based server for protein structure prediction server that is based upon HMM. SAMT08 output includes three multiple sequence alignments of homologes using different iterated search procedures, local structure features of query and the E-values for the significant PDB templates and residue-residue contact predictions (K Karplus ,2009). The server accuracy has been tested in CASP8 assessment. The results of studied genes are given in Table 3.5.

iv) Modeler

Modeller is stand alone program used for comparative homology modeling of protein three-dimensional structures (Martí-Renom *et al.*, 2002) .Modeller implements comparative homology modeling of protein structure by satisfying spatial restraints. It can be used for additional tasks, including initial modeling of loops in protein structures, optimization of different models of protein structure by generating objective function, multiple alignment of protein sequences or structures etc(Martí-Renom *et al.*, 2002 ; Sali A & T.L. Blundell, 1993).

For comparative modeling through Modeller firstly genes sequences were retrieved from the NCBI protein sequence (http://www.ncbi.nlm.nih.gov/protein). To find the homologes structure to our selected genes blastp program to search against protein databank protein (pdb) database was used. The resulting Blast hits were evaluated on the basis of low e-value and highest score and percentage query coverage. The best hit was chosen for comparative modeling.

The sequence alignment of query sequence and selected template was performed by using Align2D module in Modeller. After aligning with the help of Align2D Script file, the query and template sequences were used as input in Modeller program and 20 models were generated for each gene. Modeller derives the restraints automatically from related known structures existing in the database. 3D structures were generated by optimization of molecular probable density function. The model with the highest objective score was finally chosen.

Modeled structure contains unfavorable bond lengths, bond angles, torsion angles and contacts. Therefore, it was essential to minimize the energy to optimize local bond lengths, but energy minimization process should not move coordinates away from the real structure. Therefore, the energy minimization step is kept to a minimum. The goal of energy minimization was to relieve steric collisions and strains without significant changes in the overall structure. We optimized each model was with simulated annealing (SA) method in Modeller itself.

Accuracy of the predicted models was subjected through a series of evaluation tests. The final model is selected which showed high likelihood of correct folding and good quality. The selected models were used for further study. The results of selected genes are given in table 3.3.4 and figures are shown in Figures 3.1- 3.7.

2.4 Evaluation

In structure prediction, the evaluation is an important step meant to unravel the strengths and weaknesses of models generated during the study and to improve its effectiveness. It is in fact a review checkpoint for the project. The final model is selected which showed high likely hood of correct folding and good quality. The selected models were used for further study. The programs used for evaluation are described in following pages.

2.4.1 ProSA

The ProSA program (Protein Structure Analysis) is widely used tool for refinement and validation of experimental protein structures and in structure prediction and modeling.

ProSA server provides a web interface that enables the user to highlight potential problems in protein structures by calculating scores and energy plots. In particular, the quality scores of a protein are calculated by using all known protein structures and erroneous parts of a structure are shown and highlighted in a 3D molecule viewer (Wiederstein *et al.*, 2007)

2.4.2 RAMPAGE

RAMPAGE is a program for visualizing and assessing the Ramachandran plot of a protein structure. On the basis of high-quality protein structures and a number of parameters (such as B-factor cut off and van der Walls clashes), phi/psi plots were derived for Gly, Pro, pre-Pro and general (other) residue types, and subdivided into "favored", "allowed" and "outlier" regions. Residues in the query PDB file that fall into the "allowed" and "outlier" regions are listed, and Ramachandran plot is displayed.

2.4.3 WHAT IF

WHAT IF server provides an interactive and flexible environment for analyzing small molecules, proteins, nucleic acids, and their interactions; A relational database for protein is integrated in the program. The menu-driven operation of WHAT IF, along with the use of default values makes it very easy to use for a new user while keeping full flexibility for more research studies (Vriend, 1990).

2.4.4 PROCHECK

The PROCHECK is a program for detailed checking of stereochemistry of a protein structure. The outputs comprise a number of plots and a comprehensive residue-by-residue listing. PROCHECK give evaluate the overall quality of the structure by comparing it with well refined structures in the database and it also suggest regions that may need further checking.

2.5 Mutational Analysis of Proteins

Through text mining we modeled the reported mutation of selected proteins. The mutated structure and normal is compared using pymol (Delano W L, 2002).

2.6 Protein-protein interaction analysis and docking

Protein-Protein interaction is important for in-depth understanding of disease at system level. We used text mining and String data base (Szklarczyk1 *et al.*, 2010) which is a global resource for the finding and analysis of protein-proteins interaction to find the interacting proteins. The predicted interacting proteins are docked using GRAMM-X GRAMM v1.03 (Tovchigrechko A and Vakser AI, 2006) and Hex software (Macindoe *et al.*, 2010).

GRAMM-X use Fast Fourier Transform for the global search of the best rigid body docking conformations for protein molecules. To find the structure of a docked complex, it needs the atomic coordinates of the two molecules. The software performs an extensive 6-dimensional search through the translations and rotations of the protein molecules. The technique employed by GRAMM-X finds the area of the global minimum of

intermolecular energy for structures used for docking. The quality of the docking prediction depends on the accuracy of the structures used.

Name	URL Address
Blast	http://blast.ncbi.nlm.nih.gov/Blast.cgi
Protparam	www.expasy.ch/tools/protparam.html
ProSA	https://ProSA.services.came.sbg.ac.at/ProSA.php
RAMPAGE	http://mordred.bioc.cam.ac.uk/~rapper/rampage.php
SAM-T08	http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html
SCRATCH	http://scratch.proteomics.ics.uci.edu/
SWISS-MODEL	http://swissmodel.expasy.org/SWISS-MODEL.html
STRING	http://string-db.org/
Topmatch	http://topmatch.services.came.sbg.ac.at/
3djigsaw	http://bmm.cancerresearchuk.org/~3djigsaw/
WHAT IF	http://swift.cmbi.kun.nl/WHAT IF/

Table 2.1: Tools and databases used during present study

Chapter 3

Result and Evaluation

3. RESULTS AND EVALUATION

The chapter describes the results obtained and their evaluation done through various tools.

3.1 Primary Structure Analysis ProtParam Results

ProtParam is a tool, which allows the computation of various physical and chemical parameters for proteins. The computed parameters include the molecular weight, theoretical pI, amino acid composition (Gasteiger *et al.*, 2005). **Table 3.1** shows the calculated primary structure information of selected genes through Protparam.

Gene	No. of Amino Acids	Molecular weight	PI value	Formula	Total Atoms
RDX	583	68563.9	6.03	$C_{3026}H_{4846}N_{856}O_{927}S_{16}$	9671
ESRRB	500	55619.3	8.51	$C_{2479}H_{3937}N_{681}O_{715}S_{28}$	7840
TMPRSS3	344	37497.2	6.11	$C_{1691}H_{2630}N_{446}O_{482}S_{18}$	5267
COMT2	291	32154.8	9.30	$C_{1691}H_{2630}N_{446}O_{482}S_{18}$	4596
MYO6	1285	148713.9	8.75	$C_{6564}H_{10456}N_{1868}O_{1959}S_{57}$	20904
GIPC3	312	33981.6	5.50	$C_{1488}H_{2389}N_{423}O_{462}S_{12}$	4774

Table 3.1: Summary of primary structure information of studied genes

3.2. Secondary Structure Analysis through Scratch protein predictor

For the prediction of secondary structures Scratch protein predictor is used. We perform secondary structure analysis using it. Calculated Secondary structure information of studied genes in **Table 3.2.**

Gene Name	Domains Predicted	Domains Locations	Cysteines	Disulfide bond position
RDX	3	1 - 99 100 - 477 478 - 583	2	117,284
ESRRB	3	1 - 129 130 - 400 401 - 500	14	72,79,91,103,106,120,123,139,145,155
TMPRSS3	2	1 – 171 172 - 344	14	73,79,85,92,98,107,129,142,194,207,242,258
COMT2	2	1 - 63 64 - 291	9	21,72,134,142,260,265
MYO6	4	1 – 668,669 - 833 834 - 1127 1128 – 1285	24	236,252,278,321,362,375,442,472,587,610,691,735,790,799,817,829,1093,1101,1227,1256
GIPC3	3	1 - 99 100 - 477 478 - 583	2	117,284

Table 3.2: Calculated Secondary structure information of selected genes

3.3. Tertiary Structures Prediction

Tertiary structure is predicted by using following tools:

- SWISS-MODEL
- Modeller
- 3Djigsaw Model
- SAM-T08

3.3.1 SWISS-MODEL

SWISS-MODEL is an automated protein structure homology-modeling server, accessible through ExPASy Tools. The purpose of this server is to make Protein Modeling easier for bio-informaticiens and biochemists (Schwede *et al.*, 2003). The results of studied genes are given in **Table 3.3.**

Name of gene	Model residue range	No of residues modeled	Template	Sequence identity
RDX	3-297	294	2zpyA	100%
ESRRB	106-195	89	1lo1A	99%
TMPRSS3	75-105	30	1n7dA	38%
COMT2	83-291	208	2gpyB	22%
MYO6	2-825	823	2bkiA	98%
GIPC3	108-196	88	3ggeB	61%
HGF	35-208	173	2qj2B	100%

Table 3.3: Summary of model structures obtained through SWISS-MODELS

3.3.2 3D-JIGSAW The 3D-JIGSAW is automatic comparative modeling server for predicting the structure of query protein sequence (Bates *et al.*, 2001). The results of studied genes are given in **Table 3.4.**

Gene	Model residue range	No of residues modeled	Template	Sequence identity
RDX	1-350	350	2zpy_A	100.00
ESRRB	281-624	343	3dzy_A	34.88
TMPRSS3	132-396	264	1z8g_A	29.44
COMT2	63-266	203aa	1bht_A	100.00
MYO6	28-879	851	2bki_A	98.18
GIPC3	142-257	115	3gge_B	55.68
HGF	63-266	203	1bht_A	85.63

Table 3.4: Summary of Results obtained through 3D-JIGSAW Modeling server

SAM-T08

The SAM-T08 is web based server for protein structure prediction server that is based upon HMM. SAM-T08 output includes three multiple sequence alignments of homologs using different iterated search procedures, local structure features of query and the E-values for the significant PDB templates and residue-residue contact predictions (Kevin Karplus ,2009). The server accuracy has been tested in CASP8 assessment. The results of studied genes are given in **Table 3.5.**

Name of gene	Model residue range	No of residues modeled
RDX	1-583	583
ESRRB	1-500	500
TMPRSS3	1-344	344
GIPC3	1-312	312

Table 3.5: Summary of Results obtained through SAM-T08 Modeling server

MODELLER

MODELLER is stand alone program used for comparative homology modeling of protein three-dimensional structures. MODELLER implements comparative homology modeling of protein structure by satisfying spatial restraints. It can be used for additional tasks, including initial modeling of loops in protein structures, optimization of different models of protein structure by generating objective function, multiple alignment of protein sequences or structures etc (Martí-Renom *et al.*, 2002; Sali A & T.L. Blundell, 1993, Fiser A & Sali A., 2003).**Table 3.13** shows the results of studied genes and **figure 3.1 to figure 3.7** show the modeled structure of studied genes.

Accession No.	Protein Name	Query Coverage	Score	E-Value
2I1J_A	Moesin From Spodoptera Frugiperda Reveals The Coiled-Coil Domain	100%	681	0.0
2EMT_A	Chain A, Crystal Structure Analysis Of The Radixin Ferm Domain Complexed With Adhesion Molecule Psgl-1	53%	661	0.0
1J19_A	Chain A, Crystal Structure Of The Radxin Ferm Domain Complexed With The Icam-2 Cytoplasmic Peptide	53%	661	0.0
2D2Q_A	Crystal Structure Of The Dimerized Radixin Ferm Domain	53%	660	0.0
2D10_A	Chain A, Crystal Structure Of The Radixin Ferm Domain Complexed With The Nherf-1 C-Terminal Tail Peptide	53%	660	0.0
2ZPY_A	Chain A, Crystal Structure Of The Mouse Radxin Ferm Domain Complexed With The Mouse Cd44 Cytoplasmic Peptide	53%	660	0.0

Table 3.6: Blast search results for RDX gene

Accession No.	Protein Name	Query Coverage	Score	E-Value
1JR4_A	Chain A, Catechol O-Methyltransferase Bisubstrate-Inhibitor Complex	72%	167	4e-42
2ZLB_A	Chain A, Crystal Structure Of Apo Form Of Rat Catechol-O-Methyltransferase	72%	167	5e-42
3OE4_A	Chain A, Rat Catechol O-Methyltransferase In Complex With A Bisubstrate Inhibitor - Humanized Form.	72%	167	7e-42
3BWY	Chain A, Crystal Structure Of Human 108m Catechol O-Methyltransferase Bound With S-Adenosylmethionine And Inhibitor Dinitrocatechol.	72%	162	1e-40
3A7E_A	Structure related to 3A7E_A Chain A, Crystal Structure Of Human Comt Complexed With Sam And 3,5- Dinitrocatechol	53%	159	1e-39

Table 3.7: Blast search results for COMT2 gene.

Accession No.	Protein Name	Query Coverage	Score	E-Value
2E2R_A	Chain A, Crystal Structure Of Human Estrogen-Related Receptor Gamma Ligand Binding Domain Complex With Bisphenol A	48%	384	5e-107
1KV6_A	Chain A, Crystal Structure Of Apo Form Of Rat Catechol-O-Methyltransferase	46%	382	3e-106
1VJB_A	Chain A, Rat Catechol O-Methyltransferase In Complex With A Bisubstrate Inhibitor - Humanized Form.	46%	381	4e-106
1S9Q_A	Chain A, Crystal Structure Of Human 108m Catechol O-Methyltransferase Bound With S-Adenosylmethionine.	45%	379	1e-105
1S9P_A	Structure related to 3BWM_A Chain A, Crystal Structure Of Human Catechol O-Methyltransferase With Bound Sam And Dnc	45%	376	1e-104
2EWP_A	Structure related to 3A7E_A Chain A, Crystal Structure Of Human Comt Complexed With Sam And 3,5-Dinitrocatechol	45%	374	5e-104

Table 3.8: Blast search results for ESRRB gene.

Accession No.	Protein Name	Query coverage	Score	E-value
2BKI_A	Chain A, Myosin Vi Nucleotide-Free (Mdinsert2-Iq) Crystal Structure	66%	1779	0.0
2BKH_A	Chain A, Myosin Vi Nucleotide-Free (Mdinsert2) Crystal Structure	63%	1686	0.0
3L9I_A	Chain A, Myosin Vi Nucleotide-Free (Mdinsert2) L310g Mutant Crystal	63%	1682	0.0
2V26_A	Chain A, Myosin Vi (Md) Pre-Powerstroke State (Mg.Adp.Vo4)Adenosylmethionine.	61%	1624	0.0
2X51_A	Chain A, M6 Delta Insert1	63%	1620	0.0
2VAS_A	Chain A, Myosin Vi (Md-Insert2-Cam, Delta-Insert1) Post-Rigor State.	63%	1617	0.0

Table 3.9: Blast search results for My06 gene.

Accession No	Protein Name	Query Coverage	Score	E-Value
3GGE_A	Crystal Structure Of The Pdz Domain Of Pdz Domain-Containing Protein Gipc2	29%	110	8e-25
1KWA_A	Chain A, Human CaskLIN-2 Pdz Domain >pdb|1KWA|B Chain B, Human CaskLIN-2 Pdz Domain	25%	43.9	1e-04
1WF8_A	Chain A, Solution Structure Of The Pdz Domain Of SpinophilinNEURABINII PROTEIN	25%	37.4	0.010
1UEW_A	Chain A, Solution Structure Of The Forth Pdz Domain Of Human Atrophin-1 Interacting Protein.	17%	36.2	0.022
1V62_A	Chain A, Solution Structure Of The 3rd Pdz Domain Of Grip2	21%	35.8	0.025
1UJV_A	Chain A, Solution Structure Of The Second Pdz Domain Of Human Membrane Associated Guanylate Kinase Inverted-2	24%	35.8	0.028

Table 3.10: Blast search results for TMPRSS3 gene.

Accession No	Protein Name	Query Coverage	Score	E-Value
1Z8G_A	Chain A, Crystal Structure Of The Extracellular Region Of The Transmembrane Serine Protease Hepsin	64%	107	7e-24
1KWA_A	Chain A, Human CaskLIN-2 Pdz Domain >pdb|1KWA|B Chain B, Human CaskLIN-2 Pdz Domain	25%	43.9	1e-04
2ANY_A	Chain A, Expression, Crystallization And Three-Dimensional Structure Of The Catalytic Domain Of Human Plasma	29%	83.2	2e-16
2OQ5_A	Chain A, Crystal Structure Of Desc1, A New Member Of The Type Ii Transmembrane Serine Proteinases	36%	80.1	1e-15
1XX9_A	Crystal Structure Of The Fxia Catalytic Domain In Complex With Ecotinm	31%	78.231	5e-15
1WF8_A	Chain A, Solution Structure Of The Pdz Domain Of SpinophilinNEURABINII PROTEIN	25%	37.4	0.010

Table 3.11: Blast search results for TMPRSS3 gene.

Accession No	Protein Name	Query Coverage	Score	E-Value
3HN4_A	Chain A, Crystal Structure Of The Nk2 Fragment (28-289) Of Human Hepatocyte Growth Factor	60%	896	6e-151
1SHY_A	Chain A, The Crystal Structure Of Hgf Beta-Chain In Complex With The Sema Domain Of The Met Receptor.	32%	480	1e-135
1S15_H	Chain H, Protease-Like Domain From 2-Chain Hepatocyte Growth Factor	32%	480	1e-135
2QJ2_A	Chain A, A Mechanistic Basis For Converting A Receptor Tyrosine Kinase Agonist To An Antagonist	57%	606	1e-100
1NK1_A	Chain A, Nk1 Fragment Of Human Hepatocyte Growth Factor scatter Factor.	57%	606	2e-100
3MKP_A	Chain A, Crystal Structure Of 1k1 Mutant Of Hepatocyte Growth Factor SCATTER Factor Fragment Nk1 In Complex With Heparin.	57%	608	4e-100

Table 3.12: Blast search results for HGF gene.

Gene	Model Residue Range	No. of Residues Modeled	Template	Sequence Identity
RDX	1-583	583	2I1J_A	100%
ESRRB	1-500	500	2E2R	68%
TMPRSS3	1-344	344	1Z8G	64%
COMT2	1-291	291	2BWY	38.318
MYO6	1-1285	1285	2DFS	32.828%
GIPC3	1-312	312	1WF8	41.121%
HGF	1-723	723	3HN4	40%

Table 3.13: Results of protein structures using Modeller Program

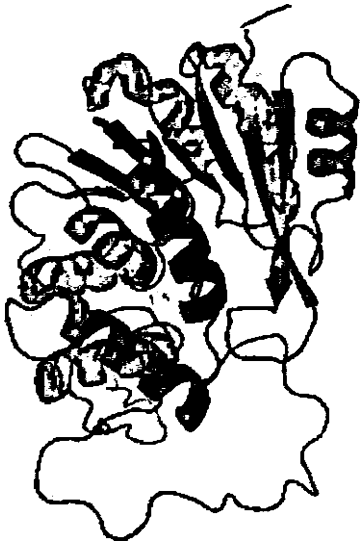


Fig 3.1: Modeller model of COMT2

Fig 3.2: Modeller model of ESRRB

Fig 3.3: Modeller model of TMPRSS3

Fig 3.4: Modeller model of COMT2

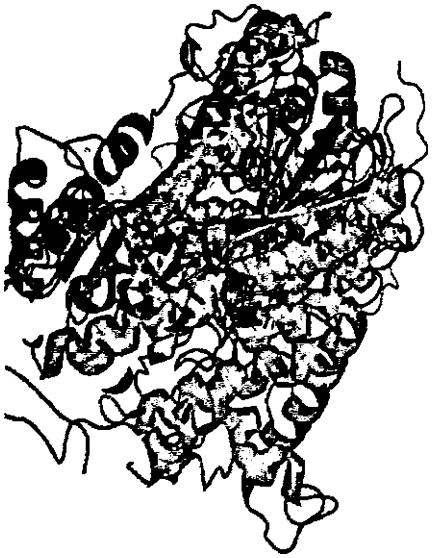


Fig 3.5: Modeller model of MY06

Fig 3.6: Modeller model of GIPC3

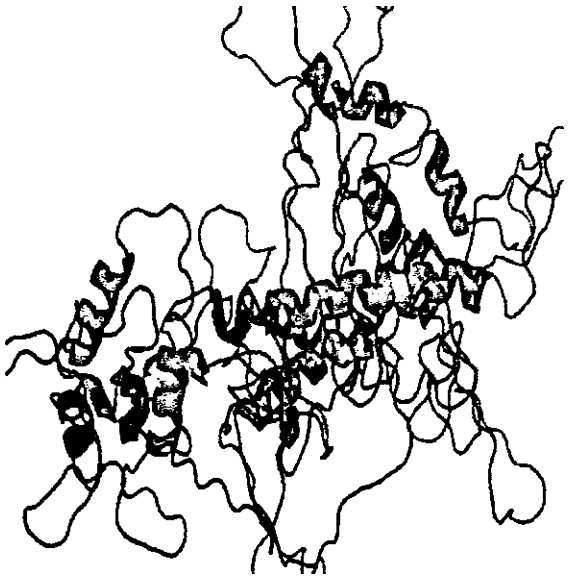


Fig 3.7: Modeller model of HGF

3.4. Protein-Protein Interaction

Protein-Protien interaction is important in-depth understanding of disease at system level. We used Text mining and String data base for protein-protein interaction analysis (Szklarczyk *et al.*, 2010). **figure 3.8 to figure 3.14** show interaction results obtained using String data base.

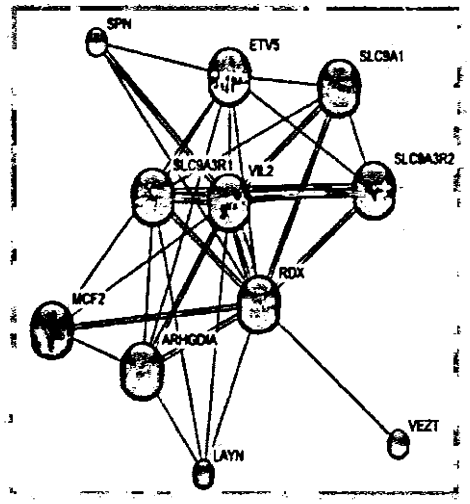


Fig 3.8: RDX Interaction network

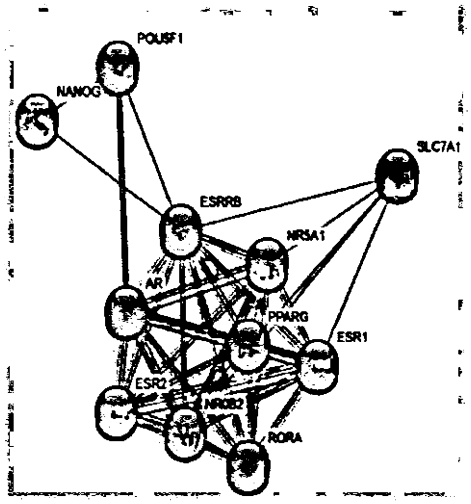


Fig 3.9: ESRRB Interaction network

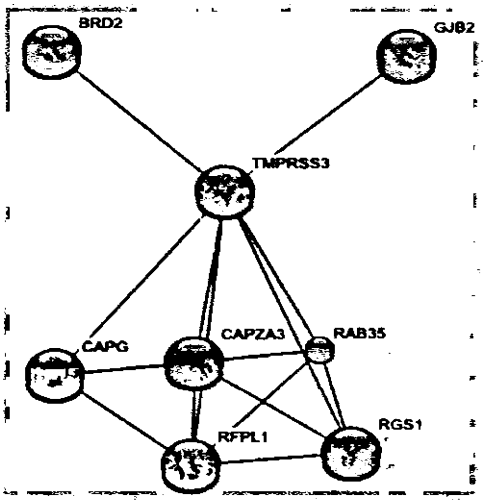


Fig 3.10: TMPRSS3 Interaction network

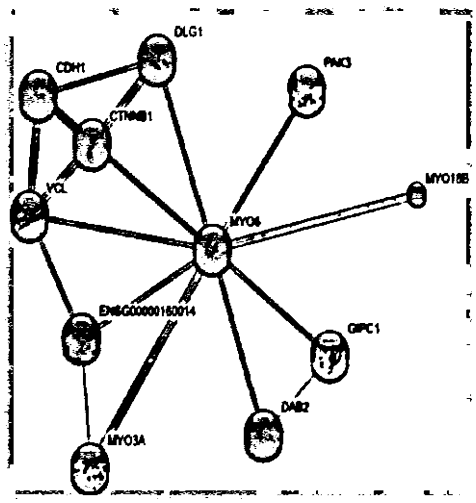


Fig 3.11: MY06 Interaction network

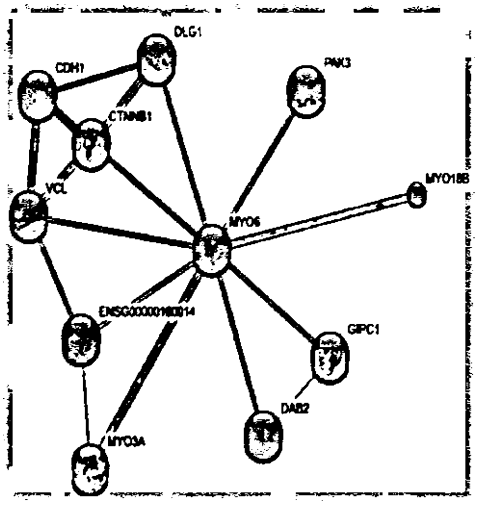


Fig 3.12: GIPC3 Interaction network

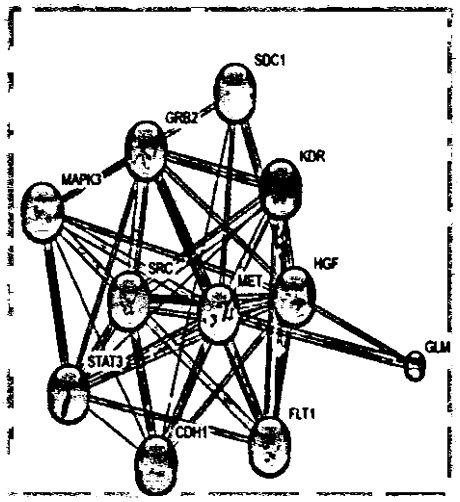


Fig 3.13: HGF Interaction network

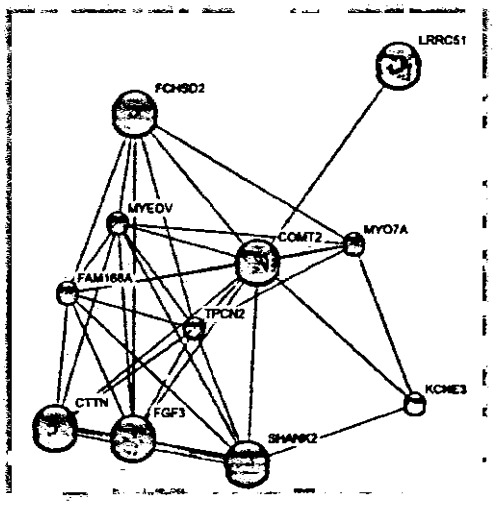


Fig 3.14: COMT2 Interaction network

3.5. Gramm-X Protein-Protein Docking

The docking of selected proteins is performed by using Gramm-X (GRAMM v1.03) **figure 3.15 to figure 3.20** shows docking results of Gramm-X. **Table 3.14** shows hydrogen bonding residues of docked protein complex.

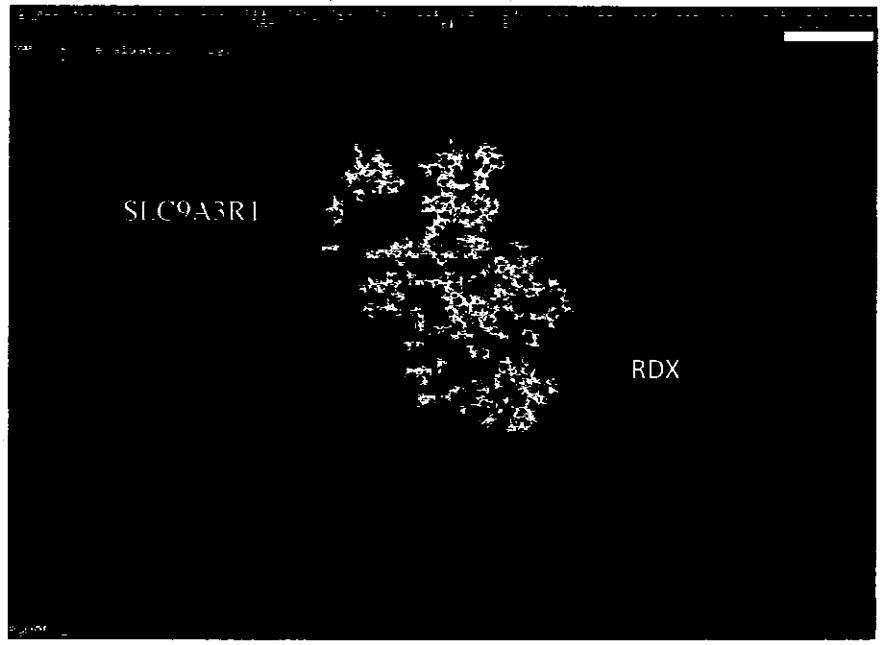


Fig 3.15: RDX docked complex showing hydrogen Bonds interaction

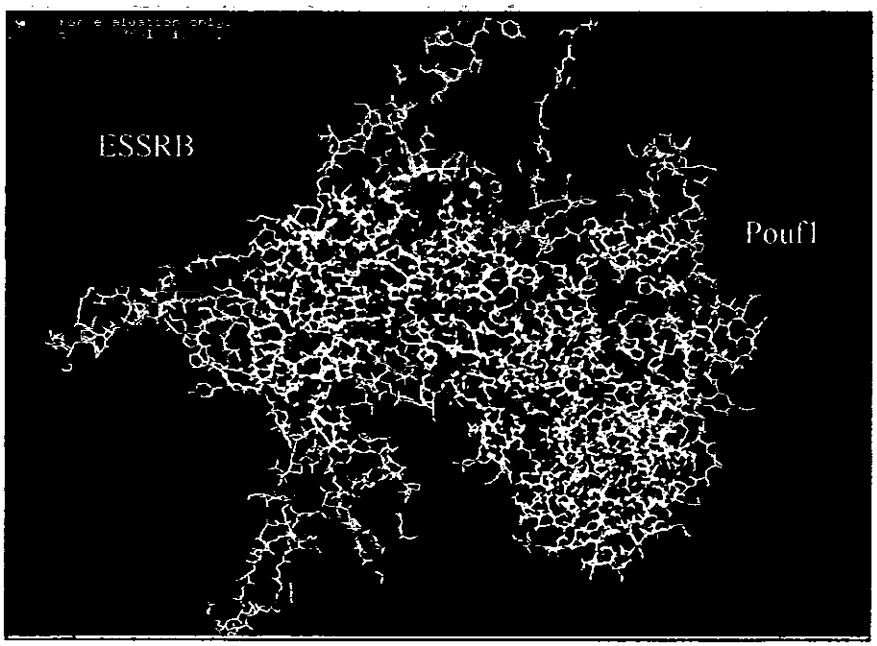


Fig 3.16: ESRRB docked complex showing hydrogen Bonds interaction

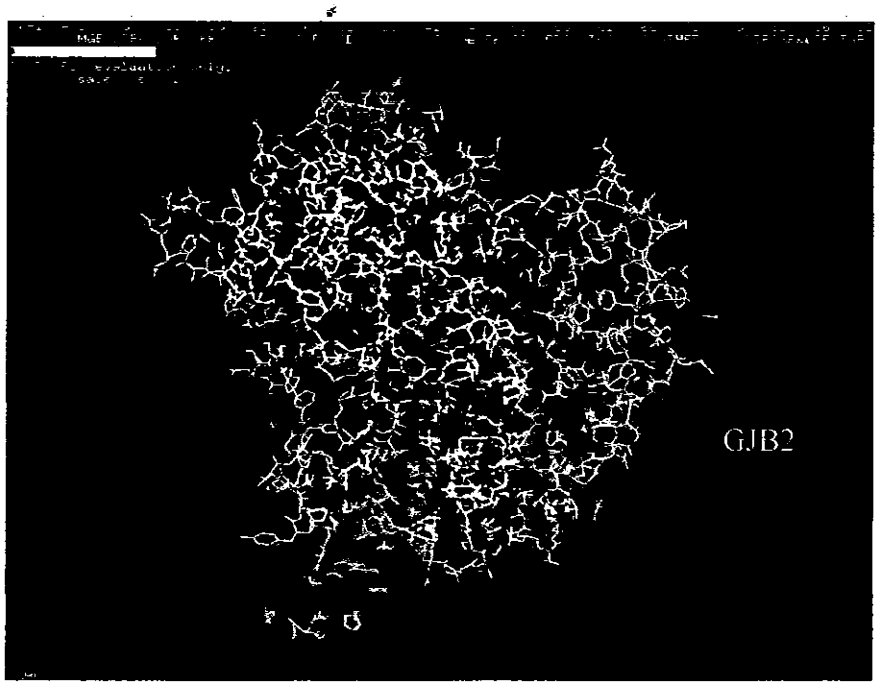


Fig 3.17: TMPRSS3 docked complex showing hydrogen Bonds interaction

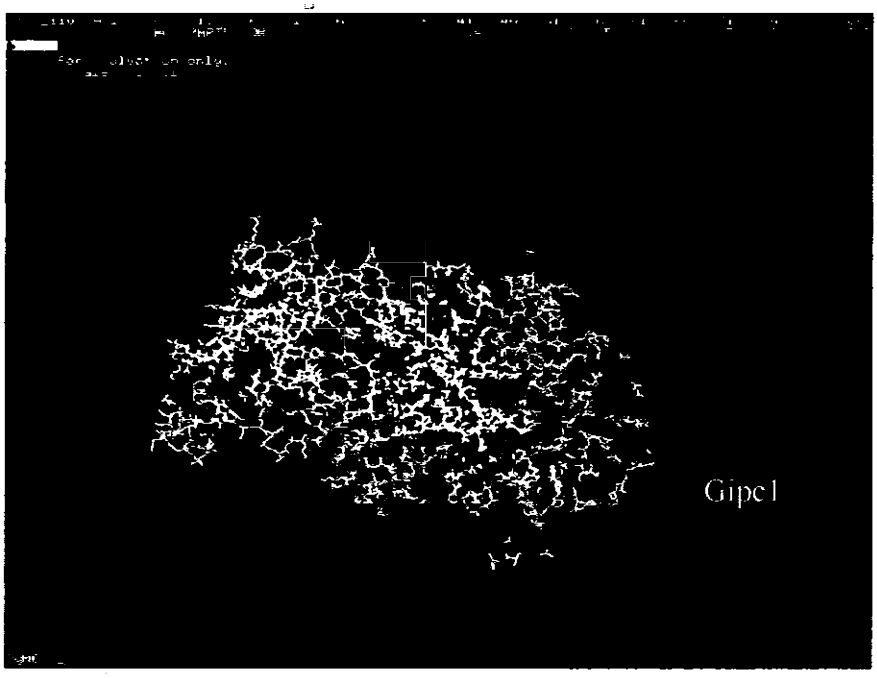


Fig 3.18: MY06 docked complex showing hydrogen Bonds interaction

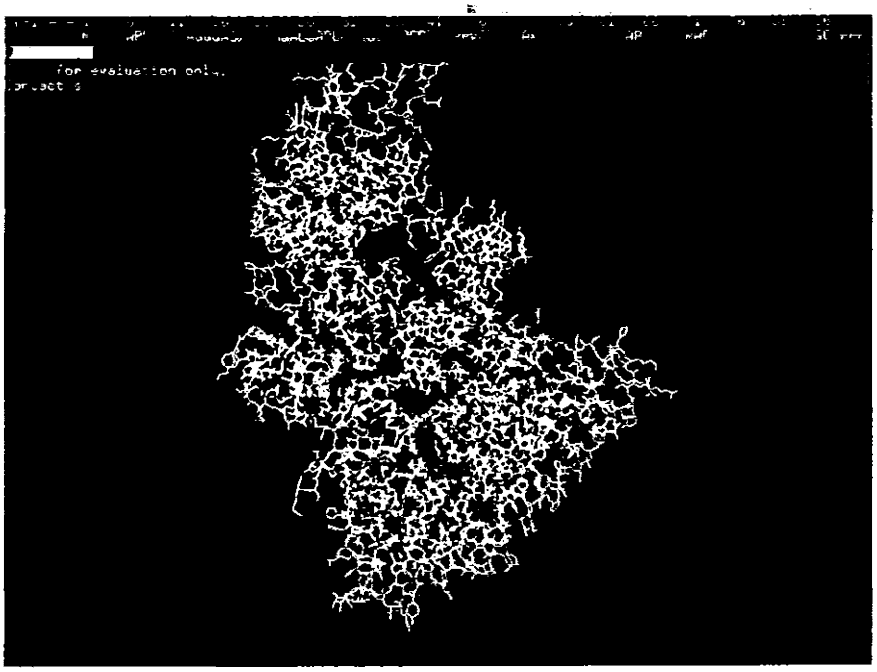


Fig 3.19: GIPC3 docked complex showing hydrogen Bonds interaction

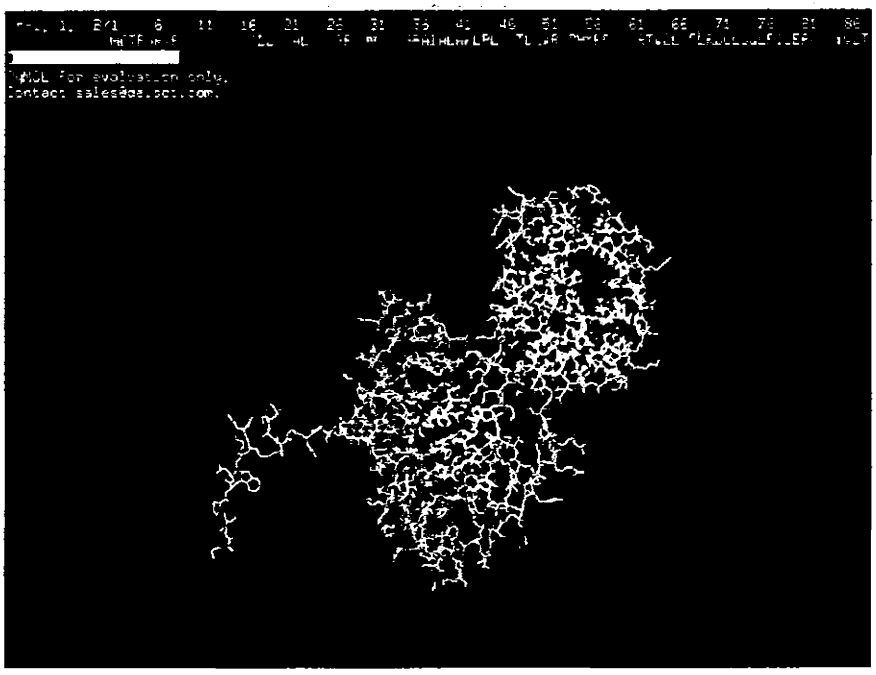


Fig 3.20: COMT2 docked complex showing hydrogen Bonds interaction

Modeled Protein Name	Interacting Protein Name	Modeled Protein Residues (Bond Distance) ↓ Interacting Protein Residues.	
RDX	SLC9A3R1	GLU`166(3.01)→ GLU`263 TYR`146 (3.39)→ASN167 HIS`161(2.47)→ LYS`162 ARG`184(3.42)→ PRO`184	GLU`159(3.44)→ GLN` 300 TYR`146 (3.23)→ SER`181 LYS`133(3.60)→ PRO`311 GLN`173(2.55) → ASN`261 ARG`370(3.21)→ ASP`185
ESRRB	Poufl	GLU`339(2.6)→ SER`55 SER`111(3.4)→ HIS`329	LEU`104(2.7)→ PRO`27 TYR`113(2.5)→ ARG`295
TMPRSS3	GJB2	VAL`108(2.4)→ HIS`67 VAL`9(3.1)→ ARG`75	TYR`296(2.7) → ASP`66 VAL`291(3.5)→ ARG`75
Comt2	LRRC51	ASP`85(3.4)→ ARG`67 ASN`62(2.1)→ SER`20	TYR`108(3.2)→ LEU`66 THR`191(2.7)→ VAL`64
My06	Gipc1	GLN`768(3.52)→ LEU`64 ASN`1165(3.18)→ ALA`53	ILE`24(3.52)→ VAL`330 THR`1135(3.36)→ GLY`28

Table 3.14: Docking Results of modeled proteins

3.6: Hex Protein-Protein Docking

Hex software (Macindoe G *et al.*, 2010) uses hashing (Bachar O *et al.*, 1993) or fast Fourier transform (FFT) correlation techniques (Katchalski-Katzir *et al.*, 1992) to find a relatively small number of putative docking complexes which is further refined and re-scored using more sophisticated techniques. **Table 3.15** shows Hex Docking results of modeled proteins.

Name of gene	E-total	E-shape
RDX	578.64	578.64
ESRRB	-615.00	-615.00
TMPRSS3	-787.04	-787.04
COMT2	-624.25	-624.25
MYO6	0.00	0.00
GIPC3	-606	-606

Table 3.15: Hex Docking results of modeled proteins

3.7: Mutational Analysis of Modeled proteins

The present study includes the modeling of mutated structures by using reported mutation of our selected genes. The normal and mutated structures are analysed using pymol (Delano W L, 2002) to see the structural effects due to mutations. **Figure 3.21 to Figure 3.29** show the mutated models of modeled proteins.

3.7.1: Mutated Models of our studied proteins

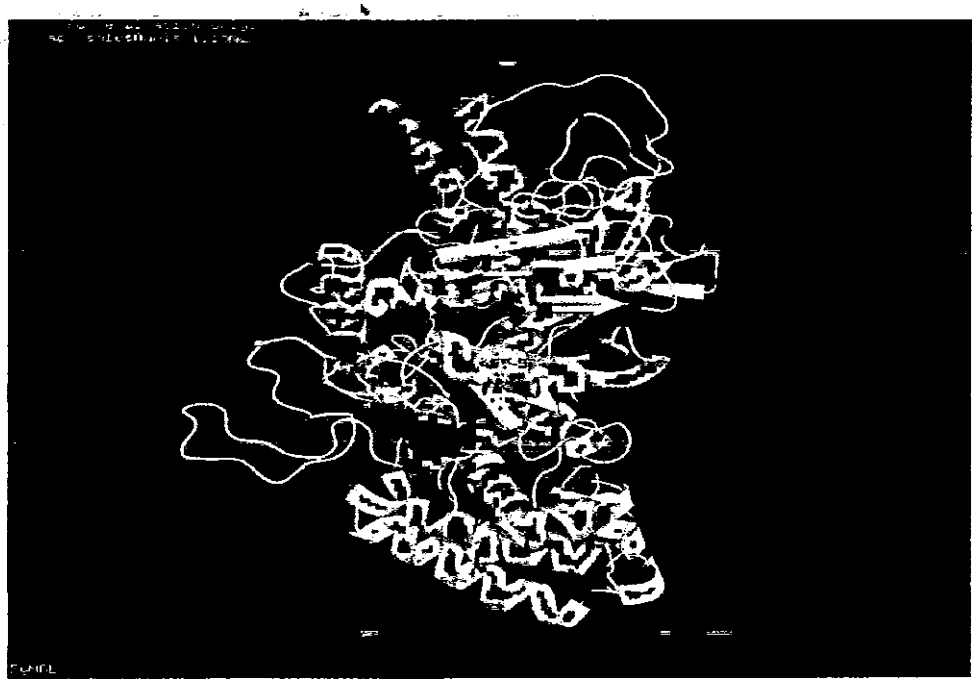


Fig 3.21: Normal structure of RDX protein with Aspartic acid at 578 position.

Fig 3.22: RDX mutated model showing mutation D578N.

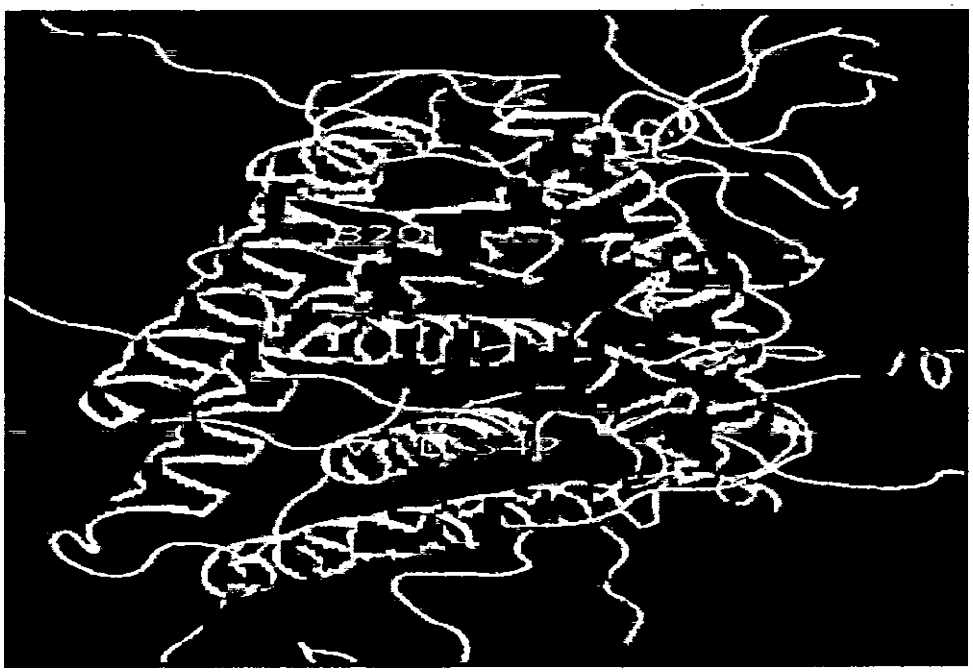


Fig 3.23: Normal structure of ESRRB protein with Leucine and Valine at positions 320 and 342 respectively

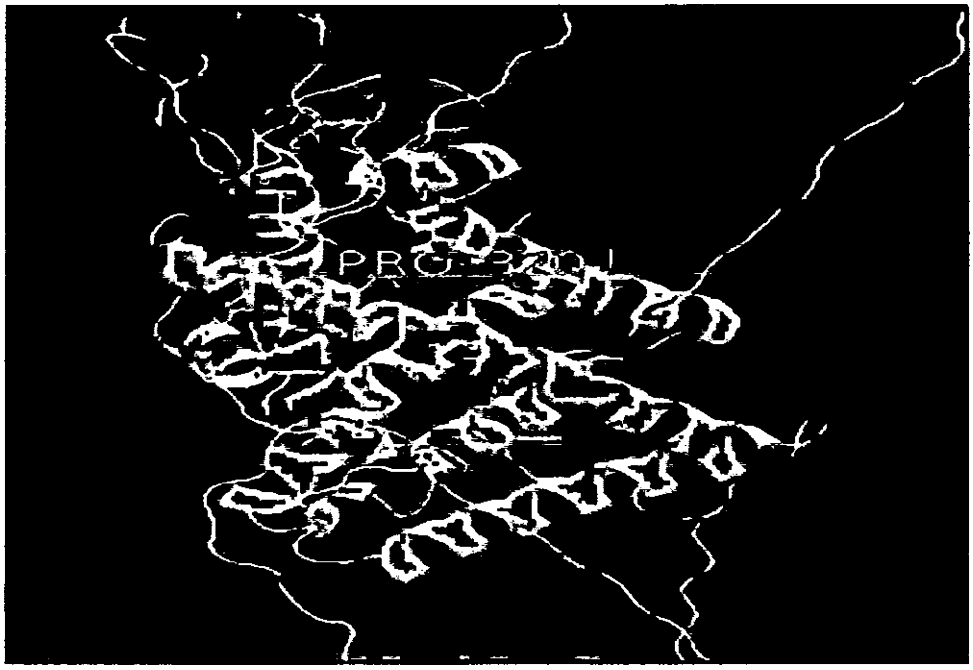


Fig 3.24: ESRRB mutated model showing mutation L320P.

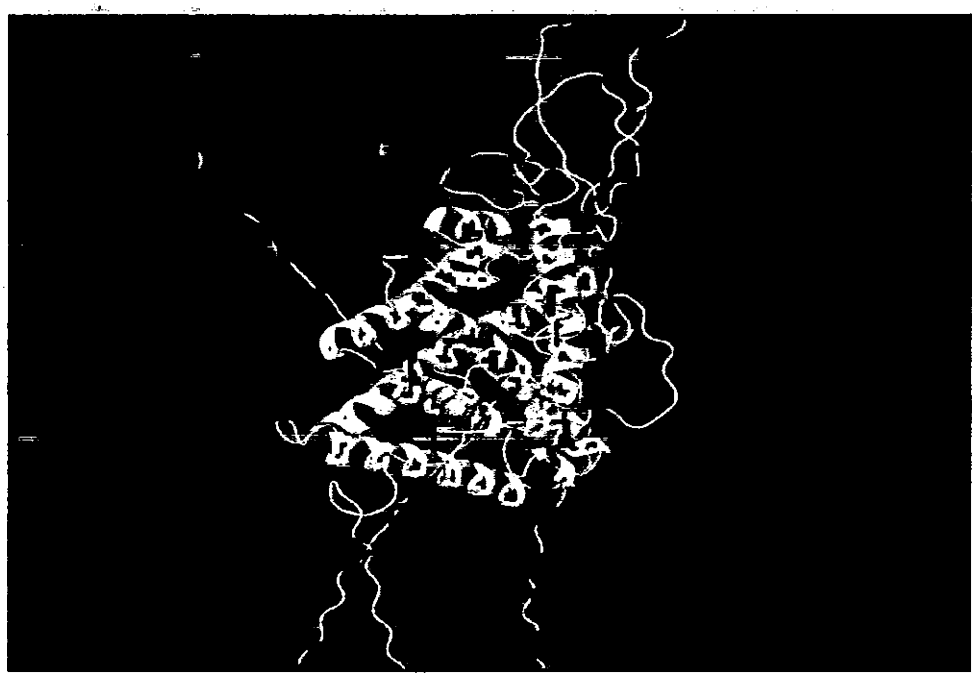


Fig 3.25: ESRRB mutated model showing mutation V342L.

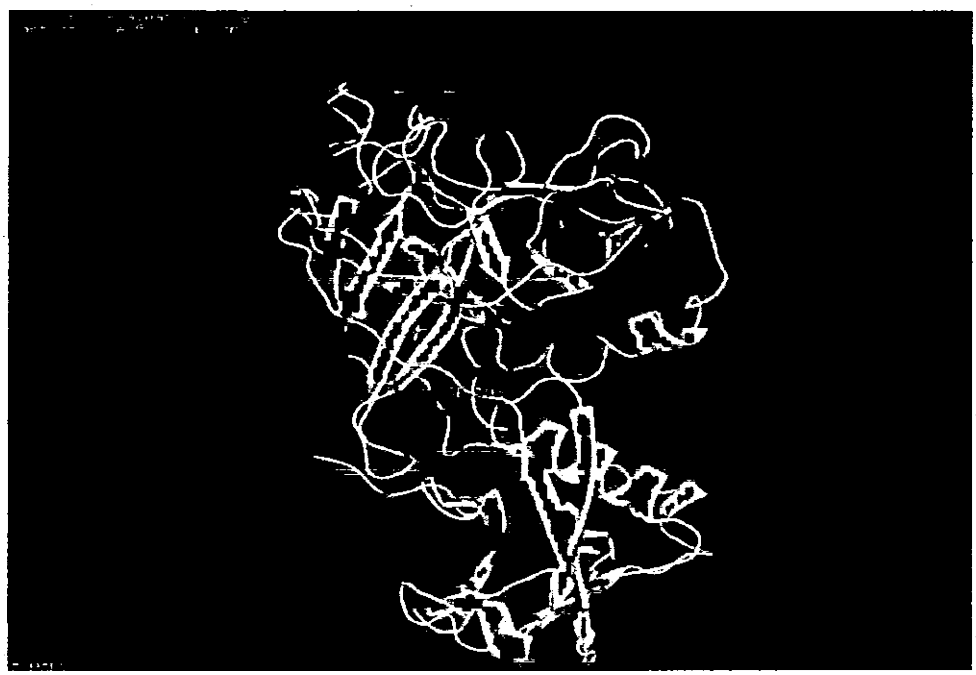


Fig 3.26: Normal TMPRSS3 protein with Aspartic acid at 103 positions.

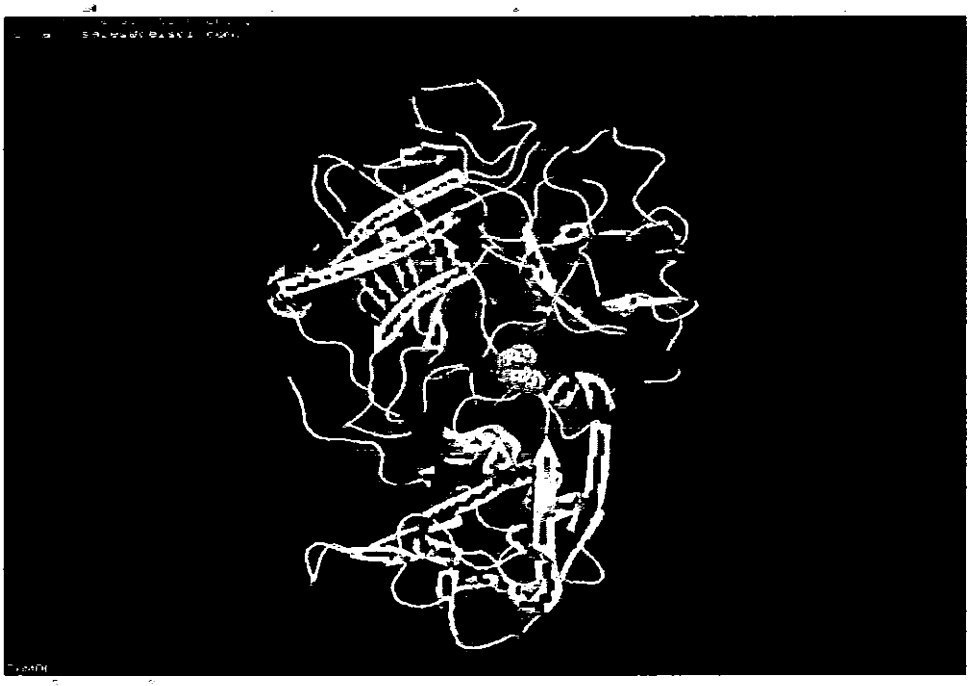


Fig 3.27: TMPRSS3 mutated model showing mutation D103G.

Fig 3.28: Normal COMT2 protein with Leucine at 16 positions.

Fig 3.29: COMT2 protein mutated model showing mutation L16P

3.8. Model Evaluation

The evaluation of modeled proteins structures is an important step of the comparative homology modeling process for two reasons. Firstly, the selection of best quality structures is important as these coordinate will be used to build more reliable models i.e x-ray and NMR. Secondly, the evaluated structure can be used to identify possible problematic regions in final protein structure.

Accuracy of the predicted models was subjected through a series of tests. Stereochemical properties of modeled proteins were evaluated through Procheck server. Proteins Backbone conformation was checked by generating PSi/Phi Ramachandran plot using Procheck and RAMPAGE Packing. Coarse Packing Quality of modeled ,proteins ,is calculated by WHAT IF packing quality control.

3.8.1. Evaluation result of modeled proteins By ProSA:

The ProSA program is an extensive used tool employed in the refinement and validation of experimental protein structures and in structure prediction (Wiederstein M *et al.*, 2007). We used ProSA for evaluation of our modeled genes. **Figure 3.30- figure 3.35** shows evaluation results of our modeled genes.

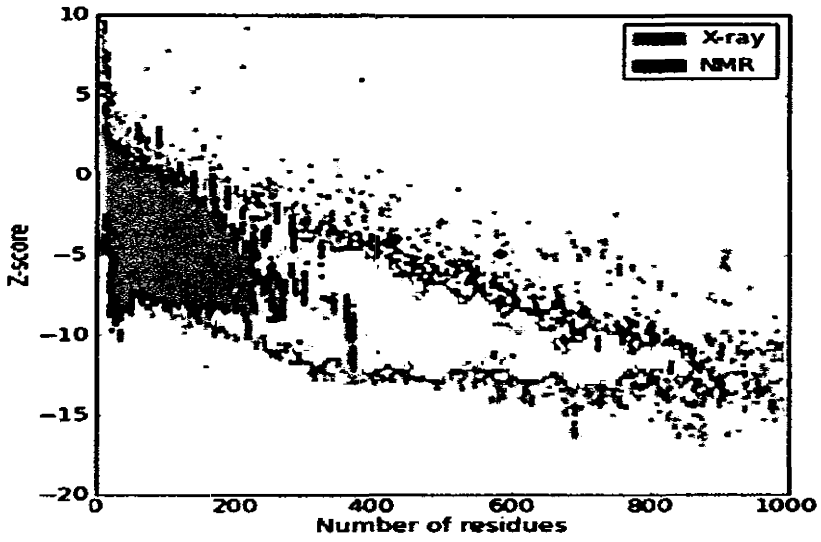


Fig 3.30: RDX Gene Z-score plot

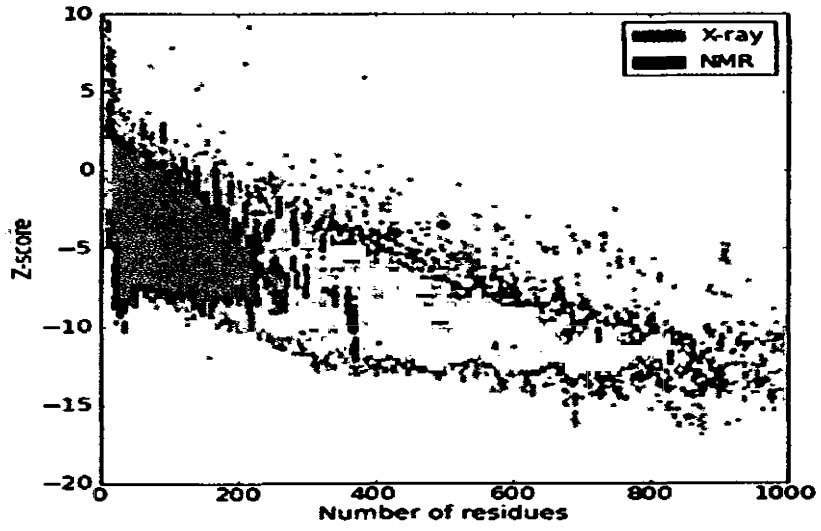


Fig 3.31: ESRRB Gene Z-score plot

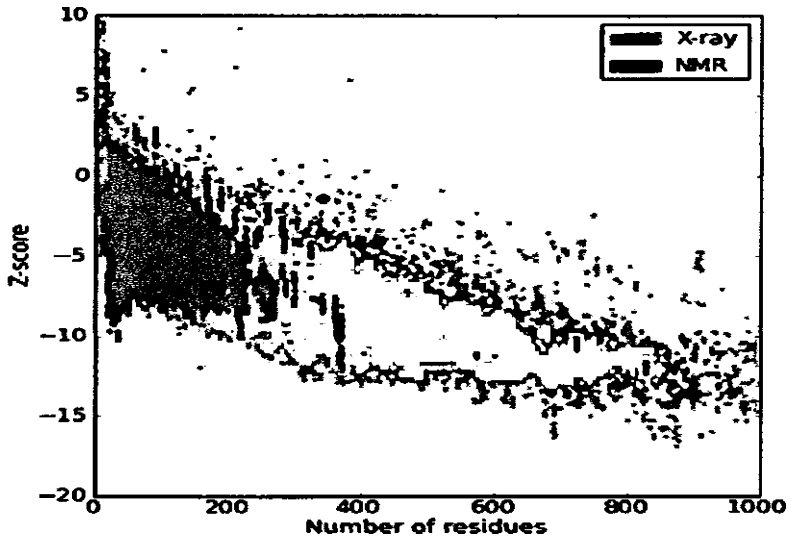


Fig 3.32: TMPRSS3 Gene Z-score plot

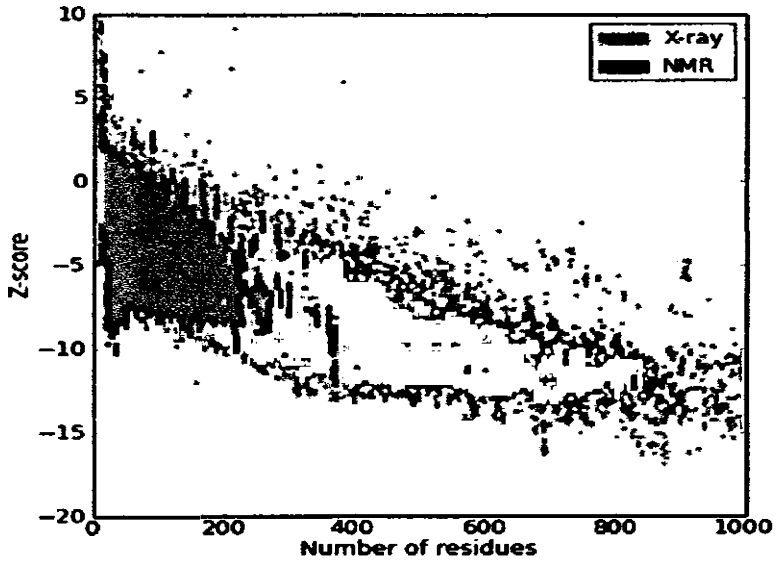


Fig 3.33: Comt2 gene Z-score plot

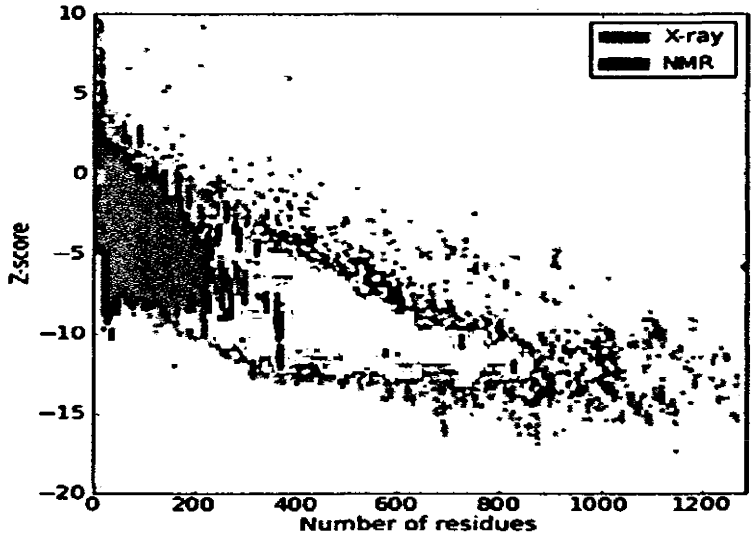


Fig: 3.34 Myo6 Z-score plot

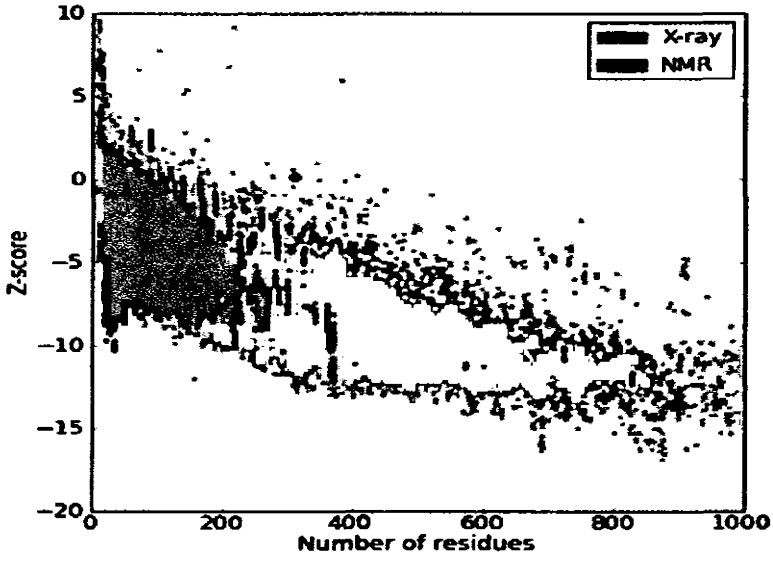


Fig: 3.35 Gipc3 Z-score plot

3.8.2. Evaluation result of modeled proteins by Rampage

Proteins Backbone conformation was checked by generating PSi/Phi Ramachandran plot using Procheck and RAMPAGE Packing. **figure 3.36-figure 3.41** show the Ramachandran plots of modeled proteins.

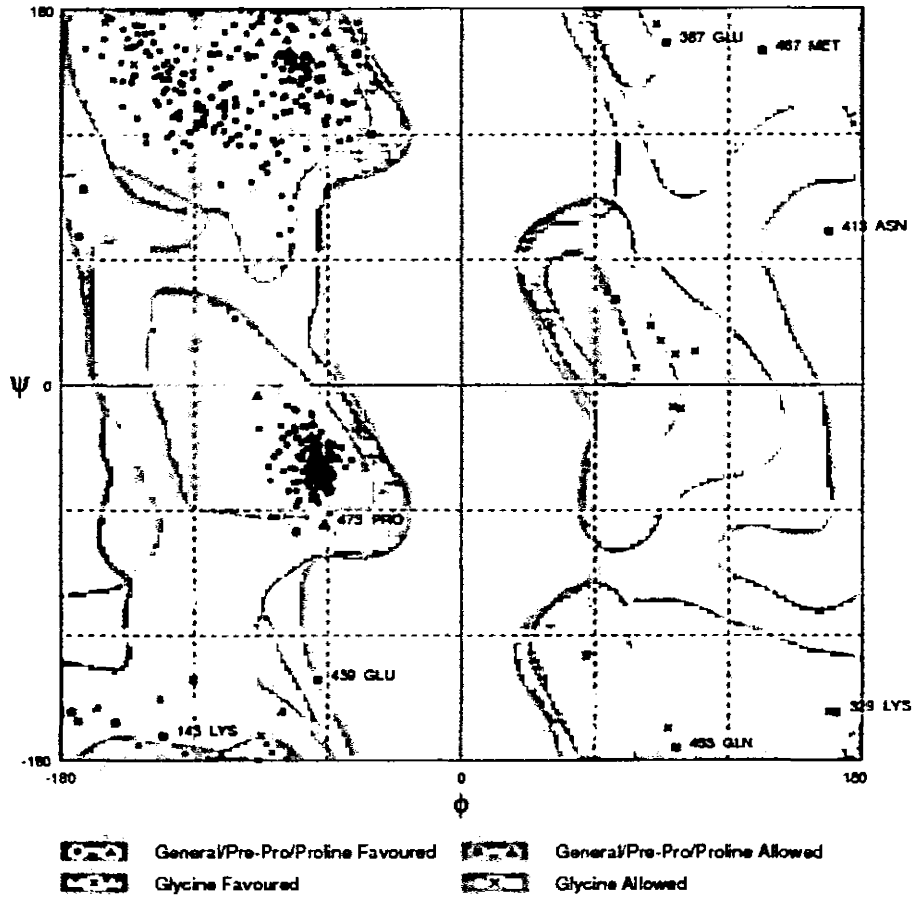


Fig 3.36: RDX gene Ramachandron Plot

Evaluation of residues:

Number of residues in favored region : 559 (96.2%)

Number of residues in allowed region : 14 (2.4%)

Number of residues in outlier region : 8 (1.4%)

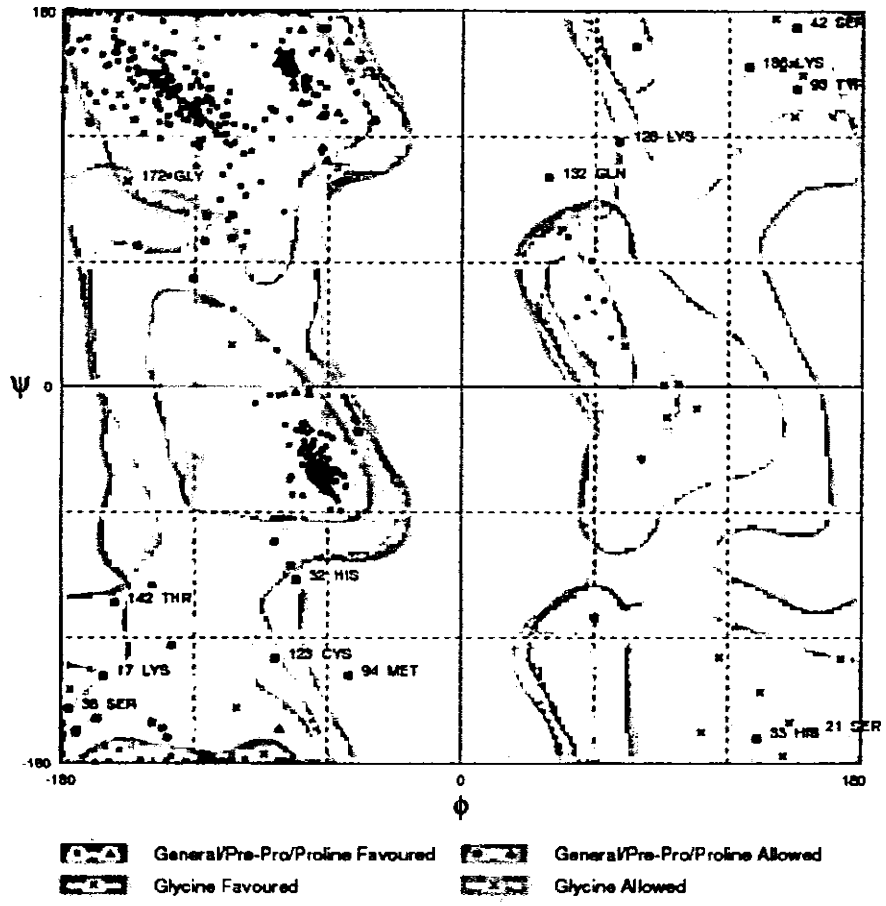


Fig 3.37: ESRRB gene Ramachandron Plot

Evaluation of residues:

Number of residues in favored region : 451 (90.6%)

Number of residues in allowed region : 33 (6.6%)

Number of residues in outlier region : 14 (2.8%)

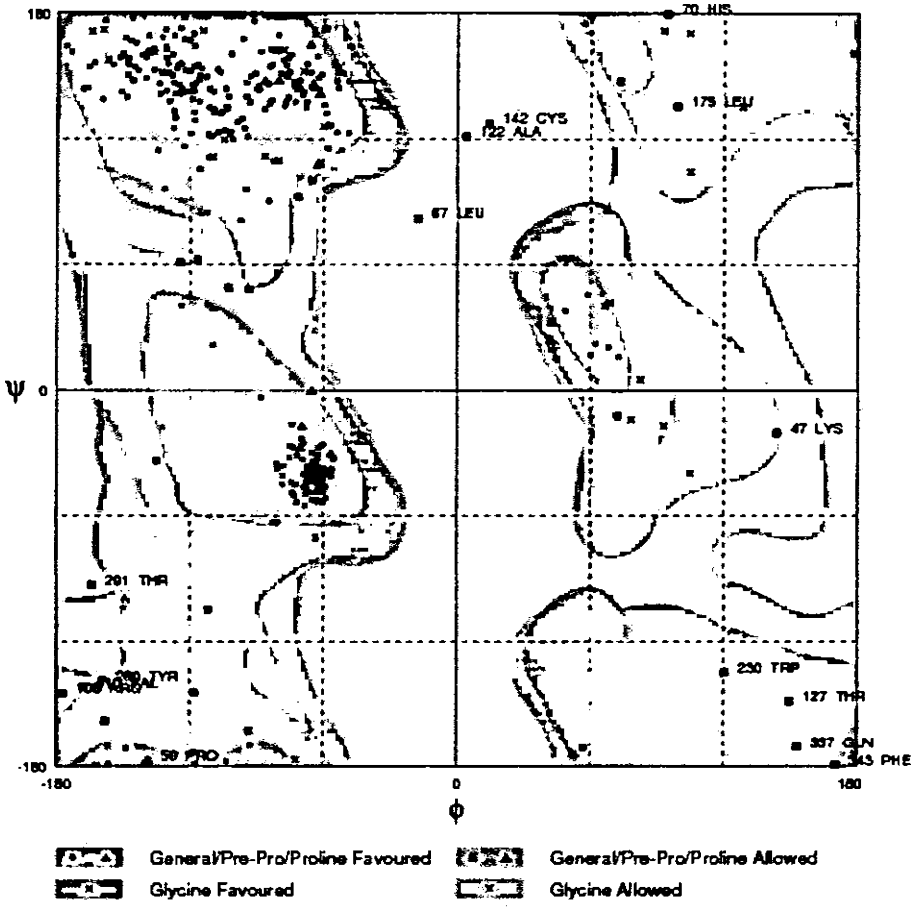


Fig 3.38: TMPRSS3 gene Ramachandron Plot

Evaluation of residues:

Number of residues in favored region : 305 (89.2%)

Number of residues in allowed region : 22 (6.4%)

Number of residues in outlier region : 15 (4.4%)

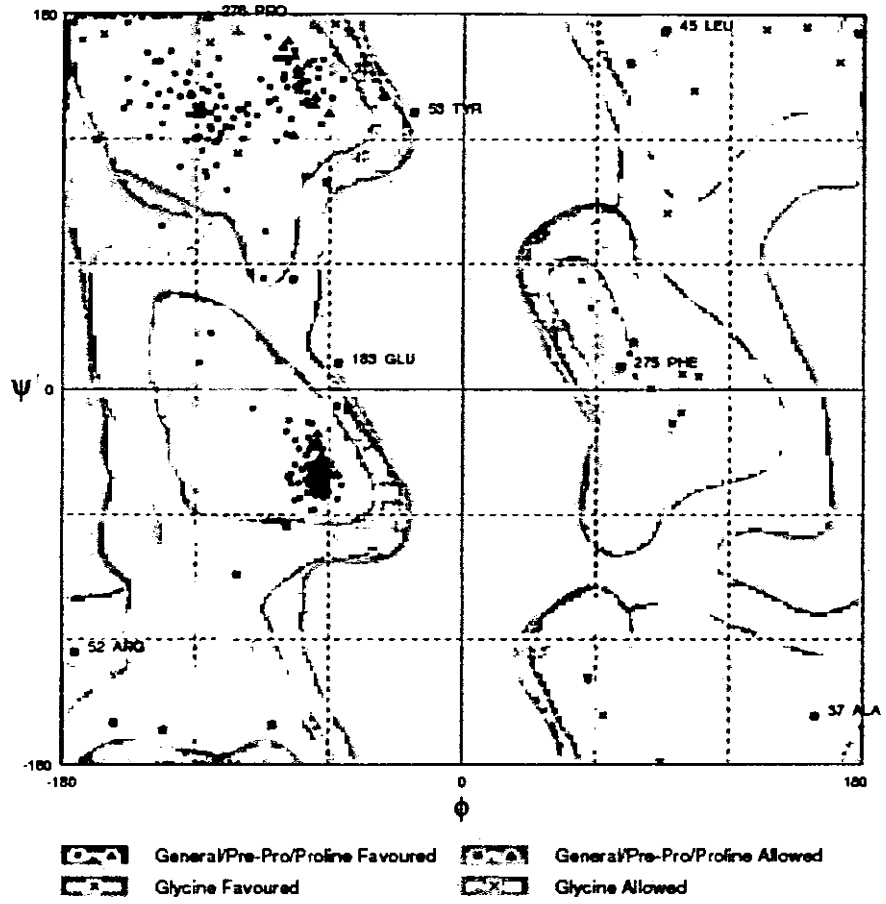


Fig 3.39: Comt2 gene Ramachandron Plot

Evaluation of residues:

Number of residues in favored region: 266 (92.0%)

Number of residues in allowed region: 16 (5.5%)

Number of residues in outlier region: 7 (2.4%)

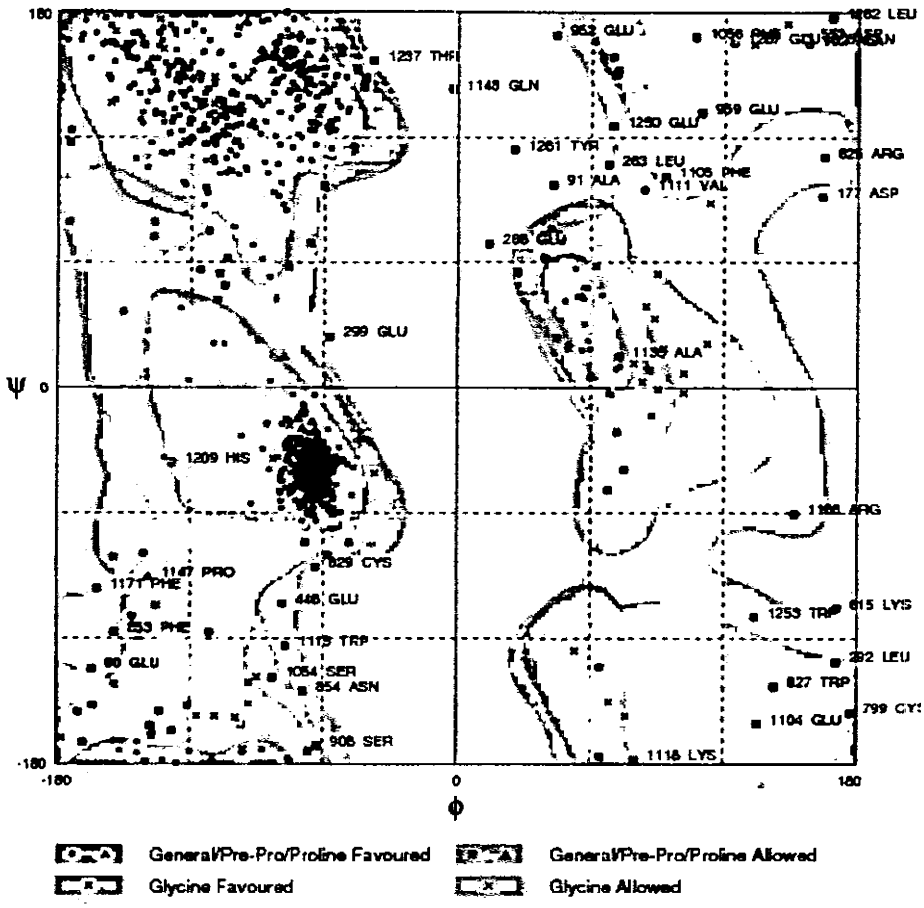


Fig 3.40: Myo6 gene Ramachandron Plot

Evaluation of residues:

Number of residues in favored region : 1166 (90%)

Number of residues in allowed region : 77 (6.0%)

Number of residues in outlier region : 40 (3.1%)

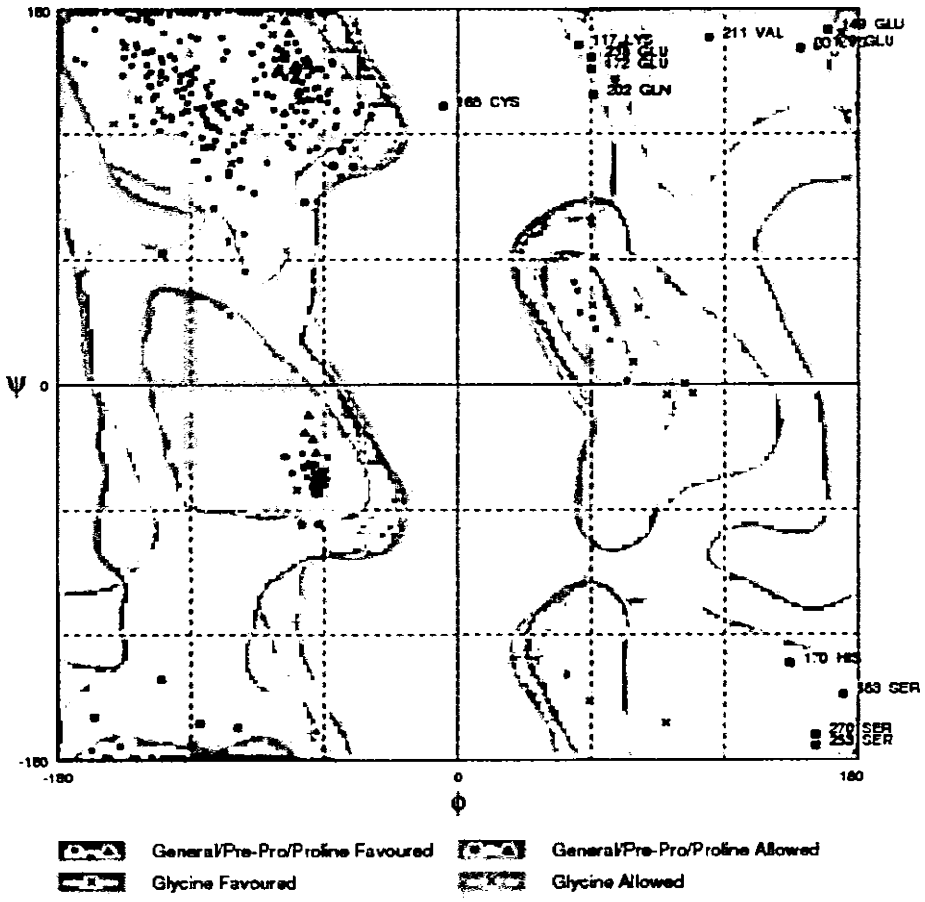


Fig 3.41: Gipc3 gene Ramachandron Plot

Evaluation of residues:

Number of residues in favored region : 275 (88%)

Number of residues in allowed region : 22 (7.1%)

Number of residues in outlier region : 13 (4.2%)

3.8.3. Evaluation through WHAT IF

WHAT IF is versatile molecular modeling program that help in analyzing small molecules, proteins and their interactions. The program has integrated relational protein structure database in it. The program is suitable for most common crystallographic work. Although there are not too many unique features in WHAT IF, the fact that everything is incorporated in one program makes it a useful tool for researchers (Vriend G, 1990). We used web version of WHAT IF for validation of our modeled structure of selected genes. The evaluation results of modeled genes are summarized in **Table 3.16.**

Gene Name	Packing quality control per amino acid	RMS-deviation in bond distances
RDX	-0.794	0.019
ESRRB	-2.390	0.021
TMPRSS3	-1.857	0.020
COMT2	1.466	0.020
MYO6	-1.298	0.019
GIPC3	-2.938	0.021
HGF	-3.094	3.149

Table 3.16: what if Server Evaluation results of selected genes

3.8.4a. Evaluation by Procheck of modeled proteins

The Procheck is a program for detailed checking of stereochemistry of a protein structure. The outputs comprise a number of plots and a comprehensive residue-by-residue listing. procheck give evaluate the overall quality of the structure by comparing it with well refined structures in the database and it also suggest regions that may need further checking (Laskowski *et al.*, 1998). Procheck evaluation results for our modeled genes are listed in **Table 3.17**.

Name Of Gene	Most Favored Regions (%Age)	Additional Allowed Regions	Generously Allowed Regions	Disallowed Regions
RDX	505(93.5%)	26(4.8%)	7 (1.3%)	2 (0.4%)
ESRRB	363(84.0%)	50(11.6%)	13 (3.0%)	6(1.4%)
TMPRSS3	242(81.5%)	38(12.8%)	10(3.4%)	7(2.4%)
Comt2	218 (88.6%)	23(9.3%)	3(1.2%)	2 (0.8%)
Myo6	1028(87.3%)	99 (8.4)	33(2.8%)	18 (1.5%)
Gipc3	242 (93.1%)	12 (4.6%)	3 (1.2%)	3 (1.2%)
HGF	497(80.7%)	83(13.5%)	25(4.1%)	11(1.8%)

Table 3.17: Procheck evaluation results of modeled Proteins

3.8.4b. G-Factors:

G-factors provide a measure of how unusual a protein property is. The values below -0.5 are considered as unusual while Values below -1.0 are considered as highly unusual. By using procheck we found the G-factor of modeled proteins. The G-factor values for studied proteins are as RDX(0.04), ESRRB(-0.32), TMPRSS3(-0.33), Comt2(-0.16), Myo6(-0.13), Gipc3(0.03), HGF(-0.67).

Chapter 4

Discussion

4. Discussion

Deafness is genetically diverse disorder and can result from environmental as well as genetic factors. In Pakistan the ratio of deafness is 1.6 percent of 1000. It is estimated that 70% of deafness results from inter cousin marriages. The genetically determined deafness can be broadly categories into two types; syndromic and non-syndromic forms. The syndromic forms of deafness include several hundred deafness syndromes.

In non-syndromic genetic deafness, autosomal recessive type is most prevalent (80%), while autosomal dominant accounts (20%), X-linked (1%), and mitochondrial (<1%) forms have also been described. Non-syndromic deafness is example of genetic heterogeneity. It is estimated that more than 70% of hereditary hearing loss is non-syndromic (Peterson MB and Willems PJ, 2006). The autosomal recessive deafness genes selected for study includes RDX, LRTOMT/COMT2, TMPRSS3, ESRRB, MYO6, GIPC3 and HGF. A search of the RCSB Protein Data Bank confirmed that the X-ray crystal structure of these genes is not publically available.

Primary and secondary analysis of studied protein is done through ProtParam and Scratch protein predictor respectively. Table 3.1 and Table 3.2 show the primary and secondry structure information of our studied genes.

Table 3.3-3.4 show the summary of structure models obtained through SWISS-MODEL, 3Djigsaw Model. The model obtained through these three programs turned to be of not full length. Another program that we used for structure prediction includes SAM-T08 which is based upon HMM. We did not get the required structures for all seven proteins as the SAM-T08 has length restrain up to 700 amino acid. We only obtained the structure of four proteins namely RDX, ESRRB, TMPRSS3 and GIPC3. The results of models obtained through SAM-T08 are listed in Table 3.5. The models obtained through Modellar are shown in figures 3.1-3.7. All models obtained through Modellar are of full length. For the selection of final structure we performed the evaluation of models obtained through SAM-T08 and Modellar.

Accuracy of predicted models was subjected through a series of tests. Stereochemical excellence of modeled proteins was evaluated through procheck program. The numerical statistics of procheck contain listing of residues in favoured, allowed and disallowed regions. The best model is supposed to has maximum no of residues in favored regions and minimum in un favorable regions. The procheck evaluation of our predicted proteins proved satisfactory level of steriochemical quality. Another parameter calculated by procheck is G-factor. The G-factor value of all predicted models except HGF protein is found more than -0.5(which is standard value for proteins). The empirical distribution of data points in protein structure is check by Ramachandran plots. The Ramachandran plots analysis show that most data points of our predicted proteins lie in favorable regions. Coarse packing quality of modeled proteins is calculated by WHAT IF packing quality control. Wheat is uses DACA for calculating packing quality per amino acid. Packing quality of our studied proteins turn out more than -5 showing satisfactory nature of predicted proteins. The ProSA test was used to evaluate the quality of consistency between the native fold of proteins and their sequence to check the energy of residue–residue interactions using a distance based pair potential. The energy is transformed to a score called Zscore. Residues with negative Z-score indicated reasonable side chain interactions. Z-score of all modeled proteins come satisfactory. All the calculated values showed high likelihood of folding thus good quality of models. The final selected models are used for further study.

For mutational analysis of our modeled proteins we used text mining to find the reported mutation of our proteins. The reported mutation was modeled by using WHAT IF web server. The mutated models were used for mutational analysis.

For RDX gene the only reported mutation associated with hearing loss is p.D578N (Shahid *et al.*, 2007). This mutation disrupts the helical structure of protein.

ESRRB protein has two reported substitutions mutation p.L320P and p.V342L. The studies of domain of ESRRB revealed that both mutations were located within the ligand-binding domain. Molecular modeling of these mutation showed that the missense mutations are likely to affect the structure and stability of these domains. In general the introduction of proline is reduces the stability of helix structure thus p.L320P mutation

might reduce the helix stability and entire ligand-binding domain of ESRRB (Collin *et al.*, 2008).For TMPRSS3 mutational analysis two reported mutations were D103G and del207C.The D103G mutation affects an aspartic acid residue of the TMPRSS3 in conserve LDLRA domain (Wattenhofer *et al.*, 2002). L16P and R167Q are the modeled mutation of Comt2 protein. In the L16P mutation the proline destabilizes the structure. R167Q mutation causes the conversion of Arginine (basic) to Glutamine (polar) amino acid. The mutated models of these proteins could be use for designing new drugs to cure deafness.

Protein–protein interactions (PPI) knowledge is necessary to understand the complex metabolic interaction networks that occur in living organisms, with the ultimate aim of designing drugs for blocking or enhancing interactions of therapeutic interest (Grosdidier *et al.*, 2009).

Docking Legand finding was done through text mining only that legend was considered that had clear evidence of its interaction with corresponding protein in autosomal recessive deafness. String data base (Szklarczyk *et al.*, 2010) was used to further validate the selection of legand. Output hits of String database are accompanied with a confidence score. The hit with highest confidence score and clear literature interaction evidence was finally chosen for docking. For docking we used GRAMM-X GRAMM v1.03 and hex software. We used the GRAMMX docking mode as generic mode to tries all ligand's positions and orientations.

Hydrogen bonding site of our docked proteins are found by using pymol (Delano W L, 2002) script files. The hydrogen bonding of docked complexes is listed in table 3.14.

For RDX and its legand SLC9A3R1(Kalay *et al.*, 2005) docked complex the most hydrogen bond forming residues lie from LYS`133 to ARG`184 suggesting possible binding pocket in this region. All the residues of RDX are found to be hydrophobic in nature thus suggesting hydrophobic nature of binding pocket.

For ESRRB hydrogen bond interaction is found within LEU`104 to GLU`339. The nature of most of these residues is found hydrophobic.

TMPRSS3 docked complex with GJB2 (Snoeckx et al., 2005) hydrogen bond interactions are found in VAL`108, TYR`296, VAL`291 and VAL`9. Here the valine and tyrosine are both hydrophobic in nature. These residues are major contributor for protein-protein interaction.

In case of Comt2 and LRRC51 docked complex ASP`85, ASN`62, THR`191 and TYR`108 are found forming hydrogen bonding, indicating the potential role of these residues in protein-protein interaction. Here Asp, Thr and Asn are hydrophilic in nature while Tyr is hydrophobic.

My06 and Gipc1 (Hertzano *et al.*, 2008) docked complex had widely distributed residues forming hydrogen bonding. The major bonding residues found includes GLN`768, ASN`116, ILE`24 and THR`1135. In this case Gln, Thr and Asn are found to be hydrophilic while only Ile is of hydrophobic in nature.

Gipc3 protein docked structure hydrogen bonding residues found include PRO`22, ARG`32, ARG`3, ASN`57, GLU`67 and PRO`73. Out of these only proline is of hydrophobic in nature. These residues are present within small range suggesting possible binding pocket in this region.

The Present study is focused on seven selected genes associated with autosomal recessive deafness. This includes the modeling of structure of these genes through comparative homology modeling. The mutational analysis, protein-protein interaction and docking of modeled proteins was also part of study.

The modeled structures could be used for deriving more refine and accurate models such as NMR and X-ray crystallographic structure of proteins. Protein–protein interactions knowledge of studied protein could be used to understand the complex metabolic interaction networks that occur in deafness. The knowledge of residues in docking could be investigated for designing new drugs for deafness.

Chapter 5

References

5. References

- Ahmed ZM., Morell RJ., Riazuddin S., Gropman A., Shaukat S., Ahmad MM., Mohiddin SA., Fananapazir L., Caruso RC, Husnain T., Khan SN., Riazuddin S., Griffith AJ., Friedman TB., Wilcox ER., (2003) Mutations of MYO6 are associated with recessive deafness, DFNB37, Am J Hum Genet, 72, 1315-1322.

- Ansar M., Amin Ud Din M., Arshad M., Sohail, M., Faiyaz-Ul-Haque M., Haque S., Ahmad W., Leal S.M., (2003) A novel autosomal recessive non-syndromic deafness locus (DFNB35) maps to 14q24.1-14q24.3 in large consanguineous kindred from Pakistan, Europ. J. Hum. Genet, 11, 77-80.

- Bates P.A., Kelley L.A., MacCallum R.M., and Sternberg M.J.E., (2001) Enhancement of Protein Modelling by Human Intervention in Applying the Automatic Programs 3D-JIGSAW and 3D-PSSM., Prpteins: Suppl, 5, 39-46.

- Bachar O., Fischer D., Nussinov R., and Wolfson H.J., (1993) A computer vision based technique for 3D sequence-independent structural comparison of proteins, Protein Eng, 6, 279–288.

- Buss F., Arden S.D., Lindsay M., Luzio J.P., Kendrick-Jones., (2001) Myosin VI isoform localized to clathrin-coated vesicles with a role in clathrin-mediated endocytosis, EMBO J, 20, 3676-3684.

- Cheng J., Randall A.Z., Sweredoski M.J., and Baldi P., (2005) SCRATCH: a protein structure and structural feature prediction server, Nucleic Acids Res, 33, 72–76.

- Collin R.W J., Kalay E., Tariq M., Peters T., van der Zwaag, B., Venselaar H., Oostrik J., Lee K., Ahmed Z.M., Caylan R Li Y., Spierenburg H.A., and 17 others (2008), Mutations of ESRRB encoding estrogen-related receptor beta cause autosomal-recessive nonsyndromic hearing impairment DFNB35, Am. J. Hum. Genet, 82, 125-138.

- Delano W L (2002).The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos, CA, USA.
- Fiser A & Sali A., (2003) ModLoop: automated modeling of loops in protein structures. Bioinformatics., 19, 2500-2501.
- Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A., (2005) Protein Identification and Analysis Tools on the ExPASy Server, Humana Press, 571-607.
- Grosdidier S., Totrov M., Fernández-Recio J., (2009) Computer applications for prediction of protein–protein interactions and rational drug design, advances and Applications in Bioinformatics and Chemistry, 2, 101–123.
- Greer J. (1981), Comparative model-building of the mammalian serine proteases, 153, 1027–1042.
- Vriend G (1990) WHAT IF: A molecular modeling and drug design program, J. Mol Graph, 8, 52-56.
- Halperin I,Ma B., Wolfson H., and Nussinov R., (2002) Principles of Docking: An Overviewof Search Algorithms and a Guide to Scoring Functions, Proteins, 47, 409-443.
- Hertzano R., Shalit E., Rzadzinska AK., Dror AA., Song L., Ron U., Tan JT., Shitrit AS., Fuchs H., Hasson T., Ben-Tal N., Sweeney HL., de Angelis MH., Steel KP., Avraham KB., (2008) A Myo6 Mutation Destroys Coordination between the Myosin Heads, Revealing New Functions of Myosin VI in the Stereocilia of Mammalian Inner Ear Hair Cells, PLoS Genet, 4.
- Jung E.J., Liu G., Zhou W., Chen X., (2006) Myosin VI is a mediator of the p53-dependent cell survival pathway, Mol. Cell. Biol. 26, 2175-2186.

- Khan SY., Ahmed ZM., Shabbir MJ., Kitajiri S., Kalsoom S., Tasneem S., (2007) Mutations of the RDX gene cause nonsyndromic hearing loss at the DFNB24 locus, Hum Mutat, 28, 417-424.

- K Karplus (2009) SAM-T08,HMM-based protein structure prediction Sciences, Nucleic Acids Research, 37, 492-497.

- Katchalski-Katzir E., Shariv,I., Eisenstein,M., Friesem,A.A., Aflalo,C., and Vakser,I.A., (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques, Proc. Natl Acad. Sci. USA, 89, 2195–2199.

- Laskowski R.A., MacArthur M.W., and Thornton J.M., (1998) Validation of protein models derived from experiment, Curr. Opin. Struct. Biol, 8, 631-639.

- Levitt M, (1992) Accurate modeling of protein conformation by automatic segment matching, J Mol Biol, 2, 507-533.s

- Marc A., arti-Renom , Stuart AC, 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.

- Masmoudi S., Antonarakis SE., Schwede T., Ghorbel AM., Gratri M., Pappasavas MP., Drira M., Elgaied-Boulila A., Wattenhofer M., Rossier C., Scott HS., Ayadi H., Guipponi M., (2001) Novel missense mutations of TMPRSS3 in two consanguineous Tunisian families with non-syndromic autosomal recessive deafness, Hum Mutat, 2,101–108.

- Martí-Renom MA., Stuart AC., Fiser A., Sánchez R., Melo F., Sali A., (2000) Comparative protein structure modeling of genes and genomes, Annu Rev Biophys Biomol Struct, .29, 291-325.

- Macindoe G., Mavridis L., Venkatraman V., Devignes M.-D., Ritchie D.W., (2010). HexServer: an FFT-based protein docking server powered by graphics processors. Nucleic Acids Research, 38, 445-449.

- Nikoletta Charizopoulou, Andrea Lelli, Margit Schraders, Kausik Ray, Michael S. Hildebrand, Arabandi Ramesh, C. R. Srikumari Srisailapathy, Jaap Oostrik, Ronald J. C. Admiraal, Harold R. Neely, Joseph R. Latoche, Richard J. H. Smith, John K. Northup, Hannie Kremer, Jeffrey R. Holt, and Konrad Noben-Trauth (2011) Gipc3 mutations associated with audiogenic seizures and sensorineural hearing loss in mouse and human, Nat Commun, 15, 2- 201.
- Pataky F., Pironkova R., Hudspeth AJ., (2004) Radixin is a constituent of stereocilia in hair cells, Proc Natl Acad Sci U S A. ,101, 2601-2606.
- Peterson MB., Willems PJ.(2006)Non-syndromic, autosomal-recessive deafness., Clin Genet: 69, 371-392.
- Phizicky EM and Fields S (1995) Protein-protein interactions: methods for detection and analysis, Microbiol Rev, 1, 94-123.
- Sali & T.L. Blundell (1993), Comparative protein modelling by satisfaction of spatial restraints, J. Mol. Biol, 234, 779-815.
- Saitoh T, Mine T, Katoh M (2002).Molecular cloning and characterization of human GIPC3, a novel gene homologous to human GIPC1 and GIPC2. Int J Oncol, 20, 577-582.
- Sippl M (1993) Recognition of errors in three-dimensional structures of proteins.Proteins . PubMed, 17, 355-362.
- Safran M., Dalah I., Alexander J., Rosen N., Iny Stein T., Shmoish M., Nativ N., Bahir I., Doniger T., Krug H., Sirota-Madi A., Olender T., Golan Y., Stelzer G., Harel A., Lancet D.,(2010) GeneCards Version 3: The human gene integrator.
- Suda J., Zhu L., Karvar S., (2010) Phosphorylation of radixin regulates cell polarity and Mrp-2 distribution in hepatocytes, Am J Physiol Cell Physiol: 300, 416-424.
- Szklarczyk D., Franceschini2 A., Kuhn3 M., Simonovic2 M., Roth2 A., Minguez4 P., Doerks4 T., Stark2 T., Muller J., Bork p., Lars J., Mering2 c.v.,

(2010) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored, Nucleic Acids Research, 39, 561-568.

- Schwede T., Kopp j., Guex N., and Peitsch Manuel C., (2003) SWISS-MODEL: an automated protein homology-modeling server, Nucleic Acids Research, 31, 3381-3385.

- Snoeckx RL., Huygen PL., Feldmann D., Marlin S., Denoyelle F., Waligora J., Mueller-Malesinska M., Pollak A., Ploski R., Murgia A., Orzan E., Castorina P., Ambrosetti U., Nowakowska-Szyrwinska E., Bal J., Wiszniewski W., Janecke AR., Nekahm-Heis D., Seeman P., Bendova O., Kenna MA., Frangulov A., Rehm HL., Tekin M., Incesulu A., Dahl HH., du Sart D., Jenkins L., Lucas D., Bitner-Glindzicz M., Avraham KB., Brownstein Z., del Castillo I., Moreno F., Blin N., Pfister M., Sziklai I., Toth T., Kelley PM., Cohn ES., Van Maldergem L., Hilbert P., Roux AF., Mondain M., Hoefsloot LH., Cremers CW., Löppönen T., Löppönen H., Parving A., Gronskov K., Schrijver I., Roberson J., Gualandi F., Martini A., Lina-Granade G., Pallares-Ruiz N., Correia C., Fialho G., Cryns K., Hilgert N., Van de Heyning P., Nishimura CJ., Smith RJ., Van Camp G., (2005) GJB2 mutations and degree of hearing loss: a multicenter study, Am J Hum Genet, 77, 945-957.

- Shahid Y. Khan, Zubair M. Ahmed, Muhammad S, Shin-ichiro Kitajiri, Saeeda K. Saba Ta, Sara S, Arabandi R,Srikumari S, Shaheen N. Khan, Richard J.H. Smith, Saima R, Thomas B. Friedman,and Sheikh R (2007) Mutations of the RDX Gene Cause Nonsyndromic Hearing Loss at the DFNB24 Locus, human mutation 28. 417-423.

- Tovchigrechko A and Vakser IA (2006) GRAMM-X public web server for protein–protein docking, Nucleic Acids Res, 1, 310–314.

- Van Camp G., Smith RJH (2009) Hereditary hearing loss. (Homepage) (Online) Available from URL: http://www.webhost.ua.ac.be/hhh/July.

- Wiederstein M., and Sippl M.J., (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins, Nucleic Acids Research, 35, 407–410.

- Wattenhofer M., Di Iorio MV., Rabionet R., Dougherty L., Pampanos A., Schwede T., Montserrat-Sentis B., Arbones ML., Iliades T., Pasquadibisceglie A., D'Amelio M., Alwan S., Rossier C., Dahl HH., Petersen MB., Estivill X., Gasparini P., Scott HS., Antonarakis SE.,(2002) Mutations in the TMPRSS3 gene are a rare cause of childhood nonsyndromic deafness in Caucasian patients, J Mol Med, 80, 124-131.

- Zhang Y (2008) Progress and challenges in protein structure prediction, Curr Opin Struct Biol, 18, 342–348.