In-Silico Protein Modelling and Drug Designing of Genes Involved in Parkinson's Disease

Researcher
Adeela Safdar
01-MSBI/FBAS/F07

Supervisor
Dr. Shaheen Shahzad
Chairperson

Department of Environmental Sciences
Faculty of Basic & Applied Sciences
International Islamic University,
Islamabad
(2012)

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

In the Name of Allāh, the Most Gracious, the Most Merciful

Department of Environmental Sciences

International Islamic University, Islamabad

Dated: 19-1-12

FINAL APPROVAL

It is certificate that we have read the thesis submitted by Ms. Adeela Safdar and it is 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

External Examiner
Dr. Sajid Rashid
Assistant Professor
Department of Bioinformatics
Quaid-e-Azam University, Islamabad

Internal Examiner
Dr. Sobia Tabassum
Assistant Professor
Department of Environmental Sciences
International Islamic University, Islamabad

Supervisor
Dr. Shaheen Shahzad
Chairperson/ Assistant Professor
Department of Bioinformatics and Biotechnology,
International Islamic University, Islamabad

Chairman & Dean
Prof. Dr. Muhammad Irfan Khan
Department of Environmental Sciences
International Islamic University, Islamabad

A thesis submitted to Department of Environmental Sciences,
International Islamic University, Islamabad as a partial
fulfillment of requirement for the award of the
degree of M.S in Bioinformatics

DEDICATION

This humble effort
The fruit of studies and thoughts
Dedicated
To the Holy Prophet
Hazrat Muhammad (P.B.U.H)

And **My Husband**
Who has been always very
Kind, encouraging and without
Whose caring support
It would not have been possible

And **My Parents** who always prayed
For me and gave me a strong
Moral support

DECLARATION

I hereby declare that the work presented in the following thesis is my own effort, except where otherwise acknowledged and that the thesis is my own composition. No part of the thesis has been previously presented for any other degree.

Date 19.1.12.

Adeela Safdar
(Adeela Safdar)

ACKNOWLEDGEMENTS

I would first like to express my humble gratefulness to Almighty Allah, The propitious and The sovereign who gave me an opportunity to flourish my thoughts and blessed me with the courage to complete this task and explore my natural abilities. All the reverence and esteem for His beloved Prophet Hazrat Mohammad (Peace Be Upon Him) the most perfect and exalted among, who enlightened the mankind on true path of life and is source of inspiration for all knowledge seekers.

The research work was accomplished under the sympathetic attitude, scholarly criticism, cheering perspective and enlightened supervision of Dr. Shaheen Shahzad, Assistant Professor (Chairperson), Department of Environmental Sciences, International Islamic University, Islamabad. I dream it my utmost pleasure in expressing my cardise gratitude with the profound benedictions to my supervisor. Her generous and expert guidance, keen interest at every step and continuous encouragement throughout this research work enabled me to achieve this goal. Moreover, I am very thankful to her for critically reading this dissertation and providing me guidance to improve it without which this work would have been impossible.

I would also like to acknowledge International Islamic University Islamabad which provided me the platform to undertake this research work and accomplish it within due course time. I owe my deepest gratitude to bioinformatics faculty for equipping me with the latest knowledge and providing me academic base to take up this work.

I cannot find words to express my adequate obligations to my husband, who has been very supportive throughout this research work. My deepest gratitude to my parents who always prayed for me to be successful in every field of life. My special thanks to all my class mates who always encouraged me and guided me. May Allah bless all these people with long, happy and peaceful lives (Ameen)

Adeela Safdar

Adeela Safdar

CONTENTS

LIST OF ABBREVIATIONS

OMIM	Online Mendelian Inheritance in Men
GBA	Glucisidase beta acid gene
NCBI	National Center Of Biotechnology Information
BLAST	Basic Local Alignment Search Tool
PDB	Protein Data Bank
ELM	Eukaryotic Linear Motif resource
SMART	Simple Modular Architecture Research Tool
HMMTOP	Hidden Markov Model TOpology prediction of Proteins
GOR4	Garnier-Osguthorpe-Robson
PARK2	Parkinson Protein 2
PINK1	PTEN Induced Putative Kinase 1
ZnFs	Zinc Fingers
NTP	Nucleotide Tri-Phosphate
SignalP-NN	SignalP-Neural Network
SignalP-HMM	SignalP-Hidden Markov Model
LRRK2	Leucine-Rich Repeat Kinase 2
SNCA	Synuclein, alpha
UCHL1	Ubiquitin Carboxyl-terminal esterase L1
Rmsd	Root Mean Square Deviation
l.b.	Lower Bound
u.b.	Upper Bound

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

During the last two decades, rapidly emerging technologies used in identification, characterization and analysis of genes have been made possible by advances in bioinformatics. Genetically inherited diseases that are caused by mutations in these genes require detailed information about the gene, its exact location in the genome, nature of mutations as well as structure and function of these genes which ultimately codes proteins.

Parkinson disease is one of the genetically inherited disease and most chronic and progressive neurodegerative diseases. The diagnosis of idiopathic Parkinson disease is clinical with manifestations including muscular rigidity, tremor, bradykinesia, and postural instability. Additional features include postural abnormalities, dystonic cramps, dysautonomia, and dementia. The disease is progressive and has onset in mid to late adulthood.

The present research work focused on two main domains; functional analysis of the genes involved in the disease to predict novel 3-dimensional models of the proteins and to study the protein-ligand interactions to develop potential drugs with the help of docking procedures. 3D models were built using homology modelling and best models were selected on the basis of lowest objective function value. Molecular docking was performed using AutoDock Vina to obtain possible orientations and conformations for the ligand at the binding sites. Binding affinities of these ligands were calculated with the selected proteins using the software. The lower the binding affinity value, the best the ligand and receptor binding.

1. INTRODUCTION

1.1 Parkinson's disease

In 1817, Parkinson disease was described as 'shaking palsy' by James Parkinson. After Alzheimer disease it is second most common neurodegenerative disorder (Polymeropoulos *et al.*, 1996).

Parkinson disease is one of the most chronic and progressive neurodegerative diseases. In Parkinson disease motor symptoms are caused as a result of loss of dopamine-producing nigro-striatal neurons in substantia nigra (Fig. 1.1). Position of substantia nigra and diminished substantia nigra was shown in figure 1.2. Studies suggested that delay in introduction of medical therapy of dopamine causes a frequent decline in the quality of life. The most common non-motor symptoms involved in Parkinson's diseases includes depression (Stuart *et al.,* 2009)

The diagnosis of idiopathic Parkinson disease with clinical manifestations including muscular rigidity, tremor, postural instability and bradykinesia. Additional features include postural abnormalities, dementia, dystonic cramps and dysautonomia. Parkinson's disease is a progressive disease and has onset in the mid of the age to late adulthood (Nussbaum and Polymeropoulos, 1997).

Many of the other diseases, non-genetic and genetic, may have parkinsonism which may be caused by dysfunction or loss of dopamine neurons in the substantia nigra of brain, but on the pathology they may or may not have Lewy bodies. Thus accurate pathological examination is very important for its diagnosis. Parkinsonism refers to any symptom that involves any types of changes in movement that are seen in Parkinson disease. Parkinsonism may be caused due other factors like stroke, meningitis, encephalitis etc.

1.2 Clinical Features

The classic idiopathic Parkinson disease has primarily clinical diagnosis, which includes resting tremor, bradykinesia, muscular rigidity, and postural instability. Additional features of the disease include dysautonomia, postural abnormalities, dementia, dystonic cramps. It is progressive disease and has an insidious onset in the midof age to late adulthood. Pathologic features of classic Parkinson disease is the presence of intracellular inclusions and Lewy bodies, in surviving neurons of various parts of brain and loss of dopaminergic neurons particularly in substantia nigra. While autosomal recessive Parkinson diseases do not show Lewy body pathology (Nussbaum and Polymeropoulos, 1997).

Other minor symptoms for the disease may include depression, difficulty in chewing and swallowing, emotional changes, speech changes, skin problems, constipation or urinary problems and sleep problems.

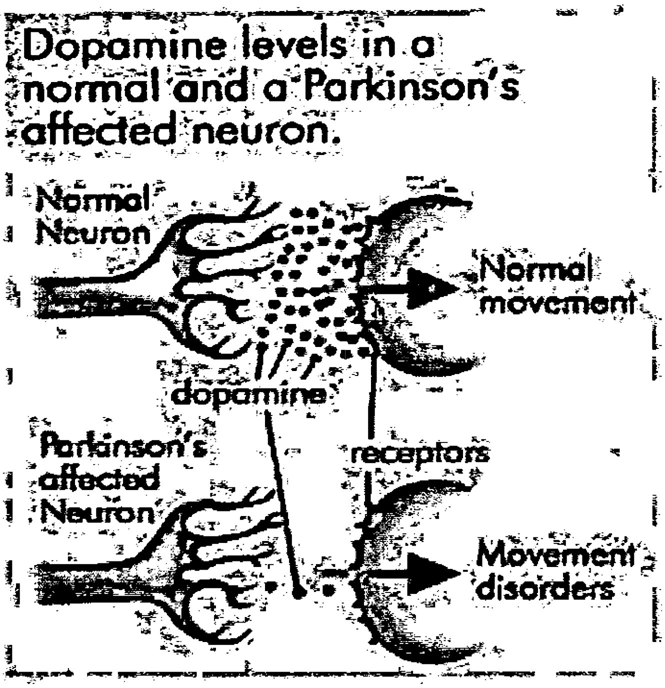

Dopamine levels in a normal and a Parkinson's affected neuron.
Normal Neuron
Normal movement
dopamine
Parkinson's affected Neuron
receptors
Movement disorders

Figure 1.1 Dopamine levels in a normal and affected Parkinson's neuron

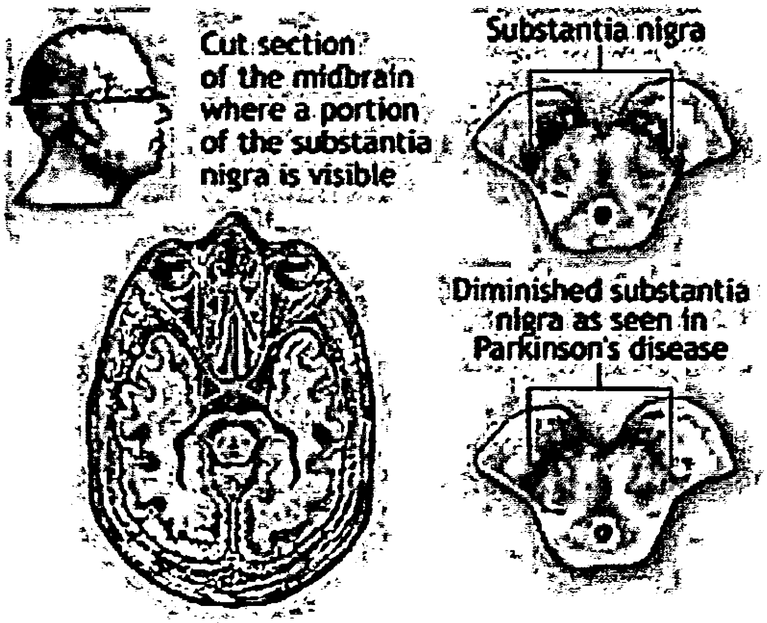

Cut section of the midbrain where a portion of the substantia nigra is visible
Substantia nigra
Diminished substantia nigra as seen in Parkinson's disease

Figure 1.2 Position of substantia nigra and diminished substantia nigra in Parkinson disease

Parkinson's disease is mostly caused at the late middle age; the average onset of the disease is 60 years but may affect the people at the age of 50 years. Some physicians also reported 5 to 10% of the early onset under the age of 40. Parkinson disease affects approximately 1% of the population, increasing to 4% to 5% of the population by the age of 85 years. It affects men slightly at higher rate than women. Parkinsonism in a rural Bavarian population was more prevalent at the age group older than 65 years. Prevalence of Parkinson disease in this population was 0.71%, drug-induced parkinsonism was 0.41%, vascular parkinsonism was 0.20%, multiple systems atrophy was 0.31%, (Trenkwalder *et al.*, 1995)

1.3 Causes of Parkinson's disease

Parkinson's disease have both environmental as well as genetic causes. The detail is as follows.

1.3.1 Initial causes

Parkinson's disease may be caused when specific nerve cells present in the brain die. These cells produce a chemical called dopamine. Dopamine is a chemical messenger, which transmits signals between corpus striatum and substantia nigra and produces smooth muscle activity. Loss of dopamine and presence of Lewy bodies causes the nerve cells of the corpus striatum to lose control on movements in normal manner (Nussbaum and Polymeropoulos, 1997)

1.3.2 Toxic causes

There are many toxins that cause Parkinson's disease or may cause symptoms that cause Parkinson's disease. These toxins include herbicides, pesticides, fungicides, insecticides or certain chemicals like metals, iron, manganese etc. These toxins can modify the structure of alpha-synuclein or interfere with the ubiquitin-proteasomal pathway. Thus it impairs the process of degradation of abnormal proteins and promotes alpha-synuclein aggregation (Seham *et al.,* 2010)

1.3.3 Genetic causes

The genetic recognition of the Parkinson disease can have two broad aspects. First recognition may be the Mendelian families, in which genes can be identified as mutations may segregate with disease or it could be present in populations where the disease is associated with more common variants. There are many genes which causes both recessive and dominant forms of parkinsonism, some of which have complex phenotypes and some overlap with sporadic Parkinson disease (Cookson *et al.,* 2010)

1.4 Types of Parkinson's disease

There are different types of Parkinson disease based on gene loci implicated in different forms of disease on Online Mendelian Inheritance in Men (OMIM).

1.4.1 Idiopathic Parkinson's disease

It is the type of Parkinson disease in which the cause of disease is not yet determined. This is the most common form of disease.

1.4.2 Parkinson disease (PARK1), familial, Type 1

It is caused by the genetic mutation in SNCA gene on chromosome 4q22.1.

1.4.3 Parkinson disease 3 (PARK3), autosomal dominant Lewy body

It is caused by the genetic mutation on chromosome 2p13.

1.4.4 Parkinson disease 4 (PARK4), autosomal dominant Lewy body

It is caused by the genetic mutation in SNCA gene on chromosome 4q22.1. This type of disease starts at the age of 45 and progresses rapidly.

1.4.5 Parkinson disease 5 (PARK5)

It is caused by the genetic mutation on chromosome 4p14.

1.4.6 Parkinson disease 6 (PARK6), autosomal recessive early-onset

It is caused by the genetic mutation on chromosome 1p36 in PINK1 gene. This condition results in the earlier on-set of the disease and fluctuates during day time.

1.4.7 Parkinson disease 2 (PARK2), autosomal recessive juvenile

This condition is inherited in recessive manner and juvenile form of disease. It is caused by mutation in PARK2 gene on chromosome 6q25.2-q27.

1.4.8 Parkinson disease 7 (PARK7), autosomal recessive early-onset

It is caused by mutation on chromosome 1p36 in gene DJ1. This condition progresses slowly and starts before the age of 40 years.

1.4.9 Parkinson disease 8 (PARK8)

It is caused by the genetic mutation in LRRK2 gene on chromosome 2q12.

1.4.10 Parkinson disease 9 (PARK9)

It is caused by the genetic mutation on chromosome 1p36 in gene ATP13A2. This type of Parkinson's disease rapidly progresses and begins after the age of 50 years.

1.4.11 Parkinson disease 10 (PARK10)

It is caused by the genetic mutation on chromosome 1p34-p32.

1.4.12 Parkinson disease 11 (PARK11)

It is caused by the genetic mutation in GIGYF2 gene on chromosome 2q37.

1.4.13 Parkinson disease 12 (PARK12)

It is caused by the genetic mutation on chromosome Xq21-q25.

1.4.14 Parkinson disease 13 (PARK13)

It is caused by the genetic mutation in HTRA2 gene on chromosome 2p12 and progresses slowly.

1.4.15 Parkinson disease 14 (PARK14)

It is caused by the genetic mutation in PLA2G6 gene on chromosome 22q13.

1.4.16 Parkinson disease 15 (PARK15)

It is caused by the genetic mutation in FBXO7 gene on chromosome 22q12-q13.

1.4.17 Parkinson disease 16 (PARK16)

It is caused by the genetic mutation on chromosome 1q32 and progresses slowly

1.4.18 Mitochondrial Parkinson's disease

This form of disease is associated to mitochondrial defects, which are the energy producing components of cells.

1.5 Association of Parkinson disease with other diseases

Depression is one of the earlier symptoms of Parkinson's disease. A study was made at America to check whether people who are taking antidepressant develop Parkinson's disease earlier than those who are not taking antidepressant before the disease was diagnosed. This risk of developing Parkinson before diagnosis was true for both men and women.

An association between Parkinson's disease and type 1 Gaucher disease has been reported. Gaucher disease is recessively inherited between families and is a disorder of glycolipid storage. GBA gene is associated with the genetic disorder of Gaucher disease. Occurrence of Parkinson's disease and Goucher disease progresses rapidly and makes resistance to conventional anti-Parkinson therapy (Varkonyi *et al.*, 2003). There were fourteen different heterozygous mutations identified in GBA gene in British population with Parkinson disease in 33 of 790 patients and 3 of 257 controls. Three novel mutations were indentified in GBA gene including L444P in 11 patients, N370S in 8 patients and R463C in 3 patients. In this research work, four patients showed the family history of this disorder and 29 patients had sporadic disease. The male to female ratio of Parkinson disease with GBA carriers was 5:2, which was significantly high. The prevalence of GBA mutations with sporadic Parkinson disease in these patients was 3.7%. This shows that in this population, for the development of Parkinson disease, mutations in the GBA gene may be the most common risk factor (Neumann *et al.*, 2009)

1.6 Genetics of Parkinson's disease

Several gene loci causing autosomal dominant forms of Parkinson's disease have been identified, which includes mutation in alpha-synuclein gene (SNCA) or triplication of SNCA on chromosome 4q22.1, mutation in UCHL1gene on chromosome 4p14, mutations in LRRK2 gene on 12q12 chromosome, mutation in GIGYF2 gene on chromosome 2q37, and mutation in HTRA2 gene on chromosome 2p12.

Several gene loci causing autosomal recessive type of early-onset Parkinson's disease has also been identified which includes mutation in parkin gene at chromosome 6q25.2-q27, mutation in DJ1 gene on 1p36, mutation in PLA2G6 gene on chromosome 22q13 and mutation in FBXO7 gene on chromosome 22q12-q13.

Mitochondrial mutations have also been reported and locus on X chromosome has also been identified causing Parkinson's disease.

Genetic causes of the occurance of complex diseases shows different spectrum of sequence variants than mutations that dominant monogenic disorders. In particular, this may include mutations which alter the gene expression. In inherited diseases, promoter mutations have been shown particularly, including neurodegenerative brain disorders (Theuns *et al.*, 2006).

Polymorphism or mutation in several genes together may have a substantial cumulative effect to the development of Parkinson disease, including HTRA2, NR4A2, LRRK2, NDUFV2, FGF20, ADH3, GBA, MAPT, and SNCA. Independently these genes may have modest effect on disease development (Abdelghaffar *et al.*, 2010).

Several loci and genes involved in development of Parkinson's disease are illustrated in table 1.1.

1.6.1 Inheritance of Parkinson's disease

There were controversies regarding the pattern of inheritance of Parkinson disease. However familial component of Parkinson's disease has been recognized.

1.6.1.1 Twin study

A pair of twins female was studied; both had a combination of anosmia and Parkinsonism. In Parkinson's disease olfactory impairment has also been seen frequent. Both twins showed early onset of disease at the age of 36 years, which is quite unusual for women. It was suggested that for genetically identified anomaly of dopamine metabolism was of causative role in the two families, previously having the same association (Kissel and Andre, 1976).

In another twin study, in the first 20 monozygotic twin pairs there was zero concordance for Parkinson disease ((Duvoisin *et al.,* 1981). There was only one monozygotic twin pair which definitely has Parkinson disease. He noted that concordance for Parkinson disease was expected from the incidence of disease and it was no more frequent in twins and therefore concluded that in the etiology of

Parkinson disease, the major factors must be nongenetic (Ward *et al.*, 1983)

1.6.1.2 Mendelian inheritance

Mendelian inheritance pattern was seen in those cases of Parkinson's disease where the disease starts before the age of 40 years in 46% of familial cases. On the bases of these results Parkinson disease was divided into etiologic groups: genetic, symptomatic, postencephalitic and idiopathic (Barbeau 1982, 1983; Pourcher, 1982, 1983).

The frequency of familial Parkinson disease was positive in 24% of the 100 consecutive cases in 6% of spouse controls. The crude segregation ratios for sibs and parents were same, and the cumulative lifetime risk was approaching 0.4 in another study of 22 non-consecutive Parkinson disease cases. This study supported a strong age factor in penetrance for autosomal dominant inheritance (Bonifati *et al.*, 1995).

1.6.1.3 Familial component

The occurrence of Parkinson disease in 772 living and deceased patients during the past 50 years was reviewed. In order to confirm whether the patients were more related to each other than random members of population over the past 11 centuries, they used genealogic information through extensive computerized database on 610,920 people of Iceland. They found that with the late-onset of the disease, a genetic component of Parkinson's disease also included the subgroup of 560 patients. These patients were more related to each other. They have not found highly penetrant

Mendelian inheritance, and between generations both early and late-onset was often skipped (Sveinbjornsdottir *et al.,* 2000).

After the completion of human genome project, the analysis and interpretation of huge biological data is being managed by the evolving science of bioinformatics. Bioinformatics is important for both management of data in medicine and modern biology.

The main tools of a bioinformatician are internet and computer software programs for the interpretation and analysis of biological data. Apart from the analysis and interpretation of genome sequence data, bioinformatics can now be used for a large number of other important tasks, including analysis of gene expression and variation, detection and prediction of gene regulation networks, prediction and analysis of protein and gene structure and function, simulation environments for whole cell modelling, complex modellng of gene regulatory dynamics and networks, and presentation and analysis of molecular pathways in order to understand gene-disease interactions. Different bioinformatics websites are being used for protein analysis and for the study of normal and disease related genes and their mutations, such as National Centre for Biotechnology Information *(www.ncbi.nlm.nih.gov)* which maintains bioinformatics databases and tools, Genebank *(www.ncbi.nlm.nih.gov/Genbank)* stores DNA sequences, Ensembl *(www.ensembl.org)* is a genome automatic annotation database, and SWISS-PROT *(www.expasy.org/sprot/)* which is an important protein sequence database for all organisms.

Genetic disorders occur as a result of mutations in certain genes, chromosomal abnormalities or many other factors which are inherited and may become lethal. Multifactorial disease is caused by many environmental and genetic factors. These genetic factors involves mutations in multiple genes interacting with environmental factors, therefore, these diseases are more difficult to analyze.

To understand the functional characteristics of proteins, 3D structural information is very important. 3D structures provides important data of proteins to understand their biological roles, their potential implications in different diseases, and for progress in drug designing (Bornot *et al.*, 2009)

Homology modeling can be perform to form the 3D structure of proteins for which the target sequence is available along with a template sequence with sequence identity >30%. The two homologous proteins will have similar structure as the protein folds are much more evolutionary conserved than their amino acid sequence.

Protein threading scans the database of known structures for unknown structure of the amino acid sequence. In each case assessment of the compatibility of the sequence to the structure is made by the scoring function, therefore predicts 3D models of the target sequence. Due to this compatibility analysis between 3D structures and linear protein sequences this method is also called 3D_1d fold recognition.

The functional characterization of proteins is a great challenge for medical, biochemical and computational sciences. Although it is now finally proved that the function of a protein can be predicted successfully through the computational approaches. Once the 3D structure is constructed the binding affinities can be identified by using docking and structure-based virtual screening of chemical libraries (Pierri *et al.,* 2010*)*

The study of genetic disorders is changing from the analysis of single gene to discovering cellular networks of genes, identification of their role in disease and understanding their complex interactions. Bioinformatics will guide to use the advantages brought by computational biology to both molecular biologists and clinical researchers. In upcoming decades the most successful clinical research team will be those who can switch between the use of sophisticated computational tools for analysis and the laboratory bench.

In the present study different bioinformatics tools were used for the protein structure prediction and drug designing of genes involved in Parkinson's disease. Prediction of the structure of some genes involved in Parkinson's disease has been made on the basis of sequence similarity. After predicting protein structures and their interacting networks, docking tools were used to discover new drug targets which can be used to develop new potential drugs for the treatment of the disease.

1.7 Objectives of the study

The main objectives of the research were to:

1. Analyze the results by using bioinformatics tools on the original gene sequence using combined strategy of candidate gene approach and genome-wide search.
2. Predict the 3D model of the original protein sequence of PARK2, PINK1, HTRA2 and ATP13A2 through homology modeling and threading.
3. Evaluation of the predicted 3D structure of the proteins.
4. Study protein-ligand interactions either to enhance protein functionality or stop its malfunctioning and develop a potential drug with the help of docking procedures.

Table 1.1 Known Genes and loci of Parkinson's disease

Parkinson Type	Gene loci	Gene
PARK1	4q22.1	SNCA
PARK4		
PARK5	4p14	UCHL1
PARK8	2q12	LRRK2
PARK11	2q37	GIGYF2
PARK13	2p12	HTRA2
PARK2	6q25.2-q27	Parkin
PARK7	1p36	DJ1
PARK6	25 cM or more centromeric to PARK7	
PARK14	22q13	PLA2G6
PARK15	22q12-q13	FBXO7
PARK3	2p13	-
PARK10	1p34-p32	-
PARK16	1q32	-
PARK12	X chromosome	-

2. MATERIALS AND METHODS

In the present study 3D model of four genes PARK2, PINK1, HTRA2 and ATP13A2 were predicted and these models were then evaluated. These four genes include PARK2, PINK1, HTRA2 and ATP13A2. Docking procedures were applied on four genes: DJ1, LRRK2, SNCA and UCHL1. For docking, protein models based on original sequence were required, in which we can insert mutations and then can find the best ligand binding positions. In the present study models were developed on the basis of similarity search and scanning existing databbases; they may not find the similar point of mutation as the models predicted in wet lab work. These protein models were taken from RCSB. Mutations were taken from OMIM (Online Mendelian Inheritance in Men). The genes selected for protein modeling and drug designing: both were taken from Gene Card and NCBI (National Center for Biotechnology Information). The tools used in this study with their web links were shown in table 2.1. The various steps involved in methodology of the present research work were shown in a form of flow diagram (Fig. 2.1).

2.1 SIMILARITY SEARCH

Similarity analysis has developed many small molecule based similarity methods that helps in mining large databases for novel molecules. Similarity searching is a powerful means of detecting distant relationships between different proteins, finds functions of novel sequenced genes and predicts new members of the gene family and thus helps to reach biologically meaningful conclusions. These methods form the basis in the field of computer-aided drug designing and protein modeling. For both BLAST and

FASTA similarity search the protein sequence were pasted on the fields given on the web portals and the server automatically found the similar sequence in the databases. For similarity search following tools were used.

2.1.1 BLAST

BLAST (Basic Local Alignment Search Tool) is an algorithm that searches local similarity regions between primary biological sequences, such as amino acid sequences of proteins or nucleotide sequences of DNA. The BLAST search enables comparison of query nucleotide and protein sequence with database of sequences or a library, and calculates statistical significance of the query sequences. It also helps to identify functional and evolutionary relationships between sequences to perform phylogenetic analysis.

2.1.2 FASTA

FASTA stands for **FAST-ALL** is a fast nucleotide comparison or fast protein comparison tool. FASTA is a high speed similarity search tool that uses substitution matrix with high level of similarity for local alignments. The high speed of this program to identify maximum matches before it attempts the more time consuming optimized search is because of the observed pattern of word hits. The speed and sensitivity is controlled by the specification of the size of the word.

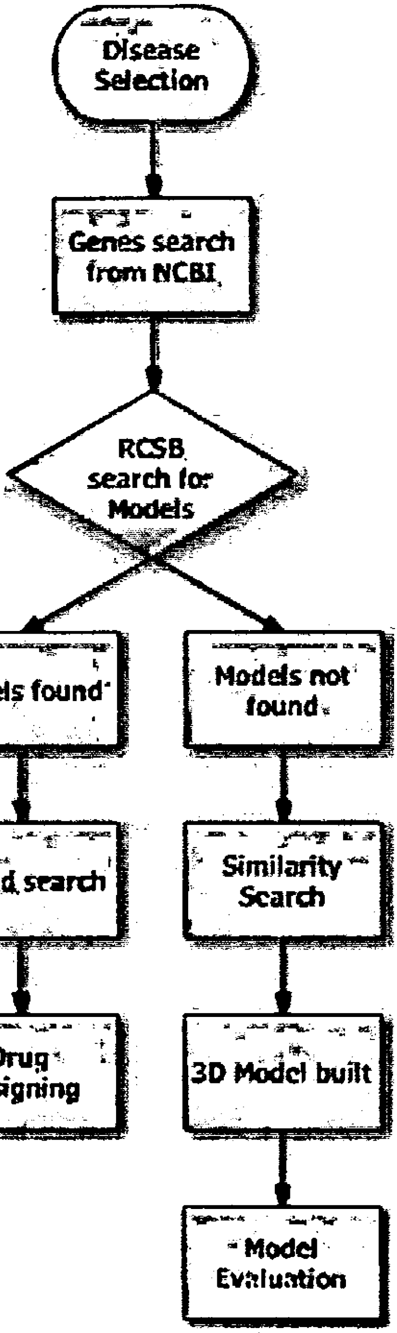


Figure 2.1 Flow chart showing various steps of methodology

2.2 PATTERN AND PROFILE SEARCH

There are many techniques of searching sequence databases to find common domains and motifs of biological significance proteins that categorize a protein into a separate family. Pattern is syntax that determines multiple combinations of possible residues within a protein sequence. Profile is the probabilistic generalization that is assign to every segment of protein that each of the 20 amino acid will occur again. The protein sequences were given to the server and the server itself performed the pattern and profile search. The PDB ids can also be used instead of the whole protein sequence to perform the search for the query sequence. Following tools were used for pattern and profile search. ELM takes sequence in FASTA format. In the SMART server we can give sequence identifier or accession number or the protein sequence itself to perform the search.

2.2.1 INTERPROSCAN

InterPro is an integrated resource for protein families, regions, domains, and sites. It combines a number of databases; ProDom, PROSITE, SUPERFAMILY, HAMAP, PRINTS, PANTHER, Pfam, PIRSF, SMART, Gene3D and TIGRFAMs. The InterPro entries are created from query protein sequence by scanning the member databases. The query sequence is compared to all the existing enteries in UniProtKB. Sequences from the same protein domain or family are grouped into unique InterPro entries having an abstract, unique accession number, and features of associated proteins, literature references and links to databases (McDowall *et al.*, 2011).

2.2.2 ELM

The ELM, Eukaryotic Linear Motif is a computational biological resource for predicting candidate functional sites in proteins independent of tertiary structure. Functional sites were identified as patterns using regular Expression rules. To reduce the false positive results context-based rules and logical filters were developed (Puntervoll *et al.,* 2003).

2.2.3 Scanprosite

ScanProsite is a tool with improved and new functions for detecting PROSITE signature matches in existing protein sequences. This web-based tool enhances the power of function prediction by detecting structural and functional intra-domain residues, based on profiles (Edouard *et al.,* 2006).

2.2.4 SMART

A Simple Modular Architecture Research Tool is used for identification, analysis and annotation of mobile domain architectures proteins and genes. The new release of SMART server contains more than 2 million protein sequences and 784 models of protein domains and 630 species (Letunic *et al.,* 2009).

2.3 POST TRANSLATIONAL MODIFICATION

Post translational modifications are chemical changes that are followed by translation. These mass modifications can be detected during analysis as glycosylation, phosphorylation, and sulfation etc. The post-translational modifications are important for the study of many diseases, such as diabetes, cancer and heart disease. SignalP was used for determining the post translation modifications.

2.3.1 SIGNALP

SignalP 3.0 server predicts the signal peptide cleavage sites in amino acid sequence of gram-negative prokaryotes and eukaryotes and gram-positive prokaryotes. By using many artificial neural networks and Hidden Markov Models, the server predicts the cleavage sites and signal and non-signal peptides in the protein sequences. The accuracy of the prediction of the cleavage site range from 6% to 17% (Jannick *et al.*, 2004). In the SignalP server protein file was given in FASTA format. SignalP server gave results in terms of scores and these scores are defined below.

2.3.1.1 Description of scores

The C-score is the cleavage score and at the cleavage site it should be significantly high.

The Y-max is the derivative of the C-score and S-score.

The S-mean is the average of S-score, between the amino acid at N-terminal and the amino acid with highest Y-max score.

The D-score is the average of S-mean and Y-max score.

2.4 TOPOLOGY PREDICTION

In membrane protein, topology prediction is the center process of the amino acid sequence to the fully folded 3D structure of protein. But generally topology is the connection between elements of secondary structure and those elements organized in 3-D structure, i.e., protein fold. For topology prediction, following tools were used. Both tools accepted the protein sequence and the server predicted the topology of various genes.

2.4.1 HMMTOP

The Hidden Markov Model for Topology Prediction is a tool that predicts the localization of transmembrane segments and topology of transmembrane proteins on the bases of the difference in amino acid distribution in different structural parts of these proteins other than by specific amino acid compositions of these parts. To search transmembrane topology with maximum likelihood among all the possible topologies of the given protein, a special hidden Markov model was developed (Tusnady and Simon, 2001).

2.4.2 TMPRED

TEMPRED is a program that finds putative transmembrane domains in query proteins and then finds the possible orientation of these domains. For scoring it uses multiple weight-matrices that are extracted by statistical analysis by using TMbase. TMbase is the collection of all annotated transmembrane proteins in Swissprot (Hofmann and Stoffel (1993).

2.5 PRIMARY STRUCTURE PREDICTION

Primary structure of biological molecules is the sequence of amino acids in the polypeptide chain which are linked through the chemical bonds called peptide bonds which is the covalent backbone of proteins. The sequence of amino acid starts from N-terminal to C-terminal amino acid of sequence that predicts the higher levels of molecular structure. ProtParam was employed to find the primary structure of proteins. Protparam also took protein sequence, sequence identifiers or Swiss-Prot/TrEMBL accession numbers and calculates the primary structure values.

2.5.1 PROTPARAM

ProtParam is a tool which uses databases of SwissProt or TrEMBL and it computes various physical and chemical parameters of the query protein stored in these databases. It computes atomic composition, molecular weight, amino acid composition, theoretical pl etc.

2.6 SECONDARY STRUCTURE PREDICTION

The secondary structure of proteins is the geometric shape of amino acids which is formed by the intramolecular and intermolecular hydrogen bonding. There are three types of secondary structure namely: alpha helix, beta sheets and random coils. Structure of the protein determines the function of protein. Structural analysis of proteins is helpful in homology modeling, drug designing and better understanding of protein-protein interactions. In the present study GOR4 was used for secondary structure prediction.

2.6.1 GOR 4

GOR4 (Garnier-Osguthorpe-Robson) uses information theory for the secondary structure prediction. GOR4 has accuracy rate of 64.4% for three state predictions. The predicted structure output is given in two formats; one is the eye friendly which is given as sequence in rows, H=helix, E=extended strand, C-coil and second at each amino acid position the probability of each predicted structure. The predicted structure is of highest probability which is compatible to at least four residues of predicted helix segment and at least two residues of predicted extended segment (Garnier *et al.,* 1996).

2.7 TERTIARY STRUCTURE PREDICTION

Tertiary structure is the combination of the secondary structure elements linked by turns and loops and forms folding of the total chain. It is the 3D structure of the entire polypeptide chain. The tertiary structure determines the function of the protein which is helpful in disease diagnosis. Tools used for tertiary structure prediction were Modeller and SAM-T08. Modeller was a four step process which gave ten models at the end predicted on the basis of similarity search, out of these ten models one was selected which had the lowest function value. SAM-T08 predicted model when the query sequence was given to the server. It could build model for the sequence length less than 700 amino acids.

2.7.1 Modeller

Modeller is a program of comparative protein three-dimensional structure modeling by satisfaction of spatial constraints. The user input an alignment of sequence which is to be modelled with known related structures and modeler automatically calculates its model. Beside from homology modeling, modeller can also perform database search, comparison of proteins structures, clustering, multiple alignments of protein sequences, optimization of various models of protein structure and de novo modelling of loops in protein structures with respect to flexibly defined objective function.

2.7.2 SAM-T08

SAM-T08 is a HMM-based protein structure prediction server that gives the final predicted 3D structure along with several useful results. Additional features include; prediction of local structure features, three multiple sequence alignments of putative homologs using different search procedures, residue-residue interaction predictions, E-values of template searches of PDB. Alignment and fold recognition was done in Protein Data Bank, and a complete three-dimensional model was created (Karplus *et al.*, 2009).

2.8 EVALUATION OF THE PREDICTED MODELS

With the increasing number of protein structures in Protein Data Bank, the understanding of the mechanism, function and evolution of the protein is also increasing. So the evaluation of the quality of protein structure is becoming very important. In response to this a large number of computer programs have been developed. In the present study two types of evaluation was performed:

1. Internal evaluation: It was done to evaluate the self-consistency of the predicted structure that the model satisfies the restraints that were used to calculate the model.
2. External evaluation: It involves the information which was not used in calculating the model.

For the evaluation of the predicted protein structures following tools were used:

2.8.1 WHATIF

WHATIF is an intelligent server for manipulating, displaying and analyzing proteins, small molecules, nucleic acids, and their interactions. WHATIF is mainly useful for validating and checking nomenclature and geometry of proteins, besides it allows more comprehensive structure evaluation. It has relational protein structure database incorporated. It has many integrated programs and the menu-driven operation of WHATIF makes it unique tool. In the present study, WHATIF was used to calculate Z-score and draw Ramachandran plots (Vriend *et al.,* 1990).

2.8.2 ProSA

PraSA is a user-friendly interface which is used frequently in protein structure validation. It displays energy plots and score that highlight potential problems in protein structure. The problematic part of the model was detected by a plot of local quality scores which were also highlighted in a 3D molecular viewer and quality scores of a protein were displayed keeping in view all known structures of protein (Wiederstein *et al.,* 2007).

2.8.3 PROCHECK

PROCHECK is software that calculates the stereochemical quality of the given protein structure, which was compared to the well-refined structures having the same resolution in the database and it also indicates its local, residue-by residue reliability. The input was a single PDB file that contains coordinates of the query protein structure. This program makes a number of plots and residue-by-residue listing (Laskowski *et al.,* 1993).

2.9 DOCKING PROCEDURES

In the field of molecular modeling, docking is the method to predict the preferred orientation of one molecule (ligand) to the second molecule (receptor), when bind with each other to form a stable compound. Docking predicts the binding affinities and activity of small molecules, thus plays an important role in rational drug designing. There are different tools used for docking procedures. In the present study AutoDock Vina was employed because it provides the easiest way to find the binding affinities in short time. It used MGLTools and Vina latest version; both of which were downloaded from the AuotDock Vina official website. But before running AutoDock Vina, a potential search was made for ligand finding. The molecular structures of the selected ligands was taken, which were in mol files. These files were converted to pdb files using Marvin Sketch. Then mutations were inserted in the selected 3D models of Parkinson disease using WHATIF server. These mutations were taken from different research papers available on the National Center of Biotechnology Information (NCBI).

2.9.1 AutoDock Vina

AutoDock Vina is a docking tool that was used for molecular docking and virtual screening with high performance and accuracy.

Table 2.1 Tools and their associated links

Name	Purpose	Web address
BLAST	Similarity search	http://blast.ncbi.nlm.nih.gov/Blast.cgi
FASTA3	Similarity search	www.ebi.ac.uk/fasta33/
Inter proscan	Pattern and profile search	www.ebi.ac.uk/Tools/InterProScan/
ELM	Pattern and profile search	http://elm.eu.org/links.html
Scan prosite	Pattern and profile search	www.expasy.ch/tools/scanprosite/
SMART	Pattern and profile search	http://smart.embl-heidelberg.de/
SignalP	Post translational modification	www.cbs.dtu.dk/services/SignalP/
HMMTOP	Topology prediction	www.enzim.hu/hmmtop/
TMpred	Topology prediction	www.ch.embnet.org/software/TMPRED_form.html
Protparam	Primary structure prediction	www.expasy.ch/tools/protparam.html
GOR4	Secondary structure prediction	http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html

Modeller 9v7	Tertiary structure prediction	http://salilab.org/modeller/
SAM-T08	Tertiary structure prediction	http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html
AutoDock Vina	Docking tool	http://vina.scripps.edu/
Whatif	Evaluation	http://swift.cmbi.kun.nl/whatif/
Prosa	Evaluation	https://prosa.services.came.sbg.ac.at/prosa.php
ProCheck	Evaluation	http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/

3. RESULTS

The main focus of the present study was to predict the 3D models of genes involved in Parkinson disease and in addition to it also to apply docking procedures on genes, whose models have been predicted previously and also to find the best ligands and develop potential drugs for the treatment of disease.

Four genes of Parkinson's disease were selected for protein modeling PARK2, PINK1, HTRA2, and ATP13A2 and genes selected for docking procedures were DJ1, LRRK2, SNCA, and UCHL1.

3.1 SIMILARITY SEARCHES

3.1.1 Basic Local Alignment Search Tool (BLAST)

Protein-Specific Iterated (PSI)-BLAST is the program that is useful in finding new members of a protein family or very distantly related proteins. BLAST results for PARK2 PNIK1, HTRA2, and ATP13A2 gene were shown in table 3.1, 3.2, 3.3 and 3.4 respectively. Sequence similarity results will be used in protein modeling; the templates were selected on the basis of maximum sequence similarity, having sequence identity >33%. The BLAST results gave the accession number of the protein with the maximum similarity, the name of the protein, its origin and expect value.

3.1.2 FASTA 3

This tool provides proteins database sequence similarity search using FASTA programs. FASTA3 results for the four genes were given in table 3.5, 3.6, 3.7 and 3.8. First column show the pdb accession number, second column shows the name of the similar protein, third column mentions from where the protein has been originated. In columns four and five percentage of sequence similarity was shown along with the E-values at the end.

Table 3.1 BLAST results for PARK2 gene

Accession No.	Protein Name	Origin of protein	E-value
1WX7|A	Chain A, Solution Structure Of The N-Terminal Ubiquitin-Like Domain In The Human Ubiquilin 3 (Ubqln3)	*Homo sapiens* (human)	0.030
1WJV|A	Chain A, Solution Structure Of The N-Terminal Zinc Finger Domain Of Mouse Cell Growth Regulating Nucleolar Protein Lyar	*Homo sapiens* (human)	2.1
2DQ7|X	Chain X, Crystal Structure Of Fyn Kinase Domain Complexed With Staurosporine	Homo sapiens (human)	9.7
2JMO|A	Chain A, Ibr Domain Of Human Parkin	*Homo sapiens* (human)	6e-40
3DPG|A	Chain A, Sgrai With Noncognate Dna Bound	*Streptomyces griseus*	6.3

Table 3.2 BLAST results for PINK1 gene

Accession No.	Protein Name	Origin of protein	E-value
2WEL|A	Chain A, Crystal Structure Of Su6656-Bound CalciumCALMODULIN-Dependent Protein Kinase 2 Delta In Complex With Calmodulin	*Homo sapiens* (human)	4e-09
1VJY|A	Chain A, Crystal Structure Of A Naphthyridine Inhibitor Of Human Tgf- Beta Type I Receptor	*Homo sapiens* (human)	3e-05

2JFM|A	Chain A, Crystal Structure Of Human Ste20-Like Kinase (Unliganded Form)	*Homo sapiens* (human)	1e-04
3GBZ|A	Chain A, Structure Of The Cmgc Cdk Kinase From Giardia Lamblia	*Giardia lamblia* ATCC 50803	3e-04

Table 3.3 BLAST results for HTRA2 gene

Accession No.	Protein Name	Origin of protein	E-value
1LCY|A	Chain A, Crystal Structure Of The Mitochondrial Serine Protease Htra2	*Homo sapiens* (human)	0.0
2PZD|A	Chain A, Crystal Structure Of The Htra2OMI PDZ DOMAIN BOUND TO A Phage-Derived Ligand (Wtmfwv)	*Homo sapiens* (human)	8e-52

Table 3.4 BLAST results for ATP13A2 gene

Accession No.	Protein Name	Origin of protein	E-value
2KIJ|A	Chain A, Solution Structure Of The Actuator Domain Of The Copper-Transporting Atpase Atp7a	*Homo sapiens* (human)	0.007
2RAR|A	Chain A, X-Ray Crystallographic Structures Show Conservation Of A Trigonal-Bipyramidal Intermediate In A Phosphoryl-Transfer Superfamily.	*Bacteroides thetaiotaomicron* VPI-5482	9.5
2O36|A	Chain A, Crystal Structure Of Engineered Thimet Oligopeptidase With Neurolysin Specificity In Neurotensin Cleavage Site	*Homo sapiens* (human)	7.4

2VOY|I	Chain I, Cryoem Model Of Copa, The Copper Transporting Atpase From Archaeoglobus Fulgidus	*Archaeoglobus fulgidus*	0.011
2DEW|X	Chain X, Crystal Structure Of Human Peptidylarginine Deiminase 4 In Complex With Histone H3 N-Terminal Tail Including Arg8	*Homo sapiens* (human)	10.0

Table 3.5 FASTA3 results for PARK2 gene

Accession No.	Protein Name	Origin of protein	Sequence identity %	Sequence similarity %	E-value
PRKN2_ HUMAN	E3 ubiquitin-protein ligase	*Homosapiens*	100.0	100	1.3e-208
A3FG77_ HUMAN	Parkin 2 OS=Homo sapiens G	*Homosapiens*	99.6	100	6.8e-208
A1IGZ9_ HUMAN	Parkin 2 OS=Homo sapiens P	*Homosapiens*	99.4	99.8	1.7e-207
B8YGJ6_ MACFA	Parkin OS=Macaca fascicula	*Macaca fascicularis* (Crab eating macaque)	97.4	98.7	8.2e-203
B9VH11_ MACFA	Parkin variant SV11bINS OS	*Macaca fascicularis* (Crab eating macaque)	92.2	93.5	2.8e-180

Table 3.6 FASTA3 results for PINK1 gene

Accession No.	Protein Name	Origin of protein	Sequence identity %	Sequence similarity %	E-value
PINK1_HUMAN	Serine/threonine-protein ki	*Homosapiens*	100.0	100	0
A5PJP5_BOVIN	PINK1 protein OS=Bos tauru	*Bos taurus* (Bovine).	86.1	91.3	2.1e-185
B5DFG1_RAT	Pink1 protein OS=Rattus norv	*Rattus norvegicus* (Rat).	82.1	92.1	4e-180
PINK1_MOUSE	Serine/threonine-protein ki	*Mus musculus* (Mouse).	81.6	91.7	9.2e-179
Q3U258_MOUSE	Putative uncharacterized p	*Mus musculus* (Mouse).	81.2	91.6	4.1e-178

Table 3.7 FASTA3 results for HTRA2 gene

Accession No.	Protein Name	Origin of protein	Sequence identity %	Sequence similarity %	E-value
HTRA2_HUMAN	Serine protease HTRA2, mito	*Homosapiens*	100.0	100.0	3.9e-165
A8K7G2_HUMAN	cDNA FLJ75762, highly simi	*Homosapiens*	99.8	99.8	1.5e-164
Q45FF7_CANFA	Protease serine 25 OS=Cani	*Canis familiaris* (Dog)	89.7	93.9	1.9e-144
HTRA2_BOVIN	Serine protease HTRA2, mito	*Bos taurus* (Bovine)	89.5	93.7	3.6e-144
B0BNB9_RAT	HtrA serine peptidase 2 OS=R	*Rattus norvegicus* (Rat)	87.3	93.0	5.8e-141

Table 3.8 FASTA3 results for ATP13A gene

Accession No.	Protein Name	Origin of protein	Sequence identity %	Sequence similarity %	E-value
AT132_HUMAN	Probable cation-transportin	*Homosapiens*	100.0	100	0
Q6S9Z9_HUMAN	Putative N-ATPase OS=Homo	*Homosapiens*	99.6	99.6	0
Q8N4D4_HUMAN	ATP13A2 protein (Fragment)	*Homosapiens*	100.0	100.0	0
A2AA78_MOUSE	ATPase type 13A2 OS=Mus mu	*Mus musculus* (Mouse)	85.4	94.7	0
AT132_MOUSE	Probable cation-transportin	*Mus musculus* (Mouse)	85.4	94.7	0

3.2 PATTERN AND PROFILE SEARCH

Pattern, families and domains of selected genes were also analyzed by using the following tools:

3.2.1 InterProscan

3.2.2 ELM

3.2.3 Scanprosite

3.2.4 SMART

3.2.1 InterProscan

InterPro is a molecular database of protein families, regions, domains, sites and repeats, in which we can scan features of known proteins and applies those features to new protein sequences.

3.2.1.1 InterProscan results of PARK2 gene

a) Ubiquitin and Ubiquitin supergroup

Ubiquitin and ubiquitin supergroup domains are highly conserved protein from protozoas to vertebrates, found in all eukaryotic cells and have 76 amino acids (Table 3.9 a). Both domains acts through its ubiquitinylation to other proteins, which is a post-translation attachment, where these change the location, function and trafficking of protein or it may destroy its target by 26S proteasome. In order to understand the specificity and regulation which are conferred upon these pathways, it is important to understand the components involved in protein ubiquitylation (E1s, E2s and E3s) is essential. Ubiquitin supergroup domain was shown in table 3.9b (Passmore and Barford, 2004).

b) Zinc finger, C6HC-type

Zinc fingers domains are small DNA-binding peptide motifs that recognize and bind to specific DNA sequences. For the construction of larger protein domains they can be used as modular building blocks (Table 3.9 c) (Klug *et al.*, 2002).

Some of these domains bind zinc and other may not, they prefer binding other metals like iron. They are mostly present in eukaryotes. Structural studies showed that they are considerable divers in terms of protein partners, affinities, and binding modes (Gamsjaeger *et al.*, 2007)

On the basis of difference in the nature and arrangement of zinc-binding residues, 14 different classes of Zn fingers have been made (Matthews *et al.* 2008)

c) Parkin and Parkin C-terminal domain

Parkin is the causative gene of AR-JP and it encodes a 52-kDa of protein which is composed of three parts: the carboxy-terminal RING-finger box, the amino-terminal ubiquitin (Ub)-like domain (Ubl) and the linker region which is the connection between two domains. The functional role of the N-terminal Ubl domain is not known whereas the C-terminal RING box serves as a recruiting motif for ubiquitin-conjugating enzymes (E2 enzymes), such as Ubc4, Ubc7, UbcH7 and UbcH8 . In the early-onset Parkinsonism the number of mutations in the parkin gene has been increased. Parkin binds through the C-terminal ring-finger domain to the E2 ubiquitin-conjugating human enzyme 8 (UbcH8). Parkin has ubiquitin-protein ligase activity in the presence of UbcH8. (Table 3.9 d,e). And unintegreted domains were shown in table 3.9f.

3.2.1.2 InterProscan results of PINK1 gene

a) Serine/threonine-protein kinase active site, Protein kinase catalytic domain and Serine/threonine-protein kinase-like domain

In most cellular activities protein phosphorylation plays a key role. Resulting in conformational change that affect protein function, it catalyses the transfer of gamma phosphates from NTP (often ATP) to one or more amino acid residues in a protein substrate side chain. The reverse process is catalyse by phosphoprotein phosphatases and protein kinases. Protein kinase is classified according to the substrate specificity into three main classes.

1. Dual specific protein kinases
2. Serine/threonine-protein kinases
3. Tyrosine-protein kinases

Protein kinases plays key roles in cellular processes. Several structures of protein kinases catalytic subunits are highly conserved. Eukaryotic enzymatic protein kinases share a conserved catalytic core with both tyrosine protein kinases and serine/threonine. There are a large number of conserved regions in this domain. In the vicinity of lysine residue there is a glycine-rich stretch of residues and this is present at the extreme of catalytic domain in N-terminal and is involved in ATP binding. The center of the domain is conserverd for aspartic acid residue which plays an important role for the catalytic activity of the enzyme. Enkaryotes, viruses and some bacteria include this catalytic domain (Table 3.10 a, b, c).

b) Protein kinase-like domain

In eukaryotic cells the phosphotransfer reaction is catalyzed by protein kinase to most fundamental signaling and regulatory processes. The core in the catalytic subunit is common to both tyrosin protein kinase and serine/treonin. The domain have catalytic apparatus in the inter-lobe cleft and it also contains nucleotide-binding site. Structurally it has functional and structural similarities to phosphoinositol phosphate kinase and ATP-grasp fold. The 3D fold of the catalytic domain and domains found in other proteins are similar (Table 3.10d). And unintegrated domains were shown in table 3.10e.

3.2.1.3 InterProscan results of HTRA2 gene

a) Peptidase S1/S6, chymotrypsin/Hap domains

Peptidases and peptidases homologues are grouped into clans and families in MEROPS databases. Based on common structural fold clans are groups of families having common ancestry.

1. Two letters were used to identify each clan; the first letter represents the catalytic type of the family which is included in the clan.
2. The first character in the peptidase families represents the catalytic type and they were identified by their catalytic type.

Sometimes the structural protein fold that identifies the clan or family may lost its catalytic activity, yet the function that recognize the protein and its binding is still present.

In viruses, bacteria and eukaryotes the proteolytic enzymes are ubiquitous, which exploit serine in their catalytic activity. On the basis of structural similarity and other functional evidence into clans, more than 20 families of serine protease have been identified. This suggests different evolutionary origins for the serine peptidase. Other then difference in evolutionary origins, the reaction mechanism is same for several peptidases. Chymotrypsin, subtilisin and carboxypeptidase C have a catalytic triad of serine as a nucleophile, histidine as a base and aspartate as an electrophile. Between families the catalytic residues have same geometric orientation and different protein folds. The clan relationship is shown by the linear arrangements of catalytic residue.

In actinomycetes, trypsin-like enzymes are found in genera Sreptomyces and Saccharopolyspora and also in fungus but the chymotrypsin family were almost totally conserved to animals. These enzymes are secreted inherently and are synthesized with signal peptide which targets these enzymes to secretory pathways. In animals these enzymes are either retained in leukocyte granules or secreted as packed vesicles for regulated secretion.

The haemophilus adhesion and penetration (Hap) family are proteins and important in the interaction with human epithelial cells. The binding domain is in C-terminal region and the N-terminal domain has serine protease activity (Table 3.11a)

b) PDZ/DHR/GLGF

In plants, bacteria, vertebrates, yeasts, and insects PDZ domains were found in diverse signalling proteins. PDZ domains may occur in one or multiple copies and are almost always found in cytoplasmic proteins. They either bind to the internal peptide sequences or the carboxyl-terminal of protein sequences. There is constitutive interaction between a PDZ domain and its target. These domains are associated with the plasma membrane which is a high concentration compartment of phosphatidylinositol 4, 5-bisphosphate (PIP2). Direct interaction between a subset of class II PDZ domains and PIP2 were found.

There are 80 to 90 amino acids in these domains, which consists of 2 alpha-helices, A and B which are arranged in globular structure and 6 beta-strands (A to F). In the elongated surface groove peptide binding of the ligand takes place as an anti-parallel B-strand which interacts with the B-strand and the B helix. At the end of a peptide the PDZ domains allows a free carboxylate group binding through a carboxylate-binding loop between the beta-A and beta-B strands (Table 3.11b).

c) Peptidase S1C, HrtA/DegP2/Q/S domain

The non-peptidase homolog and serine peptidases belong to the family S1 MEROPS peptidase and subfamily S1C.

At temperatures above 42 degrees, other members of this group are *E.coli* htrA gene product which is essential for bacterial survival and for the digestion of mis-folded protein in the periplasm. From *E-coli* mature DegP has 448 residues. Typical of a leader peptide the protein has an N-terminal sequence. Structurally bacterial HtrA belongs to the family of age-forming protease and is a serine protease.

The *E. coli* genes degQ and degS are homologues to DegP protease and encodes 455 and 355 protein residues. The DegQ protein is processed by the removal of 27-

residue N-terminal signal sequence and has the properties of serine endopeptidase. Deletion studies show that in vivo DegQ functions like DegP as a periplasmic protease (Table 3.11c).

d) Serine/cysteine peptidase, trypsin-like domain

Characteristic molecular topologies were shown by cysteine peptidases, which can be seen in their 3-D and 2-D structures. These peptidases have nucleophile which is sulphydryl group of a cysteine residue. Cysteine proteases were classified into clans and then into families on the basis of triad (Table 3.11d). Unintegrated domains were shown in table 3.11e.

3.2.1.4 InterProscan results of ATP13A2 gene

a) Cof protein

The Haloacid Dehydrogenase (HAD) superfamily includes:

- Phosphatases
- Phosphonatases
- Dehalogenases
- P-type ATPases
- Phosphomannomutases
- Beta-phosphoglucomutases

These uncharacterized proteins are involved in a various processes which may include amino acid biosynthesis or detoxification. These proteins are mostly found in bacteria and many paralogs are present in many species. For example *E. coli* has 6 of these proteins. Similarity in sequence shows that these enzymes are mostly phosphatases and works on phosphorylated sugars.

The enzyme YbiV has been found from *E. coli.* This enzyme acts as catalyst and hydrolyses sugar and inorganic phosphate from sugar phosphate and also hydrolysis ribose-5-phosphate and glucose-6-phosphate. These proteins are monomers having two domains, an alpha-beta hydrolyse and alpha-beta domains. Between the two domains at the interface the active site is present in negatively charged cavity (Table 3.12a).

b) ATPase, P-type, K/Mg/Cd/Cu/Zn/Na/Ca/Na/H-transporter

i. CATATPASE

In prokaryotes and eukaryotes P-type or E1-E2-type ATPases are from superfamily of cation transport enzymes. This group of enzymes can be divided into four major classes:

- Ca-++transporting ATPases
- K+-/Na+ and gastric H+/K+- transporting ATPases
- Plants, fungi and lower eukaryotic plasma membrane H+-transporting ATPases which are also called proton pumps.
- All bacterial P-type ATPases, except Mg++ATPasesof S.typhimurium, it has sequence which is similar to eukaryotes.

CATATPASE gives signature for the superfamily of cation-transporting ATPases and is a 6-element fingerprint which is derived from an initial alignment of 18 sequences and motifs were derived from well-conserved regions from the whole family.

ii. ATPases: Cation transporting, P-type unknown pump specificity (type V), P-type ATPase-associated region, P-type phosphorylation site

ATPases combine with ATP and hydrolyse or synthesis the transport of protons as ATPases is membrane-bound enzymes. There are different types of ATPases on the bases of their function, structure and the ions which they transport.

- F-ATPases are prime producers of ATP by using the proton gradient which is generated by phosphorylation in mitochondria, photosynthesis in chloroplast and in bacterial plasma membrane.
- V-ATPases are found in vacuoles of eukaryotes. To transport solutes and lower pH in organelles it catalyses ATP hydrolysis.
- A-ATPases functions as F-ATPases and are mostly found in Archaea.
- P-ATPases are found in bacteria, organelles and eukaryotic plasma membrane. They transport different ions across membranes. H+, K+, Mg+, Ca+, Ag+, Cu+ and Ag2+, Zn2+, Co2+, Pb2+, Ni2+, Cd2+, Cu2+.

E-ATPases are enzymes of cell-surface and they hydrolyse many NTPs, including extracellular ATP. P-type unknown pump specificity (type V) ATPases were shown in table 3.12b. ATPase-associated region were shown in table 3.12c. P-type phosphorylation site are shown in table 3.12d.

iii. ATPase_P-type: ATPase, P-type (transporting)

The P-type ATPases are family of trans-membrane transporters that acts on charged substances. During the course of reaction they form a phosphorylation intermediate. P-ATPases consists of single subunit and ion translocation pathway.

P-type ATPases are found in higher organisms in many copies. Phylogenetic analysis shows that on the bases of substrate specifications P-type ATPase subfamily is divided into many groups (Table 3.12 e).

c) Haloacid dehalogenase-like hydrolase

Structurally this family is not similar to alpha/beta hydrolase family. This family includes:

- L-2-haloacid
- Dehalogenase
- Epoxide hydrolase and
- Phosphatases

This family has two domains, one is a four helix inserted bundle, and the remaining fold is made up of the core alpha/beta domain (Table 3.12 f). And unintegrated domains were shown in table 3.12 g.

3.2.2 ELM (Eukaryotic Linear Motif Resource)

ELM is a computational resource for identifying functional sites in eukaryotic proteins. By regular expressions putative functional sites were identified and the diagrammatical representation for all the four genes respectively along with the key was shown in figure 3.1, 3.2, 3.3 and 3.4. In which the first row showed domains in the genes, second row represented globular plot, third row showed the 2D structures and fourth row gave diagrammatic view of motifs. By comparing with other databases ELM gave start and end number of the amino acid sequence representing globular, transmembrane domains and signal peptide domains (Table 3.13, 3.14, 3.15 and 3.16). It also showed description of domains, matched sequences, positions of the sequences, and amino acid numbers in Jmol, functions of the domain and cellular locations of the domains presented in table 3.17, 3.18, 3.19 and 3.20.

3.2.3 Scanprosite

Prosite consists protein families, domains and functional sites and patterns and profiles to find proteins. Patterns with a high probability of occurrence were presented diagrammatically in figure 3.5, 3.6, 3.7 and 3.8 for PARK2, PINK1, HTRA2 and ATP13A2 genes respectively. Scanprosite results showed N-glycosylation site, N-myristoylation site, and phosphorylation sites as Casein kinase II, Protein kinase C, cAMP- and cGMP-dependent protein kinase, and Tyrosine kinase in PARK2, PINK1 genes. Tyrosine kinase phosphorylation site was replaced by Amidation site in HTRA2 gene and was not present in ATP13A2 gene. These sites were shown in the first column. Second column gave the Prosite accession number, third column gave the length of the site, fourth column gave the sequence and fifth column showed the consensus patterns (Table 3.21, 3.22, 3.23 and 3.24).

a) N-glycosylation site

N-glycosylation sites are specific Asn-Xaa-Ser/Thr sequences. The presence of this consensus tripeptide sequence is not enough to suggest that asparagine is glycosylated as folds of the protein and it plays an important role in N-glycosylation regulation. N-glycosylation is inhibited due to the presence of proline between Asn and Ser/Thr. It has been studied recently that 50% of the sites that have a proline C-terminal to Ser/Thr are not glycosylated. The pattern Asn-Xaa-Cys have also been reported in few glycosylation sites.

b) N-myristoylation site

Myristate is a C14-saturated fatty acid; a large number of eukaryotic proteins on their N-terminal residue are acylated through an amide linkage by the covalent addition of C14-saturated fatty acid.

The enzyme myristoyl CoA: protein N-myristoyl transferase (NMT) is responsible for this modification. It was derived by various studies using synthetic peptides from the sequence of known N-myristoylated proteins, which are as following:

- Glycine should be at N-terminal.
- Only uncharged residues are allowed at position 2.
- Most of the residues are allowed at positions 3 and 4.
- Small uncharged residues are allowed at position 5 (Ala, Ser, Thr, Cys, Asn and Gly). Serine is preferred.
- Proline is not allowed at position 6.

c) Casein kinase II phosphorylation site

Casein kinase II (CK-2) is a protein serine/threonine kinase and it acts independently of cyclic nucleotides and calcium. CK-2 has the ability to phosphorylate many proteins. This enzyme acts on specific substrates as:

1. Ser is preferred over Thr.

2. Either Asp or Glu (an acidic residue) must be there.

3. At positions +1, +2, +4, and +5 other acidic residues increase the phosphorylation rate.

4. As the provider of acidic determinants Asp is preferred to Glu.

5. Basic residue decreases, while an acidic one will increase the phosphorylation rate at the N-terminal of the acceptor site.

d) Protein kinase C phosphorylation site

Protein kinase C shows preference for serine or threonine residues phosphorylation present near C-terminal basic residue. At the N- or C-terminal of the target amino acid the presence of other basic residues enhances phosphorylation reaction of Vmax and Km.

e) cAMP- and cGMP-dependent protein kinase phosphorylation site

Both of these kinases shows preference for the serine or threonine residues phosphorylation present near to at least two consecutive N-terminal basic residues with many exceptions.

f) Tyrosine kinase phosphorylation site

Generally tyrosine protein kinases substrates are characterized by a seven residues lysine or an arginine to the N-terminal side of the phosphorylated tyrosine. At either three or four residues to the N-terminal side of the tyrosine Asp or Glu is sometimes present with many exceptions.

g) Amidation site

The active peptides and the precursor of hormones which are amidated at C-terminal, are followed by a glycine residue and amide group is provided by this residue and basic residues Arg or Lys which acts as an active peptide precursor cleavage site. In fact all amino acids can be amidated, but charged residues such as Asp or Arg are much less reactive and neutral hydrophobic residues are good substrates such as Val or Phe. In unicellular organisms and in plants C-terminal amidation is not present.

3.2.4 SMART

Simple modular architecture research tool (SMART) is an online tool (http://smart.embl.de/) which is very useful for the annotation and identification of protein domains. In both proteins and genes, for the exploration and comparative study of domain architectures it provided a user-friendly platform (Letunic *et al.*, 2008). The diagrammatic view of the SMART result was shown in figure 3.9, 3.10, 3.11 and 3.12. SMART results also gave start and end amino acid numbers of the domains and their E-values as shown in table 3.25, 3.26, 3.27 and 3.28.

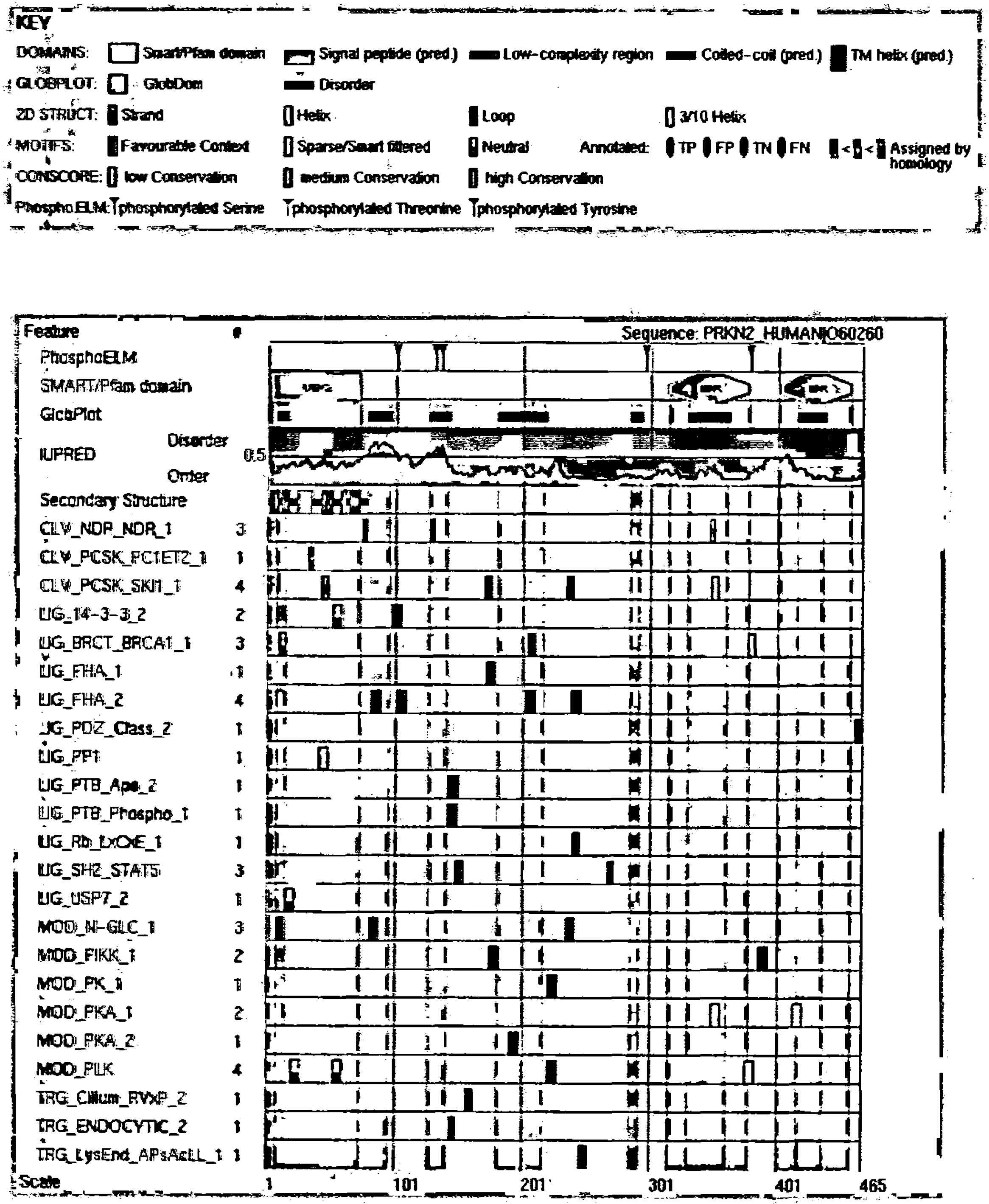


Figure 3.1 ELM result of PARK2 gene

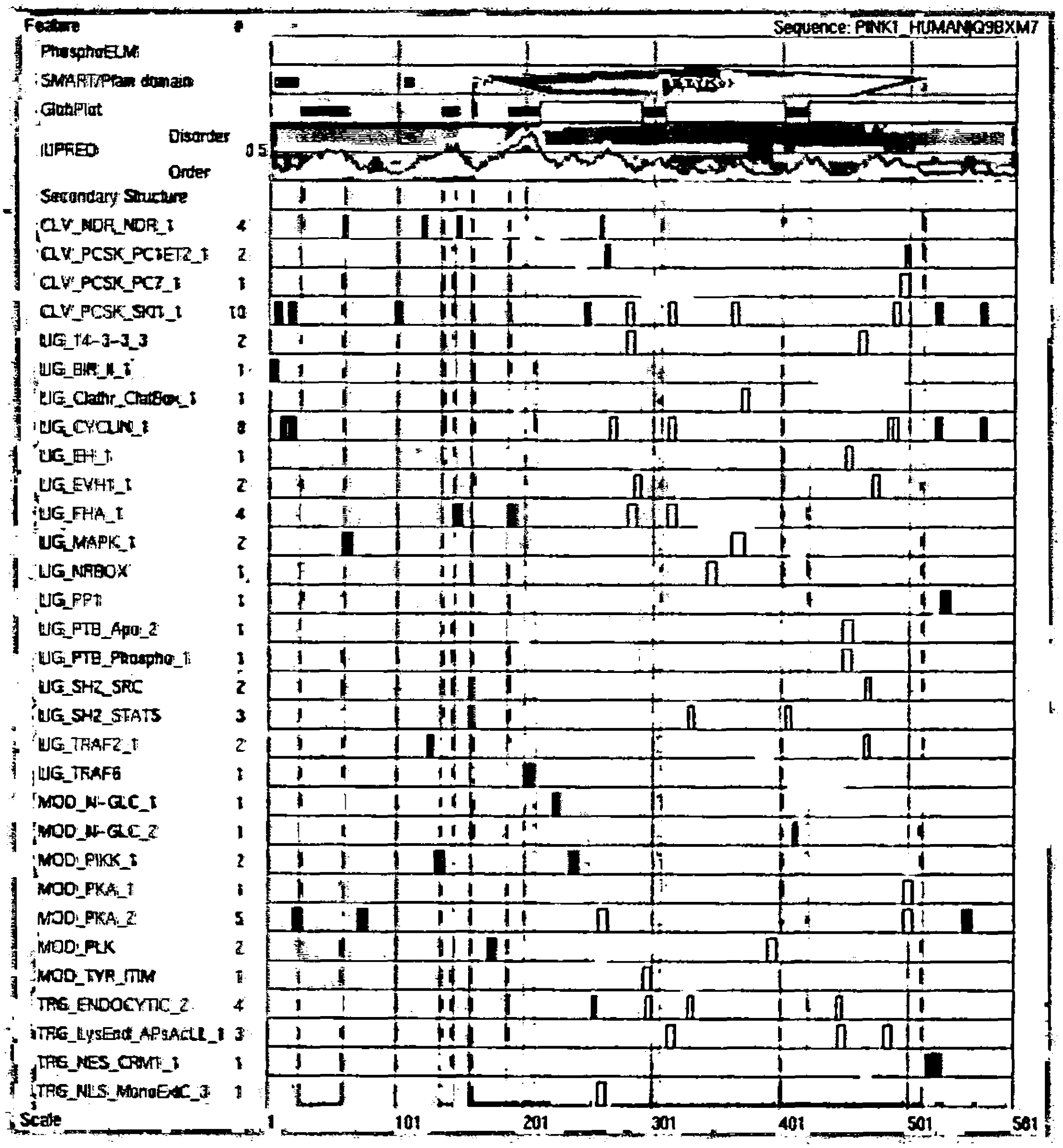


Figure 3.2 ELM result of PINK1 gene

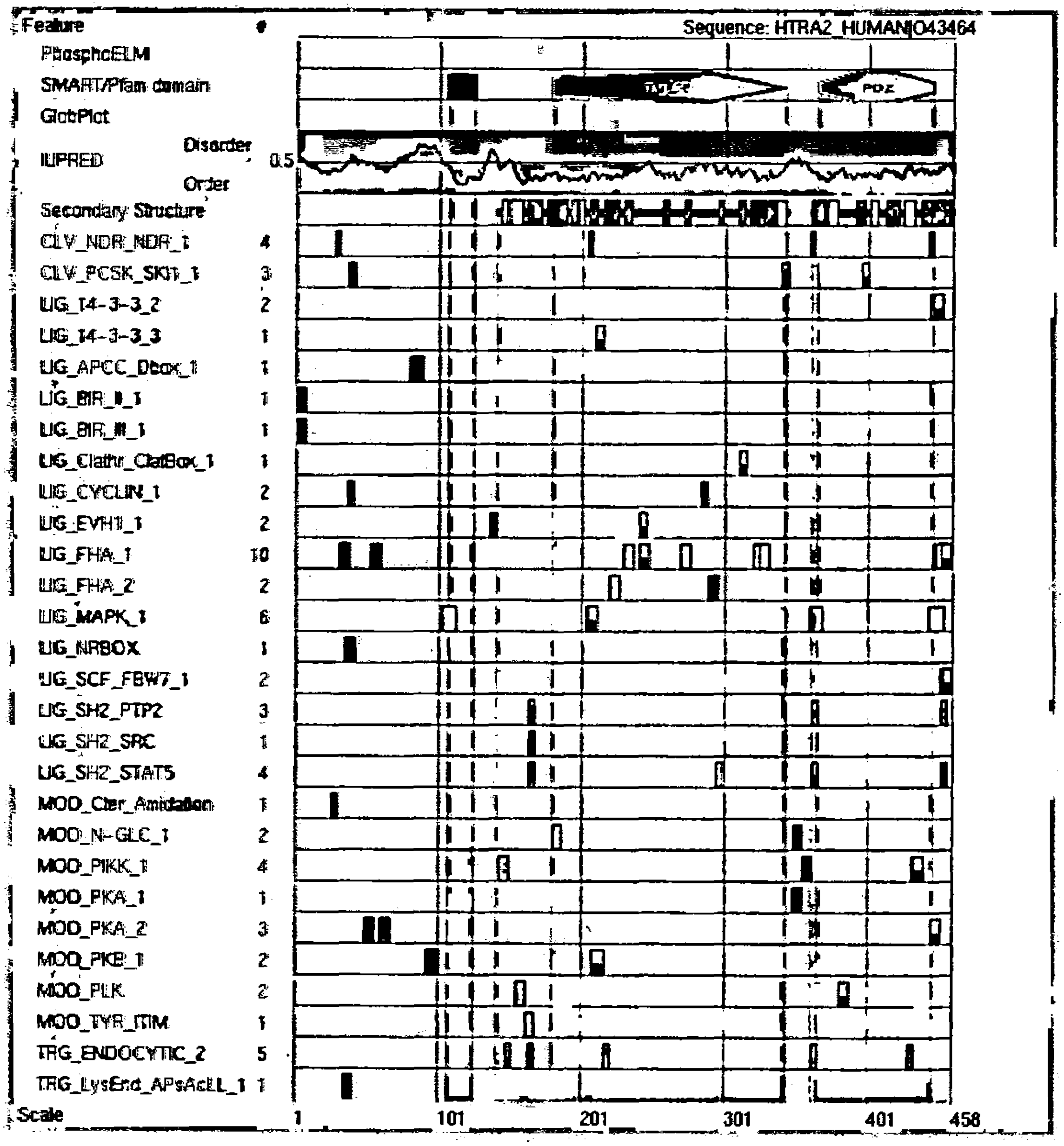


Figure 3.3 ELM result of HTRA2 gene

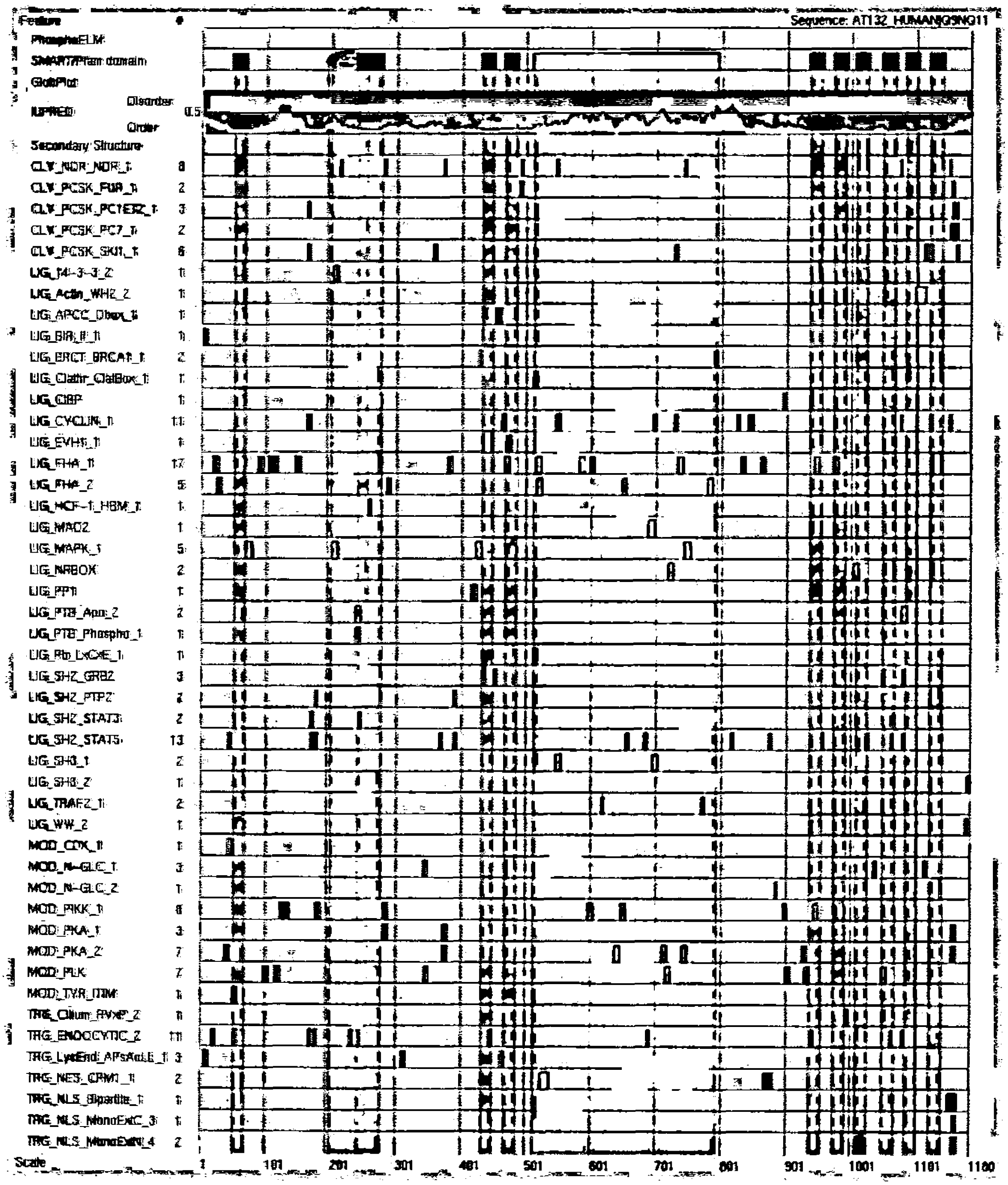


Figure 3.4 ELM result of ATP13A2 gene

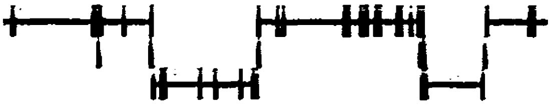


Figure 3.5 Scanprosite result of PARK2 gene

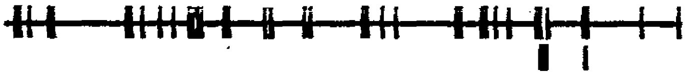


Figure 3.6 Scanprosite result of PINK1 gene

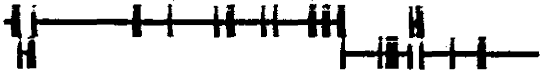


Figure 3.7 Scanprosite result of HTRA2 gene

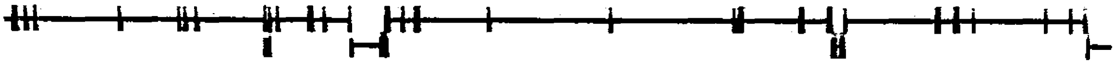


Figure 3.8 Scanprosite result of ATP13A2 gene

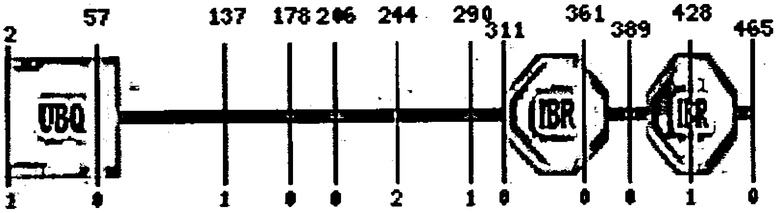


Figure 3.9 SMART result of PARK2 gene

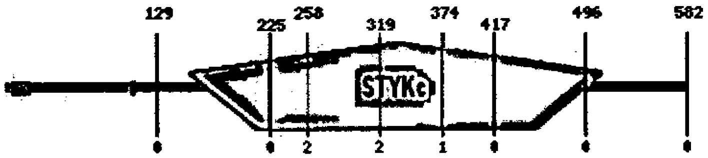


Figure 3.10 SMART result of PINK1 gene

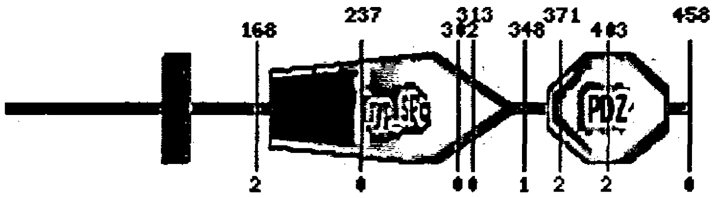


Figure 3.11 SMART result of HTRA2 gene

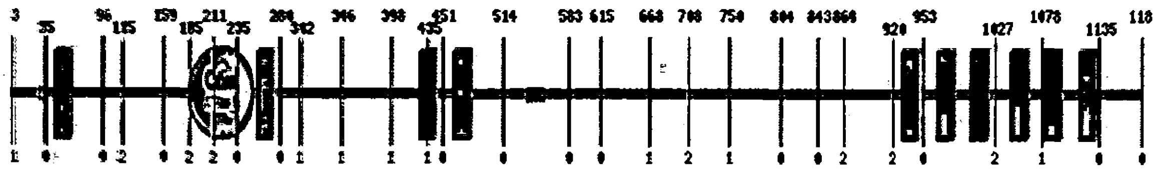


Figure 3.12 SMART result of ATP13A2 gene

Table 3.9(a) Ubiquitin domains of PARK2 gene

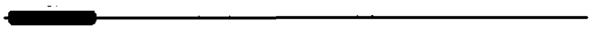
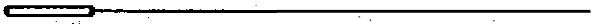
S. No.	Accession No.	Domain name	Domain
1.	PF00240	Ubiquitin	
2.	SM00213	UBQ	

Table 3.9(b) Parkin, C-terminal domains of PARK2 gene

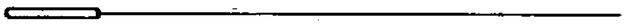
S. No.	Accession No.	Domain name	Domain
1.	PS50053	UBIQUITIN_2	

Table 3.9(c) Zinc finger, C6HC-type domains of PARK2 gene

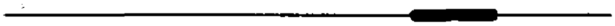
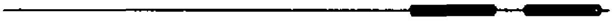
S. No.	Accession No.	Domain name	Domain
1.	PF01485	IBR	
2.	SM00647	IBR	

Table 3.9(d) Parkin domains of PARK2 gene

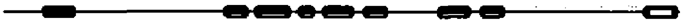
S. No.	Accession No.	Domain name	Domain
1.	PR01475	PARKIN	

Table 3.9(e) Parkin, C-terminal domains of PARK2 gene

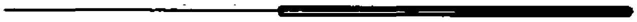
S. No.	Accession No.	Domain name	Domain
1.	PTHR116 85:SF2	parkin (ubiquitin e3 ligase prkn)	

Table 3.9(f) Unintegrated domains of PARK2 gene

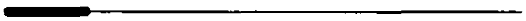
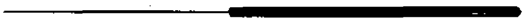
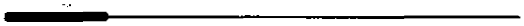
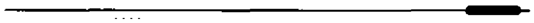
S. No.	Accession No.	Domain name	Domain
1.	G3DSA:3.10.20.90	no description	
2.	PTHR11685	ARIADNE RING ZINC FINGER	
3.	SSF54236	Ubiquitin-like	
4.	SSF57850	RING/U-box	

Table 3.10(a) The Protein kinase, catalytic domain of PINK1 gene

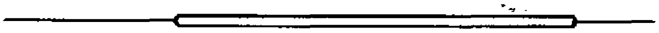
S. No.	Accession No.	Domain name	Domain
1.	PS50011	PROTEIN_KINASE_DOM	

Table 3.10(b) The Serine/threonine-protein kinase, active site of PINK1 gene

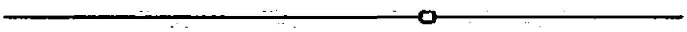
S. No.	Accession No.	Domain name	Domain
1.	PS00108	PROTEIN_KINASE_ST	

Table 3.10(c) The Serine/threonine-protein kinase-like domain of PINK1 gene

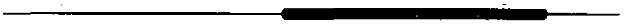
S. No.	Accession No.	Domain name	Domain
1.	PR017442	Pkinase	

Table 3.10(d) The Protein kinase-like domain of PINK1 gene

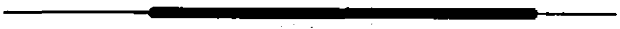
S. No.	Accession No.	Domain name	Domain
1.	SSF56112	Protein kinase-like (PK-like)	

Table 3.10(e) Unintegrated domains of PINK1 gene

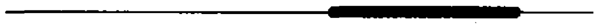
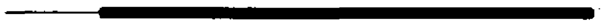
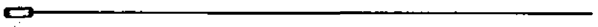
S. No.	Accession No.	Domain name	Domain
1.	G3DSA:1.10.510.10	no description	
2.	PTHR22972	SERINE/THREONINE PROTEIN KINASE	
3.	SignalP	signal-peptide	

Table 3.11(a) Peptidase S1/S6, chymotrypsin/Hap domains of HTRA2 gene

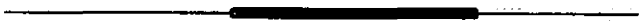
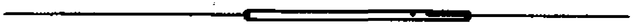
S. No.	Accession No.	Domain name	Domain
1.	PF00089	Trypsin	
2.	SM00020	Tryp_SPc	

Table 3.11(b) PDZ/DHR/GLGF domains of HTRA2 gene

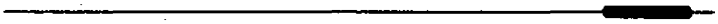
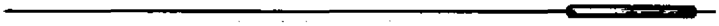
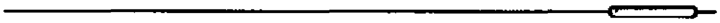
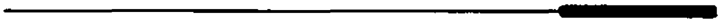
S. No.	Accession No.	Domain name	Domain
1.	PF00595	PDZ	
2.	SM00228	PDZ	
3.	PS50106	PDZ	
4.	SSF50156	PDZ domain-like	

Table 3.11(c) Peptidase S1C, HrtA/DegP2/Q/S domains of HTRA2 gene

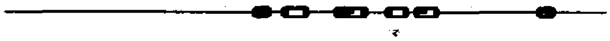
S. No.	Accession No.	Domain name	Domain
1.	IPR001940	PROTEASES2C	

Table 3.11(d) Serine/cysteine peptidase, trypsin-like domains of HTRA2 gene

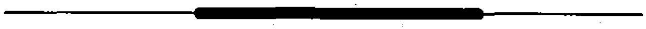
S. No.	Accession No.	Domain name	Domain
1.	IPR009003	Trypsin-like serine proteases	

Table 3.11(e) Unintegrated domains of HTRA2 gene

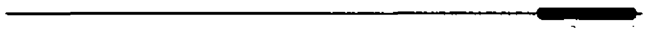
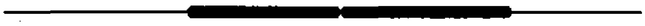
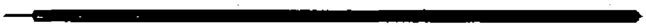
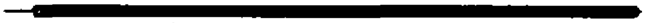
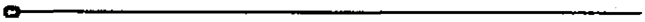
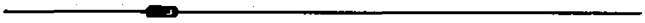
S. No.	Accession No.	Domain name	Domain
1.	G3DSA:2.30.42.10	no description	
2.	G3DSA:2.40.10.10	no description	
3.	PTHR22939	serine protease family s1c htra-related	
4.	PTHR22939:SF12	serine protease htra2	
5.	SignalP	signal-peptide	
6.	tmhmm	transmembrane_regions	

Table 3.12(a) The Cof protein domain of ATP13A2 gene

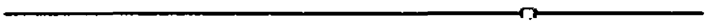
S. No.	Accession No.	Domain name	Domain
1.	IPR000150	COF_2	

Table 3.12(b) The ATPase, P-type, unknown pump specificity (type V) domain of ATP13A2 gene

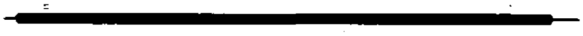
S. No.	Accession No.	Domain name	Domain
1.	IPR006544	P-ATPase-V:P-type ATPase of unknown pumps	

Table 3.12(c) The ATPase, P-type, ATPase-associated region domain of ATP13A2 gene

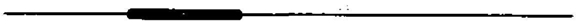
S. No.	Accession No.	Domain name	Domain
1.	IPR008250	E1-E2_ATPase	

Table 3.12(d) The ATPase, P-type phosphorylation site domain of ATP13A2 gene

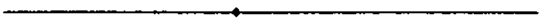
S. No.	Accession No.	Domain name	Domain
1.	IPR018303	ATPASE_E1_E2	

Table 3.12(e) The ATPase, P-type, K/Mg/Cd/Cu/Zn/Na/Ca/Na/H-transporter domain of ATP13A2 gene

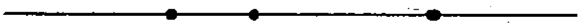
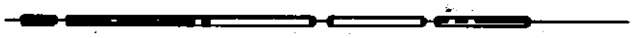
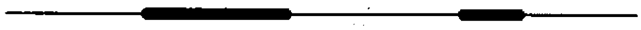
S. No.	Accession No.	Domain name	Domain
1.	PR00119	catatpase	
2.	PTHR11939	cation-transporting atpase	
3.	TIGR01494	ATPase_P-type: ATPase, P-type transporting	

Table 3.12(f) The Haloacid dehalogenase-like hydrolase domain of ATP13A2 gene

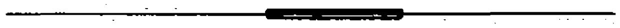
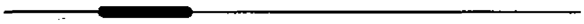
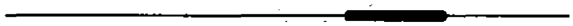
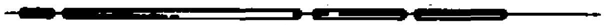
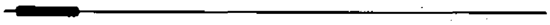
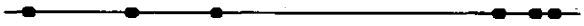
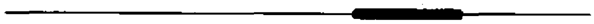
S. No.	Accession No.	Domain name	Domain
1.	IPR005834	Hydrolase	

Table 3.12(g) Unintegrated Domains of ATP13A2 gene

S. No.	Accession No.	Domain name	Domain
1.	G3DSA:2.70.150.10	no description	
2.	G3DSA:3.40.50.1000	no description	
3.	PTHR11939:SF58	N-TYPE ATPASE	
4.	PF12409	P_ATPase	
5.	tmhmm	transmembrane_regions	
6.	SSF56784	HAD-like	

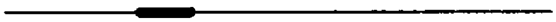
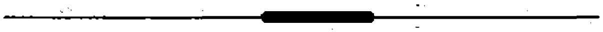
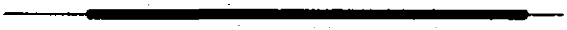
7.	SSF81653	Calcium ATPase, transduction domain A	
8.	SSF81660	Metal cation-transporting ATPase, ATP-binding domain N	
9.	SSF81665	Calcium ATPase, transmembrane domain M	

Table 3.13 ELM results showing Globular domains/ TM domains and signal peptide for PARK2 gene

Domain	Start	End
UBQ	1	72
IBR	313	377
IBR	401	457

Table 3.14 ELM results showing Globular domains/ TM domains and signal peptide for PINK1 gene

Domain	Start	End
STYKc	156	511

Table 3.15 ELM results showing Globular domains/ TM domains and signal peptide for HTRA2 gene

Domain	Start	End
transmembrane_domain	105	124
Tryp_SPc	178	342
PDZ	363	445

Table 3.16 ELM results showing Globular domains/ TM domains and signal peptide for ATP13A2 gene

Domain	Start	End
transmembrane_domain	45	67
Cation_ATPase_N	188	256
Pfam:Hydrolase	507	899
transmembrane_domain	931	953

transmembrane_domain	968	990
transmembrane_domain	1002	1024
transmembrane_domain	1044	1066
transmembrane_domain	1079	1101
transmembrane_domain	1116	1138

Table 3.17 ELM results description for PARK2 gene

ELM Name	Instances (Matched Sequence)	Positions	View in Jmol	ELM Description	Cell Compartment
CLV_NDR_NDR_1	WRK SRK	74-76 127-129	74-76 -	N-Arg dibasic convertase (nardilysine) cleavage site (Xaa-|-Arg-Lys or Arg-|-Arg-Xaa)	extracellular, Golgi apparatus, cell surface
CLV_PCSK_PC1_ET2_1	KRQ	32-34	32-34	NEC1/NEC2 cleavage site (Lys-Arg-|-Xaa)	Golgi membrane, extracellular, Golgi apparatus
CLV_PCSK_SKI1_1	RQATL RNITC	170-174 234-238	- -	Subtilisin/kexin isozyme-1 (SKI1) cleavage site ([RK]-X-[hydrophobic]-	endoplasmic reticulum, endoplasmic reticulum lumen,

				[LTKF]-|-X)	Golgi apparatus
LIG_14-3-3_2	RNDWTVQ REPQSLT	51-57 97-103	51-57 -	Longer mode 2 interacting phospho-motif for 14-3-3 proteins with key conservation RxxxS#p.	nucleus, mitochondrion, cytosol, internal side of plasma membrane
LIG_BRCT_BRCA1_1	SSHGF TSAEF	9-13 204-208	9-13 -	Phosphopeptide motif which directly interacts with the BRCT (carboxy-terminal) domain of the Breast Cancer Gene BRCA1 with low affinity	nucleus, BRCA1-BARD1 complex

Table 3.18 ELM results description for PINK1 gene

ELM Name	Instances (Matched Sequence)	Positions	View in Jmol	ELM Description	Cell Compartment
CLV_NDR_NDR_1	RRV RRA RRL	57-59 119-121 146-148	- - -	N-Arg dibasic convertase (nardilysine) cleavage site (Xaa-|-Arg-Lys or Arg-|-Arg-Xaa)	extracellular, Golgi apparatus, cell surface

CLV_PCSK_SKI1_1	RQALG
	RALLL
	RAVFL
	KNLKL
	KMLFL
		4-8
		15-19
		98-102
		520-524
		555-559
			-
			-
			-
			-
			-
				Subtilisin/kexin isozyme-1 (SKI1) cleavage site ([RK]-X-[hydrophobic]-[LTKF]-|-X)	endoplasmic reticulum,
					endoplasmic reticulum lumen,
					Golgi apparatus

LIG_CYCLIN_1	RGLQL RALLL KNLKL KMLFL	9-13 15-19 520-524 555-559	- - - -	Substrate recognition site that interacts with cyclin and thereby increases phosphorylation by cyclin/cdk complexes. Predicted protein should have the MOD_CDK site.	nucleus, cytosol

				Also used by cyclin inhibitors.	
LIG_FHA_1	LDTRRLQ	143-149	-	Phosphothreonine motif binding a subset of FHA domains that show a preference for a large aliphatic amino acid at the pT+3 position.	nucleus
LIG_MAPK_1	RRVGLGL	57-63	-	MAPK interacting molecules (e.g. MAPKKs, substrates, phosphatases) carry docking motif that help to regulate specific interaction in the MAPK cascade. The classic motif approximates (R/K)	nucleus, cytosol

Table 3.19 ELM results description for HTRA2 gene

ELM Name	Instances (Matched Sequence)	Positions	View in Jmol	ELM Description	Cell Compartment
CLV_NDR_NDR_1	RRP RRV RRY	27-29 204-206 359-361	- 204-206 359-361	N-Arg dibasic convertase (nardilysine) cleavage site (Xaa-|-Arg-Lys or Arg-|-Arg-Xaa)	extracellular, Golgi apparatus, cell surface

CLV_PCSK_SKI1_1	RALLT
	KVILG
		36-40
		395-399
			-
			395-399
				Subtilisin/kexin isozyme-1 (SKI1) cleavage site ([RK]-X-[hydrophobic]-[LTKF]-|-X)	endoplasmic reticulum,
					endoplasmic reticulum lumen,
					Golgi apparatus

LIG_14-3-3_2	RETLTLY	445-451	445-451	Longer mode 2 interacting phospho-motif for 14-3-3 proteins with key conservation RxxxS#p.	nucleus, mitochondrion, cytosol, internal side of plasma membrane
LIG_14-3-3_3	RLLSGD	209-214	209-214	Consensus derived from reported natural interactors which do not match the Mode 1 and Mode 2 ligands.	nucleus, cytosol, internal side of plasma membrane
LIG_APCC_Dbox_1	PRAQLTAVT	79-87	-	An RxxL-based motif that binds to the Cdh1 and Cdc20 components of APC/C thereby	nucleus, cytosol

				targeting the protein for destruction in a cell cycle dependent manner	

Table 3.20 ELM results description for ATP13A2 gene

ELM Name	Instances (Matched Sequence)	Positions	View in Jmol	ELM Description	Cell Compartment
CLV_NDR_NDR_1	TRK RRH RRQ RRP RRL	278-280 370-372 489-491 1072-1074 1147-1149	- - - - -	N-Arg dibasic convertase (nardilysine) cleavage site (Xaa-|-Arg-Lys or Arg-|-Arg-Xaa)	extracellular, Golgi apparatus, cell surface
CLV_PCSK_FUR_1	RLRRQ RPKRA	487-491 1150-1154	- -	Furin (PACE) cleavage site (Arg-Xaa-[Arg/Lys]-Arg-|-Xaa)	Golgi membrane, extracellular, Golgi apparatus
CLV_PCSK_PC1ET2_1	KRV KRA RF	160-162 1152-1154 1157-1159	- - -	NEC1/NEC2 cleavage site (Lys-Arg-|-Xaa)	Golgi membrane, extracellular, Golgi apparatus apparatus

CLV_PCSK_PC7_1	RLRPKRA RASKKRF	1148-1154 1153-1159	- -	Proprotein convertase 7 (PC7, PCSK7) cleavage site (Arg-Xaa-Xaa-Xaa-[Arg/Lys]-Arg-|-Xaa)	Golgi membrane, extracellular, Golgi apparatus
CLV_PCSK_SKI1_1	KRVLR KTALP RNITD KRFKQ	160-164 354-358 1109-1113 1157-1161	- - - -	Subtilisin/kexin isozyme-1 (SKI1) cleavage site ([RK]-X-[hydrophobic]-[LTKF]-|-X)	endoplasmic reticulum, endoplasmic reticulum lumen, Golgi apparatus

Table 3.21 Scanprosite results of PARK2 gene

Site	Prosite access number	Sequence Length	Sequence	Consensus Pattern
N-glycosylation site (ASN_GLYCOSYLATION)	PS00001	8-11 81-84 235-238	NSSH NATG NITC	N-{P}-[ST]-{P} N is the glycosylation site
N-myristoylation site(MYRISTYL)	PS00008	77-82 135-140	GQemNA GSpaGR	G - {EDRKHPFYW} - x(2) - [STAGCN] - {P} [G is the N -

		213-218	GAhpTS	myristoylation site]
		292-297	GCpnSL	
		319-324	GAeeCV	
		336-341	GCgaGL	
		355-360	GNglGC	
		357-362	GLgcGF	
		359-364	GCgfAF	
		450-455	GCewNR	
Casein kinase II phosphorylation site (CK2_PHOSPHO_SITE)	PS00006	83 - 86	TggD	[ST] - x(2) - [DE] [S or T is the phosphorylation site]
		103 - 106	TrvD	
		127 - 130	SrkD	
		181 - 184	ScwD	
		204 - 207	TsaE	
		218 - 221	SdkE	
		240 –243	TctD	
Protein kinase C phosphorylation site (PKC_PHOSPHO_SITE)	PS00005	127 - 129	SrK	ST] - x - [RK] S or T is the phosphorylation site
		168 - 170	TcR	
		218 - 220	SdK	
		410 - 412	TiK	
		414 –416	TtK	

cAMP- and cGMP-dependent protein kinase phosphorylation site (CAMP_PHOSPHO_SITE)	PS00004	128 - 131 348 - 351 412 –415	RKdS RKvT KKtT	[RK](2) - x - ST] S or T is the phosphorylation site
Tyrosine kinase phosphorylation sitE (TYR_PHOSPHO_SITE)	PS00007	305 –312	RilgEeq.Y	RK] - x(2) - [DE] - x(3) - Y or [RK] - x(3) - [DE] - x(2) – Y Y is the phosphorylation site

Table 3.22 Scanprosite results of PINK1 gene

Site	ScanProsite access number	Sequence Length	Sequence	Consensus Pattern
N-glycosylation site ASN_GLYCOSYLATION	PS00001	223 - 226	NISA	N - {P} - [ST] - {P} N is the glycosylation site
N-myristoylation site MYRISTYL	PS00008	10 - 15 39 - 44 105 - 110 159 - 164 165 - 170	GLqlGR GCvrGE GLglGL GQsiGK GCsaAV	G - {EDRKHPFYW} - x(2) - [STAGCN] - {P} [G is the N - myristoylation site]

		189 - 194	GLlpGR	
		307 - 312	GLghGR	
		386 - 391	GCclAD	
		408 - 413	GGngCL	
		455 - 460	GQgkAH	
Casein kinase II phosphorylation site CK2_PHOSPHO_SITE	PS00006	228 - 231	SssE	[ST] - x(2) - [DE] [S or T is the phosphorylation site]
		432 - 435	SkaD	
		465 - 468	SyqE	
Protein kinase C phosphorylation site PKC_PHOSPHO_SITE	PS00005	22 - 24	TgK	[ST] - x - [RK] S or T is the phosphorylation site
		118 - 120	SrR	
		133 - 135	TqK	
		145 - 147	TrR	
		257 - 259	TyR	
		261 - 263	SkR	
		324 - 326	TlR	
		335 - 337	SpR	
		420 - 422	TaR	
		495 - 497	SkR	
		499 - 501	SaR	
		545 - 547	TeK	

		576 - 578	SwR	
cAMP- and cGMP-dependent protein kinase phosphorylation site	PS00004	496 - 499	KRpS	[RK](2) - x - [ST] S or T is the phosphorylation site
Tyrosine kinase phosphorylation site TYR_PHOSPHO_SITE	PS00007	458 - 466	KahlEsr sY	[RK] - x(2) - [DE] - x(3) - Y or [RK] - x(3) - [DE] - x(2) - Y Y is the phosphorylation site

Table 3.23 Scanprosite results of HTRA2 gene

Site	ScanProsite access number	Sequence Length	Sequence	Consensus Pattern
N-glycosylation site ASN_GLYCOSYLATION	PS00001	181 - 184 349 - 352	NGSG NSSS	N - {P} - [ST] - {P} N is the glycosylation site
N-myristoylation site MYRISTYL	PS00008	9 - 14 22 - 27 112 - 117 191 - 196 261 - 266	GAgwSL GIrwGR GAggAV GLivTN GSpfAL	G - {EDRKHPFYW} - x(2) - [STAGCN] - {P} [G is the N - myristoylation site]

		273 - 278 286 - 291 328 - 333 353 - 358 406 - 411	GIvsSA GLpqTN GIsfAI GIsgSQ GLrpGD	
Casein kinase II phosphorylation site CK2_PHOSPHO_SITE	PS00006	221 - 224 290 - 293 383 - 386	TavD TnvE SfpD	[ST] - x(2) - [DE] [S or T is the phosphorylation site]
Protein kinase C phosphorylation site PKC_PHOSPHO_SITE	PS00005	13 - 15 142 - 144 231 - 233 322 - 324 335 - 337 357 - 359	SIR SpR TIR TmK SdR SqR	[ST] - x - [RK] S or T is the phosphorylation site
cAMP- and cGMP-dependent protein kinase phosphorylation site CAMP_PHOSPHO_SITE	PS00004	347 - 350	KKnS	[RK](2) - x - [ST] S or T is the phosphorylation site
Amidation site AMIDATION	PS00009	25 - 28	wGRR	x - G - [RK] - [RK] x is the amidation site

Table 3.24 Scanprosite results of ATP13A2 gene

Site	Prosite access number	Sequence Length	Sequence	Consensus Pattern
N-glycosylation site (ASN_GLYCOSYLATION)	PS00001	341 - 344 1033 - 1036 1110 - 1113	NESS NRTV NITD	N-{P}-[ST]-{P} N is the glycosylation site
N-myristoylation site(MYRISTYL)	PS00008	10 - 15 325 - 330 405 - 410 406 - 411 439 - 444 784 - 789 848 - 853 879 - 884 884 - 889 892 - 897 994 - 999 1014 - 1019:	GStpTG GLmpCD GGlvSS GLvsSI GTiySI GQpaSL GTvfAR GAndCG GAlkAA GIslSQ GAllSV GVqlGG	G - {EDRKHPFYW} - x(2) - [STAGCN] - {P} [G is the N - myristoylation site]
Casein kinase II phosphorylation site (CK2_PHOSPHO_SITE)		23 - 26 123 - 126 186 - 189	TsiD SqaE SllD	[ST] - x(2) - [DE] [S or T is the phosphorylation site]

	PS00006	193 - 196 205 - 208 284 - 287 517 - 520 647 - 650 779 - 782 896 - 899 1136 - 1139	ScdD SlqD TlrD TltE TqpE ThpE SqaE SvlD	
Protein kinase C phosphorylation site (PKC_PHOSPHO_SITE)	PS00005	34 - 36 278 - 280 284 - 286 292 - 294 368 - 370 402 - 404 425 - 427 1155 - 1157	SvR TrK TlR SmR ThR TaK SmK SkK	ST] - x - [RK] S or T is the phosphorylation site
cAMP- and cGMP-dependent protein kinase phosphorylation site (CAMP_PHOSPHO_SITE)	PS00004	279 - 282 370 - 373 1152 - 1155	RKqS RRhT KRaS	[RK](2) - x - ST] S or T is the phosphorylation site

Table 3.25 SMART results of PARK2 gene

Name	Begin	End	E-value
UBQ	1	72	2.95E-16
IBR	313	377	4.49E-14
IBR	401	457	1.42E-01

Table 3.26 SMART results of PINK1 gene

Name	Begin	End	E-value
low complexity	4	20	-
low complexity	105	110	-
STYKc	156	511	6.36e-13

Table 3.27 SMART results of HTRA2 gene

Name	Begin	End	E-value
transmembrane	105	124	-
Tryp_SPc	178	342	8.06e+00
PDZ	336	445	1.92e-11

Table 3.28 SMART results of ATP13A2 gene

Name	Begin	End	E-Value
low complexity	28	38	-
transmembrane	45	67	-
Cation_ATPase_N	188	256	7.39e+00
transmembrane	257	276	-
transmembrane	428	447	-
transmembrane	462	484	-
low complexity	537	556	-
transmembrane	931	953	-
transmembrane	968	990	-
transmembrane	1002	1024	-
transmembrane	1044	1066	-
transmembrane	1079	1101	-
transmembrane	1116	1138	-

3.3 Post Translation Modification

Functional diversity of the proteome increases as a result of post translation modifications by covalent addition of functional proteins or groups, degradation of entire proteins or proteolytic cleavage of regulatory subunits. SIgnalP server was used to identify the post translation modifications in different genes.

3.3.1 SignalP

SignalP 3.0 Server predicted results using neural networks and Hidden Markov Model on eukaryotes. The diagrammatical representation of SignalP-NN prediction was shown in figure 3.13, 3.14 and 3.15 for PARK2, PINK1 and HTRA2 gene. It didn't give any SignalP-NN result for ATP13A2 gene. The diagrammatical representation of SignalP-HMM prediction for all genes was given in figure 3.16, 3.17, 3.18 and 3.19 respectively. SignalP-NN measured the scores of cleavage of the amino acid sequence, their cutoff values, their positions and whether they were signal peptide or not shown in table 3.29, 3.30 and 3.31 for all genes except ATP13A2. SignalP-HMM predicted whether the protein is secretory or non-scretory, probability of signal peptide, signal anchor and the maximum cleavage site probability for all four proteins (Table 3.32, 3.33, 3.34 and 3.35).

3.4 Topology Prediction

Protein topology prediction plays an important role in the structural biology. In the present study following tools were used to find the topology of selected genes.

3.4.1 HMM top

3.4.2 TMpred

3.4.1 HMM top

HMMTOP (Hidden Markov Model for Topology Prediction) is an automated server which predicted topology of proteins and transmembrane helices (Fig. 3.20, 3.21, 3.22 and 3.23). It predicted N-terminus of the sequence, number of transmembrane helices, total entropy of model and entropy of best path (Table 3.36, 3.37, 3.38 and 3.39).

3.4.2 TMpred

TMPRED predicted membrane-spanning regions and orientation in the protein sequence. It is a statistical analysis algorithm of TMbase. A combination of several weight-matrices for scoring was used for making predictions. Probably no transmembrane protein - no possible model was found for PARK2 gene (Table 3.40, 3.41 and 3.42).

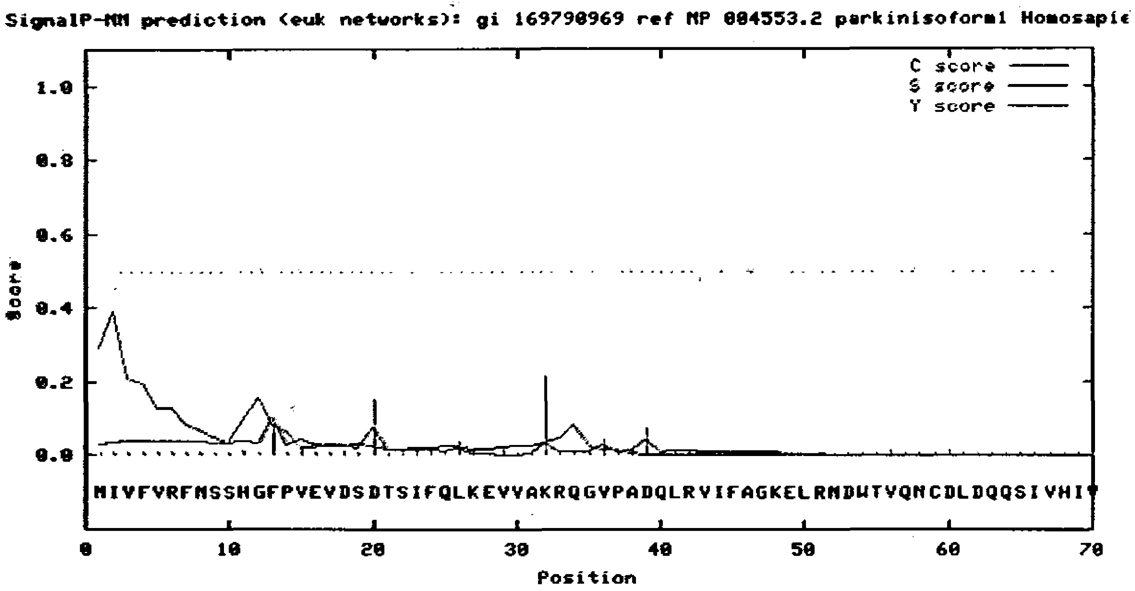


Figure 3.13 SignalP-NN prediction of PARK2 gene

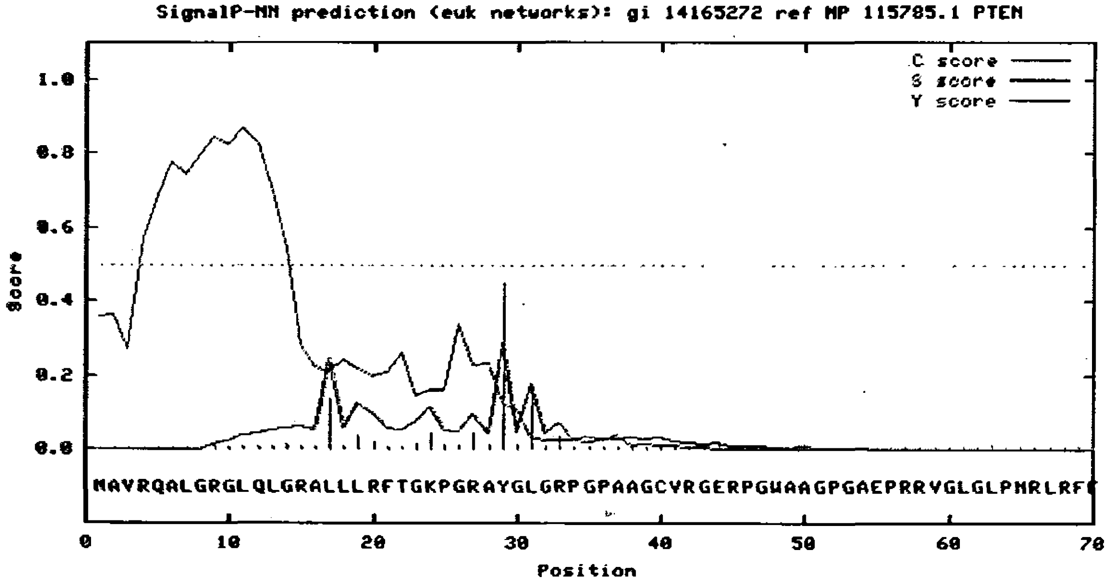


Figure 3.14 SignalP-NN prediction of PINK1 gene

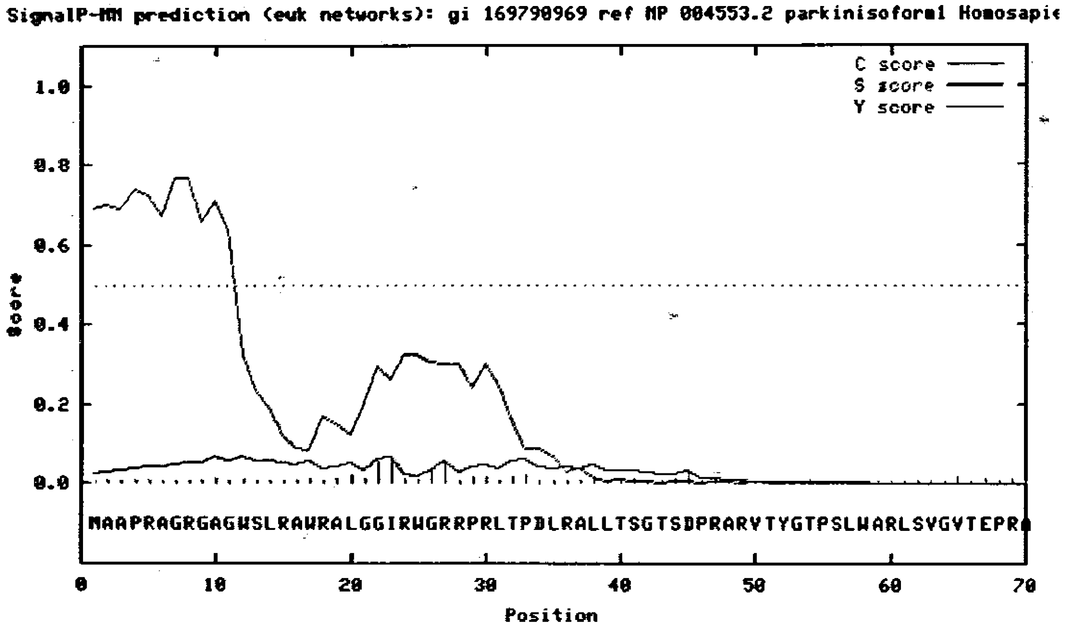


Figure 3.15 SignalP-NN prediction of HTRA2 gene

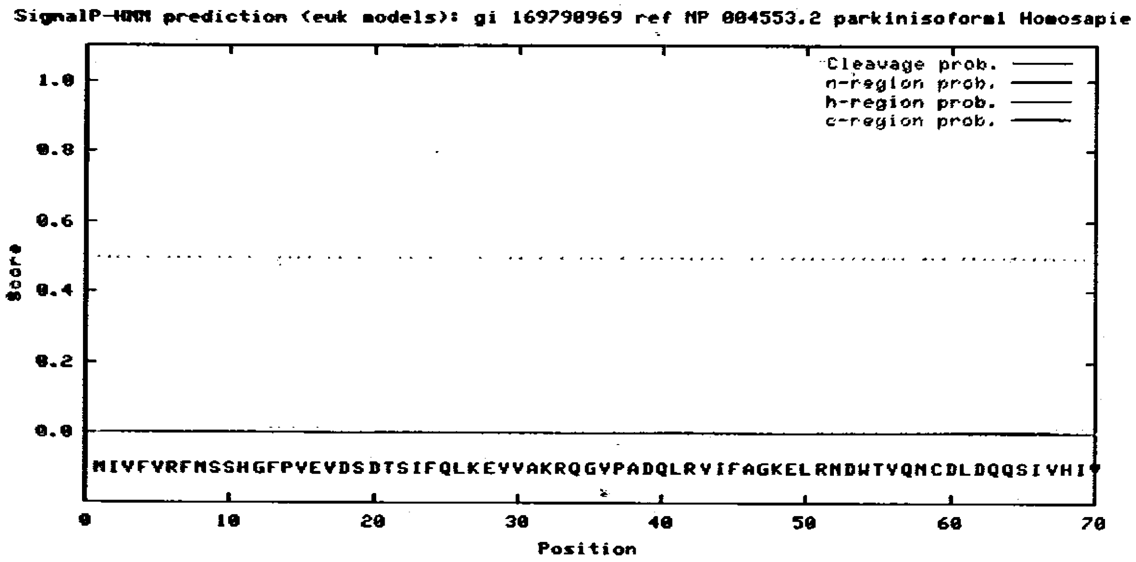


Figure 3.16 SignalP-HMM prediction of PARK2 gene

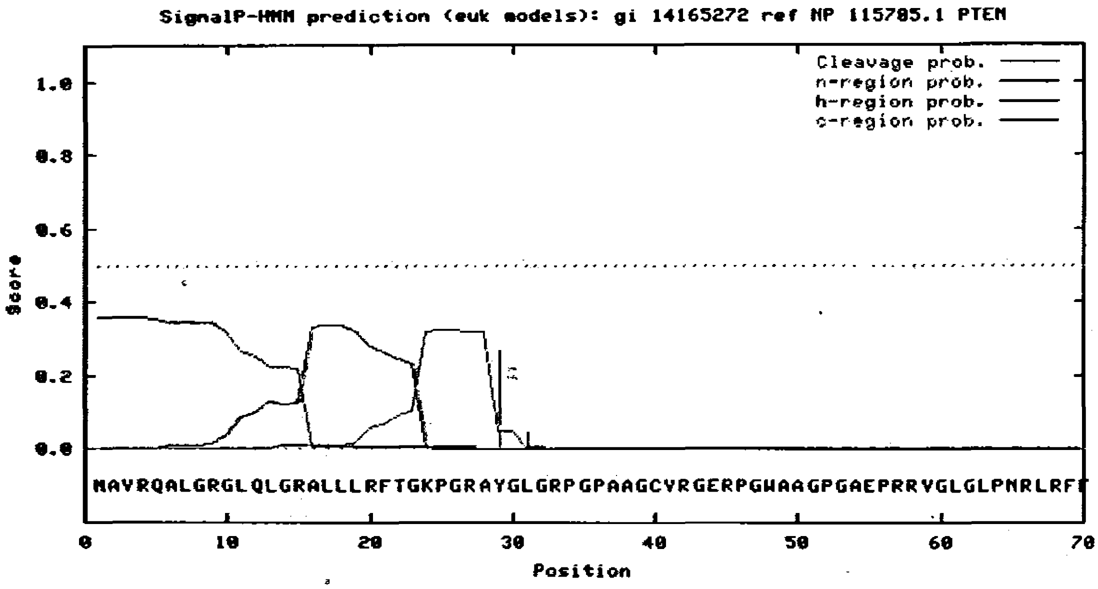


Figure 3.17 SignalP-HMM prediction of PINK1 gene

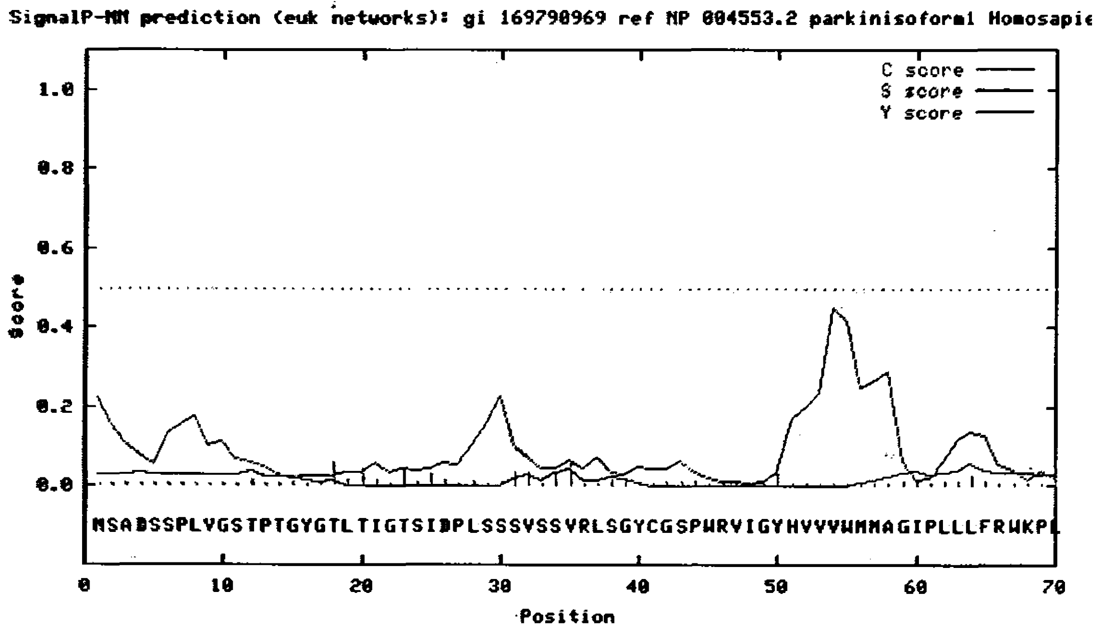


Figure 3.18 SignalP-HMM prediction of HTRA2 gene

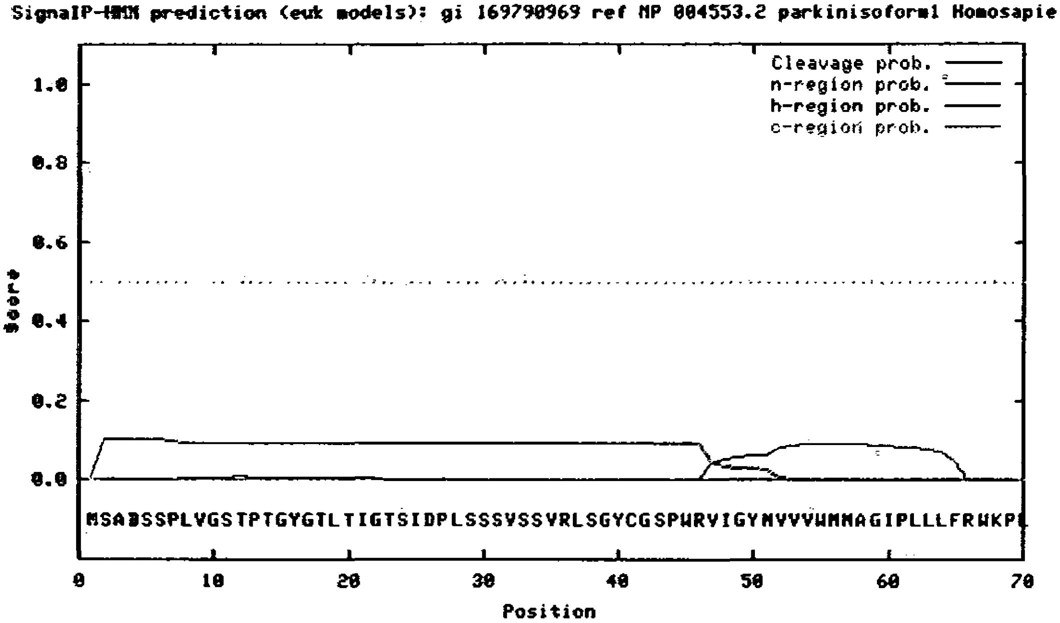


Figure 3.19 SignalP-HMM prediction of ATP13A2 gene

The best path:,

```
seq  MIVFVRFNSS HGFPVEVDSD TSIFQLKEVV AKRQGVPADQ LRVIFAGKEL   50
pred OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

seq  RNDWTVQNCD LDQQSIVHIV QRPWRKGQEM NATGGDDPRN AAGGCEREPQ  100
pred OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

seq  SLTRVDLSSS VLPGDSVGLA VILHTDSRKD SPPAGSPAGR SIYNSFYVYC  150
pred OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

seq  KGPCQRVQPG KLRVQCSTCR QATLTLTQGP SCWDDVLIPN RMSGECQSPH  200
pred OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

seq  CPGTSAEFFF KCGAHPTSDK ETSVALHLIA TNSRNITCIT CTDVRSPVLV  250
pred OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

seq  FQCNSRHVIC LDCFHLYCVT RLNDRQFVHD PQLGYSLPCV AGCPNSLIKE  300
pred OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

seq  LHHFRILGEE QYNRYQQYGA EECVLQMGGV LCPRPGCGAG LLPEPDQRKV  350
pred OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

seq  TCEGGNGLGC GFAFCRECKE AYHEGECSAV FEASGTTTQA YRVDERAAEQ  400
pred OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

seq  ARWEAASKET IKKTTKPCPR CHVPVEKNGG CMHMKCPQPQ CRLEWCWNCG  450
pred OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

seq  CEWNRVCMGD HWFDV  465
pred OOOOOOOOOO OOOOO
```

Figure 3.20 HMM top result of PARK2 gene

The best path:

```
seq  MAVRQALGRG LQLGRALLLR FTGKPGRAYG LGRPGPAAGC VRGERPGWAA   50
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  GPGAEPRRVG LGLPNRLRFF RQSVAGLAAR LQRQFVVRAW GCAGPCGRAV   100
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  FLAFGLGLGL IEEKQAESRR AVSACQEIQA IFTQKSKPGP DPLDTRRLQG   150
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  FRLEEYLIGQ SIGKGCSAAV YEATMPTLPQ NLEVTKSTGL LPGRGPGTSA   200
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  PGEGQERAPG APAFPLAIKM MWNISAGSSS EAILNTMSQE LVPASRVALA   250
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  GEYGAVTYRK SKRGPKQLAP HPNIIRVLRA FTSSVPLLPG ALVDYPDVLP   300
pred IIIIIIIIii iiiiiiiiii iiiHHHHHHH HHHHHHHHHH HHHooooooo

seq  SRLHPEGLGH GRTLFLVMKN YPCTLRQYLC VNTPSPRLAA MMLLQLLEGV   350
pred ooooooooOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

seq  DHLVQQGIAH RDLKSDNILV ELDPDGCPWL VIADFGCCLA DESIGLQLPF   400
pred OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

seq  SSWYVDRGGN GCLMAPEVST ARPGPRAVID YSKADAWAVG AIAYEIFGLV   450
pred OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

seq  NPFYGQGKAH LESRSYQEAQ LPALPESVPP DVRQLVRALL QREASKRPSA   500
pred OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

seq  RVAANVLHLS LWGEHILALK NLKLDKMVGW LLQQSAATLL ANRLTEKCCV   550
pred OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

seq  ETKMKMLFLA NLECETLCQA ALLLCSWRAA L  581
pred OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO O
```

Figure 3.21 HMM top result of PINK1 gene

The best path:

```
seq  MAAPRAGRGA GWSLRAWRAL GGIRWGRRPR LTPDLRALLT SGTSDPRARV   50
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  TYGTPSLWAR LSVGVTEPRA CLTSGTPGPR AQLTAVTPDT RTREASENSG  100
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  TRSRAWLAVA LGAGGAVLLL LWGGGRGPPA VLAAVPSPPP ASPRSQYNFI  150
pred iiiiHHHHHH HHHHHHHHHH HHoooooooo oooooooOOO OOOOOOOOOO

seq  ADVVEKTAPA VVYIEILDRH PFLGREVPIS NGSGFVVAAD GLIVTNAHVV  200
pred OOOOOOOOOO Oooooooooo ooooooHHHH HHHHHHHHHH HHHHiiiiii

seq  ADRRRVRVRL LSGDTYEAVV TAVDPVADIA TLRIQTKEPL PTLPLGRSAD  250
pred iiiiiiiiiI IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  VRQGEFVVAM GSPFALQNTI TSGIVSSAQR PARDLGLPQT NVEYIQTDAA  300
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  IDFGNSGGPL VNLDGEVIGV NTMKVTAGIS FAIPSDRLRE FLHRGEKKNS  350
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  SSGISGSQRR YIGVMMLTLS PSILAELQLR EPSFPDVQHG VLIHKVILGS  400
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  PAHRAGLRPG DVILAIGEQM VQNAEDVYEA VRTQSQLAVQ IRRGRETLTL  450
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  YVTPEVTE  458
pred IIIIIIII
```

Figure 3.22 HMM top result of HTRA2 gene

The best path:

```
seq  MSADSSPLVG STPTGYGTLT IGTSIDPLSS SVSSVRLSGY CGSPWRVIGY   50
pred OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO Ooooooooo ooooooHHHH

seq  HVVVWMMAGI PLLLFRWKPL WGVRLRLRPC NLAHAETLVI EIRDKEDSSW   100
pred HHHHHHHHHH HHHHHiiiii iiiiiiiiii IIIIIIIIII IIIIIIIIII

seq  QLFTVQVQTE AIGEGSLEPS PQSQAEDGRS QAAVGAVPEG AWKDTAQLHK   150
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  SEEAVSVGQK RVLRYYLFQG QRYIWIETQQ AFYQVSLLDH GRSCDDVHRS   200
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIii

seq  RHGLSLQDQM VRKAIYGPNV ISIPVKSYPQ LLVDEALNPY YGFQAFSIAL   250
pred iiiiiiiiii iiiHHHHHHH HHHHHHHHHH HHHooooooo oooooooooo

seq  WLADHYYWYA LCIFLISSIS ICLSLYKTRK QSQTLRDMVK LSMRVCVCRP   300
pred oooooooHHH HHHHHHHHHH HHHHHHiiii iiiiiiiiii iIIIIIIIII

seq  GGEEEWVDSS ELVPGDCLVL PQEGGLMPCD AALVAGECMV NESSLTGESI   350
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  PVLKTALPEG LGPYCAETHR RHTLFCGTLI LQARAYVGPH VLAVVTRTGF   400
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  CTAKGGLVSS ILHPRPINFK FYKHSMKFVA ALSVLALLGT IYSIFILYRN   450
pred IIIIIIIIII IIiiiiiiii iiiiiiHHHH HHHHHHHHHH HHHHHHHooo

seq  RVPLNEIVIR ALDLVTVVVP PALPAAMTVC TLYAQSRLRR QGIFCIHPLR   500
pred oooooooooo oooHHHHHHH HHHHHHHHHH HHiiiiiiii iiiiiiiIII

seq  INLGGKLQLV CFDKTGTLTE DGLDVMGVVP LKGQAFLPLV PEPRRLPVGP   550
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  LLRALATCHA LSRLQDTPVG DPMDLKMVES TGWVLEEEPA ADSAFGTQVL   600
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  AVMRPPLWEP QLQAMEEPPV PVSVLHRFPF SSALQRMSVV VAWPGATQPE   650
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  AYVKGSPELV AGLCNPETVP TDFAQMLQSY TAAGYRVVAL ASKPLPTVPS   700
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  LEAAQQLTRD TVEGDLSLLG LLVMRNLLKP QTTPVIQALR RTRIRAVMVT   750
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  GDNLQTAVTV ARGCGMVAPQ EHLIIVHATH PERGQPASLE FLPMESPTAV   800
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  NGVKDPDQAA SYTVEPDPRS RHLALSGPTF GIIVKHFPKL LPKVLVQGTV   850
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

seq  FARMAPEQKT ELVCELQKLQ YCVGMCGDGA NDCGALKAAD VGISLSQAEA   900
pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII
```

```
seq  SVVSPFTSSM ASIECVPMVI REGRCSLDTS FSVFKYMALY SLTQFISVLI   950
pred IIIIIIIIII IIIIIIIiii iiiiiiiiii iiHHHHHHHH HHHHHHHHHH

seq  LYTINTNLGD LQFLAIDLVI TTTVAVLMSR TGPALVLGRV RPPGALLSVP  1000
pred Hooooooooo oooooooHHH HHHHHHHHHH HHHHHHiiii iiiiiiiiHH

seq  VLSSLLLQMV LVTGVQLGGY FLTLAQPWFV PLNRTVAAPD NLPNYENTVV  1050
pred HHHHHHHHHH HHHHHHHooo oooooooooo oooooooooo oooooooHHH

seq  FSLSSFQYLI LAAAVSKGAP FRRPLYTNVP FLVALALLSS VLVGLVLVPG  1100
pred HHHHHHHHHH HHHHHHiiii iiiiiiiiHH HHHHHHHHHH HHHHHHHHHH

seq  LLQGPLALRN ITDTGFKLLL LGLVTLNFVG AFMLESVLDQ CLPACLRRLR  1150
pred HHoooooooo ooooooooHH HHHHHHHHHH HHHHHHHHii iiiiiiiiii

seq  PKRASKKRFK QLERELAEQP WPPLPAGPLR  1180
pred iiiIIIIIII IIIIIIIIII IIIIIIIIII
```

Figure 3.23 HMM top result of ATP13A2 gene

Table 3.29 SignalP-NN results of PARK2 gene

Measure	Position	Value	Cutoff	Signal peptide
max. C	32	0.215	0.32	NO
max. Y	13	0.105	0.33	NO
max. S	2	0.390	0.87	NO
mean S	1-12	0.152	0.48	NO
D	1-12	0.129	0.43	NO

Table 3.30 SignalP-NN results of PINK1 gene

Measure	Position	Value	Cutoff	Signal peptide
max. C	29	0.449	0.32	YES
max. Y	29	0.290	0.33	NO
max. S	11	0.871	0.87	YES
mean S	1-28	0.439	0.48	NO
D	1-28	0.364	0.43	NO

Most probably cleavage site between position 28 and 29: GRA-YG

Table 3.31 SignalP-NN results of HTRA2 gene

Measure	Position	Value	Cutoff	Signal peptide
max. C	23	0.068	0.32	NO
max. Y	12	0.072	0.33	NO
max. S	7	0.769	0.87	NO
mean S	1-11	0.706	0.48	YES
D	1-11	0.389	0.43	NO

Most likely cleavage site between position 11 and 12: GAG-WS

Table 3.32 SignalP-HMM results of PARK2 gene

Prediction	Signal peptide probability	Signal anchor probability	Max. cleavage site probability
Non-secretory protein	0.000	0.000	0.000 between position 19 and 20

Table 3.33 SignalP-HMM results of PINK1 gene

Prediction	Signal peptide probability	Signal anchor probability	Max. cleavage site probability
Non-secretory protein	0.357	0.000	0.267 between position 28 and 29

Table 3.34 SignalP-HMM results of HTRA2 gene

Measnre	Position	Value	Cutoff	Signal peptide
max. C	18	0.061	0.32	NO
max. Y	64	0.061	0.33	NO
max. S	54	0.452	0.87	NO
mean S	1-63	0.094	0.48	NO
D	1-63	0.078	0.43	NO

Table 3.35 SignalP-HMM results of ATP13A2 gene

Prediction	Signal peptide probability	Signal anchor probability	Max. cleavage site probability
Non-secretory protein	0.008	0.096	0.002 between position 17 and 18

Table 3.36 HMM top predicted results of PARK2 gene

N-terminus	Transmembrane helices	Total entropy of the model	Entropy of the best path
OUT	0	17.0277	17.0277

Table 3.37 HMM top predicted results of PINK1 gene

N-terminus	Transmembrane helices	Total entropy of the model	Entropy of the best path
IN	1	17.0137	17.0147

Table 3.38 HMM top predicted results of HTRA2 gene

N-terminus	Transmembrane helices	Total entropy of the model	Entropy of the best path
IN	2	17.0038	17.0058

Table 3.39 HMM top predicted results of ATP13A22 gene

N-terminus	Ttransmembrane helices	Total entropy of the model	Entropy of the best path
OUT	11	17.0101	17.0148

Table 3.40 TMpred transmembrane models of PINK1 gene

strongly preferred model: N-terminus inside	Amino acid length	Sequence Length	Score	Orientation
	91-111	(21)	793	i-o
	274-293	(20)	554	o-i
	436-452	(17)	668	i-o
Total Score			2015	
Alternative Model	89-111	(23)	1159	o-i
	436-452	(17)	668	i-o
Total Score			1827	

Table 3.41 TMpred transmembrane models of HTRA2 gene

strongly preferred model:N-terminus inside	Amino acid length	Sequence Length	Score	Orientation
	105-125	(21)	2399	i-o
	179-201	(23)	786	o-i
	256-279	(24)	578	i-o
	360-377	(18)	552	o-i
Total Score			4315	
Alternative Model	105-124	(20)	2453	o-i
	182-200	(19)	614	i-o
	360-377	(18)	552	o-i
Total Score			3619	

Table 3.42 TMpred transmembrane models of ATP13A2 gene

	Amino acid length	Sequence Length	Score	Orientation
Strongly preferred model:N-terminus inside	47-67	(21)	2125	i-o
	256-276	(21)	2804	o-i
	428-448	(21)	2301	i-o
	464-484	(21)	1123	o-i
	936-954	(19)	1517	i-o
	963-988	(26)	1183	o-i
	996-1025	(30)	1482	i-o
	1048-1064	(17)	793	o-i
	1081 1102	(22)	2212	i-o
	1118 1138	(21)	1660	o-i
Total Score			17200	
Alternative Model	47-67	(21)	2220	o-i
	256-276	(21)	2169	i-o
	428-447	(20)	2301	o-i
	457-479	(23)	922	i-o
	900-920	(21)	560	o-i
	936-954	(19)	1517	i-o
	963-988	(26)	1183	o-i

	996-1025	(30)	1482	i-o
	1048-1064	(17)	793	o-i
	1081-1102	(22)	2212	i-o
	1118-1138	(21)	1660	o-i
Total Score			17019	

3.5 Primary Structure Prediction

Analysis of amino acid sequence provides important information for predicted structures of proteins, which is very helpful in understanding its biochemical and cellular functions. In the present research work following tools were employed to find the primary structure of proteins.

3.5.1 ProtParam

For a given protein sequence ProtParam calculated physical and chemical parameters. The parameters includes the aliphatic index, molecular weight, extinction coefficient, theoretical pI, atomic composition, amino acid composition, estimated half-life, instability index and grand average of hydropathicity (Table 3.43, 3.44, 3.45 and 3.46).

3.6 Secondary Structure Prediction

Secondary structure is a local description of certain repeating structures; such as α helices, β strands and random coils. For the secondary structure prediction, GOR4 server was used.

3.6.1 GOR4

The accuracy of the GOR4 server ranges between 50-55%. The mostly observed random distribution of the three states in globular protein is 30% α helices, 20% β strands and 50% random coils.

3.7 Tertiary Structure Prediction

For the prediction of tertiary structure two approaches were used: one was homology modeling and the other was threading. On the basis of the templates selected through BLAST, ten models were generated for each protein sequence using Modeller 9v7 and the best one was selected on the basis of lowest objective function value shown in figure 3.32, 3.33, 3.34 and 3.35.

Tool used for threading didn't give results for amino acid sequence> 700 amino acids. Therefore it could not predict model of ATP13A2, (Fig. 3.36, 3.37 and 3.38). Percentage alignment between the target and template sequences was shown in table 3.47. The lowest objective function values of the generated models were shown in table 3.48.

Tools used for homology modeling and threading were as follow:

3.7.1 Modeller

3.7.2 SAM-T08

```
MIVFVRFNSSHGFPVEVDSDTSIFQLKEVVAKRQGVPADQLRVIFAGKELRNDWTVQNCDLDQQSIVHIV
Ccceeeecccccceeeecccccchhhhhhhhhhhcccccchhhhhhhcccccccceeeecccccceeeee
QRPWRKGQEMNATGGDDPRNAAGGCEREPQSLTRVDLSSSVLPGDSVGLAVILHTDSRKDSPPAGSPAGR
Ecchhhhcccccccccccccccccccccccccceeccccccccccccccceeeeeecccccccccccccc
SIYNSFYVYCKGPCQRVQPGKLRVQCSTCRQATLTLTQGPSCWDDVLIPNRMSGECQSPHCPGTSAEFFF
Eeeceeeeccccccccccccceeeecccccccceeeeeecccccceeeeccccccccccccccccccceee
KCGAHPTSDKETSVALHLIATNSRNITCITCTDVRSPVLVFQCNSRHVICLDCFHLYCVTRLNDRQFVHD
Ecccccccchhhhhhhhhhhccccccceeeeeccccccceeeeecccccceeeeecccccccccccccceecc
PQLGYSLPCVAGCPNSLIKELHHFRILGEEQYNRYQQYGAEECVLQMGGVLCPRPGCGAGLLPEPDQRKV
Cccceeeccccccccccchhhhhhhhhhhhhhhhhhhcceeeeeeeecccceeccccccccccccccccee
TCEGGNGLGCGFAFCRECKEAYHEGECSAVFEASGTTTQAYRVDERAAEQARWEAASKETIKKTTKPCPR
Eeecccccceeccccccccccccccceeeeecccccchhhhhhhhhhhhhhhhhhhhhhhhhcccccccc
CHVPVEKNGGCMHMKCPQPQCRLEWCWNCGCEWNRVCMGDHWFDV
Cceeeeecccccceeccceeeeeeeeeeeecceeeecccceeeec
```

Figure 3.24 Predicted secondary structure of PARK2 gene

```
MAVRQALGRGLQLGRALLLRFTGKPGRAYGLGRPGPAAGCVRGERPGWAAGPGAEPRRVGLGLPNRLRFF
cccceecchhhhhhhhhhhhcccccceeeccccccccceeccccccccccccceecccccccccccchhhh
RQSVAGLAARLQRQFVVRAWGCAGPCGRAVFLAFGLGLGLIEEKQAESRRAVSACQEIQAIFTQKSKPGP
hhhhhhhhhhhhheeeeeccccccccccchhhhhcccccchhhhhhhhhhhhhhhhhhhhhhhhhccccccc
DPLDTRRLQGFRLEEYLIGQSIGKGCSAAVYEATMPTLPQNLEVTKSTGLLPGRGPGTSAPGEGQERAPG
cccchhhhhhhhhhhhhhhcccccccceeeecccccccccccccccccccccccccccccccccccccccc
APAFPLAIKMMWNISAGSSSEAILNTMSQELVPASRVALAGEYGAVTYRKSKRGPKQLAPHPNIIRVLRA
cccccchhhheeeeeecccccchhhhhhhhhhcchhhhhhhhcccccceeeeeecccccccccccceeeeeec
FTSSVPLLPGALVDYPDVLPSRLHPEGLGHGRTLFLVMKNYPCTLRQYLCVNTPSPRLAAMMLLQLLEGV
cccccccccccccccccccccccccccccccceeeeeeeeecccceeeeeeeeeeccchhhhhhhhhhhcc
DHLVQQGIAHRDLKSDNILVELDPDGCPWLVIADFGCCLADESIGLQLPFSSWYVDRGGNGCLMAPEVST
hhhhhhhhhhcccccceeeeecccccceeeeeccccccccccceecccceeeeecccccccceeeccccccc
ARPGPRAVIDYSKADAWAVGAIAYEIFGLVNPFYGQGKAHLESRSYQEAQLPALPESVPPDVRQLVRALL
cccccceeeeecchhhhhhhhhhhhhcccccccccccccchhhhhhhhhhcccccccccccccchhhhhhhhh
QREASKRPSARVAANVLHLSLWGEHILALKNLKLDKMVGWLLQQSAATLLANRLTEKCCVETKMKMLFLA
hhhhcccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhchhhhhhhhhhhhhhhhhhcccceeecccchhhhhh
NLECETLCQAALLLCSWRAAL
Hhhhhhhhhhhhhhhccceec
```

Figure 3.25 Predicted secondary structure of PINK1 gene

```
MAAPRAGRGAGWSLRAWRALGGIRWGRRPRLTPDLRALLTSGTSDPRARVTYGTPSLWARLSVGVTEPRA
cccccccccchhhhhhhhhccccccccccccchhhhhhcccccceeeeecccccceeeeeeccccccccc
CLTSGTPGPRAQLTAVTPDTRTREASENSGTRSRAWLAVALGAGGAVLLLLWGGGRGPPAVLAAVPSPPP
ccccccccccccceecccccchhhhhhcccchhhhhhhhhhhhcceeeeeeeccccccccceeeeecccccc
ASPRSQYNFIADVVEKTAPAVVYIEILDRHPFLGREVPISNGSGFVVAADGLIVTNAHVVADRRRVRVRL
ccccchhhhhhhhhhhccccceeeeeeccccccccccceeccccccccceeeeecccceeehhhhhheeee
LSGDTYEAVVTAVDPVADIATLRIQTKEPLPTLPLGRSADVRQGEFVVAMGSPFALQNTITSGIVSSAQR
cccccceeeeeecccchhhhhhhhcccccccccccccccccccceeeeeecccccccccccccccccccc
PARDLGLPQTNVEYIQTDAAIDFGNSGGPLVNLDGEVIGVNTMKVTAGISFAIPSDRLREFLHRGEKKNS
hhhhcccccccccccchhhhhccccccceeeccccceeeeceeccccccceeeecchhhhhhhhhhcccccc
SSGISGSQRRYIGVMMLTLSPSILAELQLREPSFPDVQHGVLIHKVILGSPAHRAGLRPGDVILAIGEQM
ccccccceeeeeeeeeeecccchhhhhhhccccccccccceeeeeeeeeccccccccccccceeehhhhhh
VQNAEDVYEAVRTQSQLAVQIRRGRETLTLYVTPEVTE
Hhhhhhhhhhhhhhhhhhhhhhcccceeeeeeccceec
```

Figure 3.26 Predicted secondary structure of HTRA2 gene

```
MSADSSPLVGSTPTGYGTLTIGTSIDPLSSSVSSVRLSGYCGSPWRVIGYHVVVWMMAGIPLLLFRWKPL
cccccceeeccccccceeeeeccccccccccceeeeeccccccccceeeeeeeeeeeeeeeceeeeccccc
WGVRLRLRPCNLAHAETLVIEIRDKEDSSWQLFTVQVQTEAIGEGSLEPSPQSQAEDGRSQAAVGAVPEG
ceeeeeecccchhhhchhhhhhhhcccccchhhhhhhhhhhhcccccccccccchhhhhhhhhhhhccccc
AWKDTAQLHKSEEAVSVGQKRVLRYYLFQGQRYIWIETQQAFYQVSLLDHGRSCDDVHRSRHGLSLQDQM
chhhhhhhhhhhhhhhcchhhhhhhhhcccceeeeeeechhhheeeecccccccccchhhccccchhhhhh
VRKAIYGPNVISIPVKSYPQLLVDEALNPYYGFQAFSIALWLADHYYWYALCIFLISSISICLSLYKTRK
hhhhcccccceecccccchhhhhccccccccchhhhhhhhhcceeeeeeeeeeeecccchhhhhccccc
QSQTLRDMVKLSMRVCVCRPGGEEEWVDSSELVPGDCLVLPQEGGLMPCDAALVAGECMVNESSLTGESI
hhhhhhhheeeeeeeeeecccccccccceeecccceeecccccccccccceeeeeccceeecccccccccc
PVLKTALPEGLGPYCAETHRRHTLFCGTLILQARAYVGPHVLAVVTRTGFCTAKGGLVSSILHPRPINFK
hhhhhcccccccccccccccccccccccceeeecccccccceeeeeeecceeeccccccccceeecccccee
FYKHSMKFVAALSVLALLGTIYSIFILYRNRVPLNEIVIRALDLVTVVVPPALPAAMTVCTLYAQSRLRR
ehhhhhhhhhhhhhhhhcchhhhhhhhcccccchhhhhhcccceeeecccccccccchhhhhhhhhcccc
QGIFCIHPLRINLGGKLQLVCFDKTGTLTEDGLDVMGVVPLKGQAFLPLVPEPRRLPVGPLLRALATCHA
ccceeeccccccccceeeeeeccccccccccccceeeeeccccccccccccccccccccccchhhhhhhhh
LSRLQDTPVGDPMDLKMVESTGWVLEEEPAADSAFGTQVLAVMRPPLWEPQLQAMEEPPVPVSVLHRFPF
cccccccccccccceeeeecccchhhhccchhhhceeeeeeeccccccccccccccccccceeeeccccc
SSALQRMSVVVAWPGATQPEAYVKGSPELVAGLCNPETVPTDFAQMLQSYTAAGYRVVALASKPLPTVPS
cchhhheeeeeccccccccccccccccceeecccccccccchhhhhhhhhhhheeeeeccccccccccch
LEAAQQLTRDTVEGDLSLLGLLVMRNLLKPQTTPVIQALRRTRIRAVMVTGDNLQTAVTVARGCGMVAPQ
hhhhhhhhhhcccchhhhhhhhhhhhccccccchhhhhhhhhheeeeeecccccceeeeeeeecccccccccc
EHLIIVHATHPERGQPASLEFLPMESPTAVNGVKDPDQAASYTVEPDPRSRHLALSGPTFGIIVKHFPKL
cceeeeecccccccccccceeecccccccccccccccccceecccccccceeeeecccccceeeeeeeccccc
LPKVLVQGTVFARMAPEQKTELVCELQKLQYCVGMCGDGANDCGALKAADVGISLSQAEASVVSPFTSSM
ccccccchhhhhhccchhhhhhhhcccceeeeeeeecccccchhhhcchhhhhhhccceeccccccccee
ASIECVPMVIREGRCSLDTSFSVFKYMALYSLTQFISVLILYTINTNLGDLQFLAIDLVITTTVAVLMSR
cccccceeeeccccccccccccchhhhhhhcchhhheeeeeeeccccccchhhhhhhhhcceeeeeeeecc
TGPALVLGRVRPPGALLSVPVLSSLLLQMVLVTGVQLGGYFLTLAQPWFVPLNRTVAAPDNLPNYENTVV
cccceeecccccccccccccchhhhhhhhhheeeccccccceeeeccccccccccccccccccccccceee
FSLSSFQYLILAAAVSKGAPFRRPLYTNVPFLVALALLSSVLVGLVLVPGLLQGPLALRNITDTGFKLLL
eeccchhhhhhhhhccccccccccccccchhhhhhhhhhhhhceeeccccccccchhhhcccccchhhhh
LGLVTLNFVGAFMLESVLDQCLPACLRRLRPKRASKKRFKQLERELAEQPWPPLPAGPLR
hcccccchhhhhhhhhhhhcchhhhccccccchhhhhhhhhhhhhhcccccccccccee
```

Figure 3.27 Predicted secondary structure of ATP13A2 gene

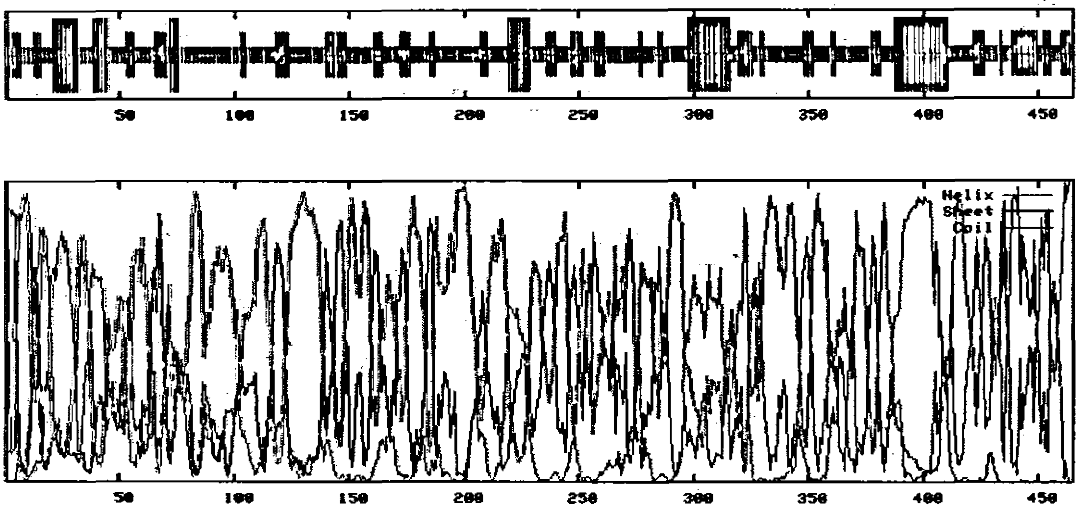


Figure 3.28 GOR4 results for PARK2 gene

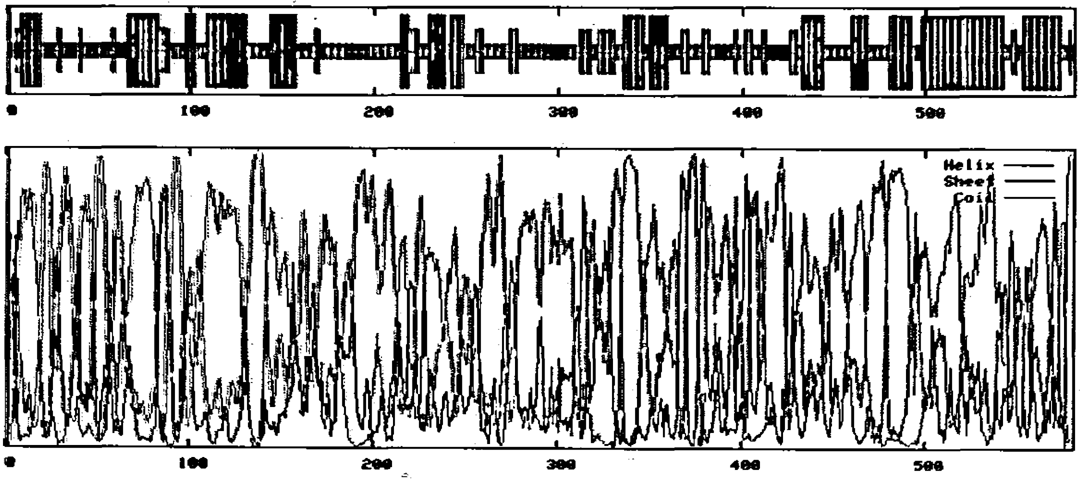


Figure 3.29 GOR4 results for PINK1 gene

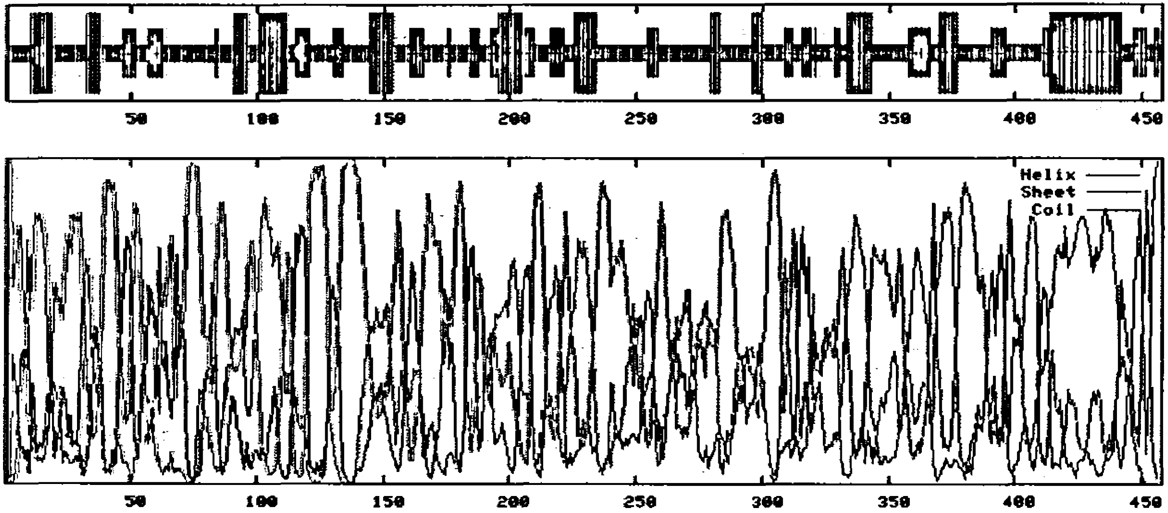


Figure 3.30 GOR4 results for HTRA2 gene

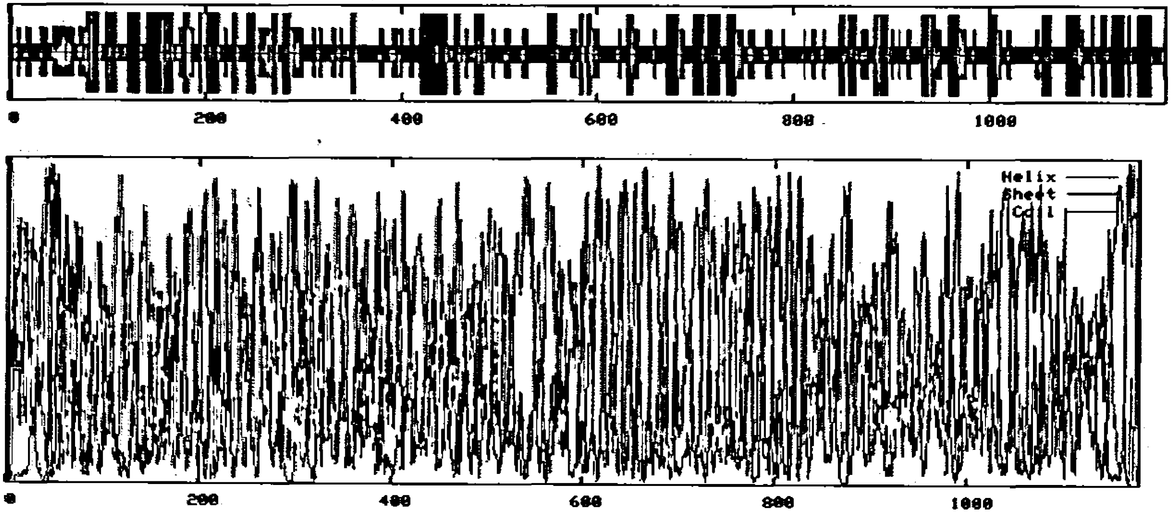


Figure 3.31 GOR4 results for ATP13A2 gene

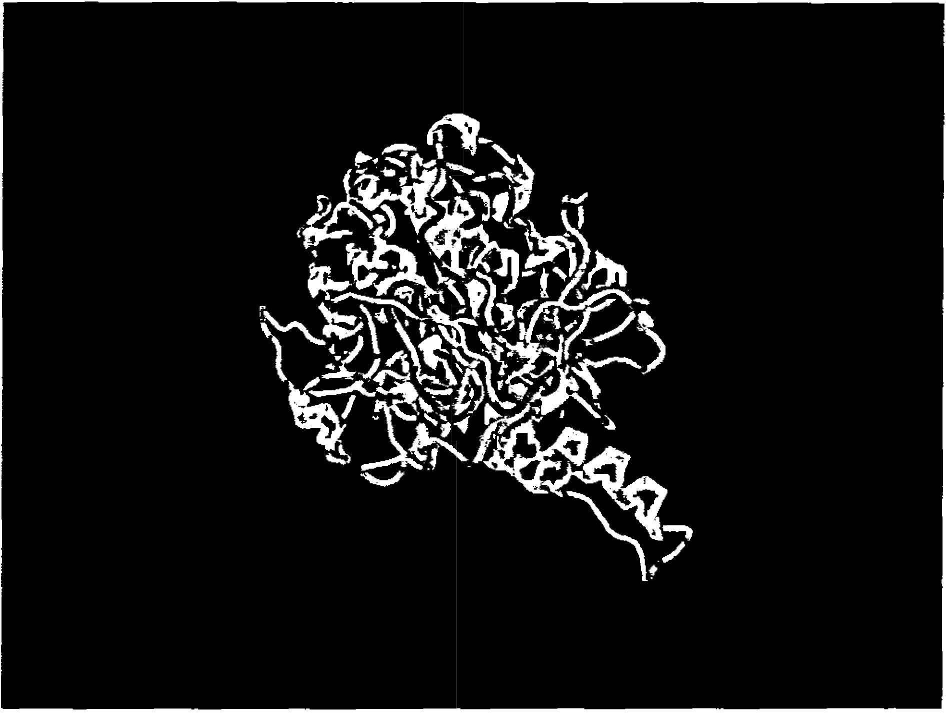


Figure 3.32 Predicted model of PARK2 gene using Modeller

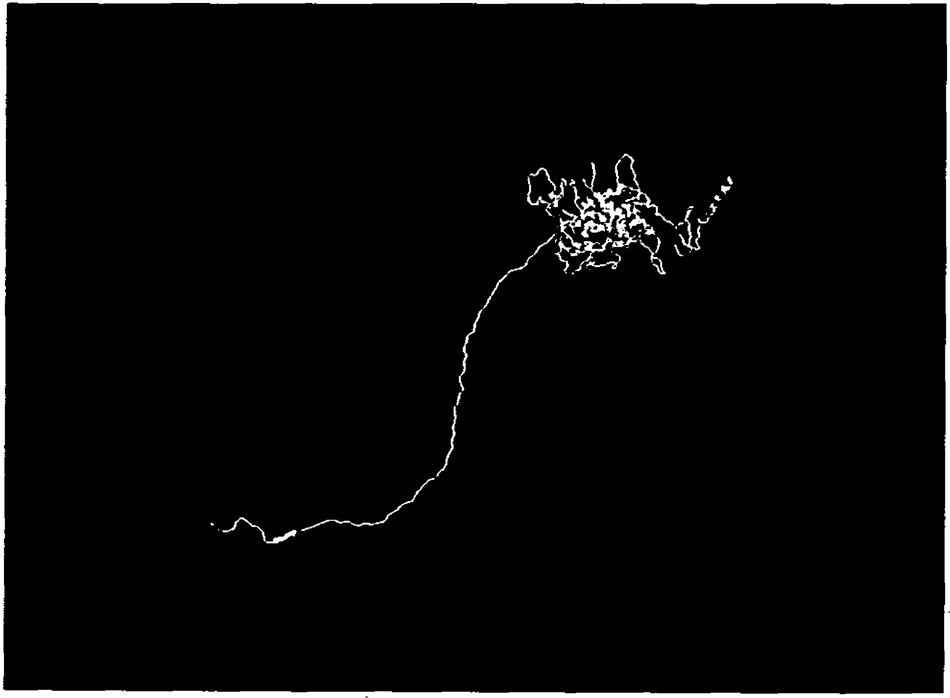


Figure 3.33 Predicted model of PINK1 gene using Modeller

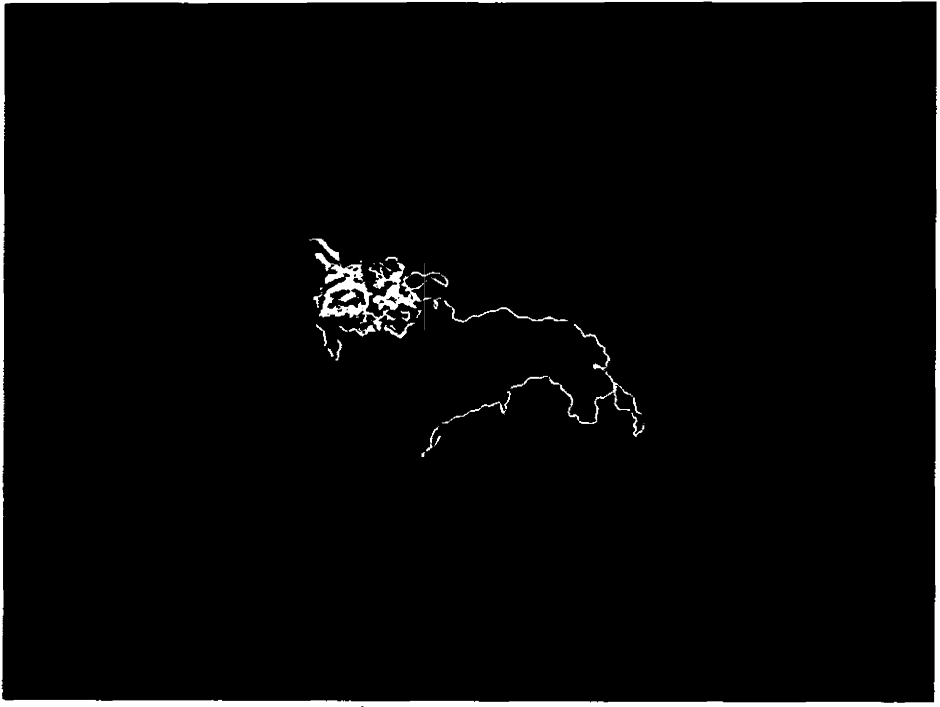


Figure 3.34 Predicted model of HTRA2 gene using Modeller

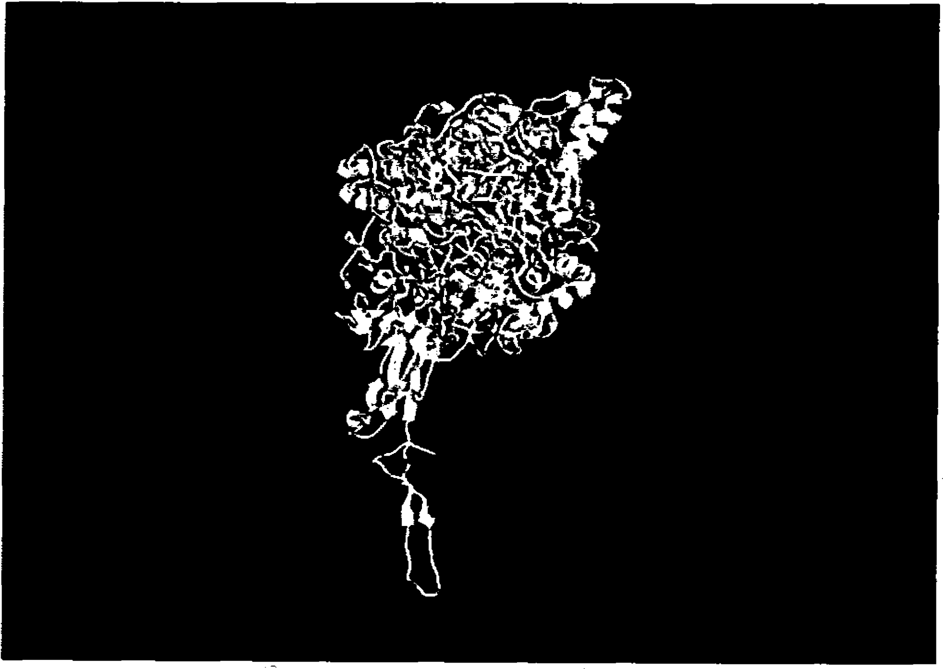


Figure 3.35 Predicted model of ATP13A2 gene using Modeller

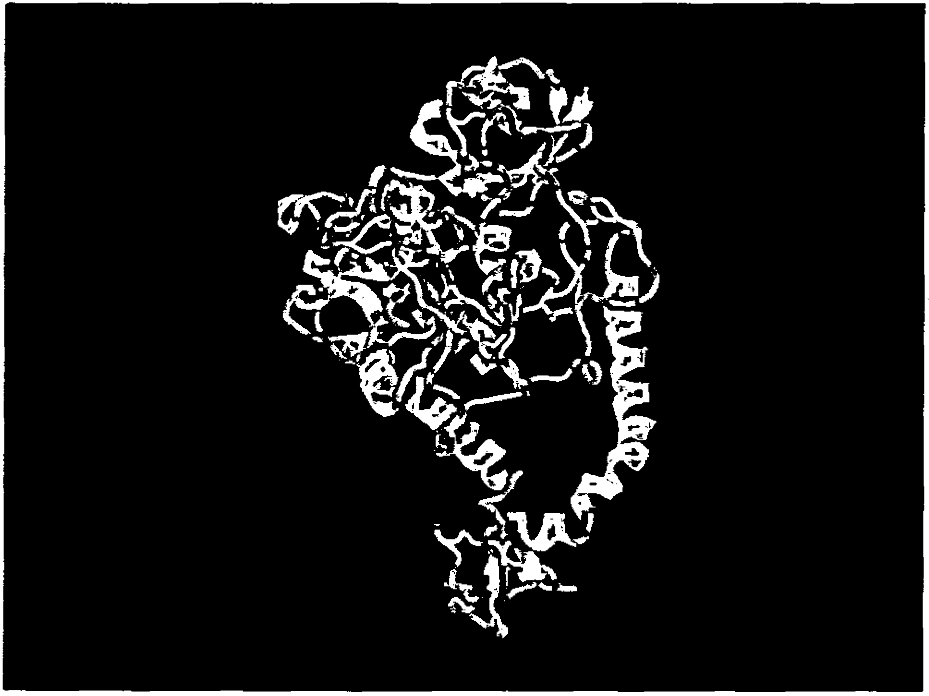


Figure 3.36 SAM-T08 predicted model of PARK2 gene

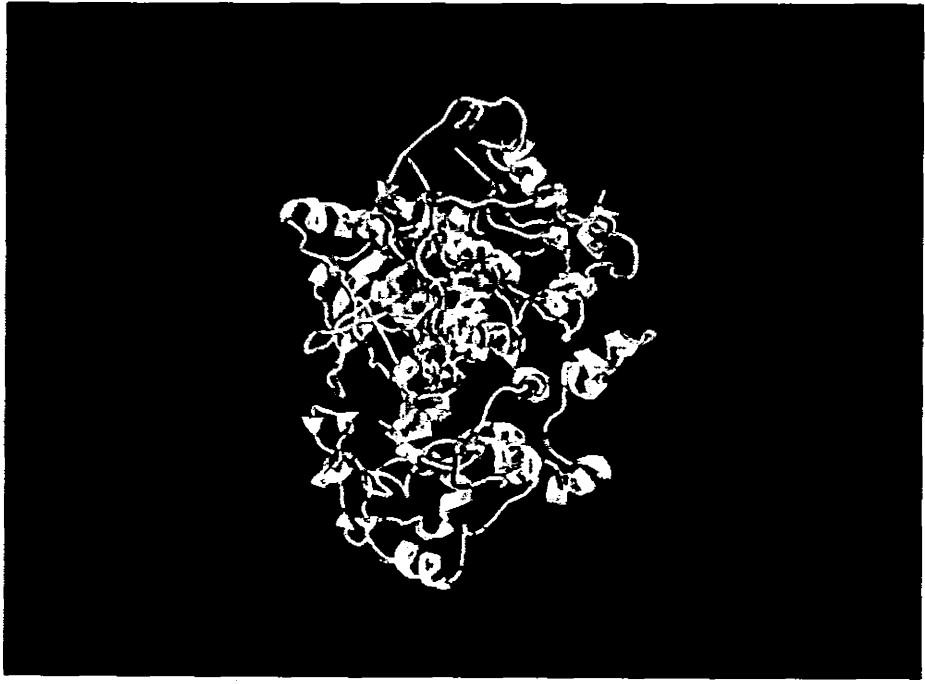


Figure 3.37 SAM-T08 predicted model of PINK1 gene

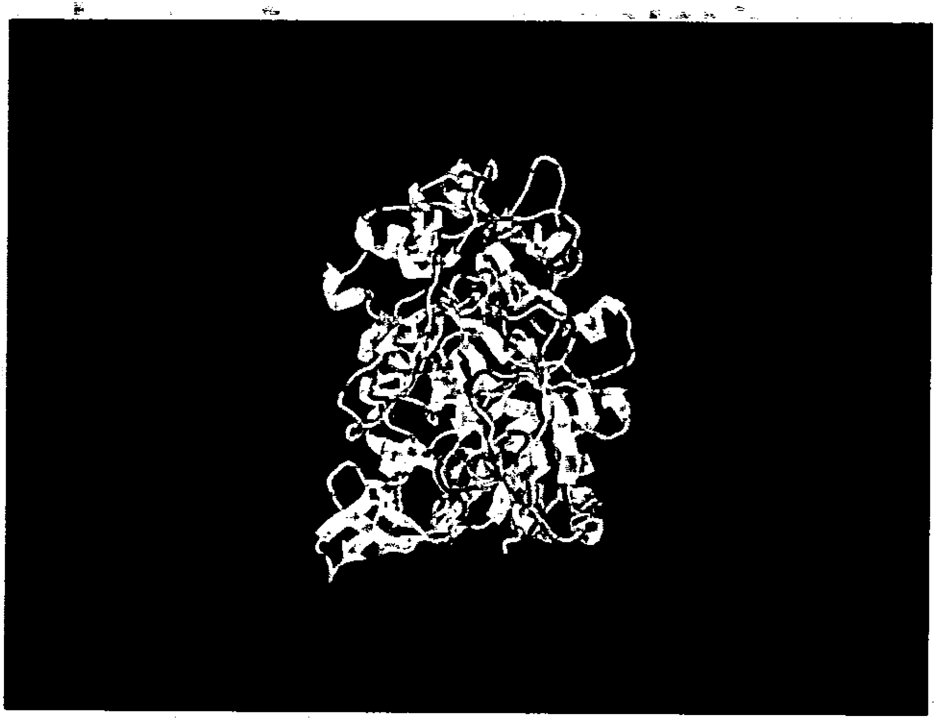


Figure 3.38 SAM-T08 predicted model of HTRA2 gene

Table 3.43 ProtParam results of PARK2 gene

Number of amino acids	465
Molecular weight	51640.6
Theoretical pI	6.71
number of negatively charged residues (Asp + Glu)	51
number of positively charged residues (Arg + Lys)	49
Formula	$C_{2228}H_{3474}N_{658}O_{676}S_{42}$
Total number of atoms	7078
The instability index	47.72

Table 3.44 ProtParam results of PINK1 gene

Number of amino acids	581
Molecular weight	62769.0
Theoretical pI	9.43
number of negatively charged residues (Asp + Glu)	46
number of positively charged residues (Arg + Lys)	66
Formula	$C_{2796}H_{4496}N_{804}O_{781}S_{28}$
Total number of atoms	8905
The instability index	48.30

Table 3.45 ProtParam results of HTRA2 gene

Number of amino acids	458
Molecular weight	48840.8
Theoretical pI	10.07
number of negatively charged residues (Asp + Glu)	40
number of positively charged residues (Arg + Lys)	50
Formula	$C_{2155}H_{3503}N_{641}O_{639}S_{7}$
Total number of atoms	6945
The instability index	43.11

Table 3.46 ProtParam results of ATP13A2 gene

Number of amino acids	1180
Molecular weight	128793.5
Theoretical pI	8.47
number of negatively charged residues (Asp + Glu)	93
number of positively charged residues (Arg + Lys)	102
Formula	$C_{5818}H_{9297}N_{1547}O_{1630}S_{56}$
Total number of atoms	18348
The instability index	50.30

Table 3.47 Percentage Alignment between target and template sequences

Tool Used	Genes Name	Templates	Identities	Template Length (amino acids)	Gene Sequence Length (amino acids)
NCBI BLAST	PARK2	1WX7|A	35%	106	465
		1WJV|A	38%	79	
		2DQ7|X	34%	283	
		2JMO|A	·98%	80	
		3DPG|A	41%	338	
	PINK1	2WEL|A	31%	327	581
		1VJY|A	34%	303	
		2JFM|A	33%	325	
	HTRA2	1LCY|A	99%	325	458
		2PZD|A	100%	113	
	ATP13A2	2KIJ|A	44%	124	1180
		2RAR|A	42%	261	
		2O36|A	38%	674	
		2VOY|I	38%	128	
		2DEW|X	35%	671	

Table 3.48 Modeller results of Generated Model

S.No.	Gene Name	Objective Number Value
1.	PARK2	22014.4062
2.	PINK1	17304.8809
3.	HTRA2	6377.9673
4.	ATP13A2	68882.0391

3.8 EVALUATION OF PREDICTED MODELS

Structural data was required in order to study the activity and function of a protein. After performing computational method of protein modeling, the validation of the predicted models was evaluated. Following different tools were used to evaluate the predicted models of the selected genes and then on the basis of different values obtained one best model was selected.

3.8.1 WHAT IF

In the field of homology modeling, drug designing, electrostatics calculations, structure validation and visualization WHAT IF provides almost 2000 options. WHAT IF structure validation gave results on the bases of Z-Score. The score showed, how well in the Ramachandran plot the backbone conformation of all residues corresponds to the known allowed areas was within expected ranges for well-refined structures. Z-Score of generated models was shown in table 3.49 and comparison of generated models and template structure was shown in table 3.50.

3.8.2 ProSA

The major problem in the theoretical and experimental models of protein was the recognition of errors. ProSA-web provided an easy interface for protein structure validation. For any input structure, ProSA calculated an overall quality score, if this score is outside the range of characteristic for native protein then the structure may contains errors. This overall model quality score was indicated by Z-Score (Fig. 3.39, 3.40, 3.41, 3.42). ProSa results were shown in table 3.51, 3.52, 3.53 and 3.54.

3.8.3 ProCheck

PROCHECK is an evaluation tool that checks the stereo-chemical quality of a protein structure. It produced a large number of plots which analyzed its overall and residue-by-residue geometry. ProCheck Ramachandran core values of generated models and their comparison with template structures were shown in table 3.55 and 3.56 respectively.

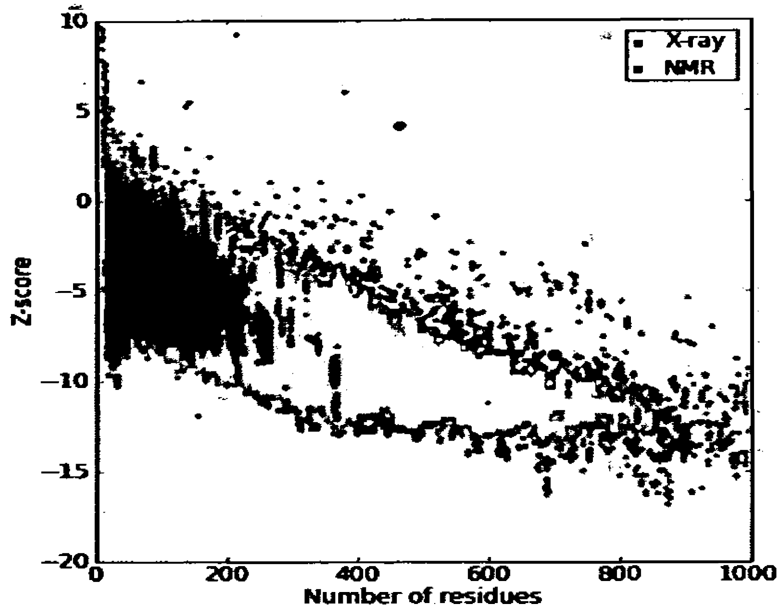


Figure 3.39 ProSa Z-Score plot for PARK2 gene

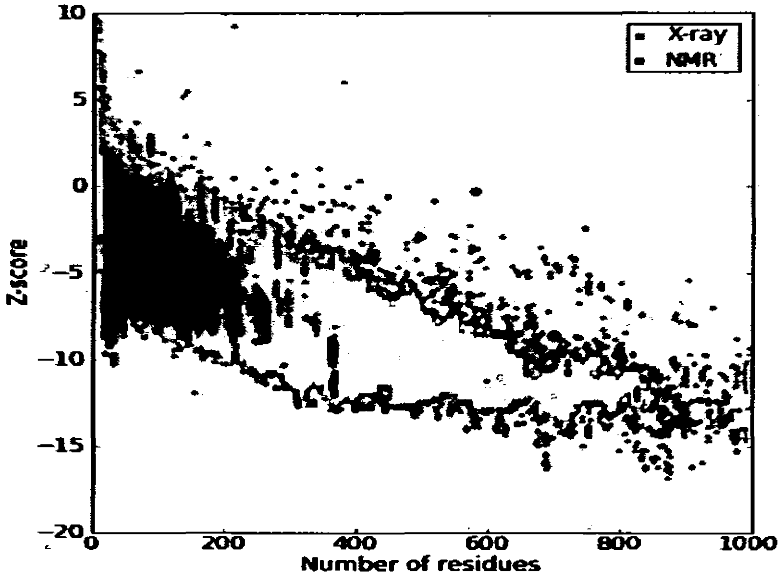


Figure 3.40 ProSa Z-Score plot for PINK1 gene

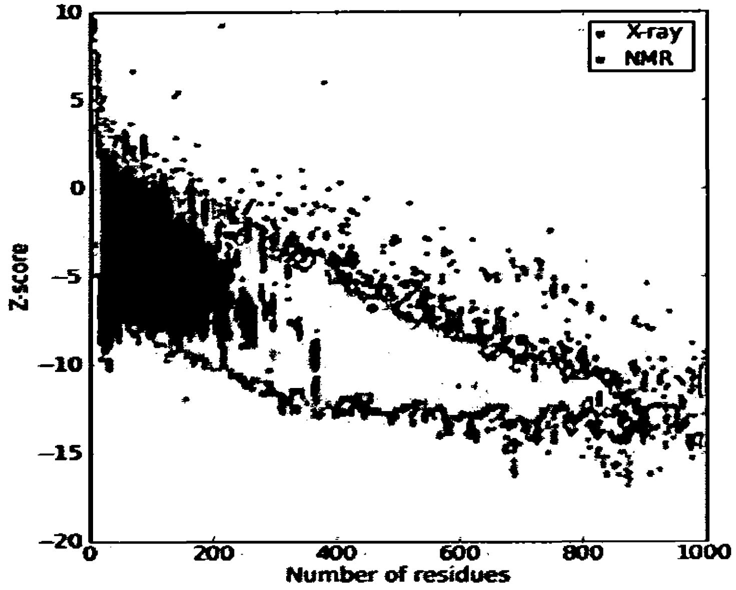


Figure 3.41 ProSa Z-Score plot for HTRA2 gene

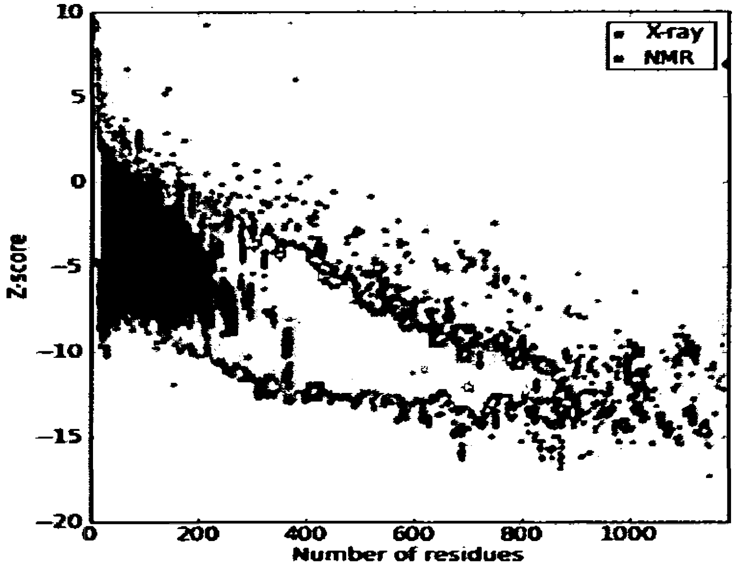


Figure 3.42 ProSa Z-Score plot for ATP13A2 gene

Table 3.49 WHAT IF Z-Score of generated Models

S.NO.	Gene Name	Z-score of Models by Modeler	Z-score of Models by SAM-T08
1.	PARK2	-3.712	-2.335
2.	PINK1	-0.687	-2.021
3.	HTRA2	-1.301	-1.610
4.	ATP13A2	-3.355	-

Table 3.50 WHAT IF Z-Score of generated Models and Template Structures

S.NO.	Gene Name	Z-score of Models by Modeler	Templates	Z-Score
1.	PARK2	-3.712	1WX7|A	-4.772
			1WJV|A	-6.202
			2DQ7|X	-3.692
			2JMO|A	-5.115
			3DPG|A	-0.630
2.	PINK1	-0.687	2WEL|A	1.045
			1VJY|A	-0.554
			2JFM|A	-2.748
			3GBZ|A	-0.408
3.	HTRA2	-1.301	1LCY|A	-2.172
			2PZD|A	-2.074
4.	ATP13A2	-3.355	2KIJ|A	-2.082
			2RAR|A	-0.426
			2O36|A	0.740
			2VOY|I	-3.233
			2DEW|X	-1.668

Table 3.51 ProSa Z-Score of generated Model and templates of PARK2 gene

Gene Name	Z-Score of model by modeler	Templates	Z-Score
PARK2	4.13	1WX7|A	-6.65
		1WJV|A	-5.19
		2DQ7|X	-7.93
		2JMO|A	-3.34
		3DPG|A	-9.14

Table 3.52 ProSa Z-Score of generated Model and templates of PINK1 gene

Gene Name	Z-Score of model by modeler	Templates	Z-Score
PINK1	-0.39	2WEL|A	-6.83
		1VJY|A	-6.84
		2JFM|A	-7.03
		3GBZ|A	-5.42

Table 3.53 ProSa Z-Score of generated Model and templates of HTRA2 gene

Gene Name	Z-Score of model by modeler	Templates	Z-Score
HTRA2	-6.91	1LCY|A	-8.47
		2PZD|A	-5.18

Table 3.54 ProSa Z-Score of generated Model and templates of ATP13A2 gene

Gene Name	Z-Score of model by modeler	Templates	Z-Score
ATP13A2	6.9	2KIJ|A	-4.15
		2RAR|A	-9.29
		2O36|A	-12.4
		2VOY|I	-4.34
		2DEW|X	-9.56

Table 3.55 ProCheck Ramachandran core value of generated Models

S.NO.	Gene Name	Ramachandran core value of Models by Modeler	Ramachandran core value of Models by SAM-T08
1.	PARK2	73.7%	87.6%
2.	PINK1	87.3%	87.8%
3.	HTRA2	90.2%	89.3%
4.	ATP13A2	72.7%	-

Table 3.56 ProCheck Ramachandran core value of generated Models and Template structures

S.NO.	Gene Name	Ramachandran core value of Models by Modeler	Templates	Ramachandran core value of templates
1.	PARK2	73.7%	1WX7|A	77.5%
			1WJV|A	75.8%
			2DQ7|X	84.0%
			2JMO|A	75.4%
			3DPG|A	90.7%
2.	PINK1	87.3%	2WEL|A	92.7%
			1VJY|A	90.7%
			2JFM|A	87.6%
			3GBZ|A	89.6%
3.	HTRA2	90.2%	1LCY|A	87.2%
			2PZD|A	91.6%
4.	ATP13A2	72.7%	2KIJ|A	87.4%
			2RAR|A	90.9%
			2O36|A	92.0%
			2VOY|I	86.7%
			2DEW|X	89.1%

3.9 Molecular Docking

3.9.1 AutoDock Vina

AutoDock Vina was used to perform the docking procedures, which is the project of The Scripps Research Institute. AutoDock Vina is an open-source program for virtual screening, molecular docking and drug discovery. This program provides multi-core compatibility, enhanced accuracy, high performance and ease of use.

Four genes of Parkinson's disease were selected, whose structure has been predicted before and their pdb files can be obtained from *www.rcsb.org*. Selected mutations were inserted in these pdb files using WHATIF tool. These mutations are shown in table 3.49. These mutations were selected for research papers present on OMIM. For these four genes, three ligands were selected to find the best conformation with these genes. Polar hydrogen atoms were added to the selected proteins, all nonpolar hydrogen atoms were merged and all ligand bonds were set to be rotatable. AutoDock Vina will calculate the binding affinities in Kcal/mol and also Root Mean Square Values (rmsd) in upper and lower bound.

Binding affinity of DJ1 gene with the three ligands Ajacine, Gallic Acid, Phenylthanoid was shown in table 3.50(a, b, c), for LRRK2 gene in table 3.51(a, b, c), for SNCA gene in table 3.52(a, b, c) and for UCHL1gene in table 3.53(a, b, c).

Table 3.57 Mutations inserted in selected Genes

Genes	Mutations inserted
DJ1	L166P
LRRK2	G2019S
SNCA	A53T
UCHL1	S18Y

Table 3.58(a) DJ1 results with ligand Ajacine

Mode	Binding affinity (Kcal/mol)	Distance from rmsd l.b.	Best mode rmsd u.b.
1.	12.8	0.000	0.000

Table 3.58(b) DJ1 results with ligand Gallic Acid

Mode	Binding affinity (kcal/mol)	Distance from rmsd l.b.	Best mode rmsd u.b.
1.	-4.3	0.000	0.000
2.	-4.3	0.007	2.402
3.	-4.0	11.955	13.931
4.	-4.0	11.950	13.848
5.	-4.0	1.331	2.818
6.	-3.9	1.406	2.185
7.	-3.7	1.290	4.042
8.	-3.7	25.675	26.594
9.	-3.6	25.630	26.476

Table 3.58(c) DJ1 results with ligand Phenylthanoid

Mode	Binding affinity (kcal/mol)	Distance from rmsd l.b.	Best mode rmsd u.b.
1.	-2.6	0.000	0.000
2.	-2.5	1.900	2.702
3.	-1.9	2.780	10.480
4.	-1.1	4.007	11.560
5.	-1.0	3.827	11.569

Table 3.59(a) LRRK2 results with ligand Ajacine

Mode	Binding affinity (kcal/mol)	Distance from rmsd l.b.	Best mode rmsd u.b.
1.	-6.6	0.000	0.000
2.	-6.2	1.484	2.255
3.	-5.8	2.658	80739
4.	-5.5	2.003	3.829
5.	-5.5	15.575	19.457
6.	-5.5	25.417	28.590
7.	-5.5	15.747	19.665
8.	-5.3	2.703	5.402
9.	-5.1	15.678	19.494

Table 3.59(b) LRRK2 results with ligand Gallic Acid

Mode	Binding affinity (kcal/mol)	Distance from rmsd l.b.	Best mode rmsd u.b.
1.	-5.7	0.000	0.000
2.	-5.6	0.303	2.411
3.	-5.4	1.311	3.848
4.	-5.0	1.380	4.054
5.	-5.0	1.452	4.590
6.	-4.9	22.240	23.536
7.	-4.8	1.491	4.400
8.	-4.8	24.849	26.031
9.	-4.8	22.243	23.534

Table 3.59(c) LRRK2 results with ligand Phenylthanoid

Mode	Binding affinity (kcal/mol)	Distance from rmsd l.b.	Best mode rmsd u.b.
1.	-7.7	0.000	0.000
2.	-7.3	2.053	3.421
3.	-7.0	1.568	2.097
4.	-6.9	1.752	2.795
5.	-6.9	3.019	9.587
6.	-6.6	13.299	18.590
7.	-6.6	12.595	16.841
8.	-6.6	2.787	9.664
9.	-6.6	3.540	7.022

Table 3.60(a) SNCA results with ligand Ajacine

Mode	Binding affinity (kcal/mol)	Distance from rmsd l.b.	Best mode rmsd u.b.
1.	-0.0	0.000	0.000
2.	-0.0	2.715	6.627
3.	-0.0	8.802	12.105
4.	-0.0	6.960	10.248
5.	-0.0	7.310	10.506
6.	-0.0	10.189	13.786
7.	-0.0	7.821	12.965
8.	-0.0	8.823	11.662
9.	-0.0	5.835	10.304

Table 3.60(b) SNCA results with ligand Gallic Acid

Mode	Binding affinity (kcal/mol)	Distance from rmsd l.b.	Best mode rmsd u.b.
1.	0.0	0.000	0.000
2.	0.0	5.236	7.105
3.	0.0	7.344	9.050
4.	0.0	6.102	7.873
5.	0.0	9.704	11.371
6.	0.0	3.465	5.706
7.	0.0	11.385	12.840
8.	0.0	4.996	6.972
9.	0.0	4.199	5.752

Table 3.60(c) SNCA results with ligand Phenylthanoid

Mode	Binding affinity (kcal/mol)	Distance from rmsd l.b.	Best mode rmsd u.b.
1.	-0.0	0.000	0.000
2.	-0.0	9.192	11.956
3.	-0.0	2.586	4.891
4.	-0.0	7.098	9.243
5.	-0.0	5.495	8.590
6.	-0.0	8.945	12.118
7.	-0.0	7.397	10.034
8.	-0.0	2.409	6.118
9.	-0.0	3.068	6.959

Table 3.61(a) UCHL1 results with ligand Ajacine

Mode	Binding affinity (kcal/mol)	Distance from rmsd l.b.	Best mode rmsd u.b.
1.	-7.2	0.000	0.000
2.	-7.1	3.673	6.755
3.	-7.1	13.448	16.910
4.	-6.8	12.501	15.824
5.	-6.7	13.068	16.195
6.	-6.7	2.773	8.596
7.	-6.7	2.915	6.486
8.	-6.7	12.843	17.889
9.	-6.6	12.795	16.659

Table 3.61(b) UCHL1 results with ligand Gallic Acid

Mode	Binding affinity (kcal/mol)	Distance from rmsd l.b.	Best mode rmsd u.b.
1.	-5.4	0.000	0.000
2.	-5.4	0.118	2.403
3.	-5.3	1.919	4.287
4.	-5.3	1.920	4.361
5.	-5.2	3.116	5.146
6.	-5.2	3.134	4.25
7.	-5.2	3.142	4.902
8.	-5.1	2.480	4.791
9.	-5.1	10.859	12.876

Table 3.61(c) UCHL1 results with ligand Phenylthanoid

Mode	Binding affinity (kcal/mol)	Distance from rmsd l.b.	Best mode rmsd u.b.
1.	-6.9	0.000	0.000
2.	-6.9	2.364	4.759
3.	-6.7	2.011	5.077
4.	-6.7	2.187	5.670
5.	-.6.	13.241	17.983
6.	-6.6	19.044	21.804
7.	-6.6	21.312	24.743
8.	-6.5	21.097	24.460
9.	-6.5	15.029	18.699

5. DISCUSSION

Proteins perform diverse functions in human body; in fact about 45% of human body is of proteins. Proteins are the main performers in the entire organism; they act as enzymes, hormones, building blocks, antibodies, source of energy and also have nutritional importance. Hundreds of thousands of protein structures have been found either through nucleic acid sequencing or direct protein sequencing. This three-dimensional model of protein determines the function of protein. Different diseases are caused when proteins are deformed or not fold correctly. These mutations cause diseases as Parkinson, cystic fibrosis, cancer, diabetes etc.

After the prediction of the 3D model of the protein the next most important feature of computational biology is the drug designing. Improvements in the structure prediction methods lead to the improved making of new drugs. Structure prediction identifies new target domains for new drugs. By knowing the structure of the target protein, full knowledge of its functional sites, specifically the diseased sites to be inactivated or activated by new target drugs can be obtained. In-silico drug designing has some key roles as it can be used for finding new antagonists or agonists for a target molecule using a series of methods either we know or do not know the structure of the target (Ekins *et al.*, 2007)

Parkinson's disease is one of the most common neurodegenerative disorders. It causes resting tremor, bradykinesia, muscular rigidity and postural instability, in addition to postural abnormalities (Stuart *et al.*, 2009). It is a non-contagious disease and may have both genetic and environmental factors involved. Many genes are involved in this disease that's why few genes were selected for protein structure prediction and drug designing.

In the present research work, functional site analysis of the selected proteins includes the prediction of domains, motifs, patterns, N-glycosylation sites, n-myristoylation sites, phosphorylation sites and amidation site. Different tools were used to predict different domains to make comparisons and then verify the results obtained. Interproscan results showed ubiquitin domain in PARK2. This is a highly conserved domain of 76 amino acids. It has seven lysins of 4-residues C-terminal tail, which forms chain with the target protein (Passmore *et al.*, 2004). Zinc finger domains were also found in PARK2, which act as building blocks for the construction of large protein domains (Klug *et al.*, 2002). In PINK1 gene the most important domains found were protein kinase catalytic domains. It is responsible for most of the phosphotransfer reactions resulting in transfer of gamma phosphates from NTP's to one or more amino acid residues, resulting in conformational change that affect protein function. Using interproscan the most important domains retrieved for HTRA2 were peptidases, which can be identified by their catalytic type, the other domain was PDZ, which were found in diverse signaling proteins. There were 80-90 amino acids in these domains, which consist of 6 beta-strands (A to F) and 2 alpha-helices, A and B arranged in globular structures. These domains allow the binding to a free carboxylate group between beta-A and beta-B strands. In ATP13A2 the most important domain retrieved was ATPases.

The domains predicted by ELM and the values of the amino acids of the domains are illustrated in figure 3.13, 3.14, 3.15 and 3.16.

ScanProsite predicted patterns of the four selected genes (PARK2, PINK1, HTRA2 and ATP13A2). Patterns are said to be the textural representation of motifs. The first pattern predicted for all the four genes was N-glycosylation sites, which are specific sequence Asn-Xaa-Ser/Thr, where *X* can be any amino acid residue except for Pro. The presence of this consensus pattern is not sufficient to conclude that aspargine is glycosylated. N-glycosylation is an important co-translational modification in endoplasmic reticulum. It is catalyzed by an enzyme called oligosaccharyl transferase (Yan *et al.,* 2005). 3 N-glycosylation sites were found in PARK2 and ATP13A2 gene, 1 in PINK2 and 2 in HTRA2.

Second pattern was N-myristoylation site, which is C14-saturated fatty acid site having N-terminal residues which are accylated with the covalent addition of C14-saturated fatty acid through an amide linkage. Total 10 N-myristoylation sites were found in PARK2, PINK1 and HTRA2 gene while 12 sites were found in ATP13A2.

Third pattern was phosphorylation sites which include Casein kinase 2, protein kinase C, cAMP and cGMP-dependent protein kinase and tyrosine kinase. Phosphorylation is an important protein post-translational modification. Identification of important phosphorylation sites is important for understanding the function of protein. Insilico prediction can help narrow down the experimental efforts and can also provide functional candidates (Du *et al.,* 2010). PARK2 found total 16 serines and therionines that are phosphorylated. PINK1 has 18, HTRA2 has 10 and ATP13A2 has 22 serines and therionines that are phosphorylated. In HTRA2 gene, 1 amidation site was also found.

The SMART server developed by Letunic *et al.*, 2008, showed the diagrammatic view of the domains and transmembranes along with the start and end values of amino acids encoding these domains and their expect value can also be predicted.

Post translation modifications are important to make the proteins functional and SignalP was used to find the cleavage site in the selected genes of Parkinson's disease which is the most important feature of SiganlP that it can predict cleavage site by using Neural Network algorithm and Hidden Markov Model. The cleavage site for PARK2 gene was between 19 and 20 amino acids, for PINK1 gene between 28 and 29 amino acids, for HTRA2 gene was between 11 and 12 amino acids and for ATP13A2 gene was between 17 and 18 amino acids.

Primary structure prediction involves the calculation of physical and chemical parameters of the four selected genes of Parkinson's disease. Secondary structure prediction is preliminary for the prediction of tertiary structure. GOR4 server not only predicted the secondary structure but it also showed the graphical representation. The mostly observed random distribution of the three states in globular protein is 30%.

Similarity search was made using BLAST and FASTA, to find the sequences having significant sequence alignment with the query sequences, which can be than used in protein modelling. Modeller 9v7 developed by Sali *et al.*, 1993 was used to build 3D structures of selected sequences. Modeller uses homology modelling approach to construct 3D models. In the first step of homology modelling, sequence alignment of the target sequence is made with the selected template sequences from BLAST, having more than 30% of sequence identity. Five templates were selected for PARK2 and ATP13A2 genes, 3 for PINK1 and 2 for HTRA2, the pdb files of these templates were taken from RCSB. For each of the target template 10 models were generated through Modeller and from these models the best model was selected having the lowest objective function value. SAM-T08 uses protein threading method to predict the 3D models of the target sequences. In protein threading, the compatibility of the target sequence to the structure has been assessed by a scoring function. But it could not predict 3D models of proteins with more than 700 amino acid residue.

Evaluation of the predicted models was made using three different tools. Predicted model quality can be measured in terms of overall greater quality factor, greater number of residues in favored regions and least z-score value. WHAT IF and ProSA calculated the model quality in terms of Z-score while ProCheck calculated the Ramachandran core quality factor for the predicted models.

Molecular docking was performed to obtain possible orientations and conformations for the ligand at the binding sites. AutoDock Vina was reported in a research paper by Trott *et al.,* 2010. Molecular docking of those proteins have been performed whose models have been predicted and their pdb files were taken from RCSB. These proteins include DJ1, LRRK2, SNCA and UCHL1. Potential ligand search was made and three ligands were selected to perform docking procedures. These ligands are ajacine, gallic acid and phenylthanoid. Mol files of these ligands were taken from ligand database and these files were converted to pdb files using Marvin Sketch, as AutoDock Vina do not accepts .mol files. Binding affinities of these ligands were calculated with the selected proteins using the software. AutoDock Vina also calculates the Root Mean

Square values in upper and lower bound. DJ1 shows high binding affinity with gallic acid i.e., -4.3kcal/mol. LRRK2 with phenylthanoid i.e., -7.7 kcal/mol. SNCA with ajacine and phenylthanoid both i.e., -0.0kcal/mol and UCHL1 with ajacine i.e., -7.2kcal/mol. The lower the binding affinity value, the best the ligand and receptor binding.

Structure based drug designing of a novel antiflaviviral inhibitor for nonstructural 3 (NS3) protein has been done using Modeller 9v7 and AUTODOCK vina. Flaviviral causes a number of diseases including encephalitis, fevers andhemorrhagic fevers. NS3 is a multifunctional protein which is involved in proteolytic processing of viral polyprotein, and C-terminal region, thus it acts as inhibitor to stop the proteolytic processing. Evaluation of the predicted models was done using PROCHECK and ProSA (Jitendra and Vinay, 2011).

Novel 3D models of proteins involved in Parkinson's disease were predicted, these proteins include PARK2, PINK1, HTRA2 and ATP13A2 and molecular docking techniques were applied to find the binding energies of three ligands with proteins DJ1, LRRK2, SNCA and UCHL1. 3D model prediction and molecular docking techniques can be applied for other genes of Parkinson disease or any other inherited disease and also these procedures can be adopted in wet labs to develop more potential drugs for Parkinson disease.

6. REFERENCES

Barbeau A and Pourcher E (1982). New data on the genetics of Parkinson's disease. Canadian Journal of Neurological Sciences 9: 53-60

Bonifati V, Fabrizio E, Vanacore N, Mari M and Meco G (1995). Familial Parkinson's disease: a clinical genetic analysis. Canadian Journal of Neurological Sciences 22: 272-279

Bornot A, Etchebest C, and Brevern A G (2009). A new prediction strategy for long local protein structures using an original description. Proteins 76(3):570-87

Cookson MR (2010). Unravelling the role of defective genes. Progress in Brain Research 183:43-57

Du P, Li T and Xu N (2010). Identifying human kinase-specific protein phosphorylation sites by integrating heterogeneous information from various sources. PLoS One 5:15411

Duvoisin RC, Eldridge R, Williams A, Nutt J and Calne D (1981). Twin study of Parkinson disease. Neurology 31: 77-80

Edouard C, Christian JAS, Alexandre G, Virginie B, Petra SLG, Elisabeth G, Amos B and Nicolas H (2006). ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Research 34: W362–W365

Ekins S, Mestres J and Testa B (2007). In silico pharmacology for drug discovery: applications to targets and beyond. British Journal of Pharmacology 152(1):21-37

Abdelghaffar H, Dakroury S, Elhak SA, Ghanem AA, and Salama MM (2010). Parkinson's disease: Is It a Toxic Syndrome? Neurology Research International

Gamsjaeger R, Liew CK, Loughlin FE, Crossley M and Mackay JP (2007). Sticky fingers: zinc-fingers as protein-recognition motifs. Trends in Biochemical Sciences 32(2):63-70

Garnier J, Gibrat JF and Robson B (1996). GOR secondary structure prediction method version IV. Methods in Enzymology R.F. Doolittle Ed. 266: 540-553

Geourjon C and Deleage G (1995). SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Computer Application Bioscience Journal 11(6):681-684

Hamza TH, Zabetian CP, Tenesa A, Laederach A, Montimurro J, Yearout D, Kay DM, Doheny KF, Paschall J, Pugh E, Kusel VI, Collura R, Roberts J, Griffith A, Samii A, Scott WK, Nutt J, Factor SA and Payami H (2010). Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. Nature Genetics 42: 781-785

Hofmann K and Stoffel W (1993). TMbase - A database of membrane spanning proteins segments. Biological Chemistry Hoppe-Seyler 374:166

Jannick DB, Henrik N, Gunnar VH and Soren B (2004). Improved prediction of signal peptides: SignalP 3.0. Journal of Molecular Biology 340:783-795

Jitendra S and Vinay R (2011). Structure based drug designing of a novel antiflaviviral inhibitor for nonstructural 3 protein. Bioinformation 6(2): 57–60

Karplus K (2009). SAM-T08: HMM-based protein structure prediction. Nucleic Acids Research 37: 492-7

Kissel P and Andre JM (1976). Maladie de parkinson et anosmie chez deux jumelles monozygotiques. Journal Genetic Human 24: 113-117 monozygotiques. Journal Genetic Human 24: 113-117

Laskowski RA, MacArthur MW, Moss DS, and Thornton JM (1993). PROCHECK: A program to check the stereochemical quality of protein structures. Journal of Applied Crystallography 26: 283-291

Letunic I, Doerks T and Bork P (2009). SMART 6: recent updates and new developments. Nucleic Acids Research 37:229-232

Matthews JM, Gamsjaeger R, Swanton MK, Kobus FJ, Lehtomaki E, Lowry JA, Kwan AH and Mackay JP (2008). Structural and biophysical analysis of the DNA binding properties of myelin transcription factor 1. Journal of Biological Chemistry 283(8):5158-67

McDowall J and Hunter S (2011). InterPro protein classification. Methods in Molecular Biology 694:37-47

Neumann J, Bras J, Deas E, Sullivan SS, Parkkinen L, Lachmann RH, Li A, Holton J, Guerreiro R, Paudel R, Segarane B, Singleton A, Lees A, Hardy J, Houlden H, Revesz T and Wood NW (2009). Glucocerebrosidase mutations in clinical and pathologically proven Parkinson's disease. Brain 132: 1783-1794

Nussbaum RL and Polymeropoulos MH (1997). Genetics of Parkinson's disease. Human Molecular Genetics 6: 1687-1691

Passmore LA and Barford D (2004). Getting into position: the catalytic mechanisms of protein ubiquitylation. Biochemical Journal. 379: 513-25

Pierri CL, Parisi G and Porcelli V (2010). Computational approaches for protein function prediction: a combined strategy from multiple sequence alignment to molecular docking-based virtual screening. Biochimica et Biophysica Acta 1804(9): 1695-712

Polymeropoulos MH, Higgins JJ, Golbe LI, Johnson WG, Ide SE, Di Iorio G, Sanges G, Stenroos ES, Pho LT, Schaffer AA, Lazzarini AM, Nussbaum RL and Duvoisin RC (1996). Mapping of a gene for Parkinson's disease to chromosome 4q21-q23. Science 274: 1197-1198

Puntervoll P, Linding R, Gemund C, Davidson CS, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A, Ferrè F, Maselli VVia A, Cesareni G, Diella F, Furga SG, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Kuster B, Citterich HM, Hunter WN, Aasland R and Gibson TJ (2003). ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Research 31(13): 3625-3630

Sali A and Blundell TL (1993). Comparative protein modelling by satisfaction of spatial restraints. Journal of Molecular Biology 234L: 779-815

Stuart H, Isaacson, Robert A and Hauser (2009). Improving Symptom Control in Early Parkinson's disease. Therapeutic Advances in Neurological Disorders 2(6): 29–41

Sveinbjornsdottir S, Hicks AA, Jonsson T, Petursson H, Guomundsson G, Frigge ML, Kong A, Gulcher JR and Stefansson K (2000). Familial aggregation of Parkinson's disease in Iceland. The New England Journal of Medicine 343: 1765-1770

Theuns J, Brouwers N, Engelborghs S, Sleegers K, Bogaerts V, Corsmit E, Pooter T, Duijn CM, Deyn PP and Broeckhoven C (2006). Promoter mutations that increase amyloid precursor-protein expression are associated with Alzheimer disease. The American Journal of Human Genetics 78: 936-946

Trenkwalder C, Schwarz J, Gebhard J, Ruland D, Trenkwalde P, Hense HW and Oertel WH (1995). Starnberg trial on epidemiology of parkinsonism and hypertension in the elderly: prevalence of Parkinson's disease and related disorders assessed by a door-to-door survey of inhabitants older than 65 years. Archives of Neurology 52: 1017-1022

Trott O and Olson AJ (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry 31(2):455-61

Tusnady GE and Simon I (2001). The HMMTOP transmembrane topology prediction server. Bioinformatics 17: 849-850

Varkonyi J, Rosenbaum H, Baumann N, MacKenzie JJ, Simon Z, Aharon-PJ and Walker JM (2003). Gaucher disease associated with Parkinsonism: four further case reports. American Journal of Medical Genetics 116: 348-351

Vriend G (1990). Whatif: A molecular modeling and drug designing program. Journal of Molecular Graphics and Modelling 29: 52-56

Ward CD, Duvoisin RC, Ince SE, Nutt JD, Eldridge R and Calne DB (1983). Parkinson's disease in 65 pairs of twins and in a set of quadruplets. Neurology 33: 815-824

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

Yan A and Lennarz WJ (2005). Unraveling the mechanism of protein N-glycosylation. The Journal of Biological Chemistry 280:3121-4

Electronic Database Information

1. BLAST: http://blast.ncbi.nlm.nih.gov/Blast.cgi
2. NCBI: www.ncbi.nlm.nih.gov
3. GeneCards: www.genecards.org
4. GenBank: www.ncbi.nlm.nih.gov/Genbank
5. OMIM: www.ncbi.nlm.nih.gov/omim
6. Ensemble: www.ensembl.org
7. RCSB: www.pdb.org/
8. PubMed: www.ncbi.nlm.nih.gov/pubmed
9. NCBI proteins: www.ncbi.nlm.nih.gov/guide/proteins/
10. KEGG LIGAND Database: www.genome.jp/ligand/
11. PubChem Structure Search: www. pubchem.ncbi.nlm.nih.gov/search/search.cgi