Combined sequence and sequence-structure based analysis of SNPs associated with genes involved in Parkinson’s disease.

A Major Project dissertation submitted

in partial fulfilment of the requirement for the degree of

Master of Technology
In
Biomedical Engineering

Submitted by

Deepak Kumar

(2K16/BME/03)

Delhi Technological University, Delhi, India

Under the supervision of

Prof. Pravir Kumar (Professor)

Department of Biotechnology
Delhi Technological University
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road,
Delhi-110042, INDIA

CERTIFICATE

This is to certify that the M. Tech. dissertation entitled **"Combined sequence and sequence-structure based analysis of SNPs associated with genes involved in Parkinson's disease"**, submitted by **Deepak Kumar (2K16/BME/03)** in partial fulfilment of the requirement for the award of the degree of Master of Engineering, Delhi Technological University (Formerly Delhi College of Engineering, University of Delhi), is an authentic record of the candidate's own work carried out by him under my guidance.

The information and data enclosed in this dissertation is original and has not been submitted elsewhere for honouring of any other degree.

Date:

Prof. Pravir Kumar (Professor)
(Project Mentor)
Department of Bio-Technology
Delhi Technological University
(Formerly Delhi College of Engineering, University of Delhi)

DECLARATION

I certified that the project report entitled **“Combined sequence and sequence-structure based analysis of SNPs associated with genes involved in Parkinson disease.”** submitted by me is in partial fulfilment of the requirement for the award of the degree of Master of Technology in Biomedical Engineering, Delhi Technological University. It is a record of original research work carried out by me under the supervision of **Prof. Pravir Kumar,** Department of Biotechnology, Delhi Technological University, Delhi.

The matter embodied in this project report is original and has not been submitted for the award of any Degree/Diploma.

Date:

Deepak Kumar
2k16/BME/03
Department of Biotechnology
Delhi Technological University
Delhi-110042

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to my supervisor and mentor, Prof. Pravir Kumar who has guided me through thick and thin. Without his valuable guidance, incessant help, calm endurance, constructive criticism, and constant encouragement the completion of this minor project would have been impossible, the congenial discussion with him made me confident and gave enough independence in performing the experiment which enabled in me a higher efficiency to complete the work.

I highly indebted to Mr. Dheeraj Kumar (Research Scholar) for his guidance and constant supervision as well as for providing necessary information regarding the tools and experiments and also for his support in completing the reports. I am highly grateful to Rohan Gupta (M.tech Bioinfo) and Rohan Singh (M.tech Bioinfo) for their valuable suggestions.

At last but never the least, words are small trophies to express my deep sense of gratitude and affection to my loving friend and my parents who gave me infinite love to go for this achievement.

Deepak Kumar

2K16/BME/03

CONTENTS

TOPIC	PAGE NO
LIST OF FIGURES	1
LIST OF TABLES	2
LIST OF ABBREVIATIONS	3
1. ABSTRACT	4
2. INTRODUCTION	5-7
3. REVIEW OF LITERATURE • SNP • Parkinson’s Disease • Neurochemical and Neuropathological features of PD	8-10
4. METHODOLOGY • SIFT • SNP & GO • PolyPhen-2 • i-Mutant 3.0 • Computational Site Directed Mutagenesis • Mutagenesis Steps	11-12
5. STASTISTICAL ANALYSIS	13-14
6. RESULTS	15-19
7. DISCUSSION	20

8. CONCLUSION 21

9. REFERENCES 22-24

LIST OF FIGURES

Description **Page no**

1. Schematic diagram of LRRK2 domain 7
2. Neuropathology of Parkinson’s disease 9
3. SIFT Interface 14
4. SNP&GO Interface 15
5. PolyPhen2 Interface 16
6. I-Mutant 3.0 Interface 17
7. Ribbon structure of rocCOR domain 18
8. Mutagenesis step 1 19
9. Mutagenesis step 2 19
10. Mutagenesis step 3 20
11. Mutagenesis step 4 20
12. Mutagenesis step 6 20
13. Mutagenesis step 8 20
14. Experimental workflow 24
15. Total energy values plot 27
16. Superimposed structure of wild and mutated LRRK2 gene 28

LIST OF TABLES

Description **Page no**

1. PD genetics Nomenclature 3
2. SIFT score 13
3. PolyPhen-2 score 14
4. SNP & GO score 14
5. Diseased SNPs 15

LIST OF ABBREVIATIONS

ARK	Ankyrin repeats
ARM	Armadillo repeats
COR	C-terminal of Roc
DNA	Deoxyribonucleic acid
GWAS	Genome Wide Association Studies
LRRK2	Leucine-rich Repeat Kinase 2
MCC	Matthew Coefficient Correlation
NAMD	Nanoscale Molecular Dynamics
NCBI	National Centre for Biotechnology Information
NPV	Negative Predictive Value
PARK	Parkinson Disease Associated Genes
PD	Parkinson’s Disease
PINK1	PTEN induced putative kinase 1
PSI-BLAST	Position-Specific Iterative Basic Local Alignment Search Tool
RI	Reliability Index
ROC	Ras of Complex Proteins
SNCA	Synuclein Alpha
SNP	Single Nucleotide Polymorphism

Combined sequence and sequence-structure based analysis of SNPs associated with genes involved in Parkinson disease.

Deepak Kumar

Delhi Technological University, Delhi, India

ABSTRACT

Mutations in SNCA, LRRK2, PINK and DJ1 plays a very important part in pathological process of Parkinson's disease therefore SNPs associated with these genes were picked for detailed examination of their unfavourable effects on human body. SNPs were taken from NCBI dbSNP and only missense and mutations with unknown significance were taken into consideration. To study the deleterious effect of these SNPs we followed sequence specific and sequence-structure specific methods in order to provide more accurate results. SIFT, PolyPhen-2, SNP & Go and iMutant3.0 were used for detection of deleterious SNPs and MD simulations were performed using NAMD to validate the results. The study suggested that V1598E and P2119L of LRRK2 gene could indirectly or directly affect the Hydrogen bonding pattern and destabilize the amino acid interactions of gene to certain extent.

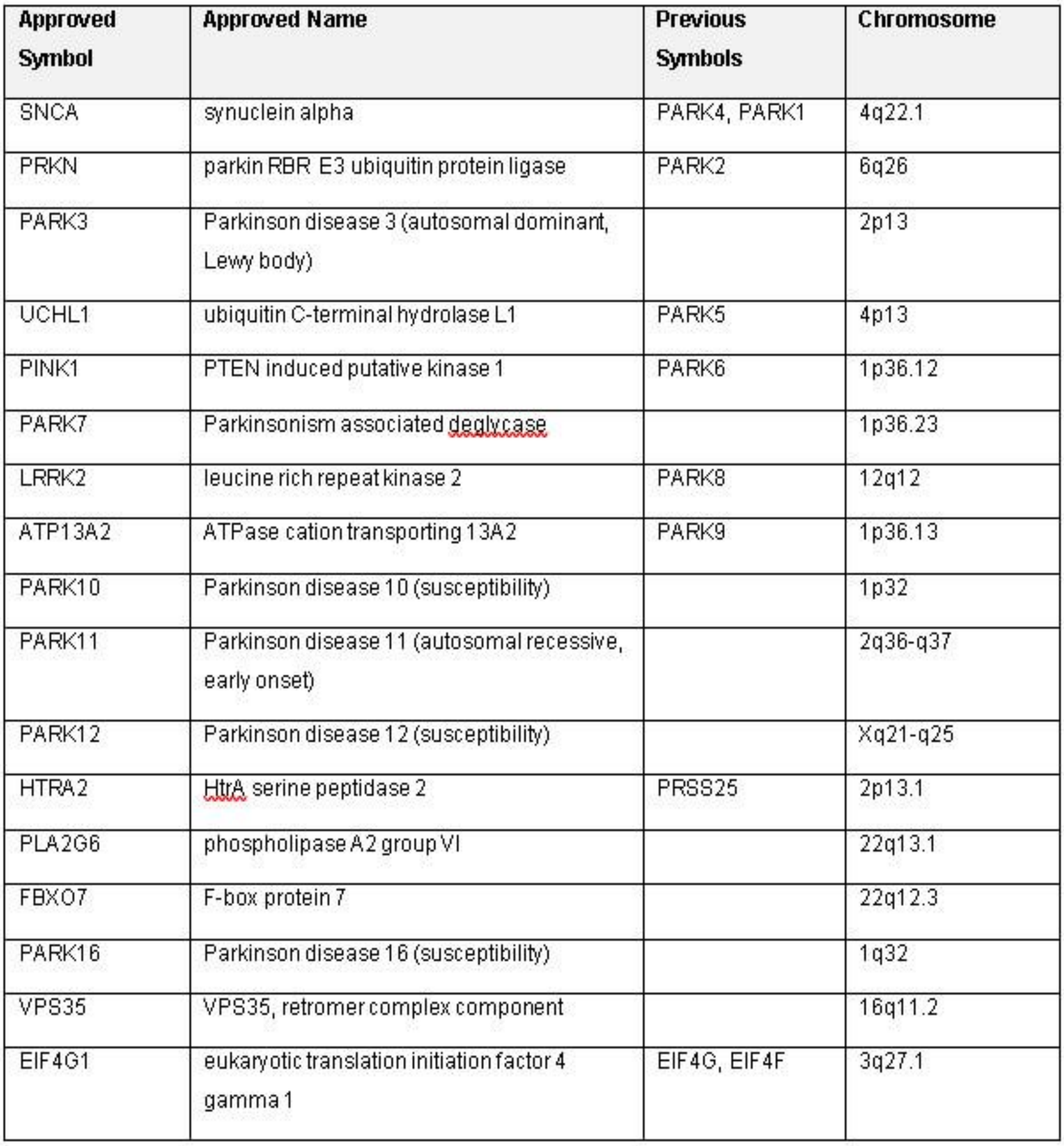
Keywords: Parkinson's Disease, SNP analysis, SNCA, LRRK2, PINK, DJ1

INTRODUCTION

A SNP known as Single Nucleotide Polymorphism is a single point mutation in the stretch of a gene. They are most commonly occurring type of mutations found in genome (approx. 90 percent of whole human DNA polymorphism in genome are SNPs). There are various publicly available online directories for SNPs for example GWAS Central, SwissVar and dsSNPs. Of all the mutations, only nonsynonymous SNPs or simply nsSNPs are of particular importance as they bring in change of amino acid residue, they are also known as missense mutations for the very same reason. Such changes in amino acid residue can result in protein instability by reducing protein dissolving ability or by altering hydrogen bonding pattern of protein.

PD is largely a neuro degenerative condition that is caused by the degeneration of dopamine producing neurons in the midbrain. The effect of PD increments with age, with 2% of people beyond 80 years old being affected making it the second most common neurodegenerative disease among human population. Currently PD genetics nomenclature includes 17 specific chromosomal locus regions that are termed *PARK* and numbered in chronological order of their identification (Table 1).

Table 1:



Approved Symbol	Approved Name	Previous Symbols	Chromosome
SNCA	synuclein alpha	PARK4, PARK1	4q22.1
PRKN	parkin RBR E3 ubiquitin protein ligase	PARK2	6q26
PARK3	Parkinson disease 3 (autosomal dominant, Lewy body)		2p13
UCHL1	ubiquitin C-terminal hydrolase L1	PARK5	4p13
PINK1	PTEN induced putative kinase 1	PARK6	1p36.12
PARK7	Parkinsonism associated deglycase		1p36.23
LRRK2	leucine rich repeat kinase 2	PARK8	12q12
ATP13A2	ATPase cation transporting 13A2	PARK9	1p36.13
PARK10	Parkinson disease 10 (susceptibility)		1p32
PARK11	Parkinson disease 11 (autosomal recessive, early onset)		2q36-q37
PARK12	Parkinson disease 12 (susceptibility)		Xq21-q25
HTRA2	HtrA serine peptidase 2	PRSS25	2p13.1
PLA2G6	phospholipase A2 group VI		22q13.1
FBXO7	F-box protein 7		22q12.3
PARK16	Parkinson disease 16 (susceptibility)		1q32
VPS35	VPS35, retromer complex component		16q11.2
EIF4G1	eukaryotic translation initiation factor 4 gamma 1	EIF4G, EIF4F	3q27.1

LRRK2 is undoubtedly the most common gene responsible for both familial and idiopathic Parkinson's disease (PD). LRRK2 is a unique multidomain structured protein having molecular weight of 286kDa (Figure 1), consisting of Ankyrin repeats (ARK), Armadillo repeats (ARM), a C-terminal of Roc (COR), leucine-rich repeats (LRR), a Ras of complex proteins (Roc), a kinase domain, and WD40 repeats [9]. LRRK2 gene gives instruction to form a protein named dardarin. It is functions in brain region and other tissues all through the body. Dardarin has a section of leucine-rich region which plays an important role in transferring of signals.

Mutations in LRRK2 are linked with Parkinson’s type 8. The most common mutation in LRRK2 is Gly2019Ser.

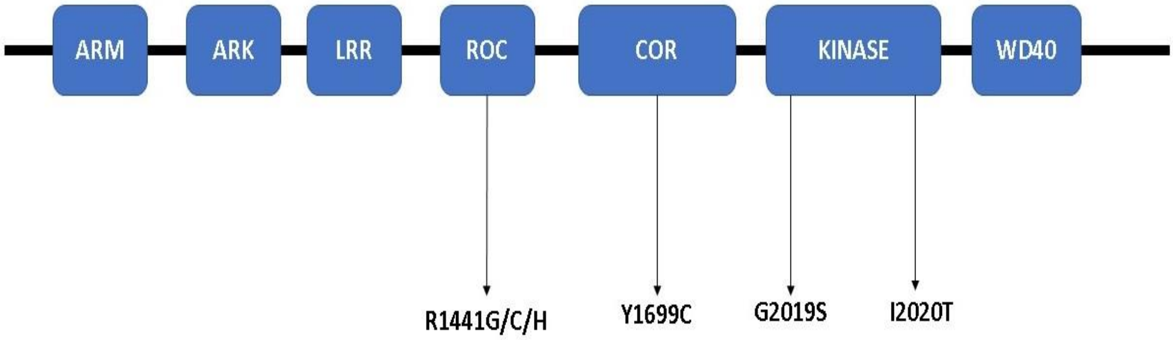


Figure 1: Schematic diagram of the domain architecture of LRRK2. Above, the most common known mutations of LRRK2 in Parkinson's disease are shown (arrows).

It’s important to identify SNPs associated with disease from the available SNP pool through experimental data but the amount of data available in database is humongous therefore it is important to carry out computational studies to help in minimizing costs and prioritise SNPs for examination. In such case subsequent studies through various independent sources can help in establishing the validity of results. In this work, we applied both sequence specific and sequence-structure specific computational approach to examine the SNPs present in SNCA, LRRK2, DJ1 & PINK1.

REVIEW OF LITERATURE

SNPs

There are various sort of mutations that alter the gene structure and function but Single nucleotide polymorphisms, simply called SNPs (snips), are the most common known type of genetic alteration in a being. SNP is basically a deviation in a single nucleotide that results in change of amino acid. For instance, a SNP may result in replacement of the nucleotide adenine (A) to nucleotide guanine (G) in a certain section of a gene/DNA.

SNPs arise throughout an individual’s DNA. They arise once every several hundred basepair on an average, that means there are about 10 million SNPs in the exclusively within human genome. They help in locating genes linked with disease by functioning as biomarkers. When SNPs arise within coding region of the gene they might affect gene’s function

Experimental studies are important to validate the identified disease linked SNPs from the pool of SNPs and to understand working role of SNPs. Although much research has been conducted on finding out the disease associated SNPs, it is hard to confirm it by following discrete studies. In this case, in-silico studies can help in saving time and costs. It also helps in analysing and ranking functionally important SNPs.

Parkinson’s Disease (PD)

idiopathic or familial parkinson's disease or simply PD is a neuro degenerative condition that result in progressive loss of the dopaminergic cells of the substantia nigra. It’s hard to differentiate Parkinson’s disease from other neuro degenerative conditions having similar clinical symptoms. Therefore, diagnosis is mainly based on history and examination of patient.

People with PD normally shows signs and symptoms of parkinsonism i.e.

1. rest tremor
2. hypokinesia (poverty of movement)
3. bradykinesia (slowness of movement)
4. rigidity
5. impaired posture
6. speech change

7. writing change

Although PD is mainly related to movement impairment, other issues such as dementia and depression may also arise.

Neurochemical and Neuropathological Features of PD

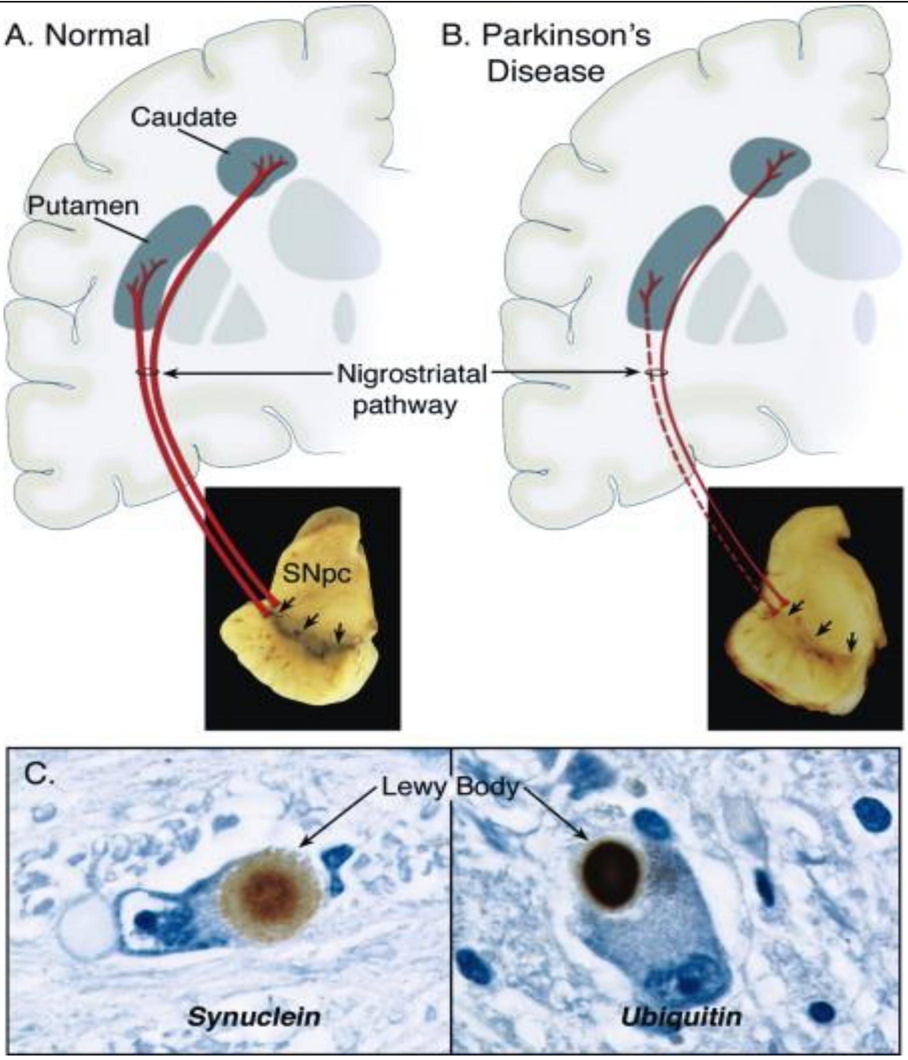


Figure 2: Neuropathology of Parkinson's Disease

(A) Pictorial description of the standard nigrostriatal bundle. It is constituted of dopamine nerve cells located in the pars compacta which is a particular region in substantia nigra. These nerve cells point to the synapse in the striatum of basal ganglia. The picture shows the normal colouring of the Synuclein pc, yield by neuro melanin within the dopaminergic neurons.

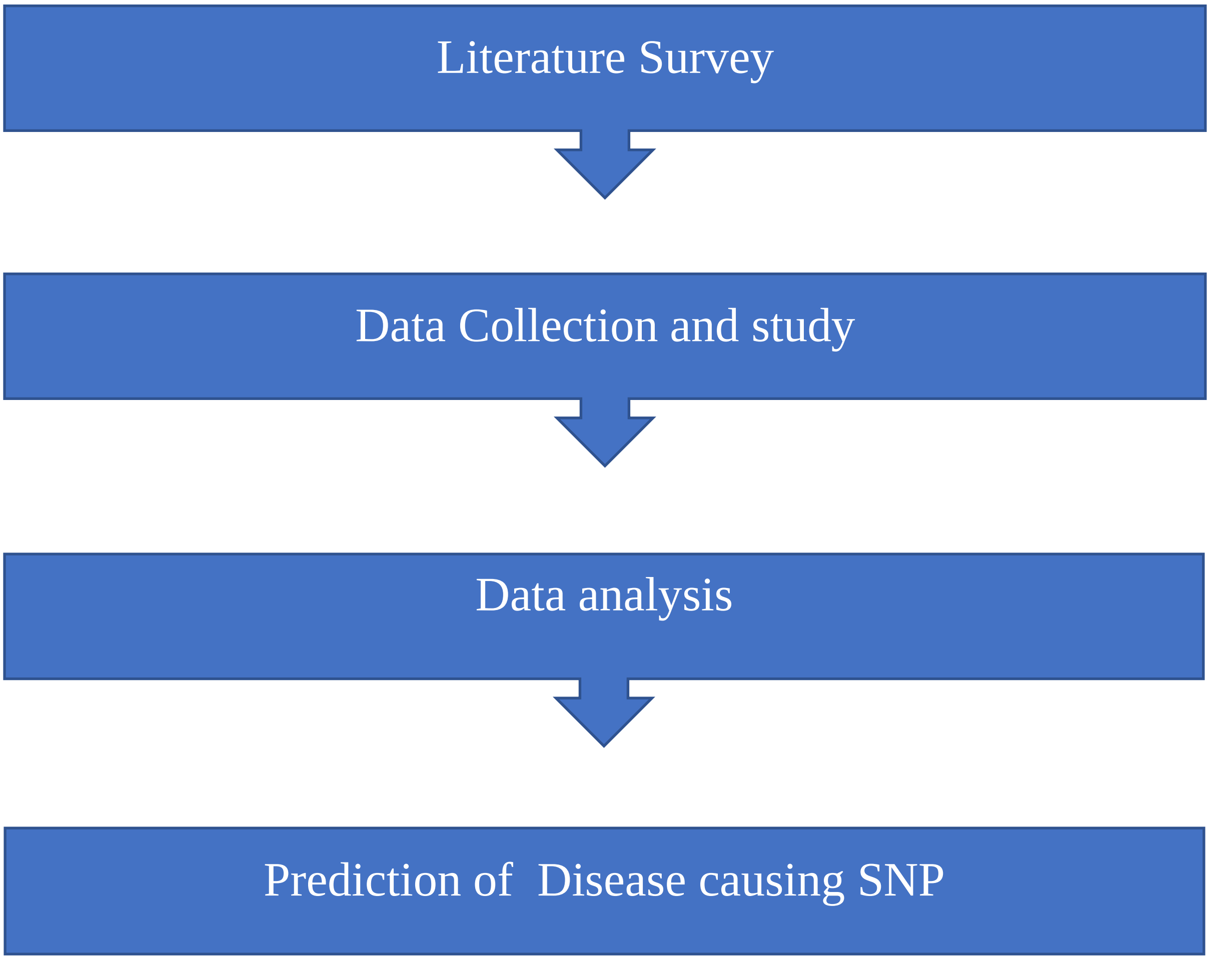
(B) Pictorial description of the standard nigrostriatal bundle. The nigrostriatal bundle undergoes deterioration in Parkinson’s disease. There is a noticeable loss of dopamine nerve cells that point to the putamen and a much smaller loss of those that point to the caudate. The figure depicts decolouration of the Synuclein pc due to the noticeable loss of dopamine nerve cells.

(C) Immuno histochemical tagging of intraneuronal elementary bodies, called Lewy bodies, in a Synuclein pc dopamine nerve cells. Tagging with an antibody against α-syn reveals a Lewy body with an intensely immunoreactive middle zone encircled by a faintly immunoreactive outer zone (left picture). On the contrary, immune-tagging with an antibody against ubiquitin produce more spread out immune reactivity within Lewy body (right picture).

OBJECTIVE

The main purpose of our research was to recognise diseased SNPs from the collection of SNPs retrieved from NCBI SNP database using various tools. Using sequence specific and sequence-structure specific method simultaneously we eliminated the shortcomings of a single method towards prediction of deleterious SNPs that might be associated with Parkinson’s disease.

WORK PLAN


Literature Survey
Data Collection and study
Data analysis
Prediction of Disease causing SNP

METHODOLOGY

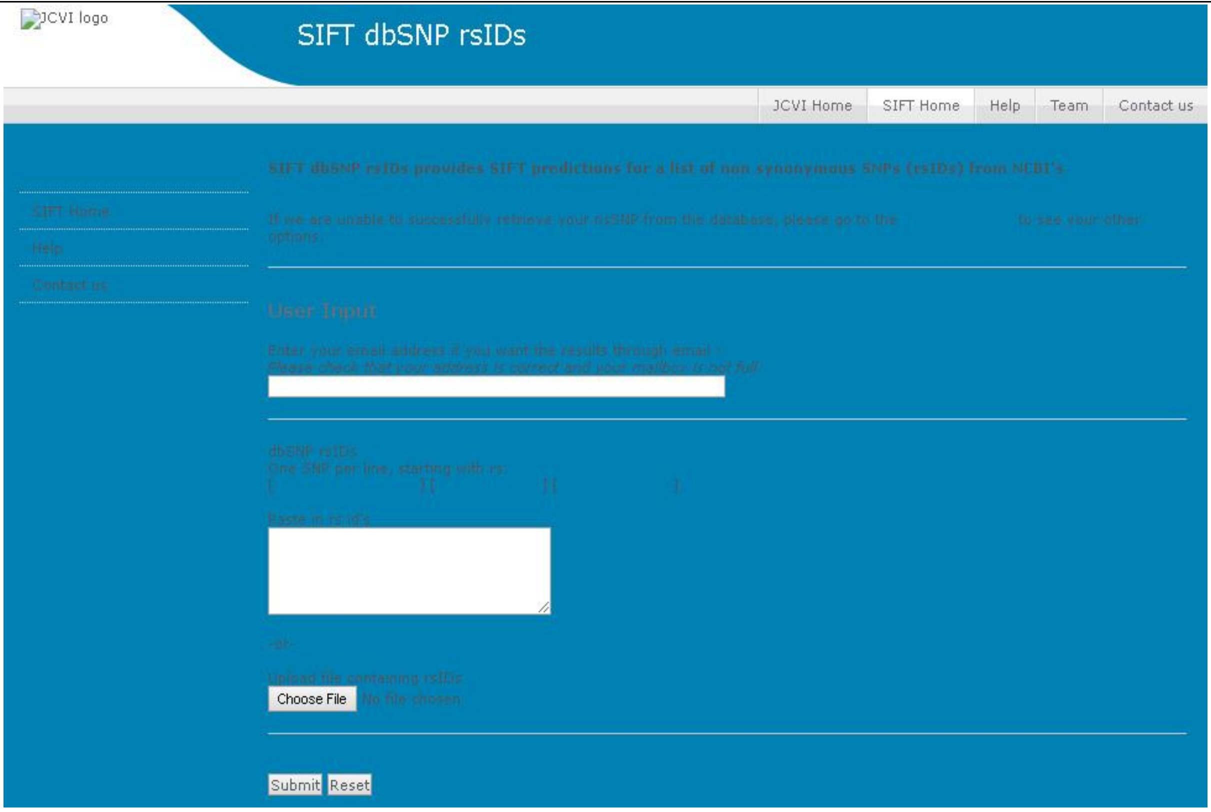
Firstly, disease related SNP sequence of SNCA, LRRK2, DJ1 and PINK1 were fetched from NCBI database of SNPs (http://www.ncbi.nlm.nih.gov./SNP/). We restricted our list to missense mutations having unknown significance.

Sequence specific and sequence-structure specific methods are the two of the most regularly used approaches towards detection of disease related SNPs via computational analysis. Sequence-structure based analysis is more precise then sequence-based analysis because it involves various effect at protein level. [29,19,25]. Whereas sequence-specific analysis fails to describe the underlying medium of how a SNP will alter the protein function. Therefore, we use both the methods to overcome the shortcomings of only using one method [30,25].

SIFT

SIFT sorts intolerant nsSNPs from tolerant nsSNPs from dbSNP database. It takes SNP id as input along with the amino acid substitution and position (http://sift.jcvi.org/www/SIFT_dbSNP.html). SIFT tool predictions are highly based on sequence homology and physiochemical properties of amino acid [21].

SIFT predicts weather an amino acid mutation will have a change on the protein functionality. It compares the conserved region of the gene with closely related sequences collected through PSI BLAST. SIFT score of greater than 0.05 is consider to be tolerant [26]. SIFT interface (Figure 3).



SIFT dbSNP rsIDs provides SIFT predictions for a list of non synonymous SNPs (rsIDs) from NCBI's

If we are unable to successfully retrieve your nsSNP from the database, please go to the to see your other options.

User Input

Enter your email address if you want the results through email :
Please check that your address is correct and your mailbox is not full

dbSNP rsIDs
One SNP per line, starting with rs:
[ ] [ ] [ ]

Paste in rs id's

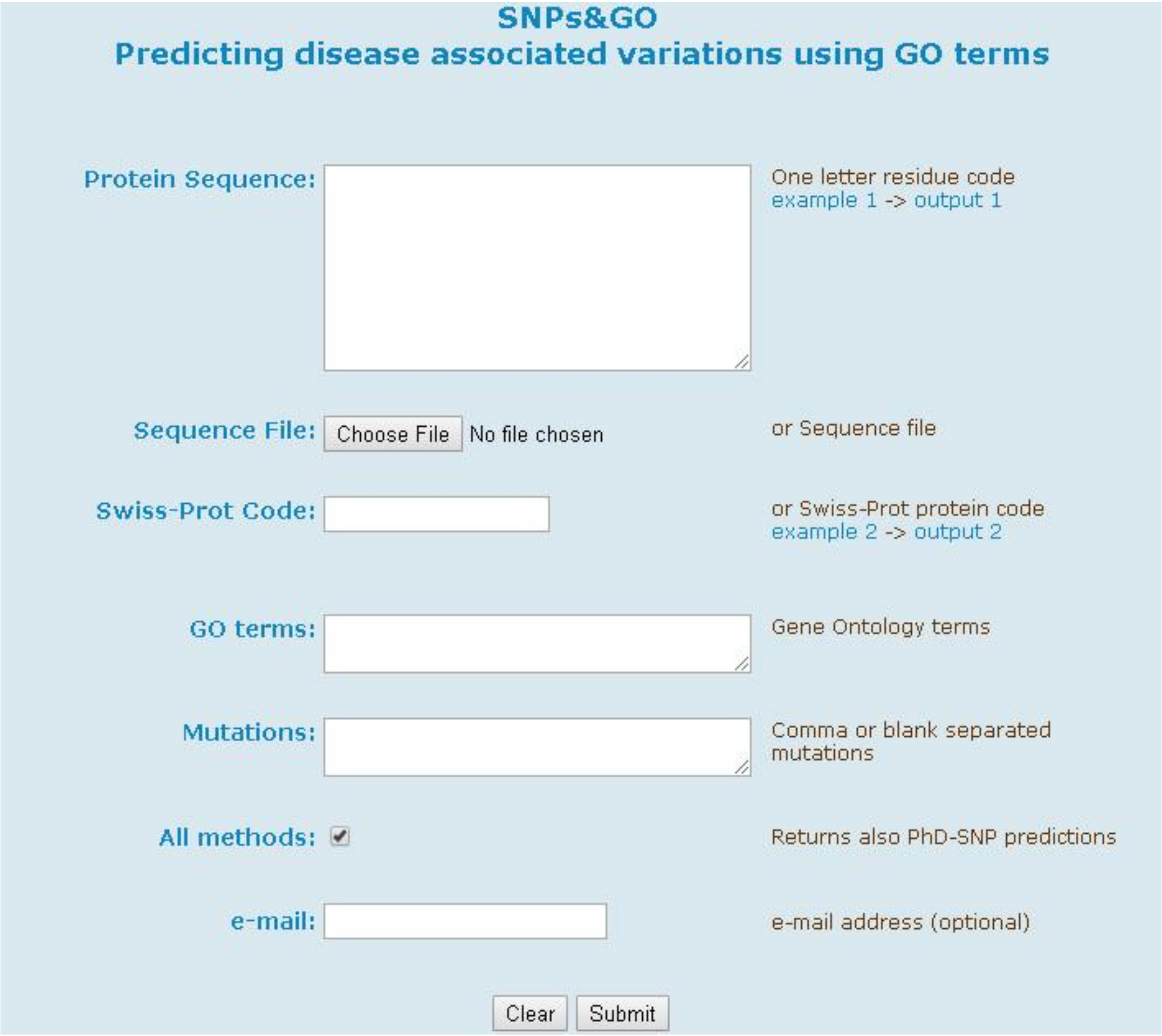
-or-

Upload file containing rsIDs
No file chosen

Figure 3: SIFT Interface

SNP&GO

SNP&GO predicts disease associated variations using GO terms. SNP&GO predicts if a given single point mutation can be classified as disease associated or not. It takes protein sequence, GO terms, amino acid substitution and position as input and return a table listing weather the mutation is Diseased or Normal along with RI (reliability index). A mutation is estimated to be damaging for a score greater than 0.05. SNP&GO interface (Figure 4)



SNPs&GO

Predicting disease associated variations using GO terms

Protein Sequence: One letter residue code
example 1 -> output 1

Sequence File: No file chosen — or Sequence file

Swiss-Prot Code: or Swiss-Prot protein code
example 2 -> output 2

GO terms: Gene Ontology terms

Mutations: Comma or blank separated mutations

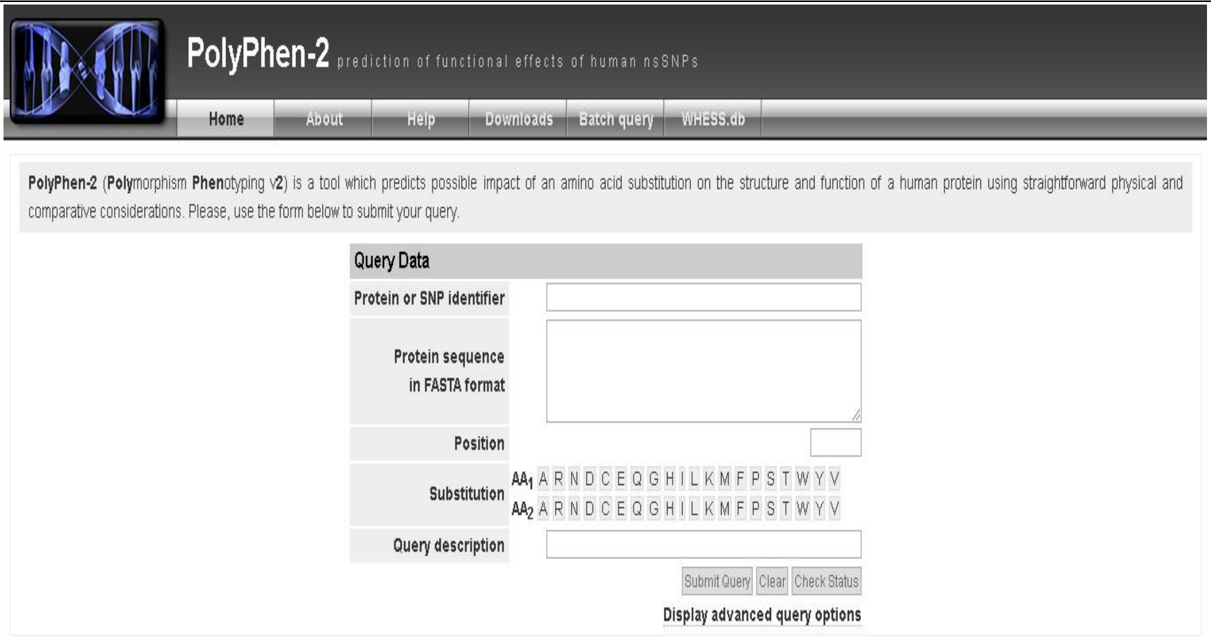
All methods: ☑ Returns also PhD-SNP predictions

e-mail: e-mail address (optional)

Figure 4: SNP & GO interface

PolyPhen-2

The polymorphism phenotyping version 2.0 predicts the potential impact of an amino acid on the structure and function of a gene (http://genetics.bwh.harvard.edu/pph2/) utilizing physical and relative factors. The outcome from the PolyPhen2 output provides a particular score that ranges from the value of 0 to 1. The 0 suggest the no deleterious effect of a SNP on protein structure while a value closer to 1 suggest that the mutation may have deleterious consequences [23, 28]. PolyPhen2 interface (Figure 5).



PolyPhen-2 (**Poly**morphism **Phen**otyping v**2**) is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations. Please, use the form below to submit your query.

Query Data

Protein or SNP identifier

Protein sequence in FASTA format

Position

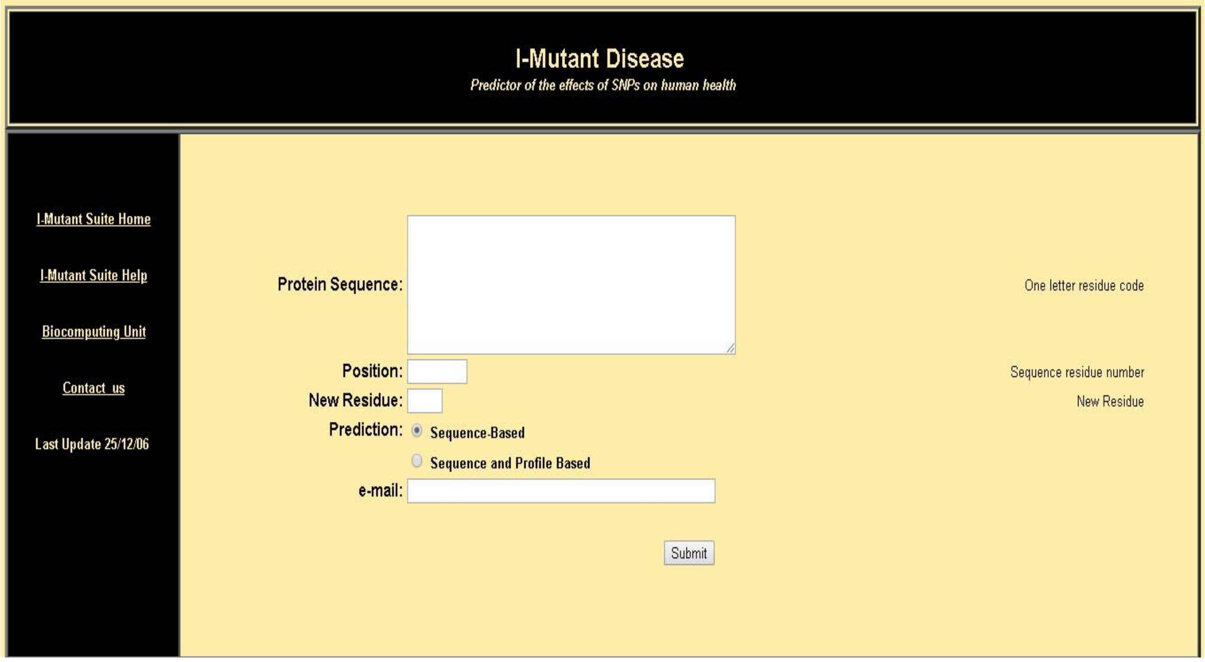
Substitution AA1 A R N D C E Q G H I L K M F P S T W Y V
AA2 A R N D C E Q G H I L K M F P S T W Y V

Query description

Figure 5: PolyPhen-2 interface

I-Mutant 3.0

I-Mutant is neural network-based web server that calculate protein stability upon single point mutation. The tool uses a archived data taken from ProTherm [1]. It had been tested to predict protein stability with 80% accuracy. Free energy changes are predicted with energy base FOLD-X tool. By coupling FOLD-X with I-Mutant, along with reliability index of later one can achieve very high accuracy of prediction (Guerois et al., 2002).



I-Mutant Disease

Predictor of the effects of SNPs on human health

Protein Sequence: One letter residue code

Position: Sequence residue number

New Residue: New Residue

Prediction: ● Sequence-Based

○ Sequence and Profile Based

e-mail:

Figure 6: I-Mutant 3.0 interface

Site-directed Mutagenesis of LRRK2

LRRK2 gene was found to be the most important gene involved in Parkinson's Disease but the complete structure of LRRK2 gene was not resolved therefore in order to study the effects of mutations we modelled LRRK2 gene from structure of rocCOR domain of Rab family protein (PDB id: 3DPU) (Figure 7) which is a microbial homologous of LRRK2 human gene.

Mutagenesis was performed using Chimera and following substitutions were made V1598E, P2119L, L119P and V366M. After mutagenesis, protein optimization, solvation and minimization were performed using charmm force field from NAMD (MD simulation) and finally total energy plot were drawn.

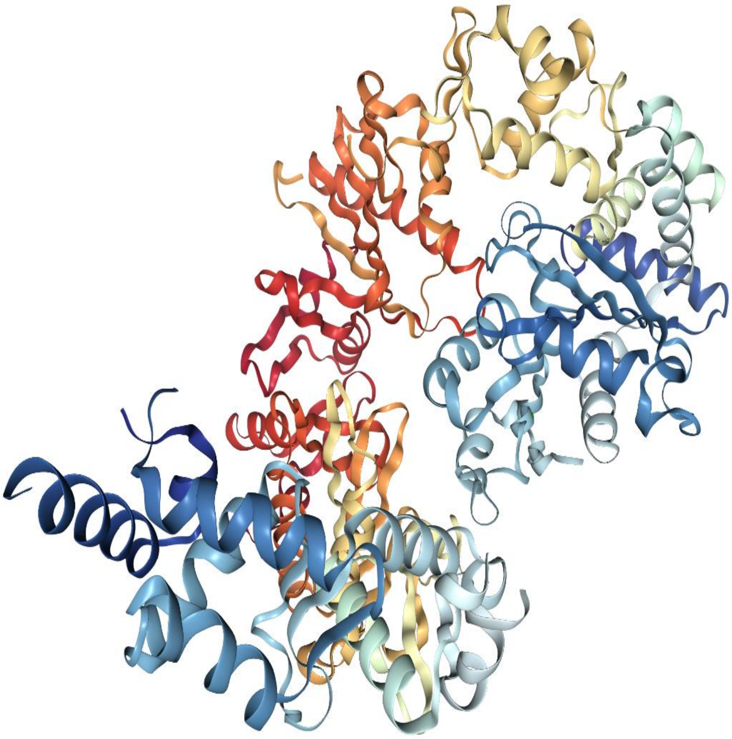


Figure 7: Ribbon structure of RocCOR domain (PDB id: 3DPU)

Mutagenesis Steps:

1. Load the wild LRRK2 gene in chimera

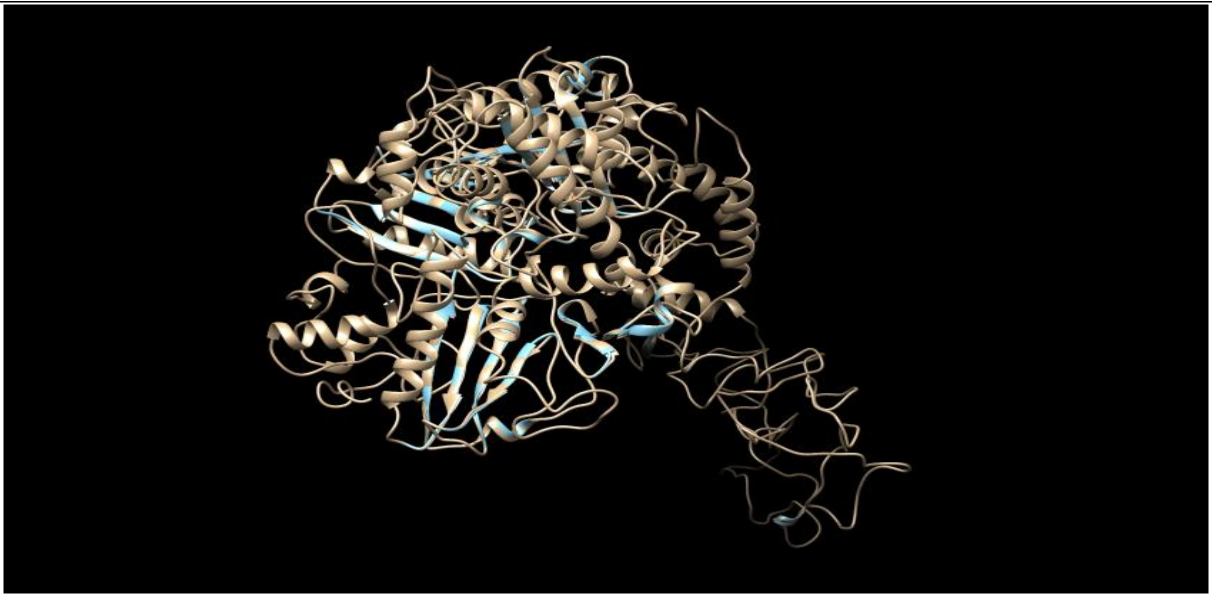
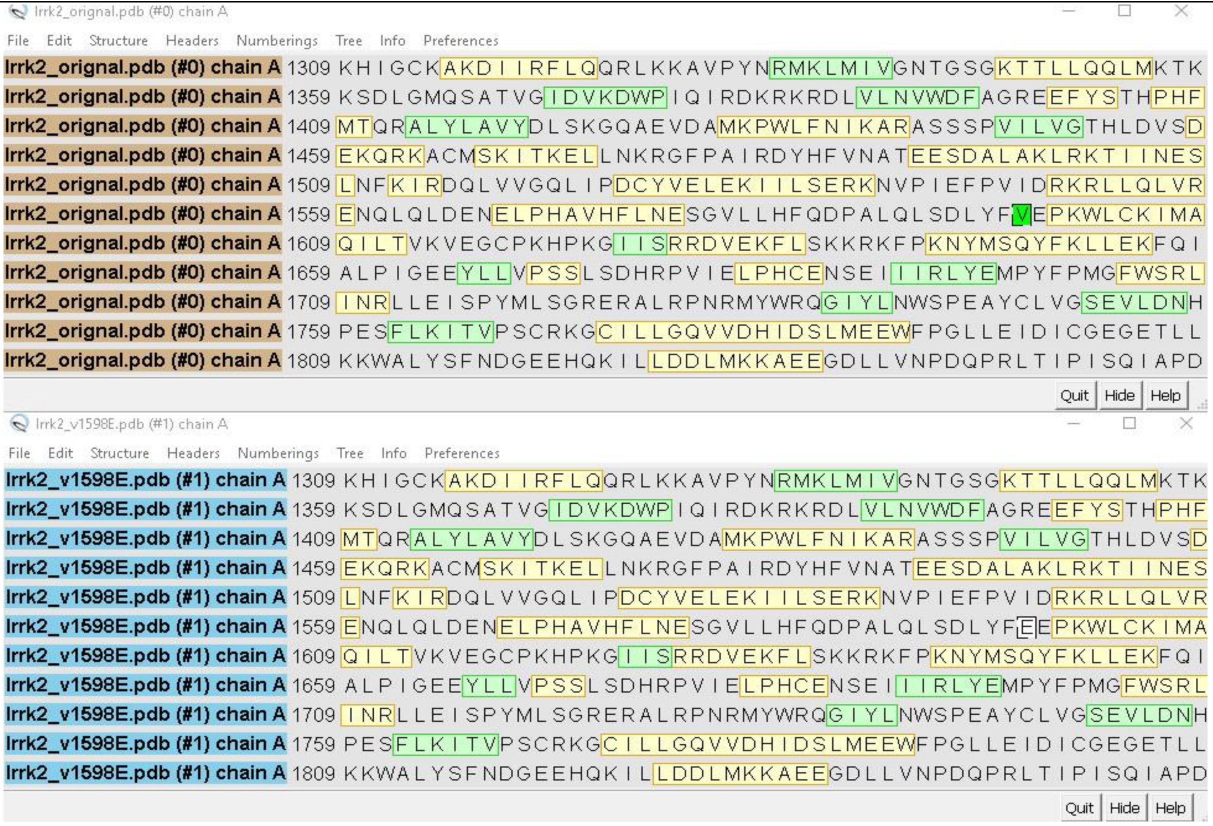


Figure 8: wild LRRK2 gene in chimera

2. Go to Favourites tab and select Sequence option from drop down menu.



lrrk2_orignal.pdb (#0) chain A
File Edit Structure Headers Numberings Tree Info Preferences
Quit Hide Help
lrrk2_v1598E.pdb (#1) chain A
File Edit Structure Headers Numberings Tree Info Preferences
Quit Hide Help

```
lrrk2_orignal.pdb (#0) chain A 1309 KHIGCKAKDIIRFLQQRLKKAVPYNRMKLMIVGNTGSGKTTLLQQLMKTK
lrrk2_orignal.pdb (#0) chain A 1359 KSDLGMQSATVGIDVKDWPIQIRDKRKRDLVLNVWDFAGREEFYSTHPHF
lrrk2_orignal.pdb (#0) chain A 1409 MTQRALYLAVYDLSKGQAEVDAMKPWLFNIKARASSSPVILVGTHLDVSD
lrrk2_orignal.pdb (#0) chain A 1459 EKQRKACMSKITKELLNKRGFPAIRDYHFVNATEESDALAKLRKTIINES
lrrk2_orignal.pdb (#0) chain A 1509 LNFKIRDQLVVGQLIPDCYVELEKIILSERKNVPIEFPVIDRKRLLQLVR
lrrk2_orignal.pdb (#0) chain A 1559 ENQLQLDENELPHAVHFLNESGVLLHFQDPALQLSDLYFVEPKWLCKIMA
lrrk2_orignal.pdb (#0) chain A 1609 QILTVKVEGCPKHPKGIISRRDVEKFLSKKRKFPKNYMSQYFKLLEKFQI
lrrk2_orignal.pdb (#0) chain A 1659 ALPIGEEYLLVPSSLSDHRPVIELPHCENSEIIIRLYEMPYFPMGFWSRL
lrrk2_orignal.pdb (#0) chain A 1709 INRLLEISPYMLSGRERALRPNRMYWRQGIYLNWSPEAYCLVGSEVLDNH
lrrk2_orignal.pdb (#0) chain A 1759 PESFLKITVPSCRKGCILLGQVVDHIDSLMEEWFPGLLEIDICGEGETLL
lrrk2_orignal.pdb (#0) chain A 1809 KKWALYSFNDGEEHQKILLDDLMKKAEEGDLLVNPDQPRLTIPISQIAPD

lrrk2_v1598E.pdb (#1) chain A 1309 KHIGCKAKDIIRFLQQRLKKAVPYNRMKLMIVGNTGSGKTTLLQQLMKTK
lrrk2_v1598E.pdb (#1) chain A 1359 KSDLGMQSATVGIDVKDWPIQIRDKRKRDLVLNVWDFAGREEFYSTHPHF
lrrk2_v1598E.pdb (#1) chain A 1409 MTQRALYLAVYDLSKGQAEVDAMKPWLFNIKARASSSPVILVGTHLDVSD
lrrk2_v1598E.pdb (#1) chain A 1459 EKQRKACMSKITKELLNKRGFPAIRDYHFVNATEESDALAKLRKTIINES
lrrk2_v1598E.pdb (#1) chain A 1509 LNFKIRDQLVVGQLIPDCYVELEKIILSERKNVPIEFPVIDRKRLLQLVR
lrrk2_v1598E.pdb (#1) chain A 1559 ENQLQLDENELPHAVHFLNESGVLLHFQDPALQLSDLYFEEPKWLCKIMA
lrrk2_v1598E.pdb (#1) chain A 1609 QILTVKVEGCPKHPKGIISRRDVEKFLSKKRKFPKNYMSQYFKLLEKFQI
lrrk2_v1598E.pdb (#1) chain A 1659 ALPIGEEYLLVPSSLSDHRPVIELPHCENSEIIIRLYEMPYFPMGFWSRL
lrrk2_v1598E.pdb (#1) chain A 1709 INRLLEISPYMLSGRERALRPNRMYWRQGIYLNWSPEAYCLVGSEVLDNH
lrrk2_v1598E.pdb (#1) chain A 1759 PESFLKITVPSCRKGCILLGQVVDHIDSLMEEWFPGLLEIDICGEGETLL
lrrk2_v1598E.pdb (#1) chain A 1809 KKWALYSFNDGEEHQKILLDDLMKKAEEGDLLVNPDQPRLTIPISQIAPD
```

Figure 9: Sequence of wild and mutated LRRK2 gene

3. Select the residue from the sequence that you want to mutate.

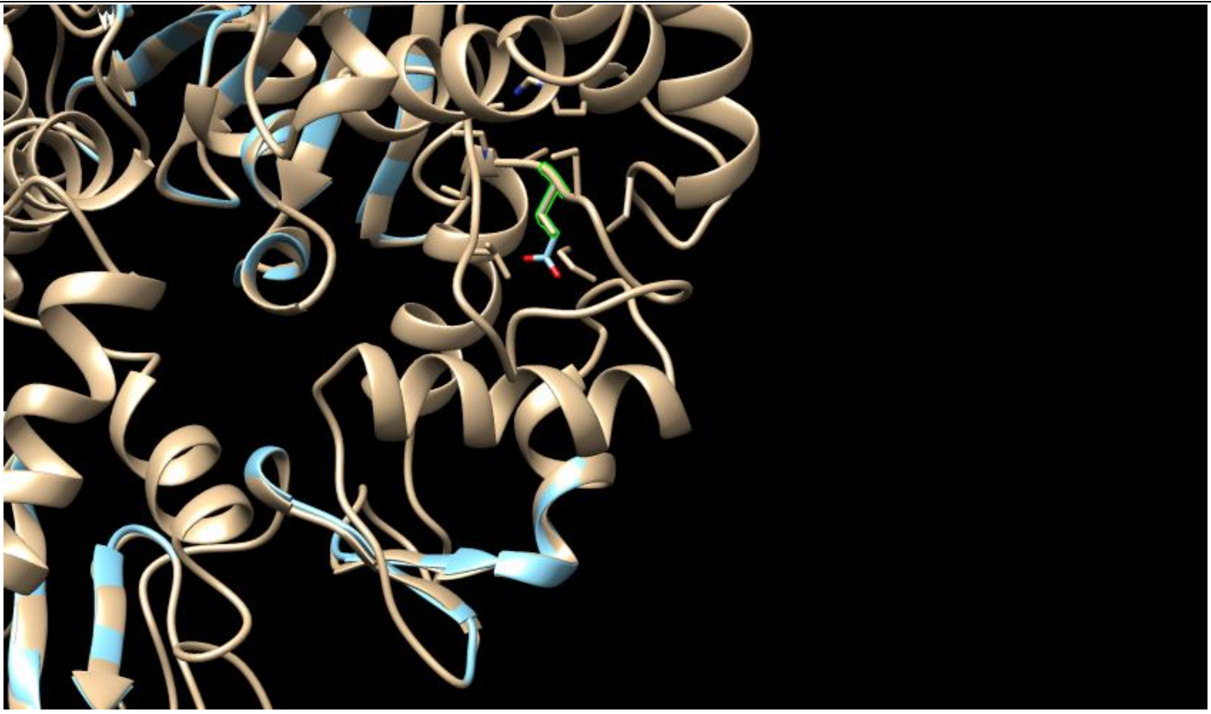
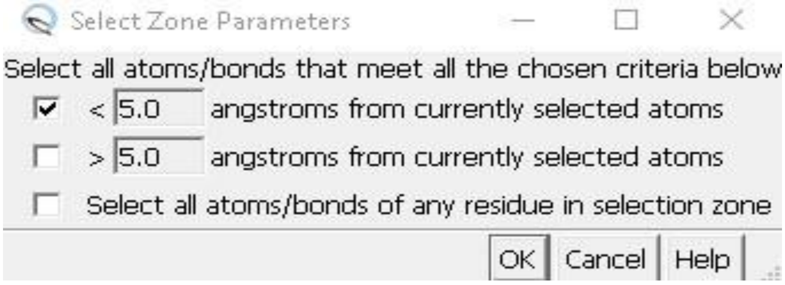


Figure 10: Selected residue highlighted in green colour.

4. Go to Select tab and choose Zone option from drop down menu. Zone should be within angstrom.



Select Zone Parameters

Select all atoms/bonds that meet all the chosen criteria below

☑ < 5.0 angstroms from currently selected atoms

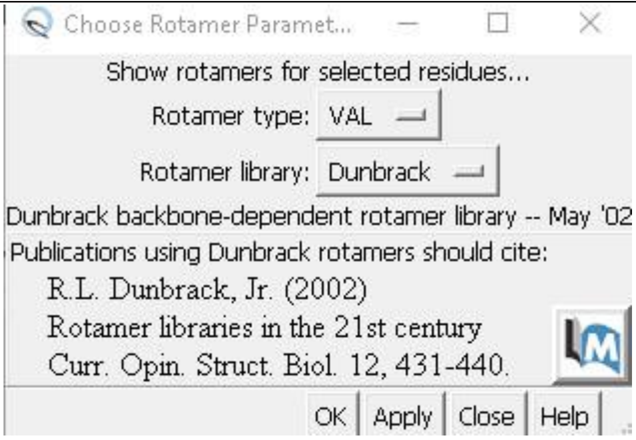
□ > 5.0 angstroms from currently selected atoms

□ Select all atoms/bonds of any residue in selection zone

Figure 11: Parameter window

5. Go to Tools > Structure Editing > Rotamers.

6. Now change the Rotamer type i.e. Amino acid (In our study we will be substituting Valine for Glutamic acid at position 1598 of LRRK2 gene)



Choose Rotamer Paramet...

Show rotamers for selected residues...

Rotamer type: VAL

Rotamer library: Dunbrack

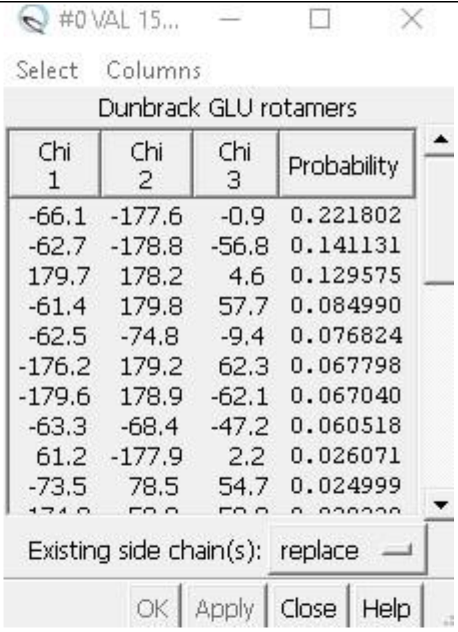
Dunbrack backbone-dependent rotamer library -- May '02

Publications using Dunbrack rotamers should cite:

R.L. Dunbrack, Jr. (2002)
Rotamer libraries in the 21st century
Curr. Opin. Struct. Biol. 12, 431-440.

Figure 12: Rotamer Parameters

8. Select the conformation with highest probability.



#0 VAL 15...

Dunbrack GLU rotamers

Chi 1	Chi 2	Chi 3	Probability
-66.1	-177.6	-0.9	0.221802
-62.7	-178.8	-56.8	0.141131
179.7	178.2	4.6	0.129575
-61.4	179.8	57.7	0.084990
-62.5	-74.8	-9.4	0.076824
-176.2	179.2	62.3	0.067798
-179.6	178.9	-62.1	0.067040
-63.3	-68.4	-47.2	0.060518
61.2	-177.9	2.2	0.026071
-73.5	78.5	54.7	0.024999

Existing side chain(s): replace

Figure 13: Dunbrack GLU rotamers

9. Now Save the newly formed structure as PDB file.

STATISTICAL ANALYSIS

In statistical analysis process we mainly refer three cross-validation methods for validating the success rate of predictor tools mentioned above. These tests are sub-sampling, independent dataset test and jack-knife test [7]. Out of all the three, only jack-knife test gives slightest random values and most unbiased according to the Equations 28 to 32 mentioned in chou,2011. Therefore, jack-knife test has been used throughout to check success rate of predictor tools [6, 4, 5, 2, 3, 13, 14, 16, 17, 15, 18, 22, 11, 12].

Prediction quality is determined using six widely known parameters viz sensitivity, precision, accuracy, specificity, Matthews correlation coefficient (MCC) and negative predictive value (NPV). In the subsequent equations true negatives, true positives, false negatives and false positives are written as *tn*, *tp*, *fn and fp* respectively.

$$\text{ACCURACY}= \frac{tp + tn}{tp+tn+fp+fn}$$

$$\text{SPECIFICITY} = \frac{tn}{fp+tn}$$

$$\text{SENSTIVITY} = \frac{tp}{tp+fn}$$

$$\text{MCC} = \frac{tp * tn-fn * fp}{\sqrt{(tp+fn)(tp+fp)(tn+fn)(tn+fp)}}$$

Sadly, these equations are hard to understand from a biologist point of view. Therefore, we considered the equations given by Chou et al. (2012). Going by these equations, the above four metrics can be written as

$$\text{ACCURACY} = 1 - \frac{N\pm + N\mp}{N+ + N-}$$

$$\text{SENSTIVITY} = 1 - \frac{N\pm}{N+}$$

$$\text{SPECIFICITY} = 1 - \frac{N\mp}{N-}$$

$$\text{MCC} = \frac{1 - \frac{N\pm N- + N\mp N+}{N+N-}}{\sqrt{}(1 + \frac{N\mp - N\pm}{N+})(1 + \frac{N\pm + N\mp}{N-})}$$

In the above equation $N+$ represent the total no. of SNPs analysed and $N-$ represent the non-synonymous SNPs. Whereas, disease predicted incorrectly as neutral are represented by $N\pm$ and $N\mp$ is wrongly predicted deleterious SNPs among non-synonymous SNPs.

Use of these matrices have been justified in various studies [2, 3, 11, 12, 15, 18].

RESULTS

The main objective of our study was to identify diseased SNPs from the pool of SNPs retrieved from NCBI SNP database using various tools. Using sequence and sequence-structure based method simultaneously we eliminated the shortcomings of a single method towards prediction of deleterious SNPs that might be associated with Parkinson’s disease. Workflow of the following study is shown in figure 14.

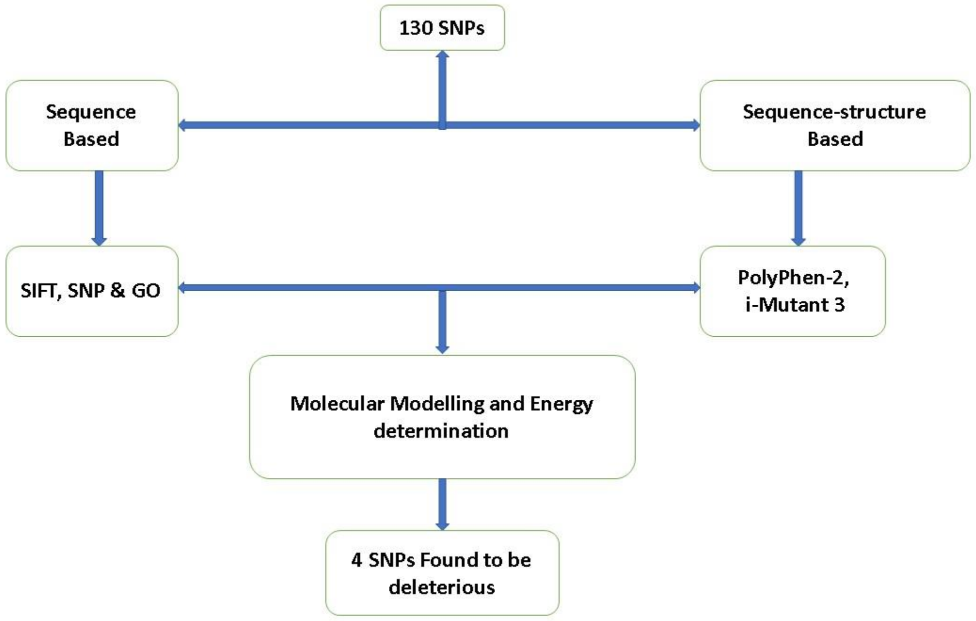


Figure 14: Workflow followed in study

The accuracy of SNP & GO has been reported to be (0.82) which is comparably good with PolyPhen-2 (0.69) and SIFT (0.65). Precision value of SNP & GO (0.90) predicted to be highest among all by Thusberg et al. (2011).

In the initial study we took 130 SNPs and analysed them with SIFT. SIFT predicted whether SNPs would have an impact on the functionality of protein by aligning the similar proteins. The output range of SIFT is from 0 to 1, where 0 represent highly deleterious SNP and 1 represents neutral SNP. The cut-off score was set to 0.05, above which an amino acid substitution is tolerated (no effect).

Among the total nsSNPs analysed, 15 nsSNPs were found to be deleterious with a tolerance index score of <u><</u>0.05. Four nsSNPs showed a highly deleterious tolerance index score of 0.00 (i.e. <0.01) (Table 2).

Table 2.

Gene Involved	SNP id	Substitution	SIFT Score
LRRK2	rs721710	V1598E	<0.01
LRRK2	rs12423862	P2119L	<0.01
LRRK2	rs33995463	L119P	<0.01
LRRK2	rs60185966	S1228I	0.01
LRRK2	rs145364431	R1728L	0.02
LRRK2	rs113065049	V366M	0.01
LRRK2	rs17519916	D944Y	0.01
LRRK2	rs72546335	S52F	0.02
LRRK2	rs72546337	I810V	0.02
LRRK2	rs74681492	P1446L	<0.01
LRRK2	rs78154388	S663P	0.02
LRRK2	rs80179604	S1228T	0.03
LRRK2	rs72547981	D2175H	0.04
LRRK2	rs111910483	L1795F	0.01
DJ1	rs886046545	G75S	0.02

Further these 15 SNPs were submitted to PolyPhen-2. PolyPhen-2 score ranges from 0 to 1. If score is <0.5 then mutation is considered benign, if it ranges between 0.5 to 0.9 then mutation is probably damaging and if score is >0.9 then mutation is possibly damaging. PolyPhen-2 predicted 12 SNPs to be deleterious, 11 belong to LRRK2 gene and 1 from DJ1 gene (rs721710, rs12423862, rs33995463, rs17519916, rs74681492, rs80179604, rs72547981, rs111910483, rs60185966, rs145364431, rs113065049 and rs886046545).

Table 3.

Gene Involved	SNP id	Substitution	PolyPhen-2 score
LRRK2	rs721710	V1598E	0.986
LRRK2	rs12423862	P2119L	1.000
LRRK2	rs33995463	L119P	0.996
LRRK2	rs17519916	D944Y	0.947
LRRK2	rs74681492	P1446L	1.000
LRRK2	rs80179604	S1228T	0.568
LRRK2	rs72547981	D2175H	0.710
LRRK2	rs145364431	R1728L	0.993
LRRK2	rs111910483	L1795F	1.000
LRRK2	rs60185966	S1228I	0.903
LRRK2	rs113065049	V366M	1.000
DJ1	rs886046545	G75S	1.000

After PolyPhen-2 SNPs were submitted to SNP&GO for further analysis. SNP&GO predicted 7 SNPs out of 12 to be deleterious (rs721710, rs12423862, rs33995463, rs60185966, rs145364431, rs113065049 and rs886046545).

Table 4.

Gene Involved	SNP id	Substitution	SNP&GO
LRRK2	rs721710	V1598E	Disease (0.781)
LRRK2	rs12423862	P2119L	Disease (0.666)
LRRK2	rs33995463	L119P	Disease (0.822)
LRRK2	rs60185966	S1228I	Disease (0.797)
LRRK2	rs145364431	R1728L	Disease (0.711)
LRRK2	rs113065049	V366M	Disease (0.729)
DJ1	rs886046545	G75S	Disease (0.984)

Finally, remaining SNPs were submitted to I-Mutant. I-Mutant checks the stability of protein by calculating change in Gibbs free energy between native and variant protein. Only those SNPs having reliability index (RI) of 5 or more were predicted to be diseased. I-Mutant predicted 4 SNPs to have deleterious effect (rs721710, rs12423862, rs33995463 and rs113065049) (Table 3). Figure 3 shows superimposed structure of native and mutated LRRK2 protein.

Table 5.

Gene Involved	SNP id	Substitution	SIFT score	SNP&GO	PolyPhen-2	I-Mutant
LRRK2	rs721710	V1598E	<0.01	Disease (0.781)	Probably Damaging (0.991)	Disease RI (5)
LRRK2	rs12423862	P2119L	<0.01	Disease (0.666)	Probably Damaging (1.000)	Disease RI (7)
LRRK2	rs33995463	L119P	<0.01	Disease (0.822)	Probably Damaging (0.996)	Disease RI (6)
LRRK2	rs113065049	V366M	0.01	Disease (0.729)	Probably Damaging (1.000)	Disease RI (6)

Additionally, wild type LRRK2 protein was mutated using UCSF Chimera (https://www.cgl.ucsf.edu/chimera/). Further mutated protein was energy minimized using NAMD (MD simulation tool) and Total energy values of native and mutated protein were plotted.

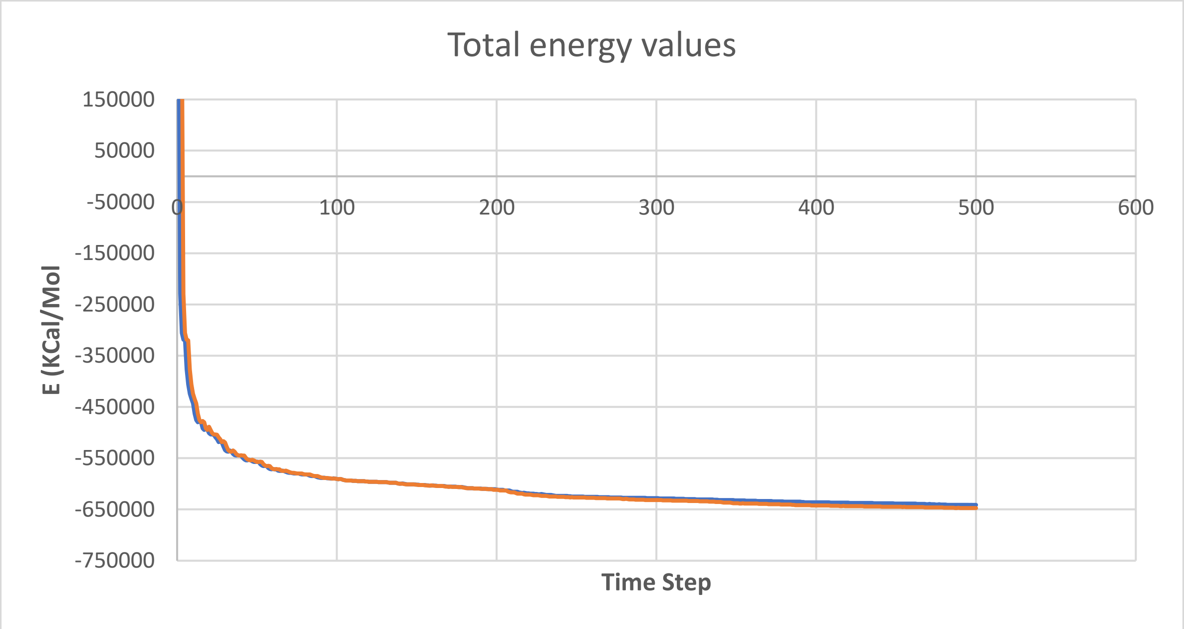


Figure 15: The relative total energy values of native LRRK2 gene vs mutated LRRK2 gene (Blue line represent mutated gene and orange represent native gene).

From the graph it can be interpreted that the total energy values of mutated protein are shifting from more negative to less negative values. Therefore, it can be said that mutations in LRRK2 gene (rs721710, rs12423862, rs33995463 and rs113065049) are indeed destabilising the protein structure and could potentially alter its function. Figure 16 Shows the superimposed structure of native and mutated LRRK2 gene.

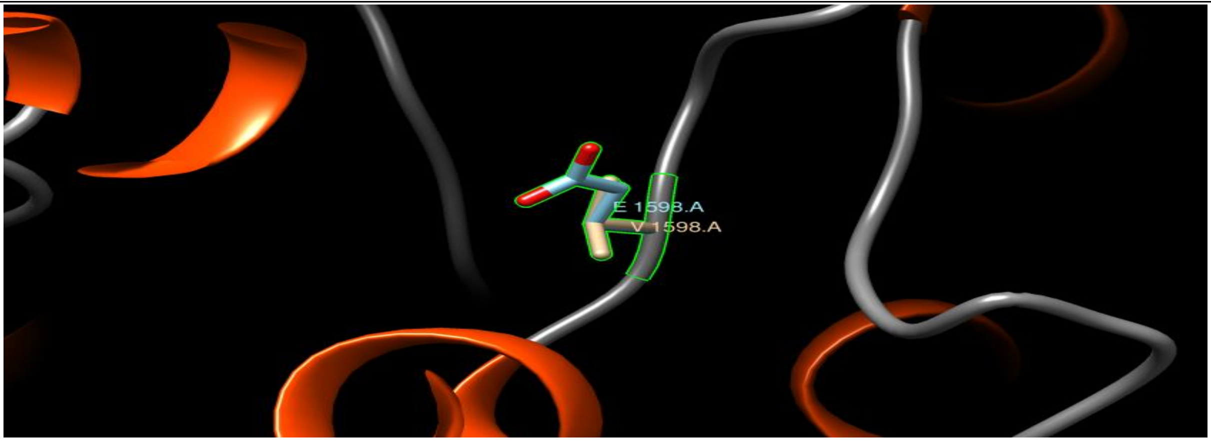


Figure 16: Superimposed structure of wild type LRRK2 and V1598E mutant, (superimposed structure of P2119L, L119P and V366M could not made because LRRK2 gene was not fully resolved).

DISCUSSION

Detection of disease causing mutations from functionally neutral mutation is important to understand the pathophysiology behind the disease. Normally each individual has around 10 million of SNPs throughout person’s entire DNA, which makes it nearly impossible to experimentally distinguish between disease causing mutation and functionally neutral mutations. Analysing vast number of SNPs through *in-vitro* methods might not be the ideal solution for researchers. In such cases Bioinformatics came to the rescue. With the vast number of tools available online we can narrow down our search to potentially harmful SNPs from the pool of SNPs available in online databases. Identification of disease causing SNPs from neutral SNPs with the help of bioinformatics tools saves a great deal of time and money. The potentially harmful SNPs detected through bioinformatics tools can be validated experimentally without having to go through each and every SNP.

In this paper, we attempted to predict the SNPs associated with Parkinson’s genes (SNCA, PINK, LRRK2 and DJ1) which could be potentially harmful. For this purpose, we adopted two methods namely sequence-based and sequence-structure based methods. Both the methods were used together in-order to minimize false positive results.

Out of the 130 missense SNPs reported in dbSNP, we found 4 SNPs to be in coding region which could affect the normal functioning of the gene. All of the 4 SNPs happen to be in LRRK2 gene. Further total energy values of native LRRK2 gene and mutated LRRK2 gene was plotted. Thus, revealing that these mutations in LRKK2 gene leads to decreased protein stability.

CONCLUSION

In the present study, we investigated the functional and structural effects of SNPs caused by the Parkinson associated genes (SNCA, LRRK2, PINK and DJ1) using different computational prediction tools. 4 SNPs were predicted to be deleterious by four different algorithms. Out of which 2 SNPs (V1598E and P2119L) of LRRK2 were found in coding region. Further, experimental studies need to be carried to better understand the role of SNPs reported in Table 3.

The *in-silico* data shows the computational approach towards identifying the disease-causing SNPs from functionally neutral SNPs which is a fast and reliable technique to analyse large number of SNPs. Also, the method used for detection of SNPs from dbSNPs is claimed to be best since it uses four different tools which eliminates the shortcomings of using a single tool for detection of SNPs (Nagamani et al., 1999).

REFRENCES

1. Bava K.A., Gromiha M.M., Uedaira H., Kitajima K., Sarai A. ProTherm, version 4.0: thermodynamic database for proteins and mutants. Nucleic Acids Res. 2004;32:D120–D121.
2. Chen W., Ding H., Feng P., Lin H., Chou K.C. iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget. 2016.
3. Chen W., Feng P., Ding H., Lin H., Chou K.C. Using deformation energy to analyze nucleosome positioning in genomes. Genomics. 2016;107:69–75.
4. Chen W., Feng P.M., Lin H., Chou K.C. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res. 2013;41:e68.
5. Chen W., Feng P.M., Lin H., Chou K.C. iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. Biomed. Res. Int. 2014:623149.
6. Chen W., Lin H., Feng P.M., Ding C., Zuo Y.C., Chou K.C. iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. PLoS One. 2012;7
7. Chou K.C. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 2011;273:236–247.
8. Chou K.C., Zhang C.T. Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 1995;30:275–349.
9. Dächsel JC, Farrer MJ. LRRK2 and Parkinson Disease. *Arch Neurol.* 2010;67(5):542–547. doi:10.1001/archneurol.2010.79
10. Guerois R., Nielsen J.E., Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J. Mol. Biol. 2002;320:369–387.
11. Jia J., Liu Z., Xiao X., Liu B., Chou K.C. iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. Anal. Biochem. 2016;497:48–56.
12. Jia J., Liu Z., Xiao X., Liu B., Chou K.C. pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. J. Theor. Biol. 2016;394:223–230.
13. Lin H., Deng E.Z., Ding H., Chen W., Chou K.C. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res. 2014;42:12961–12972.

14. Liu B., Fang L., Liu F., Wang X., Chen J., Chou K.C. Identification of real microRNA precursors with a pseudo structure status composition approach. PLoS One. 2015;10

15. Liu B., Fang L., Long R., Lan X., Chou K.C. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics. 2016;32:362–369.

16. Liu B., Fang L., Wang S., Wang X., Li H., Chou K.C. Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. J. Theor. Biol. 2015;385:153–159.

17. Liu Z., Xiao X., Qiu W.R., Chou K.C. iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition. Anal. Biochem. 2015;474:69–77.

18. Liu Z., Xiao X., Yu D.J., Jia J., Qiu W.R., Chou K.C. pRNAm-PC: predicting N(6)-methyladenosine sites in RNA sequences via physical–chemical properties. Anal. Biochem. 2016;497:60–67.

19. Mooney S.D., Krishnan V.G., Evani U.S. Bioinformatic tools for identifying disease gene and SNP candidates. Methods Mol. Biol. 2010;628:307–319.

20. Nagamani S, Singh KD, Muthusamy K. Combined sequence and sequence-structure based methods for analyzing FGF23, CYP24A1 and VDR genes. *Meta Gene*. 2016;9:26-36. doi:10.1016/j.mgene.2016.03.005.

21. Ng P.C., Henikoff S. Accounting for human polymorphisms predicted to affect protein function. Genome Res. 2002;12:436–446.

22. Qiu W.R., Xiao X., Lin W.Z., Chou K.C. iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model. J. Biomol. Struct. Dyn. 2015;33:1731–1742.

23. Ramensky V., Bork P., Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 2002;30:3894–3900.

24. S. A. de Alencar and Julio C. D. Lopes, "A Comprehensive In Silico Analysis of the Functional and Structural Impact of SNPs in the IGF1R Gene," Journal of Biomedicine and Biotechnology, vol. 2010, Article ID 715139, 8 pages, 2010.

25. Singh Kh D., Karthikeyan M. Combined sequence and sequence-structure-based methods for analyzing RAAS gene SNPs: a computational approach. J. Recept. Signal Transduct. Res. 2014;34:513–526.

26. Sherry S.T., Ward M.H., Kholodov M., Baker J., Phan L., Smigielski E.M., Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29:308–311.

27. Thusberg J., Olatubosun A., Vihinen M.m. Performance of mutation pathogenicity prediction methods on missense variants. Hum. Mutat. 2011;32:358–368.

28. Xi T., Jones I.M., Mohrenweiser H.W. Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. Genomics. 2004;83:970–979.
29. Yue P., Li Z., Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. J. Mol. Biol. 2005;353:459–473.
30. Yue P., Moult J. Identification and analysis of deleterious human SNPs. J. Mol. Biol. 2006;356:1263–1274.