Gene Prioritization by Integrated analysis of structural and functional protein-protein interaction network of neurological disorders.

A Major Project dissertation submitted

In partial fulfilment of the requirement for the degree of

Master of Technology

In

Bioinformatics

Submitted by

Yashna Paul

(2K12/BIO/24)

Delhi Technological University, Delhi, India

Under the supervision of

Dr. Yasha Hasija

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 **“Gene Prioritization by Integrated analysis of structural and functional protein-protein interaction network of neurological disorders.”**, submitted by **Yashna Paul (2K12/BIO/24)** 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 her 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:

Dr. Yasha Hasija
(Project Mentor)
Assistant Professor and Associate Head
Department of Bio-Technology
Delhi Technological University
(Formerly Delhi College of Engineering, University of Delhi)

DECLARATION

I, **Yashna Paul** hereby declare that the M. Tech. dissertation entitled **"Gene Prioritization by Integrated analysis of structural and functional protein-protein interaction network of neurological disorders"**, submitted in partial fulfilment of the requirement for the award of the degree of Master of Technology in Bioinformatics, Delhi Technological University (Formerly Delhi College of Engineering, University of Delhi), is a record of original and independent research work done by me under the supervision and guidance of **Dr. Yasha Hasija**, Assistant Professor, Department of Biotechnology, Delhi Technological University, New Delhi. The information and data enclosed in this dissertation is original and has not formed the basis of the award of any Degree/Diploma/Associateship/Fellowship or other similar title to any candidate of any university/institution.

Date:

Yashna Paul
M. Tech Bioinformatics
Department of Biotechnology
Delhi Technological University
Shahbad Daulatpur,
Main Bawana Road,
Delhi -42, India

ACKNOWLEDGEMENT

I would like to thank God for giving me the faith and perseverance to accomplish my objectives. With him by my side the whole process was easier.

I would like to express my sincere gratitude and admiration to my advisor, **Dr. Yasha Hasija**, for her guidance, support, and mentorship that has helped me to successfully complete the project.

I would also like to thank my colleagues **Imran Sidi, Neha Nagpal, Jaya Uniyal,** and **Manu Kandpal** for their time, support, and help throughout my journey. I am extremely grateful to my lab mates **Gaurav Kandoi and Isha Srivastava** for their helpful suggestions, ideas and assistance throughout my research. They have always been approachable to help me in their full capacity.

Finally, I would like to thank my family and friends for their unconditional love and support. A special thanks to my sister, **Kanika Paul,** for bearing with my work schedule, and lending me the laptop for completing my work, even when she needed it the most. To my best friend **Harsh Shinde**, thank you for your thoughtful consideration, unconditional love and support that made my work easy and pleasant. All the people mentioned above, have been a very important part during my course and especially this project. I once again thank them, from the bottom of my heart. This project would have never been the same without their indispensable support.

Yashna Paul

2k12/BIO/24

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ABBREVIATION	MEANING
XIAP	X-Linked inhibitor of apoptosis
KEGG	Kyoto Encylopedia of Genes and Genomes
3D	3-dimensional
EEG	Electroencephalograms
AD	Alzheimer’s disease
PD	Parkinson’s disease
KD	ketogenic diet
LC	locus coeruleus
WEKA	Waikato Environment for Knowledge Analysis
PPI	protein-protein interaction
PDB	Protein Data Bank
HUBBA	Hub Objects Analyser
DSS	Double Scoring Scheme
DMNC	Density of Maximum Neighbourhood Component
MNC	Maximum Neighbourhood Component
TP	True positive
FP	False positive
TN	True negative
FN	False negative
TPR	True positive rate
TNR	True negative rate
SPC	Specificity
FPR	False positive rate
PPV	Positive predicted value
ROC	Receiver operating characteristic
PCA	Principal Component Analysis
DAVID	Database for Annotation, Visualization and Integrated Discovery
MCODE	Molecular Complex Detection
ICAM1	Intercellular adhesion molecule 1
ACE	Angiotensin-1 converting enzyme
NOS3	Nitric oxide synthase 3
INSR	Insulin receptor
TGFBI	Transforming growth factor, beta induced
MMP9	Matrix metallopeptodase 9

Gene Prioritization by Integrated analysis of structural and functional protein-protein interaction network of neurological disorders.

Yashna Paul

Delhi Technological University, Delhi, India

ABSTRACT

Common neurological disorders show similar phenotypic manifestations like anxiety, depression and cognitive impairment. There is a need to identify shared genetic markers and molecular pathways that lead to these comorbid conditions. The present study aims to prioritize novel genetic markers that might increase the susceptibility of patients affected with one neurological disorder to other diseases as well. Identification of pathways involving such candidate markers that trigger similar clinical manifestations will help develop better and individualistic treatments for patients affected with neurological disorders. This systems biology study for the first time uses structural protein descriptors to analyse protein properties in a neurological protein interaction network. Results of protein prioritization by machine learning imply that XIAP (X-linked inhibitor of apoptosis), which is previously not known to be involved directly in neurological disorders, might be an important candidate for such diseases. These results are further validated when XIAP is characterized as an important hub and essential protein in neurological disease interaction network. It is also shown to be involved in more than one molecular pathways of neurological disorders. Therefore, structural and functional analysis of neurological protein-protein interaction network for prioritizing novel neurological candidates, proposes that XIAP, might be a significant genetic marker for neurological studies. This is an important finding, and results from this study can be used for further analysis and validation.

INTRODUCTION

Epilepsy is a common neurological disorder characterized by recurrent seizures. However, patients with other neurological disorders, also experience clinical manifestations similar to patients with epilepsy. The frequency of occurrence of seizures in these patients is more than that in the normal population. Apart from seizures, comorbid conditions like anxiety, depression, and cognitive impairment are common symptoms in most patients with neurological disorders. This observation implies that these neurological disorders have certain shared genetic markers and molecular pathways that lead to common clinical manifestations. There might also be genetic markers associated with one disease, the mutations in which results into over or under expression of other associated genes and interconnected molecular pathways, causing similar observable symptoms in patients with a second disease.

Keeping these observations in mind, the present study implements a strategic systems biology approach for structural and functional analysis of neurological protein interaction network. By network analysis, this study aims to identify novel putative genetic markers that are the cause of comorbid conditions in neurological disorders. The approach followed for network analysis of neurological disorders in this study is unique and novel in several ways.

Firstly, this study is targeted at the proteome level. Proteins and their interactions play a central role in the biological functioning of an organism. This provides important understanding on how a mutated gene product perturbs an interaction to control the healthy and diseased state of an organism (Chen *et al.,* 2013).

Secondly, the prioritization of novel candidates by machine learning takes into consideration the structural descriptors of proteins in an interaction network. This includes the integration of protein network topological properties and 3-dimensional protein interface structural properties of individual proteins. Studies have shown that these descriptors, individually, are optimal features of prioritizing novel candidates. The present study, however, considers both these structural descriptors together for the first time to describe the proteins under study. Structural level analysis provides important clues about the affinity and specificity of protein interaction, and hence only the proteins whose 3D structure is available, are considered (Johnson *et al*., 2013).

Third noteworthy feature of the study is the identification of protein hubs and essential proteins in the human neurological network. These proteins have highest number of interacting partners in the network, and can be thought to participate simultaneously in most neurological pathways. Therefore, important consideration could give be given to these hubs when studying disease states.

Protein prioritization results have discovered a previously not known neurological candidate, XIAP. This X-linked inhibitor of apoptosis has also been characterized as an essential hub in the disease interaction network. This study gives first clues into the direct involvement of XIAP, in human neurological networks, though previously it has been known as an agent of neuroprotection. Pathway analysis studies using DAVID (Database of Annotation, Visualization and Integrated studies), also show the involvement of this novel marker in more than one important KEGG (Kyoto Encylopedia of Genes and Genomes) neurologically associated pathways. Association of XIAP in the neurological disease network is clearly evident from the present strategic network analysis approach, and it can be proposed as a novel candidate neurological protein.

REVIEW OF LITERATURE

3.1 Aims and Objectives

- Gene prioritization my machine learning to identify novel neurological candidates in functional protein-protein interaction network that analyses protein topological properties; and integrating these properties with protein interface 3-dimensional (3D) structural properties.

- Identification of protein hubs in the functional protein-protein human interaction network that includes only those proteins whose 3D structure is available.

- Identification of protein hubs in the structurally characterized functional neurological protein-protein interaction network.

- Identification of protein hubs in a network of known neurological proteins and novel neurological candidates identified by machine learning.

- Comparison of these hubs to characterize their essentiality in different networks.

- Identification of essential proteins from the novel neurological candidates that have been identified.

- Pathway analysis study of newly identified neurological protein candidates, to analyze their involvement in different KEGG neurological pathways.

3.2 Epilepsy

Epilepsy is a common neurological disorder characterized by recurrent spontaneous seizures. Experiencing seizures more than once on account of a brain injury leads to epilepsy. It is one of the least understood diseases. It commonly begins in the childhood and in people aged over 60 years. However, epilepsy can effect anyone and can be encountered at any age. Epileptic seizures arise from within the brain. Seizures that are attributed to external factors like lack of oxygen, and consuming a lot of alcohol are not classed as epilepsy (Kendall-Taylor *et al*., 2009).

What is a seizure?

A seizure is a short episodic burst of symptoms caused by abnormal electrical activity in the brain. A seizure can last from a few seconds to a few minutes, affecting the muscles, behaviour, emotions, sensation and consciousness of the patient.

Nerve cells in the brain constantly send tiny electrical messages down the nerves to all parts of the body. Specific parts of the brain control different parts of the body and their functions. The symptoms appearing during a seizure are directly related to where in brain the abnormal burst of electrical activity occurs

Cause of epilepsy

In many cases, no cause for the seizures can be found – Idiopathic epilepsy. People suffering with idiopathic epilepsy have no other neurological (brain) condition.

Different types of epilepsy and seizures

Epileptic seizures are of two main types - generalised and partial.

Generalised seizures

When the entire or most of the brain is affected by abnormal electrical activity, the patient suffers from generalised seizures that involve much of the body in most cases.

Generalised seizures can be of various types:

- **A tonic-clonic seizure** is the most common type of generalised seizure. This type of seizure causes the entire body to stiffen, in addition to loss of consciousness. The body then shakes (convulses) due to uncontrollable muscle contractions.
- **Absence seizure** is a second type of generalised seizure that is accompanied by a brief loss of consciousness or awareness. It occurs mostly in children lasting few seconds without any convulsions.
- **A myoclonic seizure** is caused due to sudden contraction of the muscles- causing a jerk. These can affect the whole body but often occur in just one or both arms.
- **A tonic seizure** causes a brief loss of consciousness and stiffening of the body.

Partial seizures

These are focal seizures and the burst of electrical activity starts and stays only in one part of the brain. Therefore, the symptoms are also localized, depending upon which part of the brain is affected:

- **Simple partial seizures** cause muscular jerks or strange sensations in one arm or leg, odd taste, or pins and needles in one part of your body.
- **Complex partial seizures** arise from a temporal lobe or any part of the brain that causes strange behaviour for a few seconds or minutes that might include fiddling with an object, mumbling, or wandering aimlessly.

This is called a secondary generalised seizure are the ones, when a partial seizure develops into a generalized one.

Symptomatic epilepsy

Head injury, stroke, genetic syndromes, growths and tumours of the brain, infections and diseases like meningitis, encephalitis and cerebral palsy present at birth, or that develop later in life can cause epilepsy. These conditions can affect the surrounding brain cells triggering seizures.

Certain triggers make a seizures more likely at one time than another. These include stress and anxiety, antidepressants, low blood sugar level, heavy intake of alcohol, menstruation, and fever causing infections. Epilepsy is diagnosed by blood tests, brain scans and electroencephalograms (EEG- brainwave recordings). Medication is not the cure for epilepsy. Seizures can only be avoided or prevented through medication (Kojovic *et al*., 2011).

There is an accepted proposition that there is a possible link between neurobehavioral disorders and the temporal lobe or complex partial epilepsy. Also the incidence of neurobehavioral disorders is higher in patients with epilepsy than in the normal population (Kanner *et* al., 2009). Factors and mechanisms that have found to be common between epilepsy and other behavioural disorders include several common neuropathy and genetics, psychiatric illness, alterations in receptor sensitivity etc. 20-30 % of epileptic patients encounter psychiatric disturbances with a frequency of 6-12 times than that of the general population. Common psychotic conditions in epilepsy are suicidal ideation, anxiety and depression (Schmitz *et al*., 1999).

Comorbid conditions of anxiety, seizures, cognitive impairment and depression is experienced in neurological disorders like epilepsy, schizophrenia, bipolar disorder, Parkinson’s, Alzheimer’s and autism. The present study aims to study the relationship between a protein-protein interaction networks of these disorders. Previous studies have given clues that there can be shared genetic markers and molecular pathways between these disorders.

Common factors between Alzheimer's disease, Parkinson's disease and Epilepsy

Epilepsy, Alzheimer's disease (AD) and Parkinson's disease (PD) are commonly classified as neurological disorders. A remarkable reduction in LC noradrenergic neurons is an alteration observed in both AD and PD, and is associated in the early progression stages of the diseases. These diseases share common symptoms of depression and cognitive impairment attributed to aberrations in noradrenergic nervous system in the hippocampus region of the central nervous system. Also, hippocampus receives innervation from the LC (locus coeruleus) - that is severely compromised in these diseases. Therefore the hippocampus and noradrenergic neurons are important contributing factors for depression and cognition associated with these disorders, and there is a necessity to assess the relationship between them. Epilepsy is not categorized as a neurodegenerative disorder, however temporal lobe (or complex partial epilepsy) – observed in adults is related to neuronal loss in the hippocampus, and the patients exhibit comorbid conditions of cognitive impairment and depression. These conditions are attributed to the loss of noradrenergic neurons in the hippocampus, as in the case of AD and PD. There is a need to establish how loss noradrenergic neurons can contribute to comorbid symptoms in the three disorders. The genetic link that allows for common conditions needs to be established, and as of now, there are only physical evidences of these diseases being associated with each other. As another example -several anti-epileptic medications such as ketogenic diet (KD) which has improved frequency of occurrence of seizures in epileptic patients, also show a neuroprotective effects on dopaminergic neurons. These are another set of neurons that are affected in PD and cause loss of motor functions. Establishing genetic markers and mechanisms that cause similar conditions in these neurological disorders can enhance the medical treatments given to the patients (Szot P. 2012).

Autism and Epilepsy

Autism is present in up to 30% of subgroups of children with epilepsy, for example in infants and children with seizures in the first 3 years of life. Epilepsy occurs in approximately 8 to 20% of children with autism spectrum disorders with an increasing prevalence of seizures occurring into late adulthood. A major risk factor for the co-existence of epilepsy and autism is intellectual disability, with the highest risk for epilepsy or autism in either group being in those with more severe cognitive impairments. There are two distinct peaks to seizure onset in children with autism spectrum disorders one occurring early, prior to age 3 years, this is the group in which epilepsy is identified first and then autism is diagnosed and a later peak in which autism is identified first and then epilepsy occurs. There is no clear evidence that autism is caused by epilepsy, although the contribution of epilepsy and interictal epileptiform discharges to ongoing cognitive deficits in children with autism continues to be controversial and poorly understood. Children with both epilepsy and autism have increased morbidity and mortality as compared to those with only autism or epilepsy. There is evidence to suggest that when epilepsy and autism co-exist in the same person, common shared anatomical and molecular mechanisms may account for both, epilepsy and autism. To establish the shared molecular mechanisms that occur in patients with these disorders there is a need to- Identify infants with seizures at risk for autism and those with autism at risk for epilepsy, identify genetic and environmental risk factors common to epilepsy-autism, identify and develop animal models, biomarkers, and assessment tools that inform outcome in infants with epilepsy that go on to develop autism and in those with autism that go onto to develop epilepsy, explore the underlying mechanisms of convergence between autism and epilepsy, coordinate tissue and brain banking efforts in epilepsy-autism and develop treatment models behavioural and pharmacological in infants with epilepsy- autism (or at risk for autism). Therefore, putative risk factors for epilepsy in autism have been identified and these require further investigation (Bolton *et al*., 2011).

Bipolar disorder and epilepsy

Previous studies have demonstrated evidences of psychiatric comorbidity between epilepsy and bipolar disorder. The most prominent shared symptom in the two diseases is depression (Chang *et al.,* 2013). As compared to general population, in which the incidence of developing bipolar disorder is 0.07, the same in patients with epilepsy is 1.69 cases per 1000 persons-year., which is significantly high (Ettinger *et al*., 2005).

Migraine and Epilepsy

Epilepsy and Migraine are associated chronic disorders with episodic attacks. The diseases are comorbid and share overlapping pathophysiological mechanisms and common clinical features. Recently identified common genetic markers and molecular substrates for epilepsy and migraine include mutations in genes like CACNA1A, ATP1A2, SLC1A3 and POLG. However, both conditions also have several distinct and important differences. Hence, the diagnosis and treatment of each of these diseases must take into consideration a potential presence of the other (Bianchin *et al.,* 2012).

The above mentioned neurological disorders share distinct symptoms and comorbid conditions with epilepsy. A number of clinical manifestations and phenotypes are common to these disorders like anxiety, depression, stress, suicidal behaviour, and loss of memory and motor functions in some. There must be an underlying genetic answer to these common symptoms. Several molecular mechanism are commonly shared between these diseases. Neurological disorders involve complex processes and large number of genes. The present study therefore aims to identify putative shared genetic signatures between these disorders, and their potential connection with epileptic seizures. In order to find shared risk factors we have developed an integrated and systematic approach that takes into consideration protein-protein interaction properties and protein interface structural properties of the human neurological interactome. Identified risk markers will be important to determine patient prognosis for these diseases. Identifying such markers by taking into consideration the properties of known genes and proteins involved in the disorders under study is known as gene prioritization (Zhang *et al.,* 2011).

3.3 Gene Prioritization

When the normal functioning of a gene is disrupted by a disease causing aberration, the gene is called a candidate disease gene. Identification of disease specific genes is complicated by factors like gene pleiotropy, polygenic genes, influence of environmental factors, and genomic variations. To establish a link between the causal gene and the disease is expensive and time consuming experimentally. To reduce the associated costs, the candidate genes can be comprehensively prioritized before experimental testing. Computational gene prioritization involves using several associated evidences that relate each gene with the disease under study and predict potential causal links. Gene prioritization is greatly depends on the reliability and quantity of data in hand.

The approach systematically narrows down the list of genes to be tested experimentally and arranges them in the order of their likelihood of involvement in the disease. Specific relevant features and parameters like gene expression, function, pathway involved, and associated mutation effects are considered to assign the gene ‘priority. Disease genes have some characteristic features that can be used to categorize them. It has been reported that disease genes tend to interact with other disease genes. They also harbour functionally deleterious mutations. Disease genes code for proteins localized to the affected biological compartment that can be a tissue, cellular space or a specific pathway. Genes associated with a disease are longer in length and have more number of exons. In addition, they also have more orthologs and less paralogs (Bromberg Y. 2013).

Therefore, the major tasks in computationally prioritizing potential genes related to a disease phenotype is to:

1. Identify the characteristic features that form the basis to identify potent candidate disease markers.
2. Select a method for gene prioritization.

The present study involves machine learning for gene prioritization. This work specifically deals with prioritizing novel gene products, that is, proteins that are previously not known to be associated with neurological disorders. Machine learning techniques have been successfully used to find informative genes and mining critical information from raw data supplied to the machine. These prediction models have an increased interpretability and retain high accuracy to exploit the supplied data and figure out the required information effectively. A platform that can be used to apply machine learning on the dataset is WEKA (Glaab *et al*., 2012).

WEKA (Waikato Environment for Knowledge Analysis) is a software system used for data mining, developed at the University of Waikato in New Zealand. WEKA is a platform that is used to develop machine learning techniques and implement their application on real-world data mining problems. It is a compilation of machine learning algorithms that can be directly applied to a dataset for data mining tasks. WEKA can perform a wide range of statistical algorithms on the data set like data processing, classification, regression analysis, association and clustering. The results can also be graphically visualized and analysed thereafter. It is also used to develop new machine learning schemes. It is an open source software issued under the GNU General Public License (Mark *et al*., 2009).

WEKA platform gives the user a choice of various algorithms to choose from. All algorithms require a specific set of features to train upon. For the present area of study that involves identification of candidate genes that confer comorbid conditions to epilepsy and other neurological disorders, we have chosen two set of features.

These are:

1. The network properties – that define the behavior of genes and proteins in a network
2. The structural properties of protein interfaces – that takes into consideration the 3-Dimensional structure of the interface of two interacting protein partners in a protein-protein interaction network.

The present study has been comprehensively performed on the proteome level. Additionally, the identified putative protein products are also mapped to their corresponding gene markers that could be the risk factors for comorbid conditions in epilepsy and other neurological disorders under study.

3.4 Significance of network based studies

Most gene prioritization approaches are based on the assumption that genes associated with same or related disease phenotypes have shared molecular and functional mechanisms in the cell. Network based studies have been used to identify and validate novel candidate genes based upon network linkages with known disease genes. The method first constructs a gene-gene interaction network based upon genomic or proteomic data, followed by subsequent ranking of candidate genes depending upon their proximity to known disease associated genes in the network. Functional linkage networks are very helpful as they include physical (direct) interactions as well as functional (indirect) associations. Functional association data is derived from co-expression data and high throughput experiments. The goal of the present study is to exploit the functional coherence of genes involved in epilepsy and other neurological disorders under study to identify previously unknown links between these disorders that show comorbid manifestations. Of the 20 systematically assembled proteomic features for this study, 10 are network properties that are calculated for the functional protein network of the proteins involved in neurological disorders under study as well as other existing proteins with known structures in PDB (that form the unknown set), constructed using CYTOSCAPE using input protein functional interactions extracted from String Database (Linghu B *et al*., 2009).

3.5 Significance of protein-protein interactions

Proteins function by interacting with one another and also with other molecules of the cell like, DNA and RNA; and mediate vital metabolic pathways, signalling cascades, cellular processes and organismal systems. The unique function that each protein interaction confers to the system, determines its affinity and specificity. Moreover, the unique function of each interaction determines its affinity and specificity. Protein interactions therefore have a central role in the biological functioning of an organism and a perturbation of such interactions that might include gain of an inappropriate interaction or the loss of an important association controls the healthy and diseased state of an organism. Disease mutations affect the protein’s binding interface causing biochemically dysfunctional allosteric changes in the protein’s binding site. Studying protein interaction level can give insights into the molecular basis of the disease, and this information can be used to devise better methods for the prevention, diagnosis and treatment of diseases (Chen *et al*., 2013).

Protein interaction networks can also be used for evolutionary studies of individual proteins and the pathways in which the proteins are involved. Interaction maps from one specie might also have limited use to predict interactions in other species. The application of protein interaction networks that has been exploited for the present study is their use in suggesting the role and function of previously uncharacterized proteins by identifying their role in various protein complexes and pathways.

A protein-protein interaction network is composed of nodes and edges. Nodes represent individual proteins and edges represent the physical interaction between them. The topological properties of the proteins in the network essentially define the modularity of the protein interactions. The classification of proteins as hubs that have distinct properties has important implications when we relate topological properties to interacting proteins in the network. Additionally, integrating protein interface structure to topological properties of the protein, as done in the present study, helps to relate protein-protein interactions to a better extent and gives a much better criteria for gene prioritization (Gonzalez *et al*., 2012).

3.6 Topological Properties of protein interaction network

One of the basic property is the degree, and it is defines as the number of edges connected to a node. A molecule that interacts with many other molecules has a high degree. A simple path refers to a sequence of distinct but connected nodes in a network. Shortest path length between two nodes is the path that connects the nodes with minimum length (the length is measured by the number of edges that connect the nodes). The average path length that is the characteristic path length is the average of all shortest path lengths between all pairs of nodes in the network.

Average path length is an important statistical feature of the network and is used to describe the closeness of the network, which informs how quickly information is passed in the network. Highly connected nodes and essential genes are often correlated in network. Most essential genes correspond to house-keeping genes, required for the survival of the individual. Another global property of a protein network is the Betweenness centrality. It is possible to identify important nodes by specifying the way in which a node impacts the communication between two nodes. Betweenness of a node is the ratio of the number of shortest paths that pass through the node to the total number of paths that pass through the node. It is an important characteristic to classify protein hubs based upon their position in the network (Gursoy *et al*., 2008).

Protein Hubs in a protein-protein interaction (PPI) network

Hubs are defined as the proteins which have a large number of interactions in a PPI network. Hubs are of principal significance in an interaction network and greatly affect its functionality and stability. The specific recognition of interaction partners by hubs gives important clues about the structural properties of the hubs. Properties of hubs that distinguishes them from protein non-hubs includes level of intrinsic disorder, surface charge and distribution of domains as well as differences in functional domains (Patil *et al*., 2012).

Essential Interactions in the protein network

Some of the interactions in a protein network are more essential than others and hubs are important not only because they have high degree but also due to some significant interactions. An essential protein-protein interaction is the one which is indispensable to the organisms' survival. It is assumed that an essential interaction occurs between essential proteins, therefore the number of essential interactions is equal to interactions between essential proteins. Hubs have dense connection of interactions, and therefore the probability that hubs include essential interactions is more. For this specific reason the removal of hubs from a protein-protein interaction network is considered to be lethal. Also, different interactions have distinct level of importance in the network. Studying these interactions according to their significance in the network can aid for development of drug targets with increased potency (Gursoy *et al*., 2008).

In addition to protein network properties, the present study includes protein interface structural properties to characterize proteins and identify previously unknown protein markers. Protein interface property is an important characteristic, as proteins primarily interact with their interfaces. The specificity of an interaction can be attributed to the specific properties that each interface possesses. Studying the interface properties of proteins related to neurological disorders, will help explain their role in the protein interaction network and help us understand how specificity governs protein interactions. Overlapping or similar binding sites in a protein interface should have many interactions in single interface hub proteins, making those proteins important for the network. Multi-interface hub proteins have distinct binding sites for a number of protein partners.

The present study proposes a methodology that integrates protein interface 3-dimensional structural properties into neurological disease interaction networks. The neurological diseases interactions from human protein interactome are first subjected to the Network analyser plugin from CYTOSCAPE that calculates the network properties for interactions between known complexes as well as those between known and previously unknown complexes. Interface properties are then calculated for each protein individually, that takes part in the network, and that is, the interactions are replaced by interfaces, coming from known or previously unknown proteins. This study provides an analysis of neurologically related human protein interfaces as well as the topological properties of the network formed from these proteins with other proteins in the human interactome (Kar *et al*., 2009).

3.7 Interface properties and their importance

Analysing protein interfaces involved in the protein network, helps to identify the mutations occurring in the interface that might be related to specific diseases. By targeting interfaces – by altering their properties, one might be able to shut down mutated pathways, or add new and alternate interactions in the network. Assigning interfaces to protein interactions, therefore has both fundamental and practical relevance providing insights into functional specificities of the protein interactions, furthermore highlighting elements of competition as well as cooperativity amongst the interacting partners.

Therefore protein interfaces are important as they provide structural insights about the protein interactions. However a conjugative study that involves both the protein topological properties and protein interface properties has not been carried out for neurological disorders. The present study fills this void and uses a method to characterize interactions in a human neurological protein-protein interaction network using three-dimensional protein structures and interfaces to prioritize previously unknown genes that night be associated with these diseases (Johnson *et al*., 2013).

This systematic approach is utilized for identifying novel protein candidates associated with neurological disease interaction network. The identified putative markers identified in this study might be closely associated with comorbidity in these diseases. Also, they might be potential candidates for increasing the susceptibility of an individual to more than one disease through shared molecular pathways. Such findings will help in development of improvised and individualistic treatments for patients with neurological disorders.

METHODOLOGY

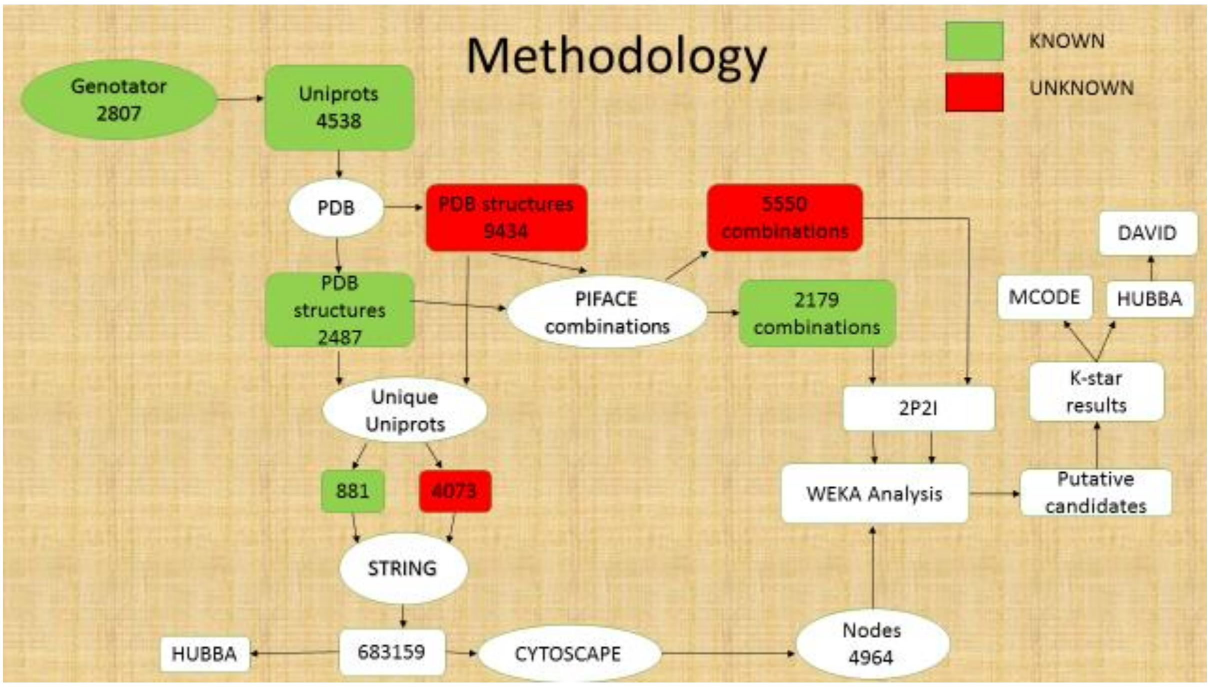

Figure 1: Methodology Flowchart

1. List of genes known to be associated with the diseases under study namely- Epilepsy, Alzheimer's, Parkinson's, Autism, Schizophrenia, Bipolar Disorder, and Migraine was taken from GENOTATOR - http://genotator.hms.harvard.edu/geno/. Total number of unique genes associated with these disorders was 2807.

   GENOTATOR is an online available real-time aggregation tool that has a multi-query engine. In automatically integrates data from 11 external clinical genetics resources to provide reliable ranking of genes in order of disease relevance. It comprehensively covers both historical genetics research and recent advancements and discoveries in disease genetics. The output is an excel sheet that consists of gene list specific to a disease (Wall *et al*., 2010).

2. These 2807 genes were found to be associated with 4538 UNIPROT Ids, that is, corresponding to 4538 identified (known) proteins.

3. Since the parameters into consideration take note of the interface structural properties of the interacting proteins, the PDB Id list of the above mentioned 4538 proteins was extracted from RCSB Protein Data Bank (PDB) - www.rcsb.org (Berman *et al*., 2000).

4. Out of the total available human protein structures (47,532) on PDB -17, 457 correspond to our list 4538 proteins associated with the group of diseases under study. Rest 30,075 is taken as the unknown set of proteins, that is, the proteins not considered to be associated with the group of diseases under study.

5. The list of 17,457 proteins included a number of structural variants associated with each PDB structure. Hence this list was then manually sorted and only the structure with the highest resolution was considered. Also, mutant and recombinant structures were avoided. Apo- structure, if available for a protein was given prime importance. From a number of structures available for each chain, a single high resolution structure was considered for a chain. The sorted list included a list of 2487 proteins, each associated with its available chain structures. It was important to include all available chain structures for each protein, as a protein interface can be formed by the combination of any of the available chains. Excluding the chain structures would mean losing out information on the protein interface structure.

6. Similarly the unknown list of 30,075 proteins was sorted, and it was reduced to 9434 proteins and their available chain structures.

a) <u>Calculation of Network properties</u>

1. Predicted protein-protein interactions for the proteins were extracted from the STRING database - http://string-db.org/. The STRING database is inclusive of protein-protein interactions that have been predicted experimentally, computationally, and those published in literature (Chaudhary *et al*., 2009).
2. A total of 683159 interactions for all the proteins was extracted. The protein-protein interaction data was used as input to build a network for CYTOSCAPE (v 3.1.0) - http://www.cytoscape.org/.

CYTOSCAPE is an open source software platform for visualizing molecular interaction data from expression profiles. The input file consists of a list of interactions, in this case- protein interactions. It can be used for visualization and analysis of network graphs involving nodes and edges. An important aspect of the software is the inclusion of its number of plugins for identifying specialized features of the network, as well as for mining important data and conclusions from the network. The input file can be in .sif and .xsls formats (Cline *et al*., 2007).

<u>The Network Analyser</u>

It is an established free open-source software platform for the analysis and visualization of molecular interaction networks. It functions as Java plugin which is well integrated into CYTOSCAPE and computes specific parameters describing network topology using efficient graph algorithms. The topological analysis of the network was carried out using this functionality of CYTOSCAPE. It is highly robust as it can help in characterization of biological networks with the help of such topological parameters. It can be used to compute two types of topological parameters, viz., simple parameters (single values) and complex parameters (distributions) on both directed (directed edges) as well as undirected networks (undirected edges).

Simple parameters includes the number of nodes, edges, self-loops, and connected components, the average number of neighbours, the network diameter, radius, density, centralization, heterogeneity, and clustering coefficient, the number of shortest paths, and the characteristic path length. Complex parameters are distributions of node degrees, neighbourhood connectivity, average clustering coefficients, topological coefficients, shortest path lengths, and shared neighbours of two nodes.

Number of connected components in undirected networks, two nodes are connected if there is a path of edges between them. Within a network all nodes that are pairwise connected form a connected component. The number of connected components indicates the connectivity of a network – a lower number of connected components suggest a stronger connectivity because many nodes are connected and form few connected components of large node size.

Shortest path parameters

The length of a path is the number of edges forming it. Two given nodes can be connected by multiple paths. The shortest path length, also called distance, between two nodes n and m is denoted by L (n,m). The network diameter is the maximum length of shortest paths between two nodes. If a network is disconnected, its diameter is the maximum of all diameters of its connected components. It can also be described as the maximum node eccentricity.

The **network radius** is the minimum among the non-zero eccentricities of the nodes in the network. The **average shortest path length**, also known as the characteristic path length, gives the expected distance between two connected nodes.

The **shortest path length distribution** gives the number of node pairs (n,m) with L(n,m) = k for k = 1,2, and so on. The network diameter and the shortest path length distribution may indicate small-world properties of the analysed network.

Degree distributions

In undirected networks, the **degree** of a node n is the number of edges linked to n. A self-loop of a node is counted like two edges for the node degree. The **node degree distribution** gives the number of nodes with degree k for k = 0, 1, and so on.

Clustering coefficient

Clustering coefficient is a ratio N / M, where N is the number of edges between the neighbours of n, and M is the maximum number of edges that could possibly exist between the neighbours of n. It always lies between 0 and 1.

The **network clustering coefficient** is the average of the clustering coefficients for all nodes in the network. Nodes with less than two neighbours are assumed to have a clustering coefficient of 0. The **average clustering coefficient distribution** gives the average of the clustering coefficients for all nodes n with k neighbours for k =2.

Parameters related to neighbourhood

The **neighbourhood** of a given node n is the set of its neighbours. The connectivity of n, denoted by k_n, is the size of its neighbourhood. The **average number of neighbours** indicates the average connectivity of a node in the network. A normalized version of this parameter is the **network density**. The density is a value between 0 and 1. It shows how densely the network is populated with edges.

Neighbourhood connectivity: The connectivity of a node is the number of its neighbours. The neighbourhood connectivity of a node n is defined as the average connectivity of all neighbours of n. The neighbourhood connectivity distribution gives the average of the neighbourhood connectivities of nodes n with k neighbours for k=0,1, and so on.

Shared neighbours P (n,m) is the number of partners shared between the nodes n and m, that is, nodes that are neighbors of both n and m. The **shared neighbours distribution** gives the number of node pairs (n,m) with P (n,m) = k for k = 1,….

Topological coefficients

The **topological coefficient** is a relative measure for the extent to which a node shares neighbours with other nodes. Mathematically, for a node n with k_n neighbours:

Where, J (n, m) is defined for all nodes m that share at least one neighbour with n. The value of J (n, m) is the number of neighbours shared between the nodes n and m, plus one if there is a direct link between n and m.

Closeness centrality is a measure of how fast information spreads from a given node to other reachable nodes in the network. It is defined as the reciprocal of shortest path length.

7. The final network had 4964 nodes and 683159 edges.
8. The functional interactions extracted from STRING, were also used as input for web based tool-HUBBA that analyses potential hubs in the network.

Hub OBjects Analyser (HUBBA) – http://www.hub.iis.sinica.edu.tw/Hubba/- is a web-based service, for exploring the essential nodes is an important work to find out what kind of roles do proteins act in a cell in biology. We identified important hubs present in our network using this online server. The interaction data from CYTOSCAPE was submitted in the web-based tool HUBBA in PSI-MITAB 2.5 format. Double Scoring Scheme (DSS) was used for topologically scoring the nodes in the network. DSS uses a parallel computation of two algorithms, viz. Density of Maximum Neighbourhood Component (DMNC) and Maximum Neighbourhood Component (MNC). The DSS logic was re-iterated to obtain most important proteins.

HUBBA explores the possibly essential proteins in the interaction network by six topology-based scoring methods and a DSS. For all the six methods applied to the protein- protein interaction dataset, DMNC was found to be the one that shares the least proteins with the other. Further, DMNC has the highest hit rate on the essential protein list. Therefore, DMNC was selected as the first method in the DSS and MNC was found to be next best method on the same criteria.

For n, most possible essential proteins are expected in the output, the 2n top ranked proteins by method A (DMNC) are selected firstly. The selected 2n proteins are further ranked by method B (MNC) and the n top ranked proteins are output. The number 2n is an empirical value for this double screening method. The main reason for selecting this scheme of scoring nodes for extracting out most relevant proteins in the network is to select methods catching diverse characters and to include most essential proteins (Lin *et al.*, 2008).

b) <u>For Structural Properties</u>

1. All protein chain combinations do not form the protein interface. Amongst a number of available chains in a protein, only a couple might be associated with forming the interface of the protein. The list of chain combinations that were involved in forming the interface of each protein, was extracted from PiFace.

   PiFace (http://prism.ccbb.ku.edu.tr/piface/index.php): It is an online protein interface property calculator tool. That allows to calculate properties of a protein interface by just submitting the PDB Id of the protein structure and its two chain Ids. Also, it allows for comparison between two interface structures. In addition to this, the online tool also allows for protein domain analysis and bulk data extraction. The clustered data contains all available PDB Ids and the chain combinations for which interface properties can be calculated (Cukuroglu *et al.*, 2014).

2. The available chain combinations were extracted from PiFace and the interface protein properties were calculated with another online available tool, called 2P2I inspector-http://2p2idb.cnrs-mrs.fr/2p2i_inspector.html. The tool characterizes protein-protein interfaces from 3D structures to calculate various physical and chemical descriptors. Input to the tool is a 4 letter PDB Id and 2 chains. PDB files can also be uploaded (Basse *et al.*, 2013).

9. The interface structural properties of 2179 proteins from the known set and 5550 proteins from the unknown set were calculated with 2P2I Inspector. These numbers indicate the number of available protein chain combinations as extracted from PiFace.

10. Interface properties that were used as structural descriptors were extracted from 2P2I, and include Total accessible surface area, gap volume, average interface accessible surface area, average neutral residues, average polar residues, average non-polar residues, average charged residues, gap volume index and interface size.

- **Total accessible surface area**: Surface area of a protein that is accessible to a solvent is called the total accessible surface area. Its unit is square angstroms.

- **Gap volume**: The gap volume gives a measure of the complementarity of the interacting surfaces. It is the volume of the gaps between two interacting surfaces.

- **Average Interface accessible surface area**: It is the average surface area exposed by the two chains in consideration.

- **Percent average neutral residues**: This is the percentage of average number of neutral amino acid residues present in the protein interface formed by the two chains under consideration.

- **Percent average polar residues**: This is the percentage of average number of polar amino acid residues present in the protein interface formed by the two chains under consideration.

- **Percent average non-polar residues:** This is the percentage of average number of non-polar amino acid residues present in the protein interface formed by the two chains under consideration.

- **Percent average charged residues**: This is the percentage of average number of charged amino acid residues present in the protein interface formed by the two chains under consideration.

- **Gap Index**: Gap index for all proteins was calculated as follows: Gap volume/total accessible surface area

- The 9th interface structural property- The **interface size** was calculated from PiFace.

11. The network properties and interface properties were combined together to prepare files for machine learning using WEKA.

a) **<u>Preparation of Training File:</u>**

1. The known set of protein properties was divided into half randomly. And same number of protein entries was taken from the unknown set.
2. This formed the training set, with 1090 known and 1090 unknown values.

b) **<u>Preparation of Test File:</u>**

Left over known proteins =1090 and unknown proteins (5550-1090) = 4460 were all included in the test file.

c) **Model Building with WEKA:**

1. The training file was used to build a model using ten-fold cross validation. Five models were build, and they are described below. Building more than one model helps explain how different classifiers make varied predictions on the training set. After ten-fold cross validation, the parameters of the built model, that describe the predictions for training set include Recall, ROC area, accuracy, precision, true positive rate and the false positive rate. All the models were individually applied on the test set to obtain the results.
    - True positive rate (TPR): Also called the sensitivity or Recall = TP/P = TP/ (TP+FN)
    - True negative rate (TNR): Also called the specificity, SPC = TN/N = TN/ (FP+TN)
    - False positive rate (FPR): Also called fall-out = FP/N = FP/ (FP+TN) = 1-SPC
    - Positive predicted value (PPV): Also called Precision = TP/ (TP+FP)
    - Accuracy = (TP+TN) / (P+N)
    - ROC area = TPR/1-SPC

    TP = True positive = correctly identified

    FP = False positive = incorrectly identified

    TN = True negative = correctly rejected

    FN = False negative = incorrectly rejected

Algorithms used for building models in WEKA

Naïve Bayes

This probabilistic classifier is based upon the Bayes theorem. Naïve Bayes classifiers perform effectively on classification tasks and are easy to use and interpret. They are very simple and work on the assumption that independent variables are also independent statistically. It is a useful means for classification when the number of input variables is high. It can sometimes outperform other sophisticated classifiers. The conditional distributions of input variables can be modelled by several methods like normal, gamma, lognormal, and Poisson.

The basic algorithm of Naïve Bayes includes calculation of the posterior probability of the event, which is prior probability of the event times the likelihood of the event to take place. A naïve Bayes classifier is trained by evaluating an approximation algorithm in a simple linear way. The classifier works by assuming that the value of a particular parameter is not related to the existence or absence of any other parameter, in a given class variable.

The advantage of using naïve Bayes classification model is that it requires only a small amount of training set data to make estimations on the parameters (that is the means and variances of the variables in the dataset) necessary for classification (Wilbur WJ. 2000).

Random Forest

This is an ensemble classifier and it can thought to be a type of nearest neighbour approach. Random forest classifiers utilize 'divide and conquer' approach to improvise on the performance of the predictor. Ensemble classifiers identify group together the "weak learners" from the dataset to form a "strong learner". Random forest begins predictions by forming "decision trees", which in terms of an ensemble classifier corresponds to the "weak learner". The algorithm of a decision tress takes the input from the top of the tree, and as the input data traverses down the tree, the data gets split into smaller data sets. The result of random forest is either the average or the weighted average of all the terminal nodes that are reached by the classifier. From a large number of parameters and features, the eligible parameters will differ from node to node. One pressure on the classifier is to make trees that are as uncorrelated as possible. This is required to reduce random forest error rate.

The runtime of a random forest classifier is quite fast, and they are equipped to deal with missing and unbalanced data. Random forest classifiers are not very useful for regression analysis as they are not capable of predicting beyond the range of the training data and they might over fit noisy datasets (Touw *et al*., 2013).

Rotation Forest

It is a method to generate ensembles of different classifiers. The method involves splitting of the feature set randomly into K subsets and application of Principal Component Analysis (PCA) on individual sets. The newly extracted feature is reassembled while retaining all other components. This is done to preserve the variability of the information in the data. The data is linearly transformed into new features. Diverse classifiers are obtained as the features are variably split leading to different rotations. The idea of the rotation approach is therefore to encourage individual accuracy and diversity simultaneously within the ensemble. Diversity and accuracy of the classifier are hence maintained. PCA involves simple rotation of the feature's coordinate axes, and the base classifier is a decision tree, and hence this ensemble classifier has been named "Rotation Forest" (Rodríguez *et al*., 2006).

K-Star

K-star is an instance based machine learning classifier that uses entropy as a distance measure. This algorithm provides a consistent approach to handle symbolic attributes, missing values and real valued attributes. It takes all possible transformation paths into consideration (Wang *et al*., 2006).

Bagging J48

J48 machine learning algorithm is a decision tree learner, version of the C4.5. Decision trees that implement J48 are built using analyses of training data and the greedy technique. Decision tree nodes evaluate the significance of all features. Input is classified in the decision tree by following a path from roots to leaves, that results in a decision of the input class. The tree building follows a top-down approach that selects the most suitable attribute each time. An information-theoretic measure calculates the classification power of each feature. The decision tree first choses a feature and then forms subsets of the training data based upon the different values of the feature that is selected. The same process gets iterated for each subset until majority of the instances belong to same class. The decision tree approach applies high accuracy set of rules, but can apply excessive rules. Therefore to gain a balance between flexibility and accuracy the trees in J48 are pruned to generalize a decision tree. The methods of pruning can be of two types: subtree raising and subtree replacement. A node is moved upwards in the direction of the root of the tress in subtree raising. In subtree replacement, nodes get replaced by leaves, therefore working backwards in the direction of the root.

Bagging

Bagging is a method to improvise classification by combining learned models from various subsets of a dataset. Application of bagging reduces over fitting and variance of the dataset. Bagging uses the instability of the classifier to perturb the training set. Hence, using the same learning algorithm, different classifiers can be produced. For example, if a training set A, has size t. For this training set, n number of new training sets Ai (t’ < t) is generated. Subsets can be generated by uniform sampling of instances from A and by replacement. Due to sampling with replacement some instances are repeated in each new training set and are called as bootstrap samples. N number of models can be fitted using n number of subsets or bootstrap samples. The output is the average of the above result (Sridhar *et al*., 2012).

12. The best results with highest precision, recall, ROC area, and accuracy are used for further analysis.

13. Using HUBBA, hub were identified for the known protein-protein interaction network. Also, hub proteins were identified for protein interaction list that included known proteins as well as novel putative protein candidates from machine learning results.

14. These three identified hub protein lists are compared to understand their essentiality in all protein networks, namely, human structural protein interaction network, neurologically associated structural protein interaction network, and neurologically associated and newly predicted protein network.

15. DAVID- http://david.abcc.ncifcrf.gov/ -analysis of hub protein list from the third network which includes neurologically associated and newly predicted protein candidates was performed, to identify common pathways between new candidates and existing KEGG-http://www.genome.jp/kegg/neurological pathways (Kanehisa *et al*., 2006).

DAVID is a web-accessible program that provides integrated information about functional genomics annotations and their graphic summaries. It contains annotated gene or protein identifiers that share categorical data information on protein domains, biochemical pathway membership, Gene Ontologies, etc. Therefore, the Database for Annotation, Visualization and

Integrated Discovery (DAVID), provides the user with a collection of data-mining tools that help in systematic integration of data from various databases. It includes functionally annotated data for a number of genomes like human, rat, fly mouse, etc. (Huang *et al*., 2009).

16. DAVID analysis of novel candidates is followed by MCODE analysis of the neurological protein interaction list that also includes the putative protein results. MCODE analysis identifies essential proteins in the network, that by definition are indispensable to an individual's survival. MCODE analysis will identify if any of the putative candidate is also an essential genetic marker.

MCODE is yet another CYTOSCAPE plugin which can be used for finding some highly interconnected regions (clusters) in a PPI network which can be either protein complexes or some functional modules. It implements the well-known automated method Molecular COmplex DEtection algorithm for delineating clusters from a network. The results can be finely tuned with a plethora of node-scoring and cluster-finding parameters. The algorithm as proposed by Bader and Hogue, works by weighting a vertex by local neighbourhood density, choosing a few seeds with high weight, and isolating the dense regions according to some threshold values. This algorithm was implemented with the help of MCODE plugin functionality for CYTOSCAPE v 3.1.0 for identification of some functional modules in our network. It consists of three stages: vertex weighting, complex prediction and optional post-processing. In the first stage, MCODE weights all the vertices based on the core clustering coefficient of vertex v which is defined to be the density of the highest k-core (a graph of minimal degree k) of the immediate neighbourhood of v (vertices connected directly to v) including v. Once the weights are computed, MCODE seeds a cluster with the highest weighted vertex and recursively moves outward from the seed vertex. A new vertex will be added to the cluster if its weight is larger than a given threshold. By such a greedy fashion, MCODE can isolate densely connected regions iteratively. In the post-processing step, MCODE filters or adds proteins based on connectivity criteria. For each vertex v, the weight w of v is:

$$w = k * d$$

Where, 'd' is the core-density of the highest k-core graph from the set of vertices including all the vertices directly connected with v and vertex v itself, defined as the ratio of the actual number of edges to possible edges between the nodes in a k-core.

The time complexity of the entire algorithm is polynomial O (nmh3) where n is the number of vertices, m is the number of edges and h is the vertex size of the average vertex neighbourhood in the input graph, G. This comes from the vertex-weighting step. Finding a k- core in a graph proceeds by progressively removing vertices of degree < k until all remaining vertices are connected to each other by degree k or more, and is thus O (n2). The highest k-core is found by trying to find k-cores from one up until all vertices have been found and cannot go beyond a number of steps equal to the highest degree in the graph. Thus, the highest k-core step is O (n3). Since this k-core step operates only on the neighbourhood of a vertex, the n in this case is the number of vertices in the average neighbourhood of a vertex, h

The strategic systems biology approach followed by applying the above methodology, which also validates machine learning results by analysing hubs and essential protein candidates, is a reliable means to prioritize novel proteins. The results of this study should be considered for further validation and analysis (Bader Hogue 2003).

RESULTS AND CONCLUSIONS

Machine Learning Results

WEKA classifiers were used to build modesls on the training set. The training set included 1090 known and equal number of unknown proteins. Known proteins are the ones that are associated with neurological disorders. The unknown list consists of all other known human proteins whose 3-Dimensional structures are available on the Protein Data Bank (PDB). There are 19 training features in both the training set and test set, for quantified description of the proteins. The features used for describing known proteins and evaluating novel candidates are an integration of protein network properties and interface structural properties. The list of protein features that the training algorithms get trained, as mentioned earlier in the methodology, are:

Protein network Properties	**Protein interface structural properties**
Average shortest path length	Total accessible surface area
Clustering coefficient	Gap volume
Closeness centrality	Average interface surface area
Eccentricity	Percent average neutral residues
Stress	Percent average polar residues
Degree	Percent average non-polar residues
Betweenness	Percent average charged residues
Neighbourhood Connectivity	Gap index
Radiality	Interface size
Topological Coefficient	

Table 1: Protein structural descriptors considered for machine learning.

Five machine learning classifiers from WEKA, were applied to the training set, to obtain five corresponding models. These include:

1. Naïve Bayes
2. Random Forest
3. Bagging with J48
4. Rotation Forest
5. K-star

After 10 fold cross-validation, the predictions of the above classifiers on the **training set** are as follows:

	Naïve Bayes	**Bagging_J48**	**Rotation Forest**	**Random Forest**	**K-star**
Precision	67.4	85.3	84.8	85.5	83.4
Recall	71.4	87.6	84.9	86.9	87.7
ROC Area	72.3	93.9	92.9	94.4	92
Accuracy	68.4	86.25	84.81	86.08	85.11

Table 2: Predictions of five machine learning classifiers on training set.

Below is a comparative analysis chart of predictions made by all the five classifiers.

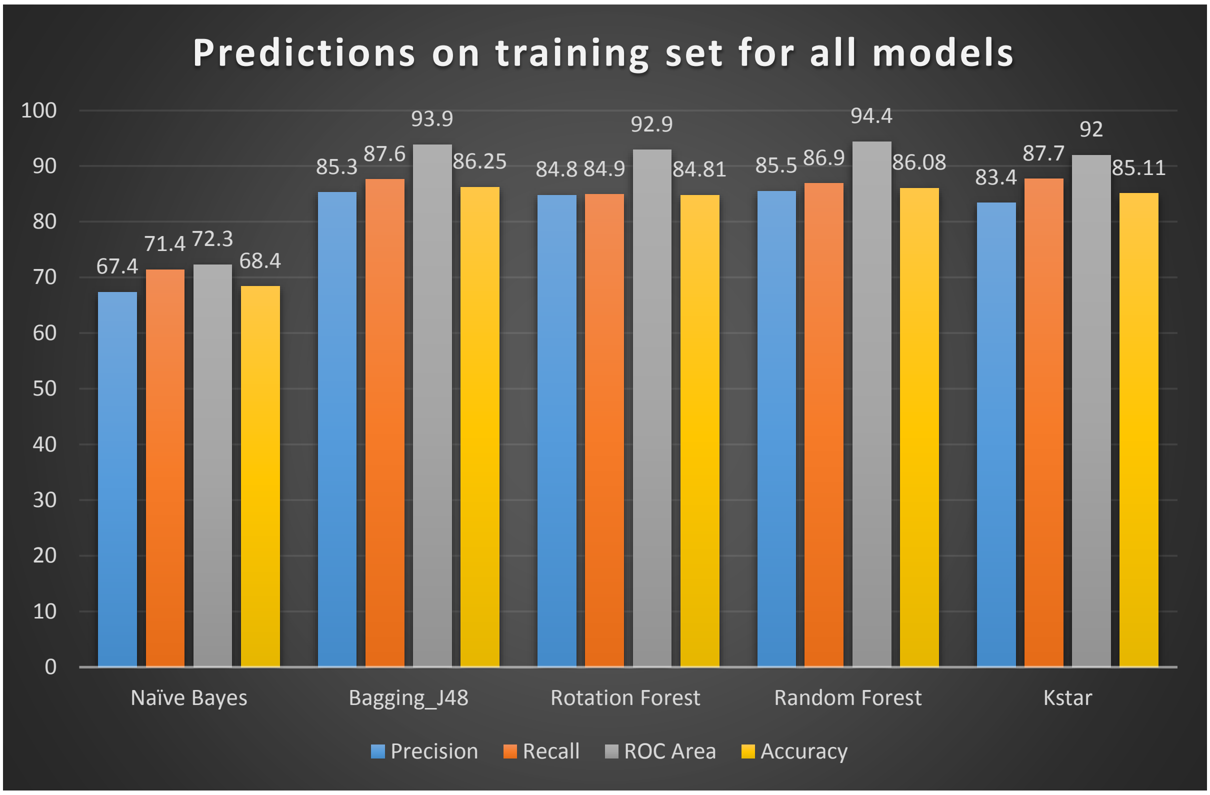

Figure 2: Graph for comparative representation for five model prediction on training dataset.

The classifier random forest has the best predictions of:

1. Precision,
2. Area under ROC curve, and
3. Accuracy

on the training data.

1. **Precision:**

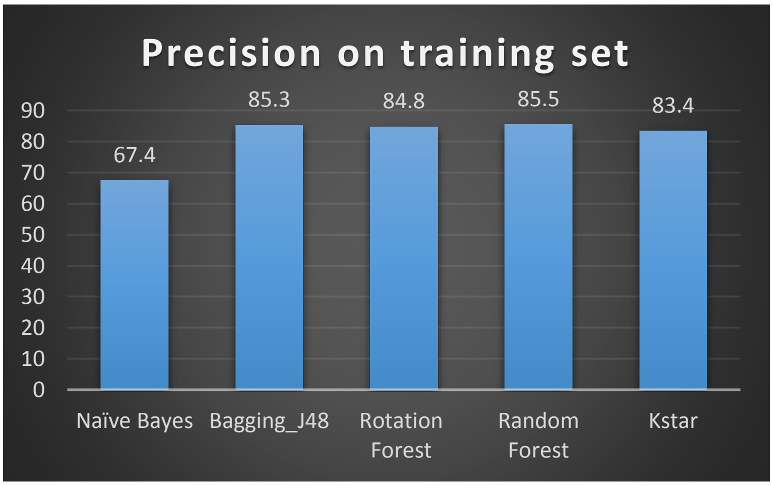

Figure 3: Comparison of precision values for five models on training set.

2. **ROC area:**

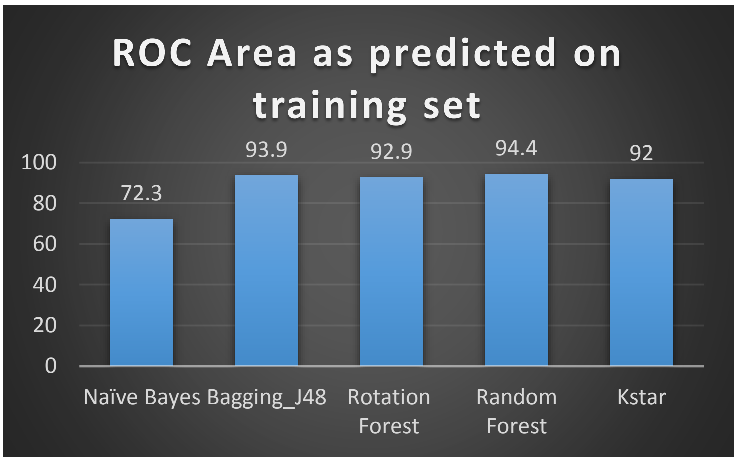

Figure 4: Comparison of ROC area values for five models on training set.

3. **Accuracy:**

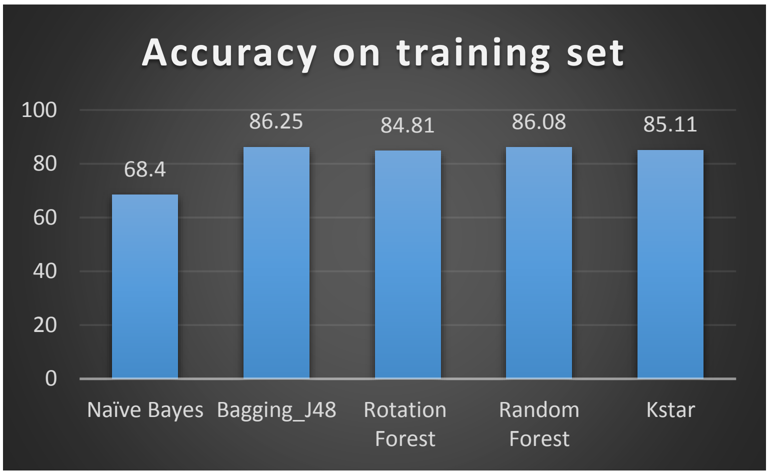

Figure 5: Comparison of Accuracy values for five models on training set.

However, the classifier K-star has the best predictions for recall on the training set

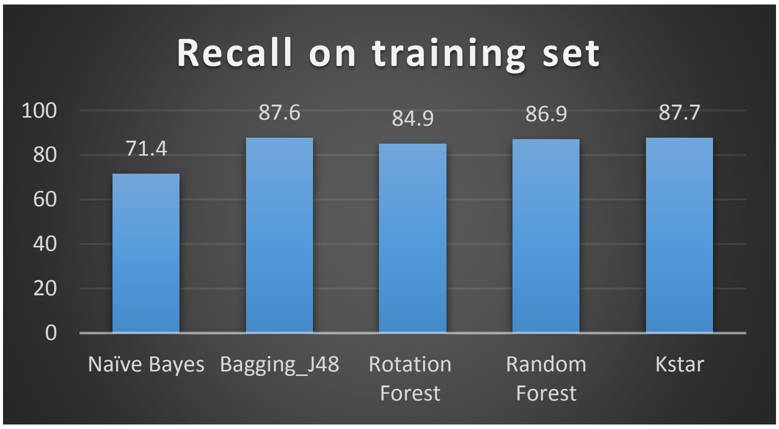

Figure 6: Comparison of recall values for five models on training set.

Precision defines the positively predicted values, and recall is the sensitivity. Precision is the number of instances that have been predicted correctly as known proteins. It is number of correct results divided by all returned results. In other words precision is the probability that an outcome picked at random is the one that is predicted correctly as the known protein. **Sensitivity/recall** is the number of known proteins that have been predicted correctly as being known. The **performance** of a single test/ classifier prediction can be calculated using the precision and the recall. The **F-score** is a single measure of the performance of the prediction, where,

$$F = 2(PRECISION*RECALL/PRECISION+RECALL)$$

Therefor the **performance** of the classifiers can be calculated as above. The results are depicted below as a graph.

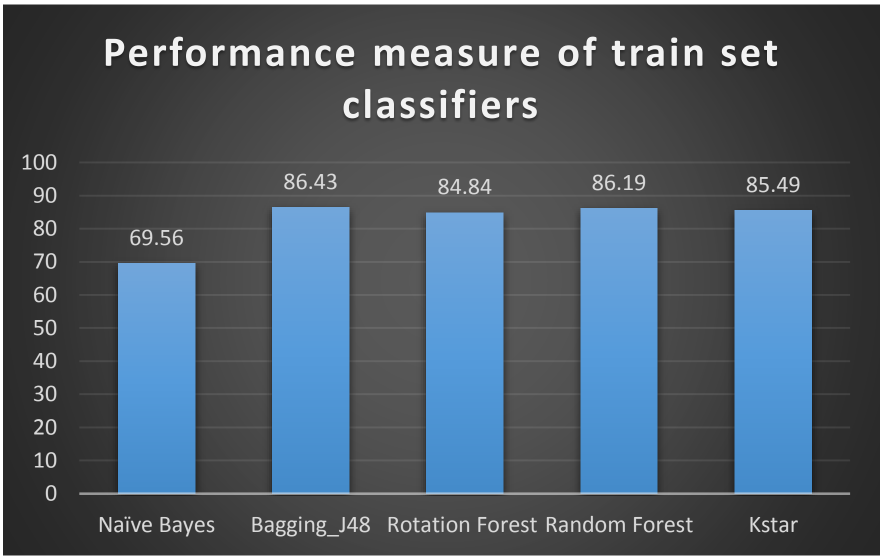

Figure 7: Comparison of performance values for five models on training set.

The graph shows that the classifier random forest has the best performance.

Accuracy is the proportion of true results, that is, the known proteins, and the unknown proteins classified as known. A hundred percent accurate result would mean that the measured values (unknown) are exactly the same as the known given values. Bagging by J48 has the best accuracy over the training set.

Receiver operating characteristic (ROC) curve is a graph that illustrates the performance of a classifier as its discrimination threshold varied. In other words it is a plot of the sensitivity versus one minus the specificity. It is a tool to select the possible optimal classifier and discard suboptimal ones. As it can be seen from the ROC curve analysis graph, random forest has the best value.

Classifiers J48 and random forest are decision tree algorithms, and represent the path followed by the algorithm for classification of the training sets as tress.

Tree diagram for J48:

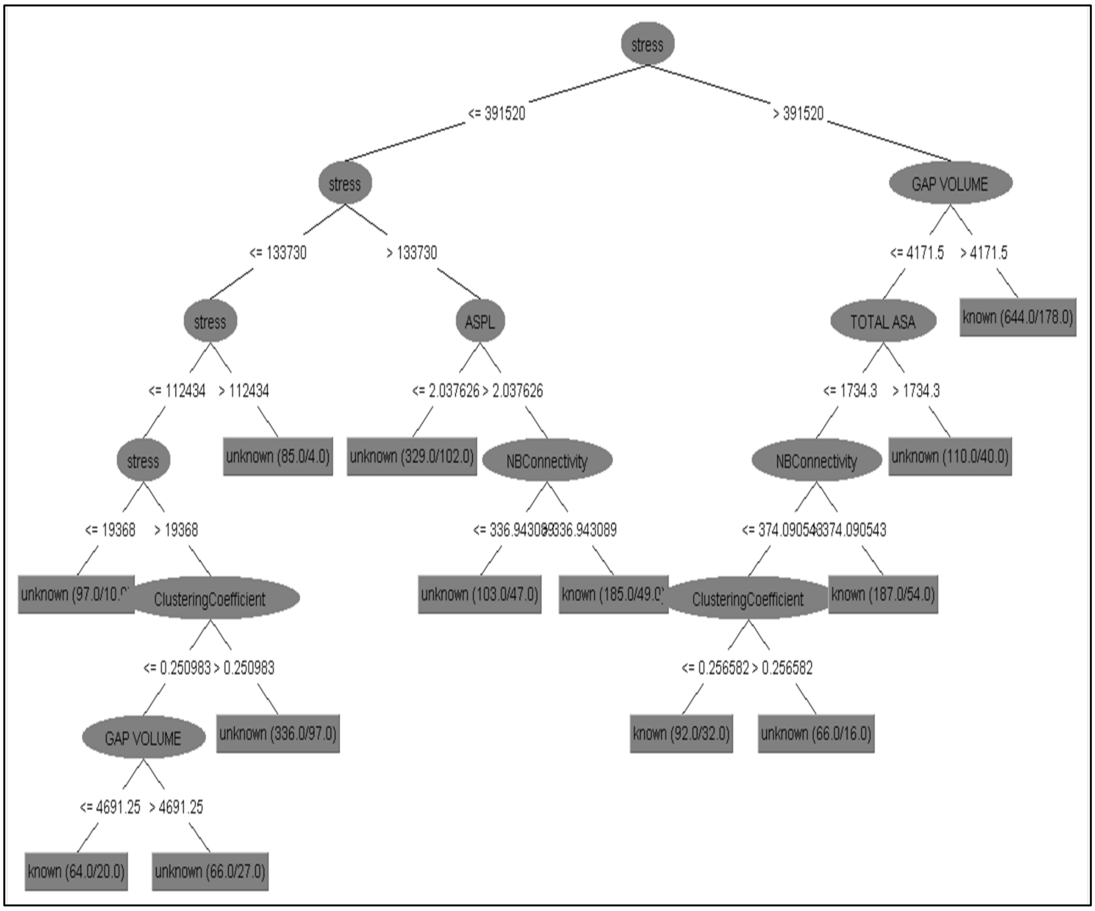

Figure 8: Decision tress representation of the J48 classifier applied by bagging on the training dataset.

It can be seen that the J48 decision tree follows a top-down approach that selects the most suitable attribute each time. The first most suitable feature selected is the stress. The decision tree measures the classification power of each feature to measure which feature will best classify the dataset. In this case it is the stress. Stress is used to form subsets of the training set depending upon the values of stress. Stress values with less than or equal to 391520 form the left subset, and the values greater than 391520 are classified on the right subset. At each node, the feature that best classifies the training dataset is used to divide the data into smaller subsets. Similar process is iterated for each subset until majority of the instances belong to same class. This approach is very accurate to make predictions on the training dataset, however it can apply excessive rules to classify the data. Hence, for a balanced model, and to generate a tree diagram, tress in J48 are pruned to form a generalized decision tress for a specific model. The above tree diagram is the pruned form of the Bagging-J48 classifier that has been used for predictions on the training set in the present study. The numeric values at each leaf is descriptive of the number of instances that reached at that node, and the second numeric values depicts the number of instances that were classified incorrectly.

.

Tree diagram for Random Forest:

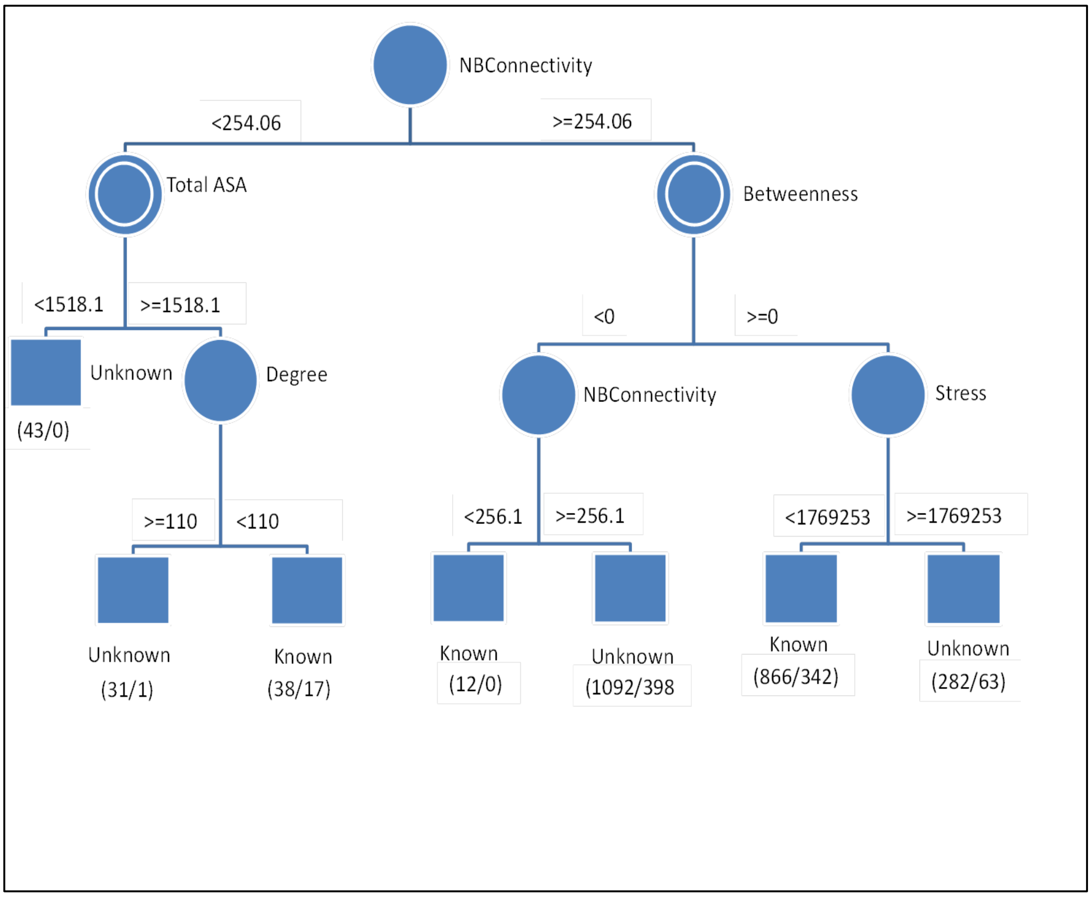

Figure 9: Decision tress representation of the random forest classifier applied on the training dataset

This decision tress takes the input from the top of the tree, and as the input data traverses down the tree, the data gets split into smaller data sets. The feature selected at each node is the strongest classifier at that node. The first feature is the neighbourhood connectivity. The result of random forest is either the average or the weighted average of all the terminal nodes that are reached by the classifier. From a large number of parameters and features, the eligible parameters will differ from node to node. The numeric values at each leaf is descriptive of the number of instances that reached at that node, and the second numeric values depicts the number of instances that were classified incorrectly.

The prediction results from the above described five classifiers in WEKA were used to predict potential candidate proteins similar to proteins involved in neurological disorders. Their predictions were applied on the test dataset individually to obtain the results. The following result predictors like precision, accuracy, recall, and area under ROC curve describe how successful the models have been to mine candidate proteins involved in neurological disorders, from a set of proteins that are previously not known to be associated with neurological disorders.

The results on the **test set** data for all classifiers is as follows:

	Naïve Bayes	Bagging_J48	Rotation Forest	Random Forest	K-star
Precision	47.1	74.1	75.7	75.3	75.7
Recall	52.1	72	68.4	71.8	76.6
ROC Area	61.6	86.2	85.6	87.3	88
Accuracy	59.72	79.88	79.76	80.43	81.84

Table 3: Predictions of five machine learning classifiers on test set.

Below is a comparative analysis chart of predictions made by all the five classifiers.

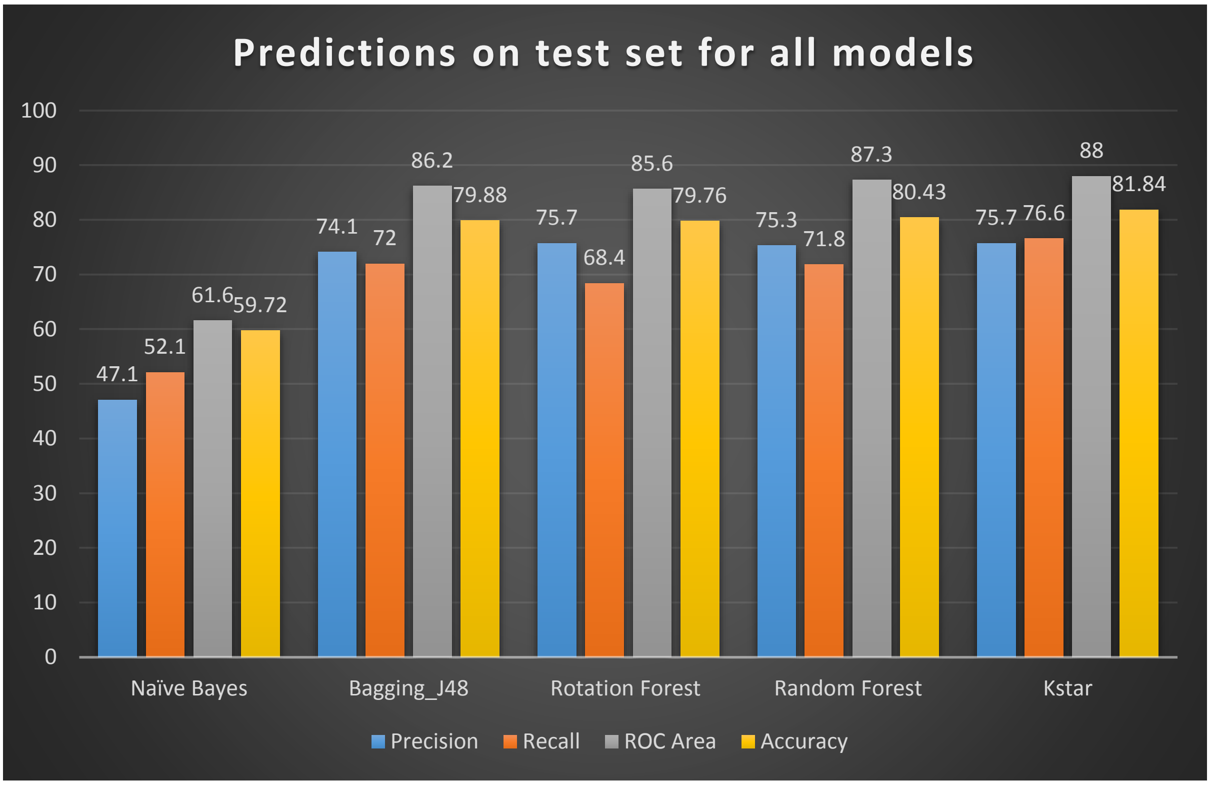

Figure 10: Graph comparative representation for five model prediction on test dataset.

Amongst the five models used for predictions on test set data, the **K-star** algorithm gives best result for precision, recall, area under ROC curve as well as for the accuracy of predictions.

1. **Precision** graph for five classifiers on the test set.

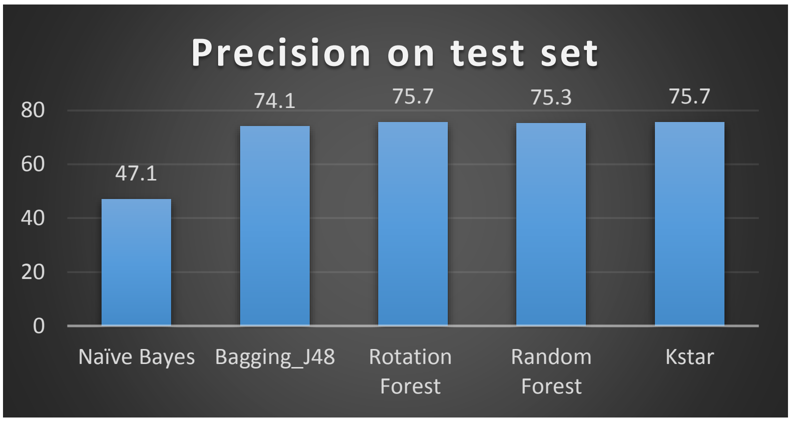

Figure 11: Comparison of precision values for five models on test set.

2. **Recall** graph for five classifiers on the test set.

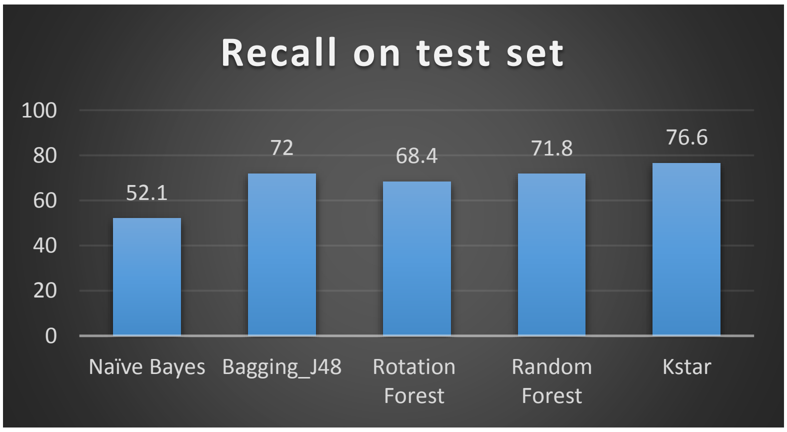

Figure 12: Comparison of recall values for five models on test set.

3. Graph for **ROC area** for models on test set.

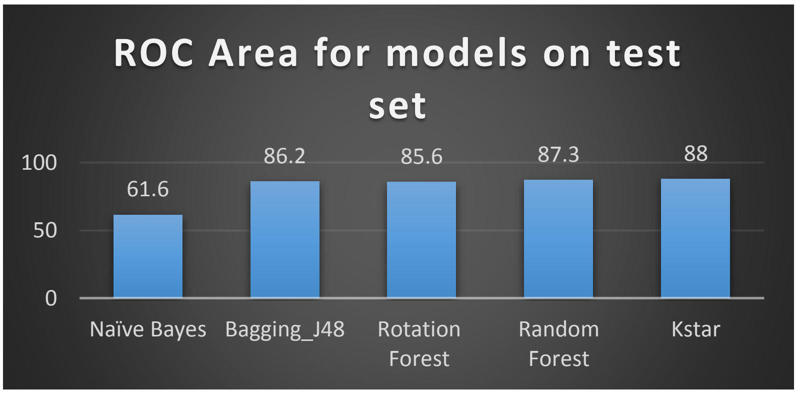

Figure 13: Comparison of ROC area values for five models on test set.

4. **Accuracy** of models on test set.

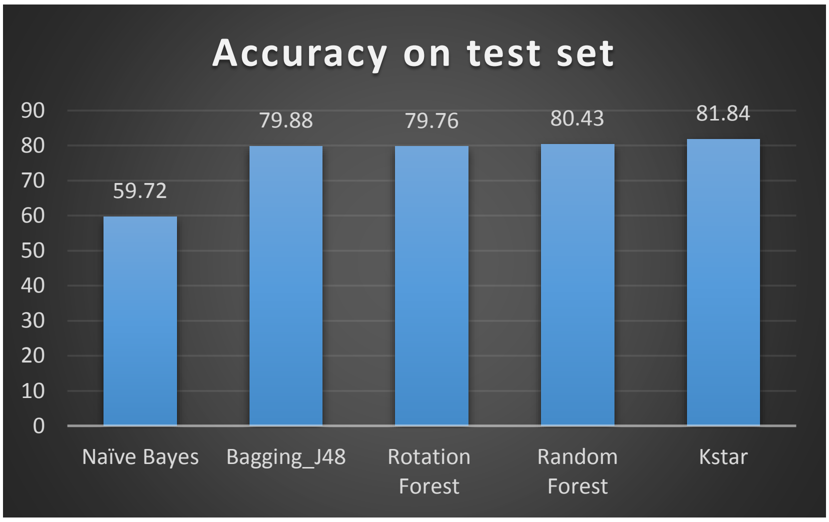

Figure 14: Comparison of accuracy values for five models on test set.

Performance/ F-score is calculated for all the models, and is depicted as bellow.

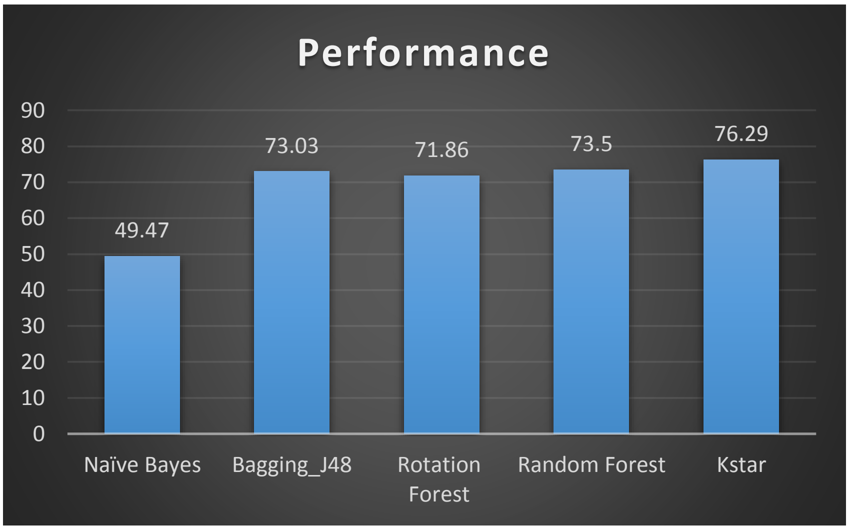

Figure 15: Comparison of performance of the classifiers on test set.

Different models might perform depending upon the dataset and the information contained. For the present test set, the model that gives best results is K-star. The best predicted putative candidate proteins from the results of this model are used for further analysis.

The table below gives the list of best 10 predictions of putative protein candidates from the unknown test set as obtained from the WEKA results of all five classifiers.

K-star	Random forest	Bagging-J48	Rotation forest	Naïve Bayes
1G82	1CC0	1S18	1ZT4	1HRK
2GJX	1CI4	1J1J	1MR1	1MR8
1MR1	1CKS	1OPL	2K03	1WNT
1NR4	1CZZ	1WMH	1L9X	1Z6X
1ZSV	1DZA	2ARY	1I3O	2B2Y
1S18	1GWQ	1I3O	1X86	2J4E
1WPQ	1HLO	2B5N	1XV9	2NN6
1CKS	1HYN	1H4O	1ZT4	1H28
1NLW	1IRJ	1YBO	2PO6	1IYI
1KN0	1KHU	2DSQ	2EWY	1Z6U

Table 4: Ten best putative candidate predictions from all five machine learning algorithms.

Many putative candidate protein predictions were found to be common in all five model results. However, they have different prediction probabilities in all results. The fact that some proteins were commonly predicted by all the classifiers increases the probability of those proteins as potential candidates, as they got mined by all classifiers. Since in certain classifier results, they might have lower prediction probabilities, they will not be taken as putative markers. This also shows that there is some common pattern in the protein feature values of the training set, that the machine learning algorithms used for classifying the proteins of the test set, and the pattern is commonly associated with these proteins that are duplicated in the results of all five classifiers. Best prediction probabilities describe how accurately the previously not considered a neurological disorder candidate protein is predicted to be associated with neurological disorders. All the analysis, hence forth consider the best 30 prediction candidates from K-star results.

Results from Hub-Object Analyser

The Hub Object analyser (HUBBA) is a web based tool that finds hub proteins from the input protein interaction data. Hub proteins have characteristic greater number of interactions in a network, than other non-hub proteins. In other words these proteins have more interaction partners, making them physiologically important for the individual. Hub proteins are essential elements, and are indispensable for an individual's survival (Batada *et al*., 2006).

Hubs are principal proteins in an interaction network and influences its stability and function. Deletion of a hub protein, or a mutation in the same, may lead to shutting down of an entire protein pathway. The specific binding of a hub protein with its protein partners is structurally very important. The web based tools is therefore used to identify such proteins in the neurological disease network. It will also be possible to identify certain features that will be specific to these hubs that possibly affect their binding affinities. Three comparative analysis of hubs has been performed in the present study. All the proteins used for developing the network throughout the study, are the ones whose 3-dimensional (3D) structure is available at the Protein Data Bank (Patil *et al*., 2010).

1. The first set of analysed hub proteins is from the human protein interaction network. All the protein interactions included are from proteins that have an available 3D structure. The HUBBA results are as follows. The hubs are ranked on the basis of increasing order of priority, marked by the colour coding.

It will be of interest to find out how many of these hubs are common to hubs in the protein interaction network of neurological disorders. That is, the second list of known proteins involved in neurological disorders is created, and their interactions extracted from STRING.

Rank	Node	Rank	Node	Rank	Node	Rank	Node
1	GSTA3_HUMAN	26	GPX1_HUMAN	51	CCR2_HUMAN	76	CERU_HUMAN
2	CSF2_HUMAN	27	TNR1A_HUMAN	52	1C07_HUMAN	77	RAGE_HUMAN
3	IL8_HUMAN	28	EDNRB_HUMAN	53	ITA4_HUMAN	78	NMS_HUMAN
4	TPA_HUMAN	29	INHBE_HUMAN	54	APAF_HUMAN	79	AREG_HUMAN
5	TIMP1_HUMAN	30	IL2RA_HUMAN	55	TNR1B_HUMAN	80	IL5_HUMAN
6	VCAM1_HUMAN	31	IBP3_HUMAN	56	PGH1_HUMAN	81	IL13_HUMAN
7	IL10_HUMAN	32	MMP3_HUMAN	57	CD38_HUMAN	82	FA5_HUMAN
8	BAX_HUMAN	33	ADML_HUMAN	58	GHRL_HUMAN	83	TIMP3_HUMAN
9	CCL2_HUMAN	34	SOMA_HUMAN	59	CXCL7_HUMAN	84	NTF4_HUMAN
10	PAI1_HUMAN	35	IL1RA_HUMAN	60	FOSB_HUMAN	85	IL9_HUMAN
11	LYAM2_HUMAN	36	FA8_HUMAN	61	CCR3_HUMAN	86	MK_HUMAN
12	NOS2_HUMAN	37	LYAM3_HUMAN	62	IL1R1_HUMAN	87	DSC3_HUMAN
13	LIF_HUMAN	38	TSP1_HUMAN	63	BDH_HUMAN	88	SSR2_HUMAN
14	SDF1_HUMAN	39	GRP_HUMAN	64	IL6RA_HUMAN	89	ITA2_HUMAN
15	HGF_HUMAN	40	MET_HUMAN	65	IL4RA_HUMAN	90	HUTH_HUMAN
16	IL1A_HUMAN	41	BCL2_HUMAN	66	ASM_HUMAN	91	ITAL_HUMAN
17	MMP2_HUMAN	42	LEPR_HUMAN	67	GNRHR_HUMAN	92	MUC1_HUMAN
18	IL4_HUMAN	43	VIP_HUMAN	68	PYY_HUMAN	93	PRLR_HUMAN
19	IGF1R_HUMAN	44	MMP1_HUMAN	69	BKRB2_HUMAN	94	PLF4_HUMAN
20	FRIL_HUMAN	45	ARY1_HUMAN	70	TLR9_HUMAN	95	UCN1_HUMAN
21	PDGFB_HUMAN	46	GA45A_HUMAN	71	ANTR2_HUMAN	96	CADM1_HUMAN
22	HMOX1_HUMAN	47	ANPRA_HUMAN	72	ALK_HUMAN	97	BKRB1_HUMAN
23	EGR1_HUMAN	48	PA24A_HUMAN	73	MOG_HUMAN	98	HMGB1_HUMAN
24	SCRB2_HUMAN	49	FOXO1_HUMAN	74	TGFB2_HUMAN	99	TKNK_HUMAN
25	IL18_HUMAN	50	RETN_HUMAN	75	ONCM_HUMAN	100	PI2R_HUMAN

Table 5: Hub proteins identified in the human structural protein interaction network.

2. The second set of analysed hub proteins comes from only the known protein candidate interactions whose 3D structures are available and that are known to be associated with neurological disorders. This step, specifically identifies the neurologically significant hub proteins, from a set of all proteins involved in neurological disorders.

Rank	Node	Rank	Node	Rank	Node	Rank	Node
1	BGH3_HUMAN	26	TPA_HUMAN	51	EDN1_HUMAN	76	FOXO1_HUMAN
2	INSR_HUMAN	27	IRS1_HUMAN	52	TTHY_HUMAN	77	PGFRB_HUMAN
3	NOS3_HUMAN	28	PA21B_HUMAN	53	SCRB2_HUMAN	78	ADML_HUMAN
4	ACE_HUMAN	29	CBP_HUMAN	54	ANF_HUMAN	79	CNR1_HUMAN
5	MMP9_HUMAN	30	PPARA_HUMAN	55	GPX1_HUMAN	80	NFKB1_HUMAN
6	PRL_HUMAN	31	BAX_HUMAN	56	FGFR2_HUMAN	81	PPAP_HUMAN
7	MK14_HUMAN	32	MBP_HUMAN	57	ANGT_HUMAN	82	CO3_HUMAN
8	ICAM1_HUMAN	33	PAI1_HUMAN	58	ESR2_HUMAN	83	MMP1_HUMAN
9	GSTA3_HUMAN	34	HGF_HUMAN	59	IL2RA_HUMAN	84	ITB2_HUMAN
10	GSTA1_HUMAN	35	SDF1_HUMAN	60	TNR1A_HUMAN	85	SOMA_HUMAN
11	GSTA4_HUMAN	36	PDGFB_HUMAN	61	TLR2_HUMAN	86	MP2K1_HUMAN
12	CASP8_HUMAN	37	NOS2_HUMAN	62	CALC_HUMAN	87	ARY1_HUMAN
13	FGF1_HUMAN	38	IGF1R_HUMAN	63	CCR5_HUMAN	88	CD38_HUMAN
14	PDE4A_HUMAN	39	MMP2_HUMAN	64	TNR16_HUMAN	89	CCL5_HUMAN
15	PERM_HUMAN	40	CCL2_HUMAN	65	MMP3_HUMAN	90	PA24A_HUMAN
16	THRB_HUMAN	41	PRGR_HUMAN	66	TSP1_HUMAN	91	GSTP1_HUMAN
17	CP2CJ_HUMAN	42	LIF_HUMAN	67	FA8_HUMAN	92	IL1R1_HUMAN
18	IL8_HUMAN	43	IL18_HUMAN	68	BCL2_HUMAN	93	CP1A1_HUMAN
19	TF65_HUMAN	44	IL4_HUMAN	69	HSP71_HUMAN	94	CTLA4_HUMAN
20	TNFL6_HUMAN	45	LYAM2_HUMAN	70	FGFR1_HUMAN	95	APAF_HUMAN
21	TLR4_HUMAN	46	FRIL_HUMAN	71	RARA_HUMAN	96	1C07_HUMAN
22	PARP1_HUMAN	47	HMOX1_HUMAN	72	LYAM3_HUMAN	97	CASP9_HUMAN
23	CRP_HUMAN	48	TKN1_HUMAN	73	CSF1R_HUMAN	98	ITA4_HUMAN
24	VCAM1_HUMAN	49	RB_HUMAN	74	MET_HUMAN	99	IL6RA_HUMAN
25	TIMP1_HUMAN	50	CATD_HUMAN	75	ERBB3_HUMAN	100	FAS_HUMAN

Table 6: Hub proteins identified in the human neurological protein interaction network

There are 21 neurologically important hub proteins that are present in the first 100 hub proteins list of the human interaction network. This shows the significance of neurological proteins in the human interactome. 21 of first 100 most important proteins in the human interactome, belongs to neurological pathways, signalling cascades, synaptic enzymes etc. This analyses proves the significance of signalling and protein interactions in the human protein network.

Five highest priority hubs as predicted by HUBBA in the neurological interaction network are as below.

UNIPROT Ids of highest priority hubs	**Gene name**	**Function**	**Disease associated**
BGH3_HUMAN	TGFBI transforming growth factor, beta-induced	The protein product of the gene acts to inhibit cell adhesion.	Migraine
INSR_HUMAN	INSR	Insulin receptor	Migraine, Alzheimer’s
NOS3_HUMAN	NOS3	Synthesizes free radical nitric oxide that acts as a mediator in several biological processes.	Migraine, schizophrenia, bipolar, epilepsy, Alzheimer’s, Parkinson’s
ACE_HUMAN	ACE	Angiotensin 1 converting enzyme	Migraine, schizophrenia, bipolar, epilepsy, Alzheimer’s, Parkinson’s, autism
MMP9_HUMAN	MMP9	Matrix metallo-peptodase 9 involved in the breakdown of extra cellular matrix	schizophrenia, bipolar, epilepsy, Migraine,

Table 7: Highest priority hub proteins identified in the human neurological protein interaction network

The uniprot Ids of the hub proteins were mapped to the corresponding genes, and these genes were found to be associated with neurological disorders under study. Some of these proteins are involved in more than one disease, validating their existence as hubs in neurological disease pathway network. Similarly other hubs predicted in the above list, can be thought as important markers involved in more than one pathway of neurological disorders.

3. The third hub proteins list is prioritized from protein interaction data of known proteins including the top best 30 putative candidates prioritized by K-star algorithm of machine learning. Therefore in addition to known neurological disease candidates, this list includes previously unknown putative neurological candidates. Their interaction data is extracted from STRING, and the same is used as input for HUBBA. This analysis informs if any of the previously unknown neurological candidates act as hub proteins in human neurological protein interaction network.

Rank	Node	Rank	Node	Rank	Node	Rank	Node
1	BGH3_HUMAN	26	PA21B_HUMAN	51	TNR1A_HUMAN	76	CSF1R_HUMAN
2	INSR_HUMAN	27	IRS1_HUMAN	52	EDN1_HUMAN	77	ERBB3_HUMAN
3	NOS3_HUMAN	28	CBP_HUMAN	53	TTHY_HUMAN	78	CNR1_HUMAN
4	MMP9_HUMAN	29	BAX_HUMAN	54	SCRB2_HUMAN	79	XIAP_HUMAN
5	ICAM1_HUMAN	30	PPARA_HUMAN	55	GPX1_HUMAN	80	PGFRB_HUMAN
6	PRL_HUMAN	31	MBP_HUMAN	56	ANF_HUMAN	81	ADML_HUMAN
7	MK14_HUMAN	32	PRGR_HUMAN	57	IL2RA_HUMAN	82	FOXO1_HUMAN
8	GSTA3_HUMAN	33	PAI1_HUMAN	58	FGFR2_HUMAN	83	CO3_HUMAN
9	GSTA1_HUMAN	34	HGF_HUMAN	59	ESR2_HUMAN	84	NFKB1_HUMAN
10	GSTA4_HUMAN	35	PDGFB_HUMAN	60	ANGT_HUMAN	85	ITB2_HUMAN
11	CASP8_HUMAN	36	TKN1_HUMAN	61	TLR2_HUMAN	86	SOMA_HUMAN
12	FGF1_HUMAN	37	SDF1_HUMAN	62	CCR5_HUMAN	87	CD38_HUMAN
13	PERM_HUMAN	38	MMP2_HUMAN	63	TNR16_HUMAN	88	ARY1_HUMAN
14	PDE4A_HUMAN	39	RNAS2_HUMAN	64	CALC_HUMAN	89	MP2K1_HUMAN
15	IL8_HUMAN	40	NOS2_HUMAN	65	MMP3_HUMAN	90	CCL5_HUMAN
16	THRB_HUMAN	41	CCL2_HUMAN	66	TSP1_HUMAN	91	IL1R1_HUMAN
17	CP2CJ_HUMAN	42	IGF1R_HUMAN	67	FA8_HUMAN	92	PA24A_HUMAN
18	TNFL6_HUMAN	43	LIF_HUMAN	68	BCL2_HUMAN	93	CTLA4_HUMAN
19	TF65_HUMAN	44	IL18_HUMAN	69	HSP71_HUMAN	94	GSTP1_HUMAN
20	TLR4_HUMAN	45	IL4_HUMAN	70	PPAP_HUMAN	95	ITA4_HUMAN
21	PARP1_HUMAN	46	LYAM2_HUMAN	71	FGFR1_HUMAN	96	APAF_HUMAN
22	CRP_HUMAN	47	HMOX1_HUMAN	72	LYAM3_HUMAN	97	CP1A1_HUMAN
23	VCAM1_HUMAN	48	FRIL_HUMAN	73	RARA_HUMAN	98	1C07_HUMAN
24	TIMP1_HUMAN	49	RB_HUMAN	74	MMP1_HUMAN	99	BRCA2_HUMAN
25	TPA_HUMAN	50	CATD_HUMAN	75	MET_HUMAN	100	CASP9_HUMAN

Table 8: Hub proteins identified in the human neurological protein interaction network that includes newly identified gene products.

Five highest priority hubs as predicted by HUBBA in the neurological interaction network are as below.

UNIPROT Ids of highest priority hubs	Gene name	Function	Disease associated
BGH3_HUMAN	TGFBI transforming growth factor, beta-induced	The protein product of the gene acts to inhibit cell adhesion.	Migraine
INSR_HUMAN	INSR	Insulin receptor	Migraine, Alzheimer's
NOS3_HUMAN	NOS3	Synthesizes free radical nitric oxide that acts as a mediator in several biological processes.	Migraine, schizophrenia, bipolar, epilepsy, Alzheimer's, Parkinson's
ACE_HUMAN	ACE	Angiotensin-1 converting enzyme	Migraine, schizophrenia, bipolar, epilepsy, Alzheimer's, Parkinson's, autism
ICAM1_HUMAN	ICAM1	Intercellular adhesion molecule 1 that encode glycoprotein found on the cell surface of immune cells.	Migraine, Alzheimer's, schizophrenia

Table 9: Highest priority hub proteins identified in the human neurological protein interaction network that includes newly identified gene products.

Out of the 30 newly added candidate proteins to the previously known protein list, 2 proteins have been categorized as hubs in the above HUBBA analysis. Also, addition of these new candidates has shuffled the known protein hub prioritization list, which is evident from comparing the list of best five hub predictions in both the cases, that is, the previously known hubs list and the previously known hubs list that includes novel candidates (Table 6 and Table 8 respectively). Therefore, the analysis informs that addition of these putative neurological candidates has added newer interactions in the protein network, which has resulted into different priority hubs than before. It is important to analyse how important these new interactions are. The significance of these interactions in network pathways needs to be further validated.

To study this, DAVID analysis of previously known hubs was compared with the DAVID analysis of the list of proteins with previously known and newly added candidates. This would give, the specific pathways in which the novel hubs are involved. The two novel hub proteins identified by HUBBA include RNAS2_HUMAN (Number 39 in Table 8, colour coded green in the table) and XIAP_HUMAN (Number 79 in Table 8, colour coded green in the network). Predictions of DAVID analysis show that RNAS2_HUMAN, is not involved in the KEGG pathways of known neurological hub proteins. However, XIAP_HUMAN is associated with the existing neurological hub proteins pathways in several ways. This protein is found to exist in several pathways like that of cancer, apoptosis, NOD-like receptor signalling pathways, small cell lung cancer pathways, and pathways for focal adhesion; along with known hub proteins of neurological disorders.

KEGG Pathways	**Genes**
Pathways in cancer	BCL2_HUMAN, CASP9_HUMAN, **XIAP_HUMAN**, PGFRB_HUMAN, BAX_HUMAN, NFKB1_HUMAN, HGF_HUMAN, MMP2_HUMAN, MET_HUMAN, IL8_HUMAN, MMP9_HUMAN, CSF1R_HUMAN, MMP1_HUMAN, FGFR2_HUMAN, RARA_HUMAN, BRCA2_HUMAN, MP2K1_HUMAN, TF65_HUMAN, CASP8_HUMAN, IGF1R_HUMAN, TNFL6_HUMAN, NOS2_HUMAN, PDGFB_HUMAN, FGFR1_HUMAN, FOXO1_HUMAN, CBP_HUMAN, RB_HUMAN, GSTP1_HUMAN, FGF1_HUMAN
Prostate cancer	BCL2_HUMAN, FGFR2_HUMAN, MP2K1_HUMAN, CASP9_HUMAN, TF65_HUMAN, PGFRB_HUMAN, NFKB1_HUMAN, IGF1R_HUMAN, PDGFB_HUMAN, FOXO1_HUMAN, FGFR1_HUMAN, CBP_HUMAN, RB_HUMAN, GSTP1_HUMAN
Cytokine-cytokine receptor interaction	CSF1R_HUMAN, CCL2_HUMAN, IL18_HUMAN, IL4_HUMAN, PRL_HUMAN, CCR5_HUMAN, SDF1_HUMAN, PGFRB_HUMAN, SOMA_HUMAN, HGF_HUMAN, IL1R1_HUMAN, TNFL6_HUMAN, PDGFB_HUMAN, MET_HUMAN, LIF_HUMAN, CCL5_HUMAN, IL8_HUMAN, TNR1A_HUMAN, TNR16_HUMAN, IL2RA_HUMAN
Apoptosis	BCL2_HUMAN, APAF_HUMAN, TF65_HUMAN, CASP9_HUMAN, **XIAP_HUMAN**, CASP8_HUMAN, TNR1A_HUMAN, NFKB1_HUMAN, BAX_HUMAN, IL1R1_HUMAN, TNFL6_HUMAN
NOD-like receptor signalling pathway	IL18_HUMAN, CCL2_HUMAN, TF65_HUMAN, **XIAP_HUMAN**, CASP8_HUMAN, IL8_HUMAN, CCL5_HUMAN, NFKB1_HUMAN, MK14_HUMAN
Melanoma	PDGFB_HUMAN, FGFR1_HUMAN, MP2K1_HUMAN, MET_HUMAN, PGFRB_HUMAN, RB_HUMAN, HGF_HUMAN, IGF1R_HUMAN, FGF1_HUMAN
Toll-like receptor signalling pathway	TLR2_HUMAN, MP2K1_HUMAN, TF65_HUMAN, CASP8_HUMAN, IL8_HUMAN, CCL5_HUMAN, NFKB1_HUMAN, TLR4_HUMAN, MK14_HUMAN
Small cell lung cancer	BCL2_HUMAN, NOS2_HUMAN, APAF_HUMAN, TF65_HUMAN, CASP9_HUMAN, **XIAP_HUMAN**, RB_HUMAN, NFKB1_HUMAN
Neurotrophin signalling pathway	BCL2_HUMAN, MP2K1_HUMAN, TF65_HUMAN, IRS1_HUMAN, NFKB1_HUMAN, TNR16_HUMAN, BAX_HUMAN, MK14_HUMAN, TNFL6_HUMAN
Epithelial cell signalling in Helicobacter pylori infection	TF65_HUMAN, MET_HUMAN, IL8_HUMAN, CCL5_HUMAN, NFKB1_HUMAN, MK14_HUMAN
Focal adhesion	BCL2_HUMAN, TSP1_HUMAN, PDGFB_HUMAN, MP2K1_HUMAN, **XIAP_HUMAN**, MET_HUMAN, PGFRB_HUMAN, HGF_HUMAN, IGF1R_HUMAN, ITA4_HUMAN
Cell adhesion molecules (CAMs)	LYAM2_HUMAN, ITB2_HUMAN, LYAM3_HUMAN, CTLA4_HUMAN, ICAM1_HUMAN, 1C07_HUMAN, VCAM1_HUMAN, ITA4_HUMAN

Table 10: Association of XIAP_HUMAN, in various neurological KEGG pathways (highlighted in green).

In a study for identifying features of cancer hubs, it was observed that hubs proteins have planar, smaller binding interfaces that are less tightly packed. This implies that hub proteins might have characteristic smaller accessible surface area and a smaller gap volume value when compared to rest of the proteins in the network. This observation suggests that XIAP_HUMAN, is a possible newly identified hub as it has a smaller interface accessible surface area as well as a smaller gap volume, compared to other putative neurological protein candidates (Kar *et al*., 2009).

Results from MCODE analysis

This is a CYTOSCAPE plugin, which is used to identify essential proteins in a protein interaction network. Essential proteins are generally products of house-keeping genes that are expressed in all cells of the body, and are required for major cellular processes like metabolism, DNA replication, maintenance of cell structure, transport processes, etc. They are crucial for an individual's survival. It has been shown that topological properties of protein-protein interaction network are useful to categorize the functionality and essentiality of proteins (Sharan *et al*., 2006).

Since we have prioritized previously unknown proteins on the basis of their topological properties and interface structure properties, this study forms a good basis to identify essential genes in a neurological protein interaction network (Kar *et al*., 2009).

MCODE is performed for two sets of protein networks. The first set includes the known proteins that are associated with neurological disorders. The second network includes previously known neurological proteins as well as 30 best predicted putative neurological proteins from K-star results. Below shown are 20 most essential proteins that have been identified by MCODE in both the lists.

KNOWN	**KNOWN WITH PUTATIVE**
GPX1_HUMAN	HSP71_HUMAN
FRIL_HUMAN	FRIL_HUMAN
CP1B1_HUMAN	CRP_HUMAN
FRIH_HUMAN	CP1B1_HUMAN
CRP_HUMAN	XIAP_HUMAN
FYN_HUMAN	FYN_HUMAN
TF65_HUMAN	FRIH_HUMAN
LEP_HUMAN	TF65_HUMAN
SDF1_HUMAN	SDF1_HUMAN
CD4_HUMAN	LEP_HUMAN
SODM_HUMAN	ICAM1_HUMAN
ICAM1_HUMAN	SODM_HUMAN
MK01_HUMAN	MK01_HUMAN
IL18_HUMAN	NOS2_HUMAN
IL1B_HUMAN	IL18_HUMAN
SCRB2_HUMAN	IL1B_HUMAN
EGFR_HUMAN	ACE_HUMAN
GSTA4_HUMAN	SCRB2_HUMAN

Table 11: MCODE analysis results for essential proteins.

Several novel high probability essential proteins are found to be associated with neurological disorders, as can be seen in list two. The protein XIAP_HUMAN, seems to be of particular importance, due to its presence as both a hub protein and an essential protein. Studies have shown that most hub proteins are essential proteins, as they have maximum number of interacting partners, and are therefore thought to be an important candidate in cellular processes. Also, studies show that most hub proteins are encoded by essential genes. For the putative neurological candidate XIAP_HUMAN, it can therefore be said that it is an essential protein that is also a hub (Kar *et al*., 2009).

DAVID analysis (Refer Appendix Table 1) on MCODE results for the protein interaction list that includes K-star results also show that the protein XIAP_HUMAN is significantly involved in the pathways and cellular processes of the body. The pathway results were similar to earlier results from DAVID analysis of the list of hub proteins with novel candidates, thus validating MCODE results. Apart from XIAP_HUMAN, other putative candidate that is categorized as essential by MCODE is RNAS2_HUMAN. However, DAVID analysis results show that this protein is not associated with pathways of neurological disorders.

Overall, the results from this study give promising novel putative neurologically related candidates, their categorization as hubs and as essential proteins, and the pathways in which these novel proteins may be involved.

DISCUSSION AND FUTURE PERSPECTIVES

The present study describes a structural systems biology approach for gene prioritization of novel protein candidates involved in neurological protein interaction networks. The method used for gene prioritization is machine learning, and the features that are used to train the machine learning algorithms are structural descriptors of proteins. Therefore, only the proteins whose 3D structure is available on the Protein Data Bank are used for the study. The structural descriptors include protein network properties, and protein interface structural properties. Various studies have shown the protein network topological properties and interface structural properties are essential features for protein prioritization (Linghu B *et al*., 2009).

This study for the first time integrates these both features for gene prioritization of novel neurological candidates. Gene prioritization for neurological disorders is an important area of research. Here, we have focussed on epilepsy, and some neurological disorders that have clinical manifestations similar to epilepsy. These include Alzheimer's, Parkinson's, Bipolar disorder, Autism, Migraine and Schizophrenia. The frequency of occurrence of seizures is more in people with these diseases than in the normal population. Also, comorbid symptoms like depression, cognitive impairment, and anxiety, are common features of these neurological disorders. Certain genetic markers and pathways have been identified that are shared in these disorders, but still a lot of scope remains. There is a need to identify markers that are important in the network of these diseases. Also, previously unknown markers that might lead be involved in shared pathways of these neurological disorders needs to be characterized.

Five machine learning algorithms were used for protein prioritization. These included Naïve Bayes, rotation forest, random forest, bagging-j48 and K-star. Best accuracy, precision and recall was obtained in the results of the K-star algorithm, making its predictions the most reliable. First thirty best predicted protein candidates that were classified by K-star as being involved in neurological disease network, were henceforth used for further study and analysis.

Hub-Object-Analyser (HUBBA) was used to identify important hubs in the neurological disease network. These are the gene products that have highest number of interactions in the network, and are thought with maximum number of pathways in the network, as they have multiple interacting partners. These results could be used for further analysing the significance of these hub candidates in neurological disease network. Some of these proteins might be shared markers in more than one disorders. Such association studies can lead to identification of genetic markers that are cause of comorbid neurological conditions. It was also identified that the highest priority hubs identified by HUBBA, are shared gene products of diseases under study, and hence it can be concluded that hub proteins identified by HUBBA are actually essential shared markers in the protein network. The identified hubs from HUBBA, can therefore be relied upon, even for the prioritized novel candidate markers.

Hubs were also analysed for novel neurological candidates from the network that included interactions from previously known and newly predicted markers. This analysis describes the essentiality of novel candidates as hubs in the neurological network. A striking finding was of a new protein hub, whose uniprot Id is XIAP_HUMAN. This protein is previously not known to behave as neurological network hub. This protein product is the X-linked inhibitor of apoptosis. In previous studies, it has been shown to be bind to caspases -3,-7,-9 to inhibit their activity. It has been recently shown to be partially associated with the pathogenesis of multiple system atrophy (Kawamoto *et al*., 2014). Studies also show that inhibition of cell adhesion by XIAP is a promising treatment for neuroprotection, by inhibiting caspases in brain injury seizures (Li *et al*., 2006).

However, the direct involvement of XIAP in neurological disorders interaction network is a new development in the present work. Also the finding that it is a hub protein, increases its probability as a potent candidate. MCODE validates the results of HUBBA, by predicting that XIAP is indeed an essential protein that might be an important component of neurological disease pathways.

To study the involvement of novel candidates in the existing neurological interaction network, DAVID analysis of the hub protein list (that included previously known and newly identified markers) was performed. Of the two novel neurological hub proteins RNAS2_HUMAN and XIAP_HUMAN, the latter is associated with the existing neurological hub proteins pathways in several ways. This protein is found to exist in several pathways like that of cancer, apoptosis, NOD-like receptor signalling pathways, small cell lung cancer pathways, and pathways for focal adhesion; along with known hub proteins of neurological disorders. However, RNAS2_HUMAN is not involved in the KEGG pathways of known neurological hub proteins.

The predictions from this study can be further used for pathway analysis of neurological disorders. As a part of future analysis, all these diseases can be individually analysed for important hubs and essential proteins. The predictions from individual disease study can then be compared with the results of this study, which would help validating the results further. XIAP is an important putative neurological marker as predicted from this study. Further validation is however needed to verify these results. Since XIAP is a hub protein, analysis can be done on how the network is affected when this hub is removed from it. The interactions that are lost can be noted. It should also be verified, if the removal of these interactions disrupts the neurological pathways. This will help establish the essentiality of this hub. This protein is known to be involved in neuroprotection, and is involved in important signalling pathways as predicted from DAVID analysis. Keeping these results in mind it can be hypothesised that the removal of this hub protein from the interaction network, would cause inappropriate interactions leading to disrupted pathways, and increase an individual's susceptibility to comorbid symptoms in neurological disorders. That is, this might be a candidate marker that makes one person affected with a neurological disorder more susceptible to seizures or other comorbid conditions, than an individual from the normal population. Further validation in this respect, still needs to be done.

If this protein is validated to be a candidate for causing neurological symptoms, this might as well be used as a marker in personalised medicine. But before this, there would be a need for genome wide association studies, to know how mutation in XIAP increases or decreases the expression of associated genes. These studies can then be linked co-expression studies of such genes in multiple neurological disease patients.

Therefore, it can be concluded that this study has analysed neurological disease interaction network from a completely new perspective that takes into consideration the structural descriptors of the proteins involved in the network. We have identified a new candidate marker protein that has shown positive association with neurological disease network, which can be validated by further analysis.

REFERENCES

1. Bader GD, Hogue CW. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*.4:2.

2. Basse MJ, Betzi S, Bourgeas R, Bouzidi S, Chetrit B, Hamon V, Morelli X, Roche P. 2P2Idb: a structural database dedicated to orthosteric modulation of protein-protein interactions. *Nucleic Acids Res*. 2013 Jan; 41(Database issue):D824-7.

3. Batada NN, Hurst LD, Tyers M. Evolutionary and physiological importance of hub proteins. *PLoS Comput Biol*. 2006 Jul 14; 2(7):e88.

4. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res*. 2000 Jan 1; 28(1):235-42.

5. Bianchin MM, Londero RG, Lima JE, Bigal ME. Migraine and epilepsy: a focus on overlapping clinical, pathophysiological, molecular, and therapeutic aspects. *Curr Pain Headache Rep.* 2010 Aug; 14(4):276-83.

6. Bolton PF, Carcani-Rathwell I, Hutton J, Goode S, Howlin P, Rutter M. Epilepsy in autism: features and correlates. *Br J Psychiatry.* 2011 Apr; 198(4):289-94.

7. Bromberg Y. Chapter 15: disease gene prioritization. *PLoS Comput Biol.* 2013 Apr; 9(4):e1002902.

8. Chang HJ, Liao CC, Hu CJ, Shen WW, Chen TL. Psychiatric disorders after epilepsy diagnosis: a population-based retrospective cohort study. *PLoS One*. 2013; 8(4):e59999.

9. Chen J, Sawyer N, Regan L. Protein-protein interactions: general trends in the relationship between binding affinity and interfacial buried surface area. *Protein Sci.* 2013 Apr; 22(4):510-5.

10. Choudhary C, Kumar C, Gnad F, Nielsen ML, Rehman M, Walther TC, Olsen JV, Mann M. Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science.* 2009 Aug 14; 325(5942):834-40.

11. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc*. 2007; 2(10):2366-82.

12. Cukuroglu E, Gursoy A, Nussinov R, Keskin O. Non-redundant unique interface structures as templates for modeling protein interactions. *PLoS One*. 2014 Jan 27; 9(1):e86738.
13. Ettinger AB, Reed ML, Goldberg JF, Hirschfeld RM. Prevalence of bipolar symptoms in epilepsy vs other chronic health disorders. *Neurology*. Aug 23 2005; 65(4):535-40.
14. Glaab E, Bacardit J, Garibaldi JM, Krasnogor N. Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. *PLoS One.* 2012; 7(7):e39932.
15. Gonzalez MW, Kann MG. Chapter 4: Protein interactions and disease. PLoS *Comput Biol*. 2012; 8(12):e1002819.
16. Gursoy A, Keskin O, Nussinov R. Topological properties of protein interaction networks from a structural perspective. *Biochem Soc Trans*. 2008 Dec; 36(Pt 6):1398-403.
17. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc.* 2009*; 4(1):44-57***.**
18. Johnson ME, Hummer G. Interface-resolved network of protein-protein interactions. *PLoS Comput Biol*. 2013; 9(5):e1003065.
19. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*. 2006 Jan 1; 34(Database issue):D354-7.
20. Kanehisa, M., et al. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*. 34, D354 (2006).
21. Kanner AM, Barry JJ, Gilliam F, et al. Psychiatric Comorbidities in epilepsy: Streamlining Recognition and Diagnosis to Improve Quality of Life. *Counselling Points*. 2009; 1:1-15.
22. Kar G, Gursoy A, Keskin O. Human cancer protein-protein interaction network: a structural perspective. *PLoS Comput Biol.* 2009 Dec; 5(12):e1000601.
23. Kawamoto Y, Ito H, Ihara M, Takahashi R. XIAP immunoreactivity in glial and neuronal cytoplasmic inclusions in multiple system atrophy. *Clin Neuropathol*.2014 Jan-Feb; 33 (1):76-83.
24. Kendall-Taylor NH, Kathomi C, Rimba K, Newton CR. Comparing characteristics of epilepsy treatment providers on the Kenyan coast: implications for treatment-seeking and intervention. *Rural Remote Health.* 2009 Oct-Dec; 9(4):1253.

25. Kojovic M, Cordivari C, Bhatia K. Myoclonic disorders: a practical approach for diagnosis and treatment. *Ther Adv Neurol Disord*. 2011 Jan; 4(1):47-62.

26. Li T, Fan Y, Luo Y, Xiao B, Lu C. In vivo delivery of a XIAP (BIR3-RING) fusion protein containing the protein transduction domain protects against neuronal death induced by seizures. *Exp Neurol*. 2006 Feb; 197 (2):301-8.

27. Lin CY, Chin CH, Wu HH, Chen SH, Ho CW, Ko MT. Hubba: hub objects analyzer—a framework of interactome hubs identification for network biology. (2008). *Nucleic Acids Res.* 1, 36.

28. Linghu B, Snitkin ES, Hu Z, Xia Y, Delisi C. Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biol.* 2009; 10(9):R91.

29. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten; the WEKA Data Mining Software: An Update; *SIGKDD Explorations,* 2009.Volume 11, Issue 1.

30. Patil A, Kinoshita K, Nakamura H. Hub promiscuity in protein-protein interaction networks. *Int J Mol Sci.* 2010 Apr 26; 11(4):1930-43.

31. Patil A, Kinoshita K, Nakamura H. Hub promiscuity in protein-protein interaction networks. *Int J Mol Sci*. 2010 Apr 26; 11(4):1930-43.

32. Rodríguez JJ, Kuncheva LI, Alonso CJ. Rotation forest: A new classifier ensemble method. *IEEE Trans Pattern Anal Mach Intell*. 2006 Oct; 28(10):1619-30.

33. Schmitz EB, Robertson MM, Trimble MR. Depression and schizophrenia in epilepsy: social and biological risk factors. *Epilepsy Res*. May 1999; 35(1):59-68.

34. Sharan R, Ideker T. Modeling cellular machinery through biological network comparison. *Nat Biotechnol*. 2006 Apr; 24 (4):427-33.

35. Sridhar M, Babu R. Evaluating the Classification Accuracy of Data Mining Algorithms for Anonymized Data. *International Journal of Computer Science and Telecommunication.* 2012 Aug; Volume 3, Issue 8.

36. Szot P. Common factors among Alzheimer's disease, Parkinson's disease, and epilepsy: possible role of the noradrenergic nervous system. *Epilepsia.* 2012 Jun; 53.

37. Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, van Hijum SA. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Brief Bioinform.* 2013 May; 14(3):315-26.

38. Wall DP, Pivovarov R, Tong M, Jung JY, Fusaro VA, DeLuca TF, Tonellato PJ. Genotator: a disease-agnostic tool for genetic annotation of disease. *BMC Med Genomics*. 2010 Oct 29; 3:50.

39. Wang H, Zheng H, Simpson D, Azuaje F. Machine learning approaches to supporting the identification of photoreceptor-enriched genes based on expression data. *BMC Bioinformatics*. 2006 Mar 8; 7:116.

40. Wilbur WJ. Boosting naïve Bayesian learning on a large subset of MEDLINE. *Proc AMIA Symp.* 2000:918-22.

41. Zhang KX, Ouellette BF. CAERUS: predicting CAncER oUtcomeS using relationship between protein structural information, protein networks, gene expression data, and mutation data. *PLoS Comput Biol*. 2011 Mar; 7(3):e1001114.

APPENDIX

Table 1:

DAVID analysis on MCODE results for the protein interaction list that includes K-star results also show that the protein XIAP_HUMAN is significantly involved in the pathways and cellular processes of the body.

KEGG Pathways	Genes	PValue
Pathways in cancer	BCL2_HUMAN, CADH1_HUMAN, IL6_HUMAN, PK3CA_HUMAN, XIAP_HUMAN, CYC_HUMAN, P53_HUMAN, GRB2_HUMAN, HGF_HUMAN, BAX_HUMAN, MMP2_HUMAN, EGFR_HUMAN, PPARG_HUMAN, AKT1_HUMAN, MET_HUMAN, IL8_HUMAN, MMP9_HUMAN, MMP1_HUMAN, FGFR2_HUMAN, TF65_HUMAN, CASP8_HUMAN, PK3CG_HUMAN, MK01_HUMAN, IGF1R_HUMAN, TNFL6_HUMAN, NOS2_HUMAN, PDGFB_HUMAN, MK03_HUMAN, RAC2_HUMAN, EGF_HUMAN, RASH_HUMAN, TGFB1_HUMAN, FGF1_HUMAN, FINC_HUMAN	3.96E-16
Melanoma	CADH1_HUMAN, PK3CA_HUMAN, P53_HUMAN, PK3CG_HUMAN, MK01_HUMAN, HGF_HUMAN, IGF1R_HUMAN, EGFR_HUMAN, PDGFB_HUMAN, MK03_HUMAN, AKT1_HUMAN, MET_HUMAN, RASH_HUMAN, EGF_HUMAN, FGF1_HUMAN	4.26E-11
Prostate cancer	BCL2_HUMAN, FGFR2_HUMAN, PK3CA_HUMAN, P53_HUMAN, TF65_HUMAN, PK3CG_HUMAN, MK01_HUMAN, GRB2_HUMAN, IGF1R_HUMAN, EGFR_HUMAN, PDGFB_HUMAN, MK03_HUMAN, AKT1_HUMAN, RASH_HUMAN, EGF_HUMAN, CREB1_HUMAN	8.72E-11
Bladder cancer	MMP1_HUMAN, CADH1_HUMAN, TSP1_HUMAN, MK03_HUMAN, P53_HUMAN, RASH_HUMAN, EGF_HUMAN, IL8_HUMAN, MK01_HUMAN, MMP9_HUMAN, MMP2_HUMAN, EGFR_HUMAN	2.25E-10
Colorectal cancer	BCL2_HUMAN, PK3CA_HUMAN, CYC_HUMAN, P53_HUMAN, PK3CG_HUMAN, GRB2_HUMAN, MK01_HUMAN, BAX_HUMAN, IGF1R_HUMAN, EGFR_HUMAN, MK03_HUMAN, RAC2_HUMAN, AKT1_HUMAN, MET_HUMAN, TGFB1_HUMAN	4.57E-10
Apoptosis	BCL2_HUMAN, PK3CA_HUMAN, CYC_HUMAN, P53_HUMAN, XIAP_HUMAN, TF65_HUMAN, CASP8_HUMAN, PK3CG_HUMAN, BAX_HUMAN, TNFL6_HUMAN, TNFA_HUMAN, AKT1_HUMAN, NGF_HUMAN, IL1B_HUMAN, TNR1A_HUMAN	7.43E-10
MAPK signaling pathway	PA24A_HUMAN, FGFR2_HUMAN, TF65_HUMAN, P53_HUMAN, HSP71_HUMAN, BDNF_HUMAN, MK01_HUMAN, GRB2_HUMAN, EGFR_HUMAN, TNFL6_HUMAN, TNFA_HUMAN, MK03_HUMAN, PDGFB_HUMAN, RAC2_HUMAN, AKT1_HUMAN, RASH_HUMAN, EGF_HUMAN, NGF_HUMAN, IL1B_HUMAN, TNR1A_HUMAN, TGFB1_HUMAN, FGF1_HUMAN, PA21B_HUMAN, MK14_HUMAN	7.56E-10
Neurotrophin signaling pathway	BCL2_HUMAN, PK3CA_HUMAN, P53_HUMAN, TF65_HUMAN, IRS1_HUMAN, BDNF_HUMAN, PK3CG_HUMAN, MK01_HUMAN, GRB2_HUMAN, BAX_HUMAN, TNFL6_HUMAN, MK03_HUMAN, AKT1_HUMAN, CALM_HUMAN, RASH_HUMAN, NGF_HUMAN, MK14_HUMAN	1.22E-09
Glioma	PK3CA_HUMAN, P53_HUMAN, PK3CG_HUMAN, GRB2_HUMAN, MK01_HUMAN, IGF1R_HUMAN, EGFR_HUMAN, PDGFB_HUMAN, MK03_HUMAN, AKT1_HUMAN, CALM_HUMAN, EGF_HUMAN, RASH_HUMAN	1.77E-09
Fc epsilon RI signaling pathway	PA24A_HUMAN, IL4_HUMAN, PK3CA_HUMAN, FYN_HUMAN, PK3CG_HUMAN, GRB2_HUMAN, MK01_HUMAN, TNFA_HUMAN, MK03_HUMAN, RAC2_HUMAN, AKT1_HUMAN, RASH_HUMAN, PA21B_HUMAN, MK14_HUMAN	2.02E-09
Cytokine-cytokine receptor interaction	CCL2_HUMAN, IL18_HUMAN, IL4_HUMAN, IL6_HUMAN, PRL_HUMAN, IFNG_HUMAN, CCR5_HUMAN, SDF1_HUMAN, HGF_HUMAN, EGFR_HUMAN, TNFL6_HUMAN, TNFA_HUMAN, IPDGFB_HUMAN, MET_HUMAN, LIF_HUMAN, LEP_HUMAN, IL1B_HUMAN, EGF_HUMAN, IL8_HUMAN, TNR1A_HUMAN, IL6RA_HUMAN, TGFB1_HUMAN, IL2RA_HUMAN	3.09E-09
Focal adhesion	BCL2_HUMAN, PK3CA_HUMAN, TSP1_HUMAN, FYN_HUMAN, XIAP_HUMAN, PK3CG_HUMAN, MK01_HUMAN, GRB2_HUMAN, HGF_HUMAN, IGF1R_HUMAN, EGFR_HUMAN, PDGFB_HUMAN, MK03_HUMAN, SRC_HUMAN, RAC2_HUMAN, AKT1_HUMAN, MET_HUMAN, RASH_HUMAN, EGF_HUMAN, FINC_HUMAN	6.22E-09
NOD-like receptor	IL18_HUMAN, TNFA_HUMAN, CCL2_HUMAN, MK03_HUMAN, IL6_HUMAN, TF65_HUMAN, XIAP_HUMAN, CASP8_HUMAN, IL1B_HUMAN, IL8_HUMAN, MK01_HUMAN, MK14_HUMAN	1.94E-08

Endometrial cancer	CADH1_HUMAN, PK3CA_HUMAN, MK03_HUMAN, P53_HUMAN, AKT1_HUMAN, RASH_HUMAN, EGF_HUMAN, PK3CG_HUMAN, GRB2_HUMAN, MK01_HUMAN, EGFR_HUMAN	3.97E-08
T cell receptor signaling pathway	IL4_HUMAN, PK3CA_HUMAN, CD4_HUMAN, FYN_HUMAN, TF65_HUMAN, IFNG_HUMAN, PK3CG_HUMAN, GRB2_HUMAN, MK01_HUMAN, TNFA_HUMAN, MK03_HUMAN, AKT1_HUMAN, RASH_HUMAN, MK14_HUMAN	1.17E-07
VEGF signaling pathway	PA24A_HUMAN, NOS3_HUMAN, PK3CA_HUMAN, MK03_HUMAN, RAC2_HUMAN, SRC_HUMAN, AKT1_HUMAN, RASH_HUMAN, PK3CG_HUMAN, MK01_HUMAN, PA21B_HUMAN, MK14_HUMAN	1.52E-07
Prion diseases	MK03_HUMAN, IL6_HUMAN, FYN_HUMAN, NOTC1_HUMAN, HSP71_HUMAN, IL1B_HUMAN, SODC_HUMAN, MK01_HUMAN, BAX_HUMAN	2.37E-07
Toll-like receptor signaling pathway	PK3CA_HUMAN, TF65_HUMAN,IL6_HUMAN, CASP8_HUMAN, PK3CG_HUMAN, MK01_HUMAN, TLR4_HUMAN, TNFA_HUMAN, MK03_HUMAN, , AKT1_HUMAN, IL1B_HUMAN, IL8_HUMAN, MK14_HUMAN	4.30E-07
Non-small cell lung cancer	PK3CA_HUMAN, MK03_HUMAN, P53_HUMAN, AKT1_HUMAN, RASH_HUMAN, EGF_HUMAN, PK3CG_HUMAN, GRB2_HUMAN, MK01_HUMAN, EGFR_HUMAN	7.18E-07
Renal cell carcinoma	PK3CA_HUMAN, MK03_HUMAN, PDGFB_HUMAN, AKT1_HUMAN, MET_HUMAN, RASH_HUMAN, PK3CG_HUMAN, GRB2_HUMAN, MK01_HUMAN, HGF_HUMAN, TGFB1_HUMAN	7.46E-07
Pancreatic cancer	PK3CA_HUMAN, MK03_HUMAN, RAC2_HUMAN, TF65_HUMAN, P53_HUMAN, AKT1_HUMAN, EGF_HUMAN, PK3CG_HUMAN, MK01_HUMAN, TGFB1_HUMAN, EGFR_HUMAN	9.76E-07
Amyotrophic lateral sclerosis (ALS)	BCL2_HUMAN, TNFA_HUMAN, GPX1_HUMAN, CYC_HUMAN, P53_HUMAN, SODC_HUMAN, TNR1A_HUMAN, BAX_HUMAN, MK14_HUMAN	6.68E-06
Chronic myeloid leukemia	PK3CA_HUMAN, MK03_HUMAN, TF65_HUMAN, P53_HUMAN, AKT1_HUMAN, RASH_HUMAN, PK3CG_HUMAN, GRB2_HUMAN, MK01_HUMAN, TGFB1_HUMAN	1.21E-05
Adherens junction	CADH1_HUMAN, MK03_HUMAN, FYN_HUMAN, INSR_HUMAN, RAC2_HUMAN, SRC_HUMAN, MET_HUMAN, MK01_HUMAN, IGF1R_HUMAN, EGFR_HUMAN	1.50E-05
GnRH signaling pathway	PA24A_HUMAN, MK03_HUMAN, SRC_HUMAN, CALM_HUMAN, RASH_HUMAN, GRB2_HUMAN, MK01_HUMAN, MMP2_HUMAN, PA21B_HUMAN, MK14_HUMAN, EGFR_HUMAN	1.68E-05
Small cell lung cancer	BCL2_HUMAN, NOS2_HUMAN, PK3CA_HUMAN, TF65_HUMAN, CYC_HUMAN, XIAP_HUMAN, P53_HUMAN, AKT1_HUMAN, PK3CG_HUMAN, FINC_HUMAN	3.07E-05
ErbB signaling pathway	PK3CA_HUMAN, MK03_HUMAN, SRC_HUMAN, AKT1_HUMAN, RASH_HUMAN, EGF_HUMAN, PK3CG_HUMAN, GRB2_HUMAN, MK01_HUMAN, EGFR_HUMAN	4.07E-05
Natural killer cell mediated cytotoxicity	TNFA_HUMAN, PK3CA_HUMAN, MK03_HUMAN, FYN_HUMAN, ICAM1_HUMAN, RAC2_HUMAN, IFNG_HUMAN, RASH_HUMAN, PK3CG_HUMAN, GRB2_HUMAN, MK01_HUMAN, TNFL6_HUMAN	4.59E-05
B cell receptor signaling pathway	PK3CA_HUMAN, MK03_HUMAN, RAC2_HUMAN, TF65_HUMAN, AKT1_HUMAN, RASH_HUMAN, PK3CG_HUMAN, GRB2_HUMAN, MK01_HUMAN	8.94E-05
Acute myeloid leukemia	PK3CA_HUMAN, MK03_HUMAN, TF65_HUMAN, AKT1_HUMAN, RASH_HUMAN, PK3CG_HUMAN, GRB2_HUMAN, MK01_HUMAN	1.14E-04
Jak-STAT signaling pathway	IL4_HUMAN, PK3CA_HUMAN, PRL_HUMAN, IL6_HUMAN, IFNG_HUMAN, AKT1_HUMAN, LIF_HUMAN, PK3CG_HUMAN, LEP_HUMAN, GRB2_HUMAN, IL6RA_HUMAN, IL2RA_HUMAN	1.84E-04
Thyroid cancer	CADH1_HUMAN, MK03_HUMAN, P53_HUMAN, PPARG_HUMAN, RASH_HUMAN, MK01_HUMAN	2.07E-04
Chemokine signaling pathway	CCL2_HUMAN, PK3CA_HUMAN, TF65_HUMAN, PK3CG_HUMAN, SDF1_HUMAN, CCR5_HUMAN, GRB2_HUMAN, MK01_HUMAN, MK03_HUMAN, RAC2_HUMAN, AKT1_HUMAN, RASH_HUMAN, IL8_HUMAN	2.41E-04
Type II diabetes mellitus	TNFA_HUMAN, PK3CA_HUMAN, MK03_HUMAN, INSR_HUMAN, IRS1_HUMAN, PK3CG_HUMAN, MK01_HUMAN	2.63E-04
Regulation of actin cytoskeleton	FGFR2_HUMAN, PK3CA_HUMAN, PK3CG_HUMAN, MK01_HUMAN, EGFR_HUMAN, PDGFB_HUMAN, MK03_HUMAN, RAC2_HUMAN, EGF_HUMAN, RASH_HUMAN, FGF1_HUMAN, THRB_HUMAN, FINC_HUMAN	8.57E-04
Metabolism of xenobiotics by cytochrome P450	GSTA1_HUMAN, CP2B6_HUMAN, CP3A4_HUMAN, GSTA3_HUMAN, GSTA4_HUMAN, CP2CJ_HUMAN, CP1B1_HUMAN	0.001001456
Aldosterone-regulated	PK3CA_HUMAN, MK03_HUMAN, INSR_HUMAN, IRS1_HUMAN, PK3CG_HUMAN, MK01_HUMAN	0.001086577

Insulin signaling pathway	PK3CA_HUMAN, MK03_HUMAN, INSR_HUMAN, IRS1_HUMAN, AKT1_HUMAN, CALM_HUMAN, RASH_HUMAN, PK3CG_HUMAN, GRB2_HUMAN, MK01_HUMAN	0.001147193
Drug metabolism	CP2A6_HUMAN, GSTA1_HUMAN, CP2B6_HUMAN, CP3A4_HUMAN, GSTA3_HUMAN, GSTA4_HUMAN, CP2CJ_HUMAN	0.001191817
Dorso-ventral axis formation	MK03_HUMAN, NOTC1_HUMAN, GRB2_HUMAN, MK01_HUMAN, EGFR_HUMAN	0.001249017
Progesterone-mediated oocyte maturation	PK3CA_HUMAN, MK03_HUMAN, AKT1_HUMAN, PK3CG_HUMAN, PRGR_HUMAN, MK01_HUMAN, IGF1R_HUMAN, MK14_HUMAN	0.001309941
Gap junction	MK03_HUMAN, PDGFB_HUMAN, SRC_HUMAN, RASH_HUMAN, EGF_HUMAN, GRB2_HUMAN, MK01_HUMAN, EGFR_HUMAN	0.001601903
Adipocytokine signaling pathway	TNFA_HUMAN, TF65_HUMAN, IRS1_HUMAN, AKT1_HUMAN, PPARA_HUMAN, LEP_HUMAN, TNR1A_HUMAN	0.001789255
Leukocyte transendothelial migration	PK3CA_HUMAN, ICAM1_HUMAN, RAC2_HUMAN, PK3CG_HUMAN, VCAM1_HUMAN, SDF1_HUMAN, MMP9_HUMAN, MMP2_HUMAN, MK14_HUMAN	0.001950913
Alzheimer's disease	TNFA_HUMAN, MK03_HUMAN, APOE_HUMAN, CYC_HUMAN, CALM_HUMAN, CASP8_HUMAN, IL1B_HUMAN, TNR1A_HUMAN, MK01_HUMAN, A4_HUMAN	0.004189101
Hematopoietic cell lineage	TNFA_HUMAN, IL4_HUMAN, CD4_HUMAN, IL6_HUMAN, IL1B_HUMAN, IL6RA_HUMAN, IL2RA_HUMAN	0.006268597
Graft-versus-host disease	TNFA_HUMAN, IL6_HUMAN, IFNG_HUMAN, IL1B_HUMAN, TNFL6_HUMAN	0.006595809
Huntington's disease	SODM_HUMAN, GPX1_HUMAN, CYC_HUMAN, P53_HUMAN, PPARG_HUMAN, BDNF_HUMAN, CASP8_HUMAN, SODC_HUMAN, CREB1_HUMAN, BAX_HUMAN	0.007964422
Endocytosis	FGFR2_HUMAN, SRC_HUMAN, MET_HUMAN, HSP71_HUMAN, RASH_HUMAN, EGF_HUMAN, CCR5_HUMAN, IGF1R_HUMAN, IL2RA_HUMAN, EGFR_HUMAN	0.009146435
Epithelial cell signaling in Helicobacter pylori infection	SRC_HUMAN, TF65_HUMAN, MET_HUMAN, IL8_HUMAN, MK14_HUMAN, EGFR_HUMAN	0.010001965
p53 signaling pathway	TSP1_HUMAN, CYC_HUMAN, P53_HUMAN, CASP8_HUMAN, BAX_HUMAN, PAI1_HUMAN	0.010001965
Fc gamma R-mediated phagocytosis	PA24A_HUMAN, PK3CA_HUMAN, MK03_HUMAN, RAC2_HUMAN, AKT1_HUMAN, PK3CG_HUMAN, MK01_HUMAN	0.010076393
Long-term depression	PA24A_HUMAN, MK03_HUMAN, RASH_HUMAN, MK01_HUMAN, IGF1R_HUMAN, PA21B_HUMAN	0.010620619
Linoleic acid metabolism	PA24A_HUMAN, CP3A4_HUMAN, CP2CJ_HUMAN, PA21B_HUMAN	0.016481296
mTOR signaling pathway	PK3CA_HUMAN, MK03_HUMAN, AKT1_HUMAN, PK3CG_HUMAN, MK01_HUMAN	0.017922723
Arachidonic acid metabolism	PA24A_HUMAN, GPX1_HUMAN, CP2B6_HUMAN, CP2CJ_HUMAN, PA21B_HUMAN	0.022929815
TGF-beta signaling pathway	TNFA_HUMAN, TSP1_HUMAN, MK03_HUMAN, IFNG_HUMAN, MK01_HUMAN, TGFB1_HUMAN	0.026585881
Allograft rejection	TNFA_HUMAN, IL4_HUMAN, IFNG_HUMAN, TNFL6_HUMAN	0.032135484
Axon guidance	MK03_HUMAN, FYN_HUMAN, RAC2_HUMAN, MET_HUMAN, RASH_HUMAN, SDF1_HUMAN, MK01_HUMAN	0.039032459
Melanogenesis	MK03_HUMAN, CALM_HUMAN, RASH_HUMAN, MK01_HUMAN, CREB1_HUMAN, EDN1_HUMAN	0.04296181
Type I diabetes mellitus	TNFA_HUMAN, IFNG_HUMAN, IL1B_HUMAN, TNFL6_HUMAN	0.047555316
Viral myocarditis	FYN_HUMAN, ICAM1_HUMAN, RAC2_HUMAN, CYC_HUMAN, CASP8_HUMAN	0.048742634
RIG-I-like receptor	TNFA_HUMAN, TF65_HUMAN, CASP8_HUMAN, IL8_HUMAN, MK14_HUMAN	0.048742634
Intestinal immune network	IL4_HUMAN, IL6_HUMAN, SDF1_HUMAN, TGFB1_HUMAN	0.069322726

Glutathione metabolism	GPX1_HUMAN, GSTA1_HUMAN, GSTA3_HUMAN, GSTA4_HUMAN	0.072745371
Retinol metabolism	CP2A6_HUMAN, CP2B6_HUMAN, CP3A4_HUMAN, CP2CJ_HUMAN	0.087169089
Cytosolic DNA-sensing pathway	IL18_HUMAN, IL6_HUMAN, TF65_HUMAN, IL1B_HUMAN	0.090951556