

# Protein Network for Associating Genes with Dementia

Brijendra Gupta

Research Scholar

Department of computer engg.  
I.I.T,(BHU),Varanasi

Ravi Bhushan Mishra

Professor

Department of computer engg.  
I.I.T,(BHU),Varanasi

## ABSTRACT

Association between causal genes and their genetic diseases is an important problem concerning human health. Linkage analysis is such a method that can identify which unknown disease genes are located in chromosomal region out of hundreds of candidate genes according to their functions, interactions, and pathways which is good identification of genes associated with general/hereditary disorders.

Here, we used method for prioritization of candidate genes of Dementia by the use of a global network distance measure, Random Walk Analysis, which detects neurological disorder been associated with distribution of sub-network among the genes.

## Keywords

Genetic diseases; Dementia; Random Walk Analysis; Neurological disorder .

## 1. INTRODUCTION

Advancements in genomic technologies have drastically boosted research on genetic Associations studies with the disease and summaries of published gene–disease associations are acknowledged either by disease or by gene, and are helpful in the study for association with the disease.

The molecular basis of causal genes is important to discover their genetic diseases. If a single nucleotide polymorphism [SNPs] in the disease is associated with a linkage interval on the chromosome are correlated with an increased susceptibility to the disease. Utility of Protein-Protein interactions to determine gene–disease associations, defines a set of candidate genes causing disease. Protein products related to the disease are also known to have same Genes that physically interact.

Several techniques uncovers gene–disease associations taking an integrative approach, leveraging Gene Ontology annotations [1-6], genes expression [7-8],protein sequences [9], biological pathways [4], Bio-text mining [10-11], and transcription factor binding sites [4] and several phenotypic traits of diseases.

Current efforts for disease-gene identification specifies candidate-gene prioritization that have been developed through computational approaches based on gene-expression data, functional annotation, or sequence-based features.[12-13]. A class of computational approaches has been proposed that reveals these two sources of information- linkage intervals and physical interaction networks to predict associated genes with diseases.

We rank all genes in all intervals instead of ranking only genes in the subinterval, related to a query disease. This allows us to find disease-causing genes present in existing disease intervals which were previously not included with the disease. We also quantified topological distribution of disease-related Proteins effects for prediction quality and are thus able to identify diseases with network-based predictions

and others for which additional information are importantly required.

Biological networks is applied to the problem of candidate-gene prioritization that includes the search for direct neighbors of other diseased genes [14]. A biological process is made up of a group of genes. If a biological process of large gene lists is abnormal in a given survey, the co-functioning genes having a higher [enriched] potential is selected as a relevant group by the high-throughput screening technologies and makes the analysis group-based view to a relevant gene instead of an individual gene-oriented analysis. Statistical methods, including Fisher's exact test, Chi-square, Hypergeometric distribution and Binomial probability lists many genes that yields a higher sensitivity [more significant P-values] and slightly enriched as well as to more specific terms. Proteins do not necessarily get into a stable physical interaction to have a specific, functional interplay: they can support some subsequent reactions in a metabolic pathway, regulating each other for transcription or post-transcription, or jointly contributing to large, structural assemblies without making direct contact. Together with physical interactions or indirect interactions constitutes the superset of 'functional protein–protein interaction' [15, 16].

Genome-wide genetic screens have emerged for browsing genes of interest networks, inspecting evidence of interaction or clustering a variety of large-scale functional association networks. Protein network information may help to connect potential hits that can serve to provide confidence, particularly if a number of hits can be observed in a densely graphed functional network. [ii] Protein network information can help in the interpretation of genomics data,[iii] Protein association networks are also surprisingly useful for the detection of disease genes, both for Mendelian and for complex diseases [17–19].

Here we focus on isolating protein–protein interaction [PPI] networks and linkage intervals to determine information which is readily extractable from them to predict gene–disease associations.

## 2. CHALLENGES

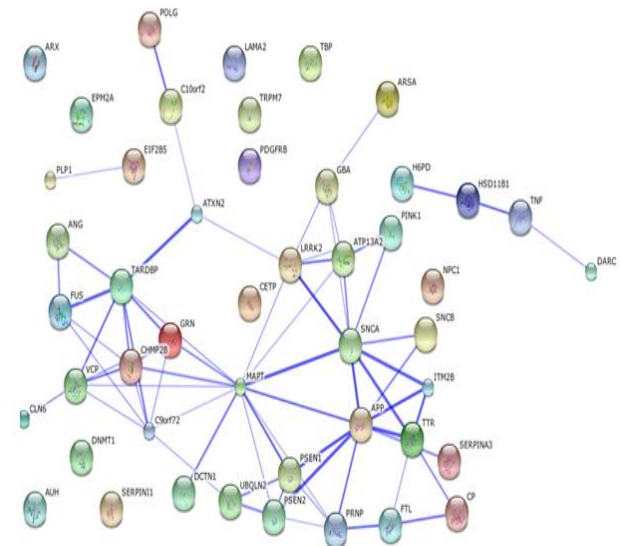
Computational methods yield several views about the utility of PPI networks to uncover gene–disease associations. But, it has several limitations:

Inspection of the list of genes is often not possible to identify the correct disease gene and the results of different computational methods for disease-gene prediction or prioritization is critical when compared. Genes in the same metabolic pathway have statistically significant genomic clusters as compared to genes chosen randomly [20]. Moreover, proteins coded by genes around disease genes in the contiguous intervals are located closely to the corresponding disease-gene family members within the PPI network than the proteins coded by randomly chosen genes.

There are no real-time databases to focus on rare disease-causing gene variants. The functional connectivity within a proteome, difficulty in remembering and comparing the algorithms and their features in a tool-by-tool manner over large number of tools and deciding, which enrichment tools among so many overwhelming choices, are most suitable for their analytic needs is a much more challenging problem. Real-time positioning of enrichment P-values, high-throughput enrichment in the data-mining environment, strong connectivity/dependency among genes, high/low frequency of annotation content and unbalanced distribution are complex. Indirect or transient types of protein interactions, gene duplications, gene re-arrangements, gene losses and assigning pairs of functionally equivalent genes is less clear..

### 3. METHODS

Associating the relationship between the quality of predictions for a disease and the biological distribution of its related proteins in the network, we obtain better results for diseases whose Proteins are located near one another in the network. We constructed a PPI network from the NCBI to observe the current performance of the network itself within genes for determining gene–disease association. Since large numbers of disease loci is not performed, therefore, We have defined artificial linkage intervals containing other genes around each of the disease genes being tested in order to simulate the loci. Diseases were associated with genes and amino acid linkage using the Online Mendelian Inheritance in Man, NCBI [21]. Online Mendelian Inheritance in Man [OMIM] database, [22] are used for genetically heterogeneous disorders, where mutations in different genes are associated with similar or even non-variant phenotypes. Diseases associated with only one gene were excluded in order to facilitate gene networks relating amino acid disorders for traits. Firstly, 48 genes were plotted ( Fig.1) and run on STRING (Fig.2)

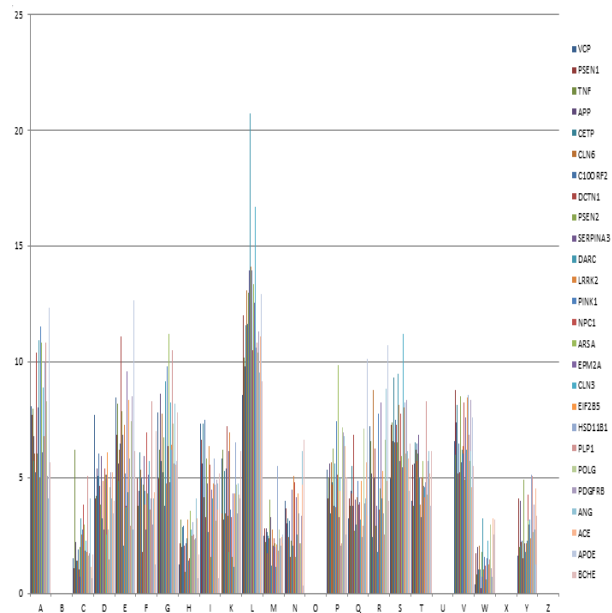


**Fig 2. 48 Genes(OMIM) graph plot against molecular pathway with functional role of Amino Acid content**

Moving ahead to relate Gene–Gene interaction in respect to Protein activities, we run the Enrichnet tools which 406 Protein activities according to gene annotation (pathway/process), but we considered 3 Genes pairing activities which are( Table 1) listed as:

**Table 2.Three Gene pairing activities with enriched Amino Acid content**

GENE PAIRING ACTIVITIES	GENE1	GENE2	GENE3	AMINO ACID RICHNESS IN GENE1	AMINO ACID RICHNESS IN GENE2	AMINO ACID RICHNESS IN GENE3
TUBULIN BINDING	MAPT	PRNP	LRRK2	Q	P	L
BETA-AMYLOID BINDING	APOE	ITIH2B	SCHE	L	L	L
ENDOPEPTIDASE ACTIVITY	PSEN1	PSEN2	ACE	L	L	L
PROTEASE BINDING	TP53	TTN	SLAVE	T	A	L
LIPID BINDING	VCP	APOE	CETP	L	L	L
HEPARIN BINDING	APP	APOE	ANG	L	L	L
KINASE ACTIVITY	PDOPRS	PINK1	TRPM7	L	L	L
ACTIN BINDING	ACE	ANG	TRPM7	L	L	L
BINDING	USQUN2	LRRK2	SP255	S	L	L
PROTEIN SERINE/THREONINE KINASE ACTIVITY	PINK1	LRRK2	TRPM7	L	L	L
PROTEIN TYROSINE KINASE ACTIVITY	PDOPRS	PINK1	LRRK2	L	L	L
PROTEIN KINASE ACTIVITY	PDOPRS	PINK1	LRRK2	L	L	L
RECEPTOR ACTIVITY	NPC1	DARC	TRPM7	L	L	L
PROTEIN HOMODIMERIZATION ACTIVITY	APOE	LRRK2	CLU6	L	L	L
CALCIUM ION BINDING	SNCA	SNCS	ARSA	A	Q	L
NUCLEIC ACID BINDING	FUS	TARDBP	FOLD	V	S	L
SEQUENCE-SPECIFIC DNA BINDING	TBP	ARX	TARDBP	Q	A	S
TRANSCRIPTION FACTOR ACTIVITY						



**Fig 1. 48 Genes(OMIM) plot against with percentage of Amino Acid content**

And the role of the enriched Amino acids for Neuropsychiatric Diseases are as:

Glycin is an inhibitory neurotransmitter in nervous system , spinal cord, brainstem and retina. It is required with glutamate for neuronal receptors activities. But, over presence has poor performance on glutaminergic activities, and may cause death due to hyperexcitability.

Proline and its derivatives are used as a catalysts in organic reactions. L- Proline are used in pharmaceutical and biotechnological applications. But, Proline is rarely found in  $\alpha$  and  $\beta$  structures and may reduce the stability of such structures.

Threonine is firstly, converted to pyruvate and undergoes metabolism and produce glycine and then converted into  $\alpha$ -ketobutyrate and enters the neurological activities pathways.

Alanine plays an important role in glucose-alanine cycle between tissues and liver. This amino acid is degraded for fuel, and has a reverse cycle for glucose formation. But, high level of alanine creates high blood pressure, cholesterol levels and links to diabetes due to increase in alanine aminotransferase.

Serine is an important participant in biosynthesis of purines and pyrimidines. It also has catalytic function of many enzymes. It is one of the three amino acid residues that are commonly phosphorylated by enzymes during cell signaling. D-serine is synthesized in brain from L- serine and serves as a neurotransmitter and binds glutamate.

Valine is an essential amino acid, and is synthesized in several steps starting from pyruvic acid. It also undergoes reductive animation with glutamate.

Glutamine is found to be useful in treatments of injuries , wound healing and cancer treatment.it is also used for muscle growth. L- glutamine is commonly added in cell-culture. But, high level of glutamine may inhibit other amino acid transport activities.

Now, in respect to the gene network out of the selected one, the mutation of the Genes persists the abnormal neurological activities which are ( Table 2) as follows:

**Table 2. Genes Mutation effects with enriched Amino Acid content**

Genes	Cytogenic Location	Mutation effects	Amino Acid Richness	Amino Acid Count	Hydrophobic value
LRRK2	12q12	Parkinson disease	L	356	1352.8
PINK1	1p36		L	81	307.8
TARDBP	1p36.2	amyotrophic lateral sclerosis, frontotemporal dementia	S	41	-32.8
VCP	9p13-p12	amyopathy with early-onset Paget disease, frontotemporal dementia	L	69	262.2
PSEN2	1q31-q42	Alzheimer disease	L	38	220.4
PRNP	20p13	Huntington disease-like syndrome, prion disease	G	45	-18
PDGFRB	5q31-q32	familial idiopathic basal ganglia calcification	L	125	475
SNCA	4q22.1	Parkinson disease, multiple system atrophy, dementia	A	18	32.4
MAPT	17q21.1	frontotemporal dementia, frontal and temporal lobes, progressive supranuclear palsy	P	97	-155.2
APOE	19q13.2	Alzheimer disease, amyloid plaques	L	41	155.8
APP	21q21.3	Alzheimer disease, dementia, hereditary cerebral amyloid angiopathy	L	49	186.2
PSEN1	14q24.3	Alzheimer disease	L	36	212.8
FYN	6q21	dementia	K	100	-390

Further, In the above table, the genes have major roles in brain function,and their mutations follow several neurological disorders.--

“LRRK2 genes” leucine-rich repeat kinase 2, gives instructions for making a protein called dardarin. dardarin protein is called a leucine-rich region, play a role such as

transmitting signals or helping to assemble the cell's structural framework [cytoskeleton] with other proteins. dardarin has an enzyme function known as kinase activity.--

PINK1 gene gives instructions for making a protein called PTEN induced putative kinase 1. It helps to protect mitochondria from malfunctioning during cellular stress. PINK1 gene mutations change or delete the kinase domain, leading to a protein functional loss. At least one mutation has the mitochondrial-targeting motif and may interrupt delivery of the protein to mitochondria. It will cause the selective death of nerve cells that distinguish Parkinson disease. The loss of these cells weakens signalling between the brain and muscles, and ultimately the brain is unable to control muscle movement.--

“TARDBP gene” TAR DNA binding protein, binds to DNA, regulates transcription, involves in processing molecules called messenger RNA [mRNA], regulates protein production and involves in neuronal and organ development.--

VCP gene ,valosin containing protein, makes an enzyme which has a wide variety of functions within cells, prevents the self-destruction of cells [apoptosis], and repaires damaged DNA.--

PSEN2 gene, presenilin 2, helps to process proteins to transmit various chemical signals from the cell membrane into the nucleus for structural framework and wide variety of functions within cells. This protein is found throughout the nervous system.--

PRNP gene, prion protein, active in the brain and several other tissues. It helps in transporting charged copper atoms [copper ions] into cells. mutations in the PRNP gene causes symptoms of Huntington disease, uncontrolled movements, emotional problems, loss of thinking ability, mental and behavioral features of prion diseases.--

PDGFRB gene, platelet-derived growth factor receptor, beta polypeptide, is important in processing the cell such as growth and division, movement, and survival. Mutations in the PDGFRB gene causes familial idiopathic basal ganglia calcification [FIBGC], i.e. abnormal deposits of calcium [calcification] in the brain, leading to movement and psychiatric problems.--

SNCA gene, synuclein, alpha [non A4 component of amyloid precursor], found at the tips of nerve cells [neurons], release chemical messengers, called neurotransmitters between neurons and is critical for normal brain function. Misfolded or excess alpha-synuclein proteins clusters together [aggregate] and impairs the function of these neurons in specific regions of the brain, interrupt the regulation of dopamine, kills neurons, weakens communication between the brain and muscles, loss of intellectual functions [dementia] and interrupts the function of the autonomic nervous system.--

MAPT gene helps in making a protein, tau. This tau is found throughout the nervous system and nerve cells [neurons] in the brain. It is involved in assembling and normalising microtubules, structuring cellular framework [the cytoskeleton]. mutations in the MAPT gene causes frontotemporal dementia, reduces tau's ability to bind to microtubules, accumulates abnormal tau in neurons and other brain cells, gradual death of cells in areas of the brain [frontal and temporal lobes] creating problems in reasoning, planning, judgment, problem-solving, movement, vision, speech, and thinking [cognition].--

APOE gene, apolipoprotein E, is a protein which combines with fats [lipids] in the body to form lipoproteins. variations of the APOE gene has a chance of developing memory loss, amyloid plaques and death of neurons.--

APP gene, amyloid precursor protein, is a protein found in many tissues and organs, including the brain and spinal cord [central nervous system].It has an important role in the formation of nerve cells [neurons] in the brain both before and after birth. Its mutation replaces the amino acid valine with the amino acid isoleucine at protein position 717 [written as Val717Ile or V717I] resulting Alzheimer disease.--

[Glu22Gln or E22Q] and [Leu34Val or L34V] cause hereditary cerebral amyloid angiopathy. The result of all of these mutations is the production of an amyloid  $\beta$  peptide which is more prone to cluster together than the normal peptide. The clustered protein forms amyloid deposits known as plaques that accumulates in the blood vessels of the brain. The amyloid plaques replaces the muscle fibers and elastic fibers which gives blood vessels flexibility, causing them weak and leads to breakage. Such a break causes bleeding [hemorrhagic stroke], brain damage and dementia. Amyloid plaques in specific parts of the brain interferes with brain function, leading to movement problems, seizures and other neurological features.--

PSEN1, presenilin 1, is involved in the development of the brain and spinal cord [central nervous system] and the survival of nerve cells [neurons]. Over production of amyloid beta peptide fragment can build up in the brain and form clumps called amyloid plaques [Alzheimer disease] leading to the death of neurons and the progressive signs and symptoms of this disorder.--

FYN-; Implicated in the cell growth, regulates intracellular calcium levels, mobilizes cytoplasmic calcium, brain development, matures brain function, regulation of axon growth, guidance, and neurite extension.

We have used Random walks to transfer annotations in networks [23]. The relationship between biological distribution of the disease d proteins and the prediction performance in the network

measures whether a set of proteins [that are associating with a given disease] is located in dense cluster in the network or is uniformly distributed.

## **STRING**

The STRING database has been designed to assemble, measure and disseminate protein–protein association

information, in a comprehensive manner. Since interactions between proteins represents as a critical component for modern biology, so STRING is not the only online resource for this topic. Apart from the primary databases, a number of resources follow a meta-analysis approach like Consensus-PathDB [24], I2D [25], VisANT [26] and, more recently, hPRINT [27], HitPredict [28], IMID [29] and IMP [30].

STRING provides a comprehensive coverage, with more than 1000 organisms, 5 million proteins and more than 200 million interactions stored, hold experimental, predicted interactions, together with interactions obtained from text mining; and includes a wealth of accessory information, such as protein structures and domains.

STRING resource includes data-sources, predicted algorithms and user-interface.

To improve the quality of links derived, we have developed an entirely new scoring scheme from co-occurrence, like co-occurrences within sentences, paragraphs and combines them by an optimized weighting scheme.

The term ‘interologs’ was coined for such pairs, a combination of the words ‘interaction’ and protein, which indicates functional linkages even in organisms where the two proteins have not fused. We search for gene families that share random similarities in their evolutions. Then, we conduct searches among genes to display a similar transcriptional response across a variety of conditions. Finally, STRING has two further sources of interactions in providing the group of associations; these are text-mining and interaction transfer in between organisms. Interactions specified semantically using Natural Language Processing [31] conserves genomic neighborhood clustering of genes to infer functional linkage, STRING now uses this type of arrangement in its neighborhood algorithm as well ‘Ortholog’ [32] complemented by associations that are predicted computationally, typically for STRING, using a number of prediction algorithms [33,34]. First, we conduct systematic search for genes that are found closely within chromosomes, which is a good sign for functional linkages. Second, we search for probabilities where genes have joined to encode a single fusion

## **INTEGRATION OF PROTEIN STRUCTURES**

### **1- 3D structure information.**

### **2- protein structures**

serves as an evidence for interaction themselves, when more than one distinct peptide chain are found in the structure. Therefore, a stable and reliable protein–protein interaction is considered.

STRING relies on an extended version of the COGs [‘Clusters of Orthologous Groups’ [35]], maintained at

The eggNOG database [36]. Some other databases use STRING networks for further computations/ annotations, [STITCH, [37]], or by using kinase–substrate predictions [NetworKIN, [38]]. STRING has also been integrated as NeAT [Network Analysis Tools, [39]]. It follows high-throughput genomics, proteomics and bioinformatics scanning approaches. It allow researchers to measure the changes and regulate genome-wide genes alongwith certain biological conditions, generate large ‘interesting’ gene lists as their final outputs and plays successful role in gene functional analysis of large gene collections for various high-throughput biological studies.

## **ENRICHNET**

Annotation databases, such as Gene Ontology [GO] [40], collects biological knowledge in a gene-to-annotation format, are very suitable for high-throughput scanning for the enrichment analysis.

Assessment of acknowledged functional associations between these genes/proteins, pathways, cellular processes and complexes.

The significance of the overlap IS evaluated between gene/protein sets using a statistical test like Fisher's exact test. EnrichNet is a web-application, integrates information from the known interactions network structure between proteins.

#### Benefits:

EnrichNet scores functional associations between all gene/protein sets, even if the no genes/proteins are shared. It accounts implicit functional information on genes/proteins alongwith missing annotations taken from molecular interaction data

EnrichNet provides a ranking table of putative association for network structure of physical interactions between the gene/protein sets, and provides an interactive network visualization. It computes tissue-specific association scores in the human species Using tissue-specific gene expression data.

EnrichNet provides a choice between six different pathway/process databases [KEGG, BioCarta, WikiPathways, Reactome, , NCI Pathway DB, Gene Ontolgy,] and enables the user to upload self-defined human gene/protein interaction networks.

common human microarray platforms are ENTREZ gene identifiers and probe identifiers . The maximum size limit for the target gene/protein set is 1000, however for specific results, it is recommended to use dataset sizes range from 10 and 500. The user can map the gene/protein list and pathways from the STRING database or upload a self-defined interaction network in ENSEMBL gene format.

The results includes an interactive ranking table that shows which pathways have been prioritized for a putative functional association alongwith the user-defined gene sets in terms of a network-based association scoring [the Xd-score] and a classical overrepresentation-based scoring using Fisher's exact test.

#### 3.1.1 Uploading your own interaction network data

For using a pre-defined human protein-protein interaction network, the user only needs to input the gene/protein identifiers of interest into the text box.

#### 3.1.2 Ranking of pathways in terms of network interconnectivity with mapped gene/protein sets

EnrichNet pathway rankings and network visualizations are generated using the list of known human mutated genes as input, the pathway definitions is from the KEGG database and a protein-protein interaction network is from the STRING 9.0 database [using interactions with combined score  $\geq 600$ ]. In the following, a general description of the pathway ranking results given by EnrichNet will be provided.

The output generated by EnrichNet from a user-defined gene/protein set is a ranking of pathways for a selected database [the Kyoto Encyclopedia of Genes and Genomes [KEGG], Reactome, WikiPathways, BioCarta, NCI Pathway Interaction DB, InterPro or Gene Ontology] in terms of its estimated functional association with the user's defined gene/protein set. This quantitative association estimation maps both the user-defined dataset and the pathway datasets over a molecular interaction network and compute network-based gene/protein set enrichment scores using the automatic procedure.

We find compute graph visualization option, a visualization network of the user-defined gene set and the mapped pathway will be generated and displayed together. The first column shows Visualization of sub-networks representing pathway/gene set associations. The second column visualises the network similarity scores [Xd-scores], which evaluate the

network interconnectivity between the user-defined gene/protein set and the cellular pathways/processes mapped onto the molecular interaction network. The third column shows classical overlap-based enrichment analysis score, the gene set overlap between the pathway and the user-defined gene set, measured by the Fisher exact test. The fourth column shows the number of genes/proteins in the uploaded and mapped user-defined gene/protein set The fifth column shows the mapped pathway and The sixth column shows their intersection. Final column shows the corresponding XD-score difference when restricted to the molecular interaction network to different human tissue-types, and a bar plot and a table of tissue-specific XD-scores for each pathway is displayed by clicking on the corresponding hyperlinks.

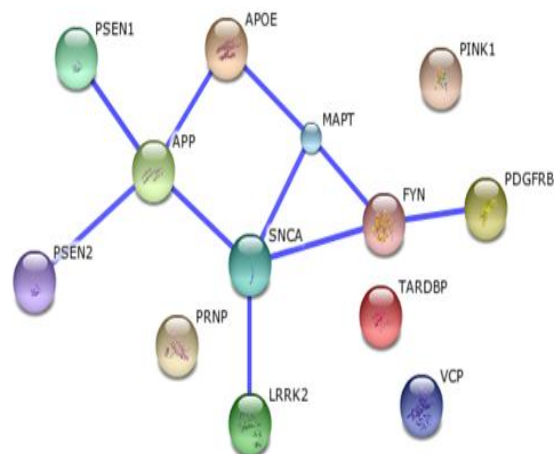
#### 3.1.3 Visualization of sub-networks representing pathway/gene set associations

When a gene/protein set with classical enrichment analysis methods is analyzed, the user only obtains a ranking list of pathways/processes as output, which indicates that pathways are most likely to have a functionally associated with user's gene/protein of interest. To explore in-depth for a gene set/pathway association, EnrichNet visualizes a network for each pathway that appears in the ranking, accessible via the compute graph visualization link in the first column that contains a graph visualization of the user's gene/protein set [blue nodes], the selected pathway [red nodes] and their shared nodes in a molecular interaction network--

To facilitate the analysis of densely clustered graph regions with nodes for large gene/protein sets, we click on the node for a certain gene/protein, a list will be displayed which provides information on the network topology of the node [e.g. the degree, clustering coefficient, and various node centrality measures] and contains a hyperlink to an external functional annotation database giving biological information about the selected or gene/protein.

## 4. RESULTS

We evaluate the Gene- Gene interaction with these Genes, with 0.9 probability and custome Genes selection, we observe the following result:



**Fig 3. 13 mainly involved Genes(OMIM) graph plot against molecular pathway with functional role of Amino Acid content through RWR implemented by STRING**

So, with the graphical view( Fig 3), we find that 13 Genes are mainly interacted for the Dementia. But, here the main role is the presence of enriched amino acid whose mutation are responsible for the effects. Similar genomic context in different organism suggest a similar function of the proteins[Neighborhood]. Proteins that are fused in few genomes are very likely to be functionally [Fusion-fission events]. Proteins that have a similar function in the same metabolic pathway, have similar phylogenetic profile[Occurrence].

Observed patterns of similar expression of genes Predicts association between genes[Coexpression]. We used STRING database. No single method was superior for all of the genes to be tested, our RWR method has outperformed all other methods on average. Graph partitioning is a better technique for predicting gene–disease associations because it uncovers functional modules in PPI networks, and similar diseases are caused by proteins that bears similar biological processes.

## 5. CONCLUSIONS

‘previous-knowledge’-targeted on experimental studies over individual genes may perform better than historical training data. In order to simulate the real-life search, genes are connected to each other by multiple paths, distinguished genes are connected to the real disease genes by single pathway only. networks persists all aspects of human health and bears a network approach in the analysis of cellular functions of genes and gene products, rather than a list of ‘‘disease genes,’’so it will be necessary for the understanding of disease mechanisms and proteins mutations in phenotypically similar diseases to form highly interlinked subnetworks within the larger protein interaction networks and deliver accurate performance.

## 6. REFERENCES

- [1] Franke, L. *et al.* [2006]. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.*, **78**, 1011–1025.
- [2] Gaulton, K.J. *et al.* [2007] A computational system to select candidate genes for complex human traits. *Bioinformatics*, **23**, 1132–1140.
- [3] Perez-Iratxeta, C. *et al.* [2007] Update of the G2D tool for prioritization of gene candidates to inherited diseases. *Nucleic Acids Res.*, **35**, W212–W216.
- [4] Aerts, S. *et al.* [2006] Gene prioritization through genomic data fusion. *Nat. Biotechnol.*, **24**, 537–544.
- [5] Sam, L. *et al.* [2007] Discovery of protein interaction networks shared by diseases. In *Pacific Symposium on Biocomputing*. World Scientific, Singapore, pp. 76–87.
- [6] Radivojac, P. *et al.* [2008] An integrated approach to inferring gene-disease associations in humans. *Proteins*, **72**, 1030–1037.
- [7] Karni, S. *et al.* [2009] A network-based method for predicting disease-causing genes. *J. Comput. Biol.*, **16**, 181–189.
- [8] Ma, X. *et al.* [2007] CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics*, **23**, 215–221.
- [9] George, R.A. *et al.* [2006] Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res.*, **34**, e130.
- [10] Ozgur, A. *et al.* [2008] Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, **24**, i277–i285.
- [11] van Driel, M.A. *et al.* [2006] A text-mining analysis of the human genome. *Eur. J. Hum. Genet.*, **14**, 535–542.
- [12] Perez-Iratxeta, C., Bork, P., and Andrade, M.A. [2002]. Association of genes to genetically inherited diseases using data mining. *Nat. Genet.* **31**, 316–319.
- [13] Lo´pez-Bigas, N., and Ouzounis, C.A. [2004]. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res.* **32**, 3108–3114.
- [14] Oti, M., Snel, B., Huynen, M., and Brunner, H.G. [2006]. Predicting disease genes using protein–protein interactions. *J. Med. Genet.* **43**, 691–698.
- [15] Huynen, M., Snel, B., Lathe, W. 3rd and Bork, P. [2000] Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.*, **10**, 1204–1210.
- [16] Eisenberg, D., Marcotte, E.M., Xenarios, I. and Yeates, T.O. [2000] Protein function in the post-genomic era. *Nature*, **405**, 823–826.
- [17] Lee, I., Blom, U.M., Wang, P.I., Shim, J.E. and Marcotte, M. [2011] Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.*, **21**, 1109–1121.
- [18] Moreau, Y. and Tranchevent, L.C. [2012] Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Genet.*, **13**, 523–536.
- [19] Piro, R.M. and Di Cunto, F. [2012] Computational approaches to disease-gene prediction: rationale, classification and successes. *FEBS J.*, **279**, 678–696.
- [20] Lee, J.M., and Sonnhammer, E.L.L. [2003]. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.* **13**, 875–882.
- [21] McKusick, V. [2007] Mendelian inheritance in man and its online version, OMIM. *Am. J. Hum. Genet.*, **80**, 588–604.
- [22] Hamosh, A., Scott, A.F., Amberger, J., Bocchini, C., Valle, D., and McKusick, V.A. [2002]. Online mendelian inheritance in man [OMIM], a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **30**, 52–55.
- [23] Kohler, S., Bauer, S., Horn, D. and Robinson, P.N. [2008] Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.
- [24] Kamburov, A., Pentchev, K., Galicka, H. Wierling, C., Lehrach, H. and Herwig, R. [2011] ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res.*, **39**, D712–D717.
- [25] Niu, Y., Otasek, D. and Jurisica, I. [2010] Evaluation of linguistic features useful in extraction of interactions from PubMed; application to annotating known, high-throughput and predicted interactions in I2D. *Bioinformatics*, **26**, 111–119.
- [26] Hu, Z., Hung, J.H., Wang, Y., Chang, Y.C., Huang, C.L. Huyc k, M. and DeLisi, C. [2009] VisANT 3.5: multi-scale

network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Res.*, 37, W115–W121.

- [27] Elefsinioti, A., Sarac, O.S., Hegele, A., Plake, C., Hubner, N.C., Poser, I., Sarov, M., Hyman, A., Mann, M., Schroeder, M. et al. [2011] Large-scale de novo prediction of physical protein-protein association. *Mol. Cell. Proteomics*, 10, M111 010629.
- [28] Patil, A., Nakai, K. and Nakamura, H. [2011] HitPredict: a database of quality assessed protein-protein interactions in nine species. *Nucleic Acids Res.*, 39, D744–D749.
- [29] Balaji, S., McClendon, C., Chowdhary, R., Liu, J.S. and Zhang, J. [2012] IMID: integrated molecular interaction database. *Bioinformatics*, 28, 747–749.
- [30] Wong, A.K., Park, C.Y., Greene, C.S., Bongo, L.A., Guan, Y. and Troyanskaya, O.G. [2012] IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. *Nucleic Acids Res.*, 40, W484–W490.
- [31] Saric, J., Jensen, L.J., Ouzounova, R., Rojas, I. and Bork, P. [2006] Extraction of regulatory gene/protein networks from Medline. *Bioinformatics*, 22, 645–650.
- [32] Walhout, A.J., Sordella, R., Lu, X., Hartley, J.L., Temple, G.F., Brasch, M.A., Thierry-Mieg, N. and Vidal, M. [2000] Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science*, 287, 116–122.
- [33] Skrabanek, L., Saini, H.K., Bader, G.D. and Enright, A.J. [2008] Computational prediction of protein-protein interactions. *Mol. Biotechnol.*, 38, 1–17.
- [34] Harrington, E.D., Jensen, L.J. and Bork, P. [2008] Predicting biological networks from genomic data. *FEBS Lett.*, 582, 1251–1258.
- [35] Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D. and Koonin, E.V. [2001] The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, 29, 22–28.
- [36] Jensen, L.J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T. and Bork, P. [2008] eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.*, 36, D250–D254.
- [37] Kuhn, M., von Mering, C., Campillos, M., Jensen, L.J. and Bork, P. [2008] STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.*, 36, D684–D688.
- [38] Linding, R., Jensen, L.J., Ostheimer, G.J., van Vugt, M.A., Jorgensen, C., Miron, I.M., Diella, F., Colwill, K., Taylor, L., Elder, K. et al. [2007] Systematic discovery of in vivo phosphorylation networks. *Cell*, 129, 1415–1426.
- [39] Brohee, S., Faust, K., Lima-Mendez, G., Sand, O., Janky, R., Vanderstocken, G., Deville, Y. and van Helden, J. [2008] NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. *Nucleic Acids Res.*, 36, W444–W451.
- [40] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. et al. (2000) Gene ontology: tool for the unification of biology. The GeneOntology Consortium. *Nat. Genet.*, 25, 25–29.