BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 9(19)

Protein interaction network for Alzheimer's disease using computational approach

V Srinivasa Rao[1]*, K Srinivas[1], GN Sunand Kumar[1] & GN Sujin[2]

[1]Department of CSE, VR Siddhartha Engineering College, VR Siddhartha Engineering College, Kanuru, Vijayawada; [2]Department of CSE, Mahatma Gandhi Institute of Technology, Mahatma Gandhi Institute of Technology, Gandipet, Hyderabad; V Srinivasa Rao – Email: drvsrao9@gmail.com; *Corresponding author

Received November 10, 2013; Accepted November 25, 2013; Published December 06, 2013

Abstract:
Alzheimer's disease (AD) is the most common form of dementia. It is the sixth leading cause of death in old age people. Despite recent advances in the field of drug design, the medical treatment for the disease is purely symptomatic and hardly effective. Thus there is a need to understand the molecular mechanism behind the disease in order to improve the drug aspects of the disease. We provided two contributions in the field of proteomics in drug design. First, we have constructed a protein-protein interaction network for Alzheimer's disease reviewed proteins with 1412 interactions predicted among 969 proteins. Second, the disease proteins were given confidence scores to prioritize and then analyzed for their homology nature with respect to paralogs and homologs. The homology persisted with the mouse giving a basis for drug design phase. The method will create a new drug design technique in the field of bioinformatics by linking drug design process with protein-protein interactions via signal pathways. This method can be improvised for other diseases in future.

Key Words: Alzheimer's disease, protein network, protein interactions, protein interaction network, PPI for Alzheimer's, PPI network, Alzheimer's disease network.

Background:
Deciphering the structure and dynamics of complex network of protein-protein interactions is among the essential objectives of systems biology for understanding many aspects of living systems in depth **[1].** The construction of protein interactome was supported by ongoing experimental and computational techniques. The number of experimentally supported PPIs for model organisms has been increasing in recent years as evident from the large protein-protein interaction (PPI) databases. The experimentally identified PPIs are mined and stored in open source databases. Currently, the experimental techniques for the massive characterization of PPI networks still have several drawbacks **[2].** First, there is surprisingly low convergence rate between the results of similar kind of experiments. Second, experimental techniques like yeast two hybrid often produce a large number of false positives with an estimated percentage of 10% in some cases. Third, experimental approaches are still unable to reach a high-throughput state since the inherent drawbacks of the methodologies are only allowing them to test a fraction of all possible pairs of proteins. Finally, these limitations of experimental techniques arise from their experimental nature itself. However, appropriate care has been taken in the construction phase as we considered only the experimentally reviewed ones.

Alzheimer's disease (AD) is an irreversible, progressive brain disorder that slowly destroys memory and thinking skills and the ability to carry out the simplest tasks **[3].** Alzheimer's disease is the most common cause of dementia among the older people. Dementia is the loss of cognitive functioning like thinking, remembering, reasoning etc to an extent that it interferes with a person's daily activities. Plaques and tangles in the brain are the major causes for Alzheimer's disease and the third being the loss of connections between nerve cells (neurons) in the brain **[3].** Molecular Changes in the deoxyribonucleic acid (DNA) of Alzheimer's Patient's Brain

gives the initial information about the severity of the disease. The reasons for choosing AD for this study are two-fold **[3]:** first, the lack of food and drug administration (FDA) approved drugs to treat AD today, in spite of decades of research on the disease's molecular mechanisms; second, the wealth of biomedical research articles published for AD studies can make validations of our approach less challenging. Biological networks capture a variety of molecular interactions and in particular, protein–protein interaction networks facilitate the understanding of pathogenic mechanisms that trigger the onset and progression of diseases **[4].** Protein interaction networks present gene products that physically interact with each other to accomplish particular cellular functions, such as metabolism, cell cycle control, and signal transduction **[5].** Advanced network based approaches are becoming particularly important to identify pathways or functional modules that may indicate potential therapeutic target(s) **[6]**. Recently, network theory is making an important contribution in topological study of biological networks, such as protein-protein interaction (PPI) networks **[6].** A PPI network can be described as a complex network of proteins joined by interactions. Proteins are represented as nodes in such a graph; two proteins that interact physically are represented as adjacent nodes connected by an edge. In general, an average of five interaction partners per protein has been calculated by Piehler. J **[5].**

Most biological processes can hardly be understood without a comprehensive analysis of a large number of molecular components and interactions **[1].** From the simple system to complex ones, the interactions between different molecules usually determine the resulting phenotype. This is the case with cellular proteins, which rarely work in isolation but are frequently involved in pathways and interaction networks. The eventual perturbation of these networks can lead to disease or even death **[1].** So, the knowledge of protein-protein interactions can greatly contribute to the understanding of living systems in general and pathology in particular. In recent years, identifying candidate genes of complex diseases was mainly based on biochemical networks such as metabolic networks **[7],** transcriptional regulatory networks **[8],** and protein-protein interaction networks (PPINs) **[9].** An understanding of the basic biochemistry of the key interactions in AD may provide a framework needed to develop drugs for curing AD. Moreover, interacting proteins have been shown to have a tendency of sharing similar functions and causing the same disorder **[10-11].** The objective of the present study is to construct the current experimentally supported network of direct human protein interactions, explore it for potential target proteins. At one end, the UniProt Knowledge base (UniProtKB) **[12]** was taken as the reference set of nodes that the network can have. Then we performed text mining on the PPI databases, i.e Human protein reference database (HPRD) **[13],** InAact molecular interaction database (IntAct) **[14],** The molecular interaction database (MINT) **[15],** Database of interacting proteins (DIP) **[16],** Systems biology of the innate immune response (INNATEDB) **[17],** bio-molecular interaction network database (BIND) **[18]** and biological general repository for interaction datasets (BioGRID) **[19],** for direct interactions between the reference proteins. We analyzed the network for the prioritized proteins among the reference protein set.

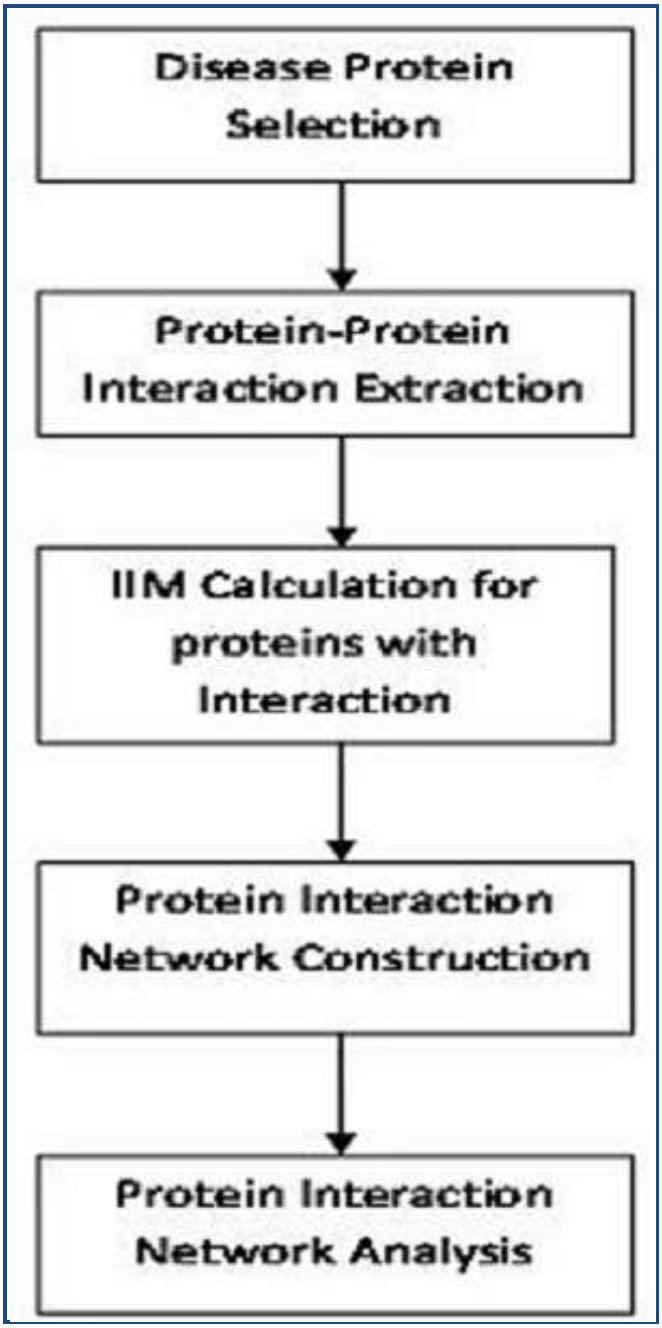

Figure 1: Workflow of the proposed method

Methodology:

We used experimentally validated PPI information to obtain the neighbors for each seed group. Our analysis includes only direct interactions identified either by biochemical experiments or by two-hybrid studies. In this method, integrated interlogous dataset (formerly OHPID **[20]**) was taken as the basis for interaction data. The dataset includes the interactions taken from different databases that includes IntAct **[14],** HPRD **[13]**, BioGrid **[14]**, MINT **[15]**, DIP**[16]**, INNATEDB **[17]** and BIND **[18].** The dataset has a massive set of 8,46,116 interactions, of which 4,90,600(58%) were source interactions and 3,70,002(42%) were the predicted ones. In these interactions, 1,73,338(20%) interactions were related to Human. Out of 1,73,338 interactions,1,20,030(69%) were from source and 59,741(31%) were predicted interactions and all the interactions were considered in the construction of network.

Second, 136 Reviewed disease proteins (seed) were taken as input after performing text mining on Uniprot etc and the proteins can be accessed from supplement file1. For these proteins, an IIM (In-Direct Interaction Matrix) was calculated using the IIM algorithm which takes the input proteins and produces the interactions upto the required cycle length. The workflow for the foresaid method was given in the **(Figure 1).** After taking all the 1412 interactions into consideration, the interactions were then converted into SIF (Simple Interaction Format), which specifies the nodes and interactions. Then the SIF file was loaded into the Cytoscape **[21],** a tool used to visualize molecular interaction networks. In **(Figure 2)**, the nodes are shown in green color and edges are shown in black color. Whenever the resultant dataset contains a protein-protein interaction between the protein A and protein B, the generated network depicts an edge between two nodes A and B. The network thus constructed can be seen in **(Figure 2).**

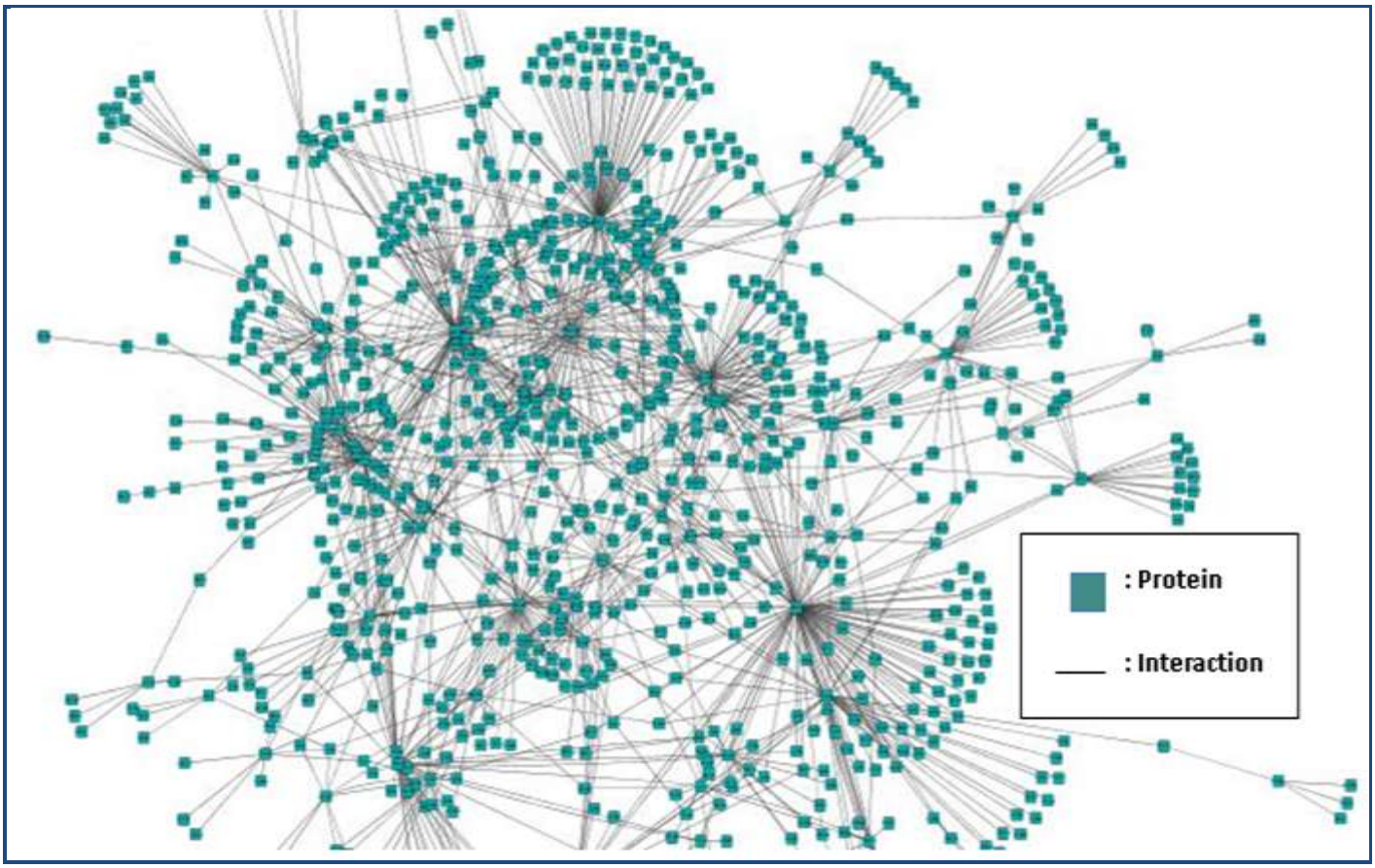

Figure 2: Protein-protein interaction network for Alzheimer's disease

Results & Discussion:

The novelty of our approach was instead of merging PPI information for any protein identifier stored in source databases; we have mined PPIs exclusively between reference proteins. The interactions which were detected by the algorithm were given in the supplement file2. The PPI network has 73 proteins with at least one characterized partner. The network was further investigated for individual protein-protein interactions of corresponding proteins for further research. The cycle level was fixed at three as there will not be any interactions beyond this level as seen from the resultant dataset. The confidence score for each protein was also calculated using the equation **(Please see supplementary materials for equation and explanation).**

The top five high scored ranked proteins are identified as APP-P05067, SP1-P08047, GSK3B-P49841, PSEN1-P49768 and UBB-P0CG47 respectively. These proteins with highest number of interactions are involved in essential biological processes. These results will be helpful in drug target identification. The network analysis was further performed using the Cytoscape tool for important network properties. The network properties were recorded **(Please see supplementary materials for network properties)**

We resolve the issue of PPI redundancy which originates from combination of records of multiple databases at different levels. We comment on the proteins represented with high degree in interactome network; We resolve the orphan proteins inclusion (no direct PPIs with reference proteins) as a result of their interaction at different cycle level; Dataset preparation has augmented additional interactions with the application of IIM algorithm.

The prediction performance of this method depends heavily on the topology of the network and the quality of protein-protein interactions with respect to size and reliability **[22].** The network structure follows a scale-free property with few hubs and the majority of proteins involving in small number of interactions and some interactions needed further verification. The

homology of top high ranked proteins was persisted with mouse (MUS MUSCULUS). The protein APP (UniProt identifier: P05067) was the protein identified with the largest number of interactions in the constructed network as conformed by Jiao Li *et al* **[23].** To assess the reliability of this network, the interactions of P05067 (Amyloid beta A4 protein) were compared with the results of STRING **[24]** database. The results show that IIM algorithm successfully detected 80% of interactions when compared with STRING. However, out of 80% of the interactions detected successfully, 20% interactions need further experimentation for validation purpose. The results were recorded in the **Table 1 (see supplementary material)**.

Conclusion:

We have provided a novel method, which will extract the direct protein-protein interactions from integrated databases referring to manually reviewed UniProtKB proteins. We suggest that this PPI network has to expand to its maximum potential with support of more reviewed proteins and their potential interactions. Now, this set of protein interactions may trigger text mining efforts for identification of any novel disease proteins and their interactions. The method already provides a new technique for investigation of important biological processes and molecular functions in the context of drug research. This method can be improvised for other diseases in future.

References:

[1] Kalpa MI *et al*. *BMC Syst Biol*. 2013 **7**: 96 [PMID: 24088582]
[2] Fiona Browne *et al. Advances in Artificial Intelligence.* 2010 **2010**: 924529
[3] Alzheimer's Disease Fact Sheet. http://www.nia.nih.gov /Alzheimers/Publications/adfact.htm.
[4] Jaeger S & Aloy P, *IUBMB Life*. 2012 **64**: 529 [PMID: 22573601]
[5] Piehler J, *Curr Opin Struct Biol*. 2005 **15**: 4 [PMID: 15718127]
[6] Goni J *et al. BMC Syst Biol*. 2008 **2:** 52 [PMID: 18570646]
[7] Ravasz E *et al. Science.* 2002 **297**: 1551 [PMID: 12202830]
[8] Lee TI *et al. Science.* 2002 **298**: 799 [PMID: 12399584]
[9] Han JD *et al. Nature*. 2004 **430**: 88 [PMID: 15190252]
[10] Wang X *et al*. *Nat Biotechnol*. 2012 **30**: 159 [PMID: 22252508]
[11] V Srinivasa Rao & K Srinivas, *Journal of Bioinformatics and Sequence Analysis*. 2011 **3**: 89
[12] Chen C *et al. Nucleic Acids Res*. 2013 **41**: D43 [PMID: 23161681]
[13] Peri S *et al. Nucleic Acids Res*. 2004 **32**: D497 [PMID: 14681466]
[14] Hermjakob H *et al. Nucleic Acids Res.* 2004 **32** :D452 [PMID: 14681455]
[15] Chatr-aryamontri A *et al. Nucleic Acids Res.* 2007 **35**: D572 [PMID: 17135203]
[16] Xenarios I *et al. Nucleic Acids Res.* 2002 **30:** 303 [PMID: 11752321]
[17] Lynn DJ *et al*. *Mol Syst Biol*. 2008 **4**: 218 [PMID: 18766178]
[18] Bader GD *et al*. *Nucleic Acids Res.* 2001 **29**: 242 [PMID: 11125103]
[19] Stark C *et al. Nucleic Acids Res.* 2006 **34**: D535 [PMID:16381927]
[20] Brown KR & Jurisica I, *Bioinformatics* 2005 **21**: 2076 [PMID: 15657099]
[21] Shannon P *et al. Genome Res.* 2003 **13**: 2498 [PMID: 14597658]
[22] Lage K *et al. Nat Biotechnol*. 2007 **25**: 309 [PMID: 17344885]
[23] Li J *et al. PLoS Comput Biol.* 2009 **5:** e1000450 [PMID: 19649302]
[24] Von Mering C *et al. Nucleic Acids Res.* 2005 **33**: D433 [PMID: 15608232]

Edited by P Kangueane
Citation: **Rao *et al.*** Bioinformation 9(19): 968-972 (2013)

BIOINFORMATION

open access

Supplementary material:

The confidence score for each protein was also calculated using the following equation:

$score_n = \frac{1}{N}\sum_{i=1}^{3}\frac{p_i}{i}$	Where n is protein number from 1 to N N is number of proteins i is cycle level from 1 to 3 p_i is number of proteins with cycle level i

The network properties were recorded as follows:

Network Charateristic	Value
Number of Nodes	969
Number of Edges	1412
Network	un-directed
Clustering Coefficient	0.041
Connected Components:	21
Network Diameter	10
Network Radius:	1
Network Centralization	0.084
Average Number of Neighbors:	2.603
Network Density	0.003
Network Heterogeneity	2.517
Isolated Nodes:	0
Multi-edge Node Pairs	73

Table 1: Comparison of the interactions detected using IIM with the interactions detected using STRING for the protein P05067 (APP)

S.No	Gene Symbol	Uniprot Protein id	String	IIM
1	KAT5	Q92993	✓	✓
2	NAE1	Q13564	✓	✓
3	ITM2B	Q9Y287	✓	✓
4	APBB1	000213	✓	✓
5	BACE1	P56817	✓	×
6	APOE	P02649	✓	✓
7	APBA1	Q02410	✓	✓
8	PSEN1	P49768	✓	✓
9	ATP6VOA4	Q9HBG4	✓	×
10	TGFB1	P01137	✓	✓