Graph theoretic network analysis reveals protein pathways underlying cell death following neurotropic viral infection

Sourish Ghosh, G. Vinodh Kumar, Anirban Basu & Arpan Banerjee

Complex protein networks underlie any cellular function. Certain proteins play a pivotal role in many network configurations, disruption of whose expression proves fatal to the cell. An efficient method to tease out such key proteins in a network is still unavailable. Here, we used graph-theoretic measures on protein-protein interaction data (interactome) to extract biophysically relevant information about individual protein regulation and network properties such as formation of function specific modules (sub-networks) of proteins. We took 5 major proteins that are involved in neuronal apoptosis post Chandipura Virus (CHPV) infection as seed proteins in a database to create a meta-network of immediately interacting proteins (1st order network). Graph theoretic measures were employed to rank the proteins in terms of their connectivity and the degree up to which they can be organized into smaller modules (hubs). We repeated the analysis on 2nd order interactome that includes proteins connected directly with proteins of 1st order. FADD and Casp-3 were connected maximally to other proteins in both analyses, thus indicating their importance in neuronal apoptosis. Thus, our analysis provides a blueprint for the detection and validation of protein networks disrupted by viral infections.

Metabolic functions are outcomes of interactions among various cellular proteins. An emerging concept in the field of proteomics is that the understanding of these interactions is critical for elucidating the mechanism of metabolic functions1,2. However, parsing interactions important for certain functions or a disease involves analyzing huge interactomes containing information about a large number of genes and proteins along with their interacting partners. Mathematical modelling has been instrumental in analyzing these huge datasets and systematically understanding the interplay between various proteins and the metabolic functions involved1-5. Recent technical developments consider the huge protein interactome as a complex graph wherein individual proteins are nodes of the graph and the interactions are modelled as the edges6. Graph theoretic analysis provides an efficient handle to decipher various aspects of proteins in a network that interact with a specific functional objective. For example, is one protein more important than others, does a group of protein exhibit more interactions (densely connected) than other groups, do some proteins act as hubs through which majority of interactions are routed? Graph theory provides several parameters to study properties of constituent proteins in an interactome: degree centrality, clustering, betweenness, shortest path, modularity, etc., each of which may be meaningful for understanding function7-9. Based on a hypothesis about the operational structure of the interactome, researchers can decide upon what parameters to investigate.

Modularity quantifies how the nodes of a network are interacting among each other to form “hubs”10. Hubs or modules are closely interacting group of nodes with more connections within the module and sparse connections between modules. Real world networks such as the Internet, power grids, brain
Chandipura Virus (CHPV) a member of the Rhabdoviridae family, has been ranked among the emerging viruses in the Indian subcontinent. CHPV was first identified in two patients in the year 1965 from the Chandipura village in Maharashtra (India)\textsuperscript{15}. The first major outbreak took place in 2003 and resulted in death of 183 children. This was followed by sporadic attacks every year. Presently CHPV has a case-by-case fatality rate of around 55–77%\textsuperscript{16–18}. The virus has been reported to cause encephalitis along with neurodegeneration leading to death. Common symptoms which have been diagnosed are high grade fever, vomiting, altered sensorium, generalized convulsions, decerebrate posture and coma. CHPV, being an arbovirus with sand flies (Phlebotomus sps.) as the carrier (vector), enters the host system through the skin, penetrating into the circulatory system of the body (which is also referred to as peripheral circulatory system). CHPV is cleared off the peripheral circulatory system by the host immune system within a couple of days post infection as observed in a mouse model\textsuperscript{19–20}. But this virus finds a safe place to replicate in the brain.

In an earlier article some of us have shown in a mouse model that CHPV induces neuronal death through a Fas-mediated extrinsic apoptosis pathway\textsuperscript{17}. From there we identified 5 proteins pertaining to the extrinsic apoptotic pathway. However, from this analysis we did not get the information about all the proteins that may be involved in the apoptotic process following CHPV infection.

In this article we identified a large number of proteins (from an online database) that interact with the five proteins whose expressions were monitored in the earlier wet-lab experiment of CHPV infection\textsuperscript{17}. The resultant network of proteins constituted a “1st order interactome”\textsuperscript{14}. Furthermore, we estimated a “2nd order interactome” by identifying the proteins that were directly interacting with the 1st order interactome. We calculated the modularity of 1st and 2nd order interactomes and degree centrality of individual proteins. These two measures quantified both a global measure of segregation of network and an individual connectivity measure of candidate proteins. 2nd order connectome results were used to test the robustness of 1st order connectome results and address the issue of predictive validity of the model. Together they revealed the protein-protein network configuration underlying neuronal apoptosis following CHPV infection. The issue of face validity was addressed by comparisons of empirical measures with those computed on simulated random networks. The methods and results obtained here provide an operational blueprint for understanding the pivotal dependencies of the virus within the host system and will help in the conceptualization and design of effective therapeutics.

**Results**

From our results in an earlier study\textsuperscript{17} we concluded that CHPV induces neuronal apoptosis through Fas-mediated extrinsic apoptotic pathway with the involvement of the following five proteins: Fas, Fas-associated Death Domain (FADD), Caspase-8 (Casp-8), cleaved Caspase-3 (Casp-3), and X-linked inhibitor of apoptosis (XIAP) (Fig. 1, see also Methods section for more details). These 5 proteins were inserted as inputs to STRING 9.1 online database (http://string-db.org/) for extraction of the 1st and 2nd order interactomes. The 1st order interactome contained 26 proteins while the 2nd order contained 71 proteins (Fig. 1b,c). The names of each protein from 1st and 2nd order interactomes are presented in Table 1.

The MATLAB-based Visual Connectome Toolbox\textsuperscript{21} was used for graph-theoretic analysis of 1st and 2nd order interactome data. We computed the degrees of freedom (degree centrality) for each protein in the 1st & 2nd order networks. Subsequently, we arranged them in a descending order (Table 2). From Table 2, we observed that in both 1st and 2nd order networks FADD and Casp-3 are the common members among the top 5 proteins having highest degree centrality values. Mutual cross-validation of results from 1st and 2nd order network analysis confirms that FADD and Casp-3 are dominant players in apoptotic pathway underlying CHPV infection in neurons. Modularity determines how well a network can be divided into subgroups (hubs). Generally the modularity score ranges between \([-0.5, 1]\) with more modular networks having a positive score. A more randomly assigned network will have a modularity score of approximately zero. We computed modularity scores of both 1st and 2nd order networks. The modularity score of the 1st order network was 0.36 while the 2nd order was 0.41. Theoretically, due to the random partitioning of nodes into modules to initiate the graph theoretic algorithm, the results may vary trial to trial unless the modular structure is significantly unambiguous. In our data set the modularity score remained unchanged in all 50 repetitions of the analysis. Additionally, we evaluated the significance of the estimated modularity score by comparing with the modularity scores of a random network with an identical number of nodes. We start with an adjacency matrix with all values set to zero for a given number of nodes. Then we randomly assigned a value 1 in upper diagonal matrix locations. Finally, symmetric locations in lower diagonal positions are assigned values 1 to design the adjacency matrix for which network metrics are computed. Diagonal elements were always assigned a value 0 to avoid self-connections. The mean modularity score of a random network (50 repetitions) with 26 nodes was 0.13, whereas for a random network with 71 nodes, the score was 0.09. In both cases the estimated modularity values of the empirical networks were statistically significant at Bonferroni corrected p < 0.05.
(χ² = 20.67, df = 1 for 1st order and χ² = 58.40, df = 1 for 2nd order interactome). In case of the 1st order network, our analysis indicated the presence of 4 modules whereas in case of 2nd order network 12 modules were identified.

Using the Ci scores from Table 1 and Fig. 2, we color coded each module in Fig. 1(b,c). Modules 2 and 4 of the 1st order interactome and module numbers 3 and 2 of 2nd order interactome, respectively were presented in identical colors because they have multiple common members. The common members of module number 2 from 1st order and 3 from 2nd order are Casp-8, Tnfrs10b, Cflar, Fas, FADD, TRADD. Module 4 from 1st order and 2 from 2nd order has Casp-9, Casp-7, XIAP, Apaf-1 and Diablo. Extraction of a consistent network structure from the analysis of 1st order and 2nd order interactomes provides confidence about the biological relevance of the key modules. Table 3 lists the UniProt IDs of all proteins identified in the 1st and 2nd order interactomes.

Discussion

In this report we propose an analysis framework to compute the modular structure of a complex protein-protein interaction network (interactome). The choice of seed proteins: Fas, Fas associated Death Domain (FADD), Caspase-8, Caspase-3 and X-linked Inhibitor of Apoptosis Protein (XIAP) for the construction of the interactome was guided from our previous experimental findings. These were apoptotic proteins over-expressed in mouse neurons following Chandipura Virus infection. We used the STRING 9.1 database to compute the first order interactome. There are currently several bioinformatics toolboxes available, each with their own set of unique controls. We chose STRING 9.1 because it was the only method to the best of our knowledge that allowed us to prune networks based on a statistical confidence level. However, it is pertinent to note that the database used to extract the interactome will immensely influence the estimation of any functional modular structure. A study comparing the interactomes extracted from several data sets may potentially help future research in terms of data interpretation.

Next, we computed the graph theory metrics: Modularity Score (Q), Community Structure (Ci) and Degree Centrality (Z) to infer further about the protein-protein interactions underlying apoptosis. To establish the predictive validity of our analysis, we constructed a second order interactome based on secondary interacting partners of the seed proteins using the STRING 9.1 database (at 95% confidence) and re-calculated the graph theory metrics. The consistent presence of key protein assemblies in the first order and second order interactomes provides confidence regarding the robustness of our approach. Finally, we compared the closeness of modularity and degree centrality computed in empirical networks with that obtained for simulated random networks. Since, no modular structures are expected
| Protein Name                  | Protein Name                  | Module 1st | 2nd | Protein Name                  | Module 1st | 2nd | Protein Name                  | Module 1st | 2nd |
|------------------------------|------------------------------|-----------|-----|------------------------------|-----------|-----|------------------------------|-----------|-----|
| Caspase-3                    | Casp3                        | 1         | 2   | Caspase-7                    | Casp7      | 4   | Cyclin-dependent kinase inhibitor 1A (P21) | Cdkn1a    | 1   |
| Caspase-8                    | Casp8                        | 2         | 3   | Direct IAP binding protein with low pl | Diablo     | 4   | Forkhead box O4               | Foxo4      | 1   |
| Fas                          | Fas                          | 2         | 3   | Proto-oncogene tyrosine-protein kinase | Fyn       | 2   | cAMP responsive element binding protein 1 | Creb1      | 10  |
| Fas-associated Death Domain  | Fadd                         | 2         | 3   | Cylindromatosis              | Cyld      | 3   | Transformed mouse 3T3 cell double minute 2 | Mdm2       | 10  |
| X-linked inhibitor of apoptosis protein | Xiap                      | 4         | 2   | Cluster of Differentiation 40 | Cd40      | 3   | Cyclin-dependent kinase inhibitor 1B | Cdkn1b     | 1   |
| Fas Ligand                   | Fasl                         | 2         | 3   | TNF receptor-associated factor 3 | Traf3      | 8   | Forkhead box O1              | Foxo1      | 1   |
| Tumor Necrosis Factor (TNF) receptor-associated factor 2 | Traf2                     | 3         | 3   | Ubiquitin-c                  | Ubc       | 10  | Tuberculosis 2               | Tsc2       | 1   |
| Tumor necrosis factor receptor type 1-associated DEATH domain | Tradd                      | 3         | 3   | Toll-like receptor adaptor molecule 1 | Ticami     | 8   | Mechanistic target of rapamycin (serine/threonine kinase) | Mtor       | 1   |
| Receptor-interacting serine/threonine-protein kinase 1 | Rpk1                        | 3         | 3   | Death domain-containing protein | Cradd     | 3   | Phosphatase and tensin homolog | Pten       | 1   |
| CASP8 and FADD-like apoptosis regulator | Cflar                      | 2         | 3   | Inhibitor of nuclear factor kappa-B kinase subunit gamma | Ikbkg     | 3   | RPTOR independent companion of MTOR, complex 2 | Rictor     | 1   |
| Tumor necrosis factor receptor superfamily, member 10b | Tnfrsf10b                  | 2         | 3   | TNF receptor-associated factor 1 | Traf1      | 8   | SMT3 suppressor of mif two 3 homolog 1 | Sumo1      | 6   |
| B-cell receptor associated protein 31 | Bcap31                     | 2         | 3   | Toll-like receptor 4           | Tlr4       | 8   | Mitogen-activated protein kinase kinase kinase 5 | Map3k5     | 1   |
| Baculoviral IAP repeat-containing protein 2 | Birc2                     | 4         | 8   | Mitogen-activated protein kinase kinase 7 | map3k7    | 8   | DNA methyltransferase 1-associated protein 1 | Dmap1      | 10  |
| Baculoviral IAP repeat-containing protein 3 | birc3                      | 4         | 8   | Profilin1                    | Pfn1       | 10  | Hemeo-domain interacting protein kinase 1 | Hipk1      | 10  |
| Gelsolin                     | Gsn                          | 1         | 10  | B cell leukemia/lymphoma 2     | Bcl2       | 1   | Alpha thalasemia/mental retardation syndrome X-linked homolog | Atrx       | 10  |
| DNA fragmentation factor subunit alpha | Dffa                       | 1         | 2   | BCL2-like 11 (apoptosis facilitator) | bcl2l11    | 1   | V-real avian reticuloendotheliosis viral oncogene homolog A | Rela       | 9   |
| DNA fragmentation factor subunit beta | Dttb                      | 1         | 2   | Signal transducer and activator of transcription 3 | Stat3      | 8   | FBJ osteosarcoma oncogene | Fos        | 9   |
| Apoptotic protease activating factor 1 | Apaf1                      | 4         | 2   | Furin                        | Furin      | 10  | Adiponectin, CIQ and collagen domain containing | Adipoq     | 5   |
| Nerve Growth Factor          | Ngf                          | 1         | 11  | Transient receptor potential channel, subfamily Y, member 1 | Trpvl      | 8   | Leptin                       | Lep        | 9   |
| RAC-alpha serine/threonine-protein kinase | Akt1                       | 1         | 1   | Fibroblast growth factor receptor substrate 2 | Frs2       | 3   | Interleukin-6               | Il6        | 9   |
| Death-associated protein 6   | Daxx                         | 2         | 10  | Neurotrophic tyrosine kinase, receptor, type 2 | Ntrk2      | 3   | Interleukin-1a              | Ili1a      | 9   |
| Tumor Necrosis Factor        | Tnf                          | 3         | 9   | Nerve growth factor receptor (TNFR superfamily, member 16) | Ngfr       | 8   | Peroxisome proliferative activated receptor, gamma, coactivator 1 alpha | Pparg-c1a  | 7   |
| Tumor necrosis factor receptor superfamily member 1A | Tnfrsf1a                  | 3         | 3   | Neurotrophic tyrosine kinase, receptor, type 1 | Ntrk1      | 8   | Caspin                       | Clspn      | 10  |
| Caspase-9                    | Casp9                        | 4         | 2   | src homology 2 domain-containing transforming protein C1 | Shc1       | 8   |                                  |            |     |

Table 1. Protein names, community structure value (Ci) score of the 1st and 2nd order interactome. The protein names for the table were arranged according to the chronology in which they have been queried from the STRING 9.1 database. The first 5 are the proteins whose expressions were monitored empirically the next 21 were the primary interacting partners. The next 45 secondary interacting partners were added to the list.
in a random network, this addressed the issue of face-validity, that is, whether the method is successful in extracting meaningful information and helped us control false positives.

Modularity score (Q) of a network ranges between \([-0.5, 1]\) with negative Q scores signifying random interactions within the network. As the within group interactions increase, the network starts to become more modular and the Q value shifts more towards the positive side nearing to 1. For every network there exists an optimal Q value beyond which the modularity score cannot be enhanced even if we increase the number of modules. In our case we have determined the Q values of 1st order and 2nd order interactome are 0.3911 and 0.4716, respectively. These scores were stable across 50 independent runs. The community structure also remained unchanged. These two findings give us the confidence to state that protein-protein interactions are indeed highly modular due to their inherent biological properties. Hence it is pertinent that the interactive nodes of both the networks have been classified into a maximum number of possible modules. Next we focus on each module to decipher their biological significance.

In the first order interactome, 4 interactive modules were identified (Fig. 1b). We could clearly characterize that all proteins segregated in separate modules on the basis of their functional role in the apoptotic process. Module 2 and 3 consist of all the proteins which are mostly known as death domain (DD) and death-inducing signalling complex (DISC). Proteins like FADD, TRADD, Cflar RIPK1, Daxx, Bcap31, Tnfrs1a & 10b have been reported to contribute the DD\(^{22,23}\), while Caspase 8 and FADD forms the DISC\(^{24,25}\). Other members like Fas (Module 2) and TNF (Module 3) are commonly known as the initiators of the death process. Module 3 consists of proteins that are co-stimulators of tumor necrosis factor (TNF) induced cell death whereas Module 2 consists of proteins that contribute to both Fas and TNF pathways. Module 4 is a heterogeneous group that consists of both apoptotic activators and inhibitors that belong to caspase group. XIAP has been previously reported both in our previous report and other researchers to be 1b & Birc3 antagonist\(^{17}\) while Birc2 & Birc3 is well known to be in association with TNF to combat the apoptosis signalling\(^{26,27}\). Surprisingly, TNF and Birc2 and Birc3 were not in the same module in our 1st order interactome. Other apoptotic activators of module 4 are Apaf-1 and Diablo along with the caspases like Casp 9 & 7. Overall this module represents proteins that are affecting the intermediate phase of apoptosis before the appearance of the final executioner of the apoptotic pathways. Module 1 is a classical cluster consisting of the close interactors of Casp3, the final executioner of the apoptotic pathway. This module consists of some of the targets of Casp3 which gets cleaved in order to bring about various changes in the cellular environment and help in completion of the apoptotic process. Both Dffa and Dffb are cleaved by Casp3 to effect the DNA fragmentation\(^{28}\) while Gsn cleavage brings about morphological changes to the cell during apoptosis\(^{29}\). Ngf\(^{30}\) has been previously reported to be closely associated with Casp3. Akt-1 activation in response to cytokine receptor signalling has been associated with anti-apoptotic processes\(^{31}\). In our analysis we observed that although Akt-1 is linked with other modules, its association with Caspase-3 is strong, and as a result Akt-1 has been grouped in Module 1. However, the scenario changes drastically once we enhance the network including the primary interactors of each of the proteins in the 1st order interactome model to develop the 2nd order interactome.

The 2nd order interactome segregated into 12 modules, among which 7 were larger groups, each containing 6 or more members while the rest were smaller groups with single nodes (Fig. 1c). The module configurations of the 2nd order interactome clearly indicate that most members of module 3 and 2 are also present in module 2 & 4 of the 1st order interactome, respectively. Module 3 in 2nd order interactome consists of Casp-8 and FADD, key players of the DD and DISC processes. Module 2 is now an integrated assembly formed from nodes of module 1 and 4 of 1st order interactome and consists of proteins taking part in the intermediate stage and the final execution of apoptosis. A closer look at the 2nd order interactome reveals the 4 major groups apart from 2 and 3. Modules 1, 8 and 9 have been built around few of the major anti-apoptotic proteins of the 1st order interactome for example Akt1, Birc 2, Birc3, Traf2 and TNF. We observed that in 1st order interactome TNF and Traf2 were included within module 3 whereas in 2nd order interactome TNF and Traf2 were placed in modules 8 and 9 respectively. TNF has been earlier reported to be involved in activation of apoptotic pathways\(^{32,33}\). But from our analysis we propose TNF may have some anti-apoptotic function based on its interactions with cytokines IL-1a and IL-6, that have been reported to be involved in cell survival\(^{34,35}\). The modules 8, 10 and 11 being influenced by the anti-apoptotic proteins form a significant part of this network that was not so prominent in the 1st order interactome. Other modules such as 4, 5, 6, 7 and 12 although consisting of fewer members in the context of our study, have the potential to embark into larger modules if an even bigger network is considered. This is simply because these modules consist of very important proteins that have been known to play pivotal roles in apoptosis.

Degree centrality is simply defined as the interaction score of a particular node within a network. The more interactions a node has within a group of nodes which are mutually interacting among each other, the higher its chance will be to form a module. Hence the community structure formation largely depends upon degree centrality of the nodes within a complex network. Casp3 and FADD were ranked among top 5 proteins when nodes of 1st and 2nd order interactomes were sorted in terms of degree centrality (Table 2). This signifies the pivotal role played by these two proteins in apoptosis and also gives us confidence to interpret the biological significance of modules from graph-theoretic measures. In Table 2 we see an interesting pattern. Nodes in the 1st order interactome that have positive degree centrality scores remained to be in the positive side in the 2nd order interactome. However, degree centrality of nodes that had 0 or negative values in the 1st order connectome either enhanced or got depreciated in 2nd
order. In order to explain this pattern we have to carefully analyze both the interactome models. Nodes having positive scores in the 1st order connectome interact not only maximally within their modules but also with other nodes in different modules. Hence, with the increase in number of interacting partners in the 2nd order interactome, the overall connectivity is enhanced for the constituent nodes. For example, Akt1 in the 1st order interactome interacts with several nodes of different modules but not consistently within one module. However, in the 2nd order interactome, the degree centrality of Akt1 increased and creation of a separate module involving Akt1 was observed\(^{36,37}\). Other nodes like TNF and Traf2 that were in one module in 1st order, increased their interactive partners and gained entry to bigger modules in 2nd order interactome. Nodes that have a predominant role to play in apoptosis maintained their modules and their degree centrality scores across both 1st and 2nd order interactome models.

In conclusion, we have outlined a robust method for studying the interactome underlying apoptosis following CHPV infection. This method may be used to study other metabolic pathways in order to yield important information about the strategic proteins of a specific network and the functionally important modules within the network. In the future, therapeutic targeting of particular proteins in case of various disease conditions needs to be investigated.

**Methods**

**Empirical data.** In an earlier study\(^{17}\), samples of Chandipura Virus was inoculated into Balb/c mouse intraperitoneally (i.p.) post-natal 10 days, at a plaque forming unit (pfu/ml) of \(3 \times 10^5\). The animals developed CHPV related symptoms of hind limb paralysis, high grade fever and severe weight loss, within 72–96 hours post infection leading to death. From immunoblotting and immunostaining analyses performed on the extracted brain tissue, we found over-expression of 6 proteins of the extrinsic apoptosis pathway: Fas, FADD (Fas-associated Death Domain), Caspase-8, Caspase-3 and XIAP (Poly ADP Ribose Polymerase-1). Our results were further validated using RNAi studies, ELISA assays and flow-cytometric analyses\(^{17}\). Table 3 enlists the protein names with their corresponding Uniprot IDs.

**Generation of meta-network.** STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is an open datasource providing information about protein-protein interactions based on experimental data, computational prediction methods and public database\(^{38}\). STRING 9.1 database contains information about more than 5.4 million proteins and >1100 organisms\(^{39}\). The database has two modes of applications: Protein-mode (for protein interactions) and COG-mode (for gene interactions). STRING imports protein association information from databases of physical interaction and curated biological pathway knowledge (MINT, HPRD, BIND, DIP, BioGRID, KEGG, Reactome, IntAct, EcoCyc, NCI-Nature Pathway Interaction Database, GO). Protein/genes are queried to the STRING database which as an output that represents the associations in the form of a graph network with nodes (proteins/genes) and edges (interactions). The edges are weighted, integrated and a confidence score is assigned to each of them based upon the evidence of the association obtained from experimental data, computational prediction and public data collection methods. Based on these edges are assigned various shades of color (blue)\(^{38}\). The prediction methods generally used in determining the interactions are:

**Neighbourhood.** This method of prediction utilizes the theory that protein interactions validated in case of one or more species is predicted to carry more weightage and confidence score.

**Gene Fusion.** Proteins fused in one genome are likely to be functionally linked and hence carry stronger association.

**Co-occurrence.** Occurrence of two proteins within the same metabolic pathway is predicted to functionally linked with each other. Hence their co-occurrence strengthens their confidence score.

**Co-expression.** Simultaneous expression of two proteins is also predicted to have strong interaction between them.

**Generation of 1st order interactome**
The 5 proteins identified through molecular analyses were queries in the STRING 9.1 that produced 26 interacting partners as an output from the *Mus musculus* database. The STRING 9.1 software defines significance of the interactions between various queried proteins in terms of confidence score. This confidence score is an empirical score defined by the number of citations and experimental evidence for a particular interaction. The highest (0.95) confidence score in the database, that defines the significance of interactions between various queried protein was chosen to extract interactomes in this study. Furthermore, we limited the number of interacting partner to 1000 in the provision for maximum interacting partners using active prediction methods as neighbourhood, gene fusion, co-occurrence and co-expression.
In order to investigate the structure of an even larger network we queried for interacting partners of all the 26 proteins obtained from the previous analysis. The 2nd order connectome in Fig. 1b was generated from STRING 9.1 database using same confidence score (0.95) as for the 1st order connectome and limiting to 1000 interacting partners.

**Graph theoretic analysis.**

The adjacency matrices for graph theoretic analysis were created from 1st and 2nd order interactomes. Visual Connectome analysis tool box in MATLAB was used to compute the modularity score and the degree centrality of all the nodes.

**Degree Centrality.**

Degree centrality is the property that defines the connectivity of particular node with other nodes of the same network. This means the higher number of connections of a particular node with other nodes in a network, higher is its degree centrality. The node with the highest degree centrality is the one through which maximum edges pass.

Degree centrality of a vertex \( v \), for a given graph \( G = (V,E) \) with \(|V|\) vertices and \(|E|\) edges is defined as

\[
Z(v) = \text{deg}(v)
\]

**Modularity.**

Modularity score is used to measure the community structure within a network. The value of modularity ranges between \([-0.5, 1)\) with 0 and negative values meaning a network with randomly assigned edges to positive values indicating highly communal structure. In a given graph \( G(V,E) \) which can be partitioned into two membership variables \( s \). If a node \( v \) falls into community 1 then \( s_v = 1 \) or else \( s_v = -1 \). An adjacency matrix may be denoted by \( A \), which says \( A_{vv} = 1 \) means there is a

| Protein   | Z score | Protein   | Z score | Protein   | Z score |
|-----------|---------|-----------|---------|-----------|---------|
| Casp3     | 1.9518  | Akt1      | 3.1326  | Ntrk1     | 0       | Ntrk2   | −0.4472 |
| Casp8     | 1.569143| Casp8     | 2.0533  | Shc1      | 0       | Fas     | −0.5019 |
| Casp9     | 1.224745| Ngf       | 1.7889  | Mdm2      | 0       | Fasl    | −0.5019 |
| Casp7     | 1.224745| Fadd      | 1.7339  | Sumol     | 0       | Tradd   | −0.5019 |
| Fadd      | 0.998545| Casp3     | 1.6202  | Map3k5    | 0       | Tnfsf10b| −0.5019 |
| Diablo    | 0.612372| Tnf       | 1.3618  | Dmap1     | 0       | Apaf1   | −0.5401 |
| Fas       | 0.427948| Il6       | 1.3618  | Hipk1     | 0       | Rela    | −0.3447 |
| Traf2     | 0       | Ripk1     | 1.0951  | Atrx      | 0       | Lep     | −0.5447 |
| Tradd     | 0       | Casp9     | 0.9001  | Adipoq    | 0       | Il1a    | −0.5447 |
| Ripk1     | 0       | Birc2     | 0.7771  | Ppurgc1a  | 0       | Cdkn1a  | −0.7627 |
| Tnf       | 0       | birc3     | 0.7771  | Clspn     | 0       | Foxo4   | −0.7627 |
| Tnfsf11a  | 0       | Cd40      | 0.7771  | bc1i11    | −0.0545 | Cylb    | −0.8213 |
| Fascl     | −0.14265| Tlr4      | 0.7771  | Stat3     | −0.0545 | Traf3   | −0.9991 |
| Ciflar    | −0.14265| Traf2     | 0.4563  | Foxo1     | −0.0545 | Bcap31  | −1.1407 |
| Tnfsf10b  | −0.14265| Tsc2      | 0.2996  | Traf3     | −0.111  | Cadd    | −1.1407 |
| Difb      | −0.24398| Mtor      | 0.2996  | Map3k7    | −0.111  | Ceeb1   | −1.1802 |
| Difb      | −0.24398| Xiap      | 0.18    | Ciflar    | −0.1825 | Difb    | −1.2601 |
| Ngrf      | −0.24398| Casp7     | 0.18    | Ikbkg     | −0.1825 | Difb    | −1.2601 |
| Akt1      | −0.24398| Diablo    | 0.18    | Fyn       | −0.4086 | Ticam1  | −1.8872 |
| Xiap      | −0.61237| Tnfsf11a  | 0.1369  | Bcl2      | −0.4086 |        |         |
| Birc2     | −0.61237| Fos       | 0.0908  | Cdkn1b    | −0.4086 |        |         |
| birc3     | −0.61237| Gsn       | 0       | Pten      | −0.4086 |        |         |
| Gsn       | −0.9759 | Daxx      | 0       | Rictor    | −0.4086 |        |         |
| Apaf1     | −1.22474| Ubc       | 0       | Furin     | −0.4472 |        |         |
| Bcap31    | −1.28384| Pfn1      | 0       | Trpv1     | −0.4472 |        |         |
| Daxx      | −1.28384| Ngrf      | 0       | Fas2      | −0.4472 |        |         |

Table 2. The protein names were arranged in decreasing order according to their respective degree centrality (Z) scores for both 1st (a) and 2nd order (b) interactome.

**Generation of 2nd order network**

In order to investigate the structure of an even larger network we queried for interacting partners of all the 26 proteins obtained from the previous analysis. The 2nd order connectome in Fig. 1b was generated from STRING 9.1 database using same confidence score (0.95) as for the 1st order connectome and limiting to 1000 interacting partners.

**Graph theoretic analysis.** The adjacency matrices for graph theoretic analysis were created from 1st and 2nd order interactomes. Visual Connectome analysis tool box in MATLAB was used to compute the modularity score and the degree centrality of all the nodes.

**Degree Centrality.** Degree centrality is the property that defines the connectivity of particular node with other nodes of the same network. This means the higher number of connections of a particular node with other nodes in a network, higher is its degree centrality. The node with the highest degree centrality is the one through which maximum edges pass.

Degree centrality of a vertex \( v \), for a given graph \( G = (V,E) \) with \(|V|\) vertices and \(|E|\) edges is defined as

\[
Z(v) = \text{deg}(v)
\]
The connection between nodes \( v \) and \( w \) and \( A_{vw} = 0 \) when there are no interactions. Modularity \((Q)\) is then defined as the fraction of edges that fall within community 1 or 2, minus the expected number of edges within communities 1 and 2 for a random graph with the same node degree distribution as the given graph.

The expected number of edges will be calculated using the concept of Configuration Models\(^4\). The configuration model is a randomized representation of a particular graph. Given a network with \( n \) nodes, where each node \( v \) has a node degree \( k_v \), the configuration model intercepts each edge into two halves, and then each half edge is defined as a stub, that is rewired randomly with any other stub in the network even allowing self loops. Hence even though the node degree distribution of the graph remains intact, the configuration model results in a completely random network. Let the total number of stubs be

\[
I_n = \sum_{v} k_v - 2m
\]

If two nodes \( v \) and \( w \) with node degrees \( k_v \) and \( k_w \), respectively are these nodes, then

\[
\text{Expectation of full edges between } v \text{ and } w = \frac{\text{Full edges between } v \text{ and } w}{\text{Total number of rewiring possible}}
\]

Total number of rewiring possible \( = \) Total number of stubs left after choosing a particular stub

\[
= I_n - 1 = I_n = 2m
\]

---

**Figure 2.** Representative plots for Community Structure \((Ci)\) Vs protein node numbers were plotted in this figure for 1\(^{st}\) (a) and 2\(^{nd}\) (b) order interactomes. The Ci value from each analysis was obtained from running the codes for 50 times. Thereafter the mean Ci values corresponding to the mean Modularity score \((Q)\) for each protein was plotted against the corresponding protein node number.
Expected number of full edges between \( v \) and \( w = k_v \times k_w / l_n = k_v \times k_w / 2m \) (5)

Modularity score is calculated as

\[ Q = \frac{1}{2m} \sum_{vw} \left( A_{vw} - \frac{k_v \times k_w}{2m} \right) (s_v s_w + 1) / 2 \] (6)

The above equation is valid for two-community structure and can be generalized into \( c \)-community structure.

| Protein Symbol | UniProt ID | Protein Symbol | UniProt ID |
|----------------|------------|----------------|------------|
| Casp3          | P70677     | map3k7         | Q923A8     |
| Casp8          | O89110     | Pfn1           | P62962     |
| Fas            | P25446     | Bcl2           | P10417     |
| Fadd           | Q61160     | bcl2l11        | O54918     |
| Xiap           | Q69898     | Stat3          | P42227     |
| Fasl           | P41047     | Furin          | P23188     |
| Traf2          | P39429     | Trpv1          | Q704Y3     |
| Tradd          | Q3U0V2     | Fsr2           | Q8CI80     |
| Ripk1          | Q60855     | Ntrk2          | P15209     |
| Cllar          | O35732     | Ngfr           | Q8CFT3     |
| Tnnsf10b       | Q9Q2ZM4    | Ntrk1          | Q3UF87     |
| Bcap31         | Q61335     | Shc1           | P98083     |
| Birc2          | Q62210     | Cdkn1a         | P39689     |
| birc3          | O08863     | Foxo4          | Q9WYH3     |
| Gen            | P13020     | Creb1          | Q01147     |
| Difa           | O54786     | Mdm2           | P23804     |
| Difb           | O54788     | Cdkn1b         | P46414     |
| Apaf1          | O88879     | Foxo1          | Q9R1E9     |
| Ngf            | P01139     | Tsc2           | Q7T2T21    |
| Akl1           | P31750     | Mtor           | Q9JLN9     |
| Daxx           | O35613     | Pten           | O08586     |
| Tnf            | P06804     | Rictor         | Q6Q86      |
| tnrsf1a        | P25118     | Sumo1          | P63166     |
| Casp9          | Q8C3Q9     | Map3k5         | Q14AY4     |
| Casp7          | P97864     | Dmap1          | Q9F44      |
| Diablo         | Q9JIQ3     | Hpk1           | O88904     |
| Fyn            | P39688     | Attr           | Q61687     |
| Cyld           | Q8O7Q2     | Rela           | Q04207     |
| Cd40           | P27512     | Fos            | P01101     |
| Traf3          | Q60803     | Adipoq         | Q60994     |
| Ubc            | P0CG50     | Lep            | P41160     |
| Ticam1         | Q8OUF7     | Il6            | P08505     |
| Cradd          | O88843     | Il1a           | P01582     |
| Ilkhg          | Q8VC91     | Ppargc1a       | O70343     |
| Traf1          | P39428     | Clspn          | Q80YR7     |
| Tlr4           | Q9QUK6     |                |            |

Table 3. The table enlists the Uniprot identification numbers for all the proteins which were used in our analysis.
$Q = \frac{1}{2m} \sum_{vw} \left( A_{vw} - \frac{k_v * k_w}{2m} \right) \delta (c_v, c_w) \tag{7}$

Modularity Optimization. The (6) can be re-written as:

$$Q = \frac{1}{4m} s^T B s$$

where $s$ is column vector whose elements are $s_v$ and $B$ is a symmetric matrix

$$B_{vw} = A_{vw} - \frac{k_v * k_w}{2m} \tag{9}$$

$B_{vw}$ is also referred to as the modularity matrix which will be having elements whose rows and columns sum upto 0, so that it always has an eigenvector $(1, 1, 1..)$ with eigen value 0. The algorithm that we used, initially divided the network into two communities and in further iterations the community structure is subdivided. For a group $g$ of size $n_g$ we can express the contribution to modularity as

$$\Delta Q = \frac{1}{2m} \left[ \frac{1}{2} \sum_{v,w \in g} B_{vw} (s_v s_w + 1) - \sum_{v,w \in g} B_{vw} \right]$$

which simplify to:

$$\Delta Q = \frac{1}{4m} \left[ \sum_{v,w \in g} B_{vw} - \delta_{vw} \sum_{k \in g} B_{vw} s_v s_w \right]$$

that can be expressed as

$$\Delta Q = \frac{1}{4m} s^T B^g s \tag{11}$$

where $\delta$ stands for Kronecker $\delta$ symbol and $B^g$ represents the $n_g \times n_g$ matrix with vertices $v, w$ in a particular group $g$ having values of

$$B^g = \left[ \sum_{v,w \in g} B_{vw} - \delta_{vw} \sum_{k \in g} B_{vw} \right] \tag{12}$$

Certainly (8) and (11) are similar and therefore spectral approach was applied to the generalized modularity matrix to maximize the values of $\Delta Q$. $\sum_{k \in g} B_{vw}$ for a complete network happens to be a symmetric matrix and thus (11) turns to nothing but (8). Once $\Delta Q$ is almost 0 for an indivisible network, then further subdividing beyond this point will not contribute to the increase in modularity value $Q$. This can be used to terminate community structure division.

The algorithm ran with the following theory: The modularity matrix, (9) was constructed for both interactomes and found the most positive eigenvalue and the corresponding eigenvector in each case. The algorithm divided the network into two parts depending upon the signs of the elements of the corresponding vectors, and then subdividing using the generalized modularity matrix (12). In the process $\Delta Q$ comes to 0 or negative at any stage of subdivision the algorithm left subgraph undivided. Hence, the algorithm would end at a certain point when the optimal network has been estimated. In order to fine tune this method of community structure optimization further, the Visual Connectome toolbox that we employed uses Kernighan-Lin algorithm.

References
1. Beltrao, P., Bork, P., Krokan, N. J. & van Noort, V. Evolution and functional cross-talk of protein post-translational modifications. *Mol Syst Biol* 9, 714 (2013).
2. Kim, J. & Copley, S. D. Inhibitory cross-talk upon introduction of a new metabolic pathway into an existing metabolic network. *Proc Natl Acad Sci USA* 109, E2856–2864 (2012).
3. Cusick, M. E., Klitgord, N., Vidal, M. & Hill, D. E. Interactome: gateway into systems biology. *Hum Mol Genet* 14 Spec No. 2, R171–181 (2005).
4. Meyer, M. J., Das, J., Wang, X. & Yu, H. INstruct: a database of high-quality 3D structurally resolved protein interactome networks. *Bioinformatics* 29, 1577–1579 (2013).
5. Stojimirovic, A. & Yu, Y. K. Information flow in interaction networks. *J Comput Biol* 14, 1115–1143 (2007).
6. Mason, O. & Verwoerd, M. Graph theory and networks in Biology. *IET Syst Biol* 1, 89–119 (2007).
7. Christensen, C., Thakar, J. & Albert, R. Systems-level insights into cellular regulation: inferring, analysing, and modelling intracellular networks. *IET Syst Biol* 1, 61–77 (2007).
8. Chang, X., Xu, T., Li, Y. & Wang, K. Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of ‘date’ and ‘party’ hubs. *Sci Rep* 3, 1691 (2013).
9. Rubinov, M. & Sporns, O. Weight-conserving characterization of complex functional brain networks. *Neuroimage* 56, 2068–2079 (2011).
Graph theoretic network analysis reveals protein pathways underlying cell death following neurotropic viral infection. Sci. Rep. 5, 14438; doi: 10.1038/srep14438 (2015).

Acknowledgement
The study was supported by NBRC Core funds and by grants, (BT/PR/7907/MED/29/702/2013) and a Tata Innovation Fellowship (BT/HRD/35/01/02/2014) to Anirban Basu from the Department of Biotechnology (DBT), Ministry of Science and Technology Government of India; Ramalingaswami fellowship, DBT (BT/RLF/Re-entry/31/2011) and Innovative Young Bio-technologist Award (IYBA), DBT (BT/07/IYBA/2013) to Arpan Banerjee(AB). AB also acknowledges the support of Centre of Excellence in Epilepsy and MEG (BT/01/COE/09/08/2011) from DBT.

Author Contributions
S.G., A.Basu and A.Banerjee conceived the study; S.G., G.V.K and A.Banerjee prepared the figures; S.G. and A.Banerjee wrote the manuscript; S.G., G.V.K., A.Basu and A.Banerjee reviewed the manuscript.

Additional Information
Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Ghosh, S. et al. Graph theoretic network analysis reveals protein pathways underlying cell death following neurotropic viral infection. Sci. Rep. 5, 14438; doi: 10.1038/srep14438 (2015).
