k-Shell decomposition reveals structural properties of gene co-expression network for neurodevelopment

A. Ercument Cicek[1,2]

[1] Department of Computer Engineering, Faculty of Engineering, Bilkent University, Ankara, Turkey

[2] Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

Correspondence: cicek@cs.bilkent.edu.tr

Abstract: Neurodevelopment is a dynamic and complex process, which involves interactions of thousands of genes. Understanding the mechanisms of brain development is important for uncovering the genetic architectures of neurodevelopmental disorders such as autism spectrum disorder and intellectual disability. BrainSpan dataset is an important resource for studying the transcriptional mechanisms governing neurodevelopment. It contains RNA-seq and microarray data for 13 developmental periods in 8-16 brain regions. Various important studies used this dataset; in particular to generate gene co-expression networks. The topology of the BrainSpan gene co-expression network yielded various important gene clusters, which are found to play key roles in diseases. In this work, we analyze the topology of the BrainSpan gene co-expression network using the k-shell decomposition method. k-Shell decomposition is an unsupervised method to (1) decompose a network into layers (shells) using the connectivity information and (2) to detect a nucleus which is central to overall connectivity. Our results show that there are 267 layers in the BrainSpan gene co-expression network. The nucleus contains 2584 genes, which are related to chromatin modification function. We compared and contrasted the structure with the

2015), (iii) intellectual disability (Gudenas et al., 2015; Riazuddin et al., 2016) and (iv) Parkinson's disease (Liscovitch and French, 2014).

Despite its central role in neurodevelopmental disorder research, which uses the connectivity patterns and topological properties of the BrainSpan co-expression network, to the best of our knowledge, there is no detailed work on analyzing the structural properties of the BrainSpan network itself. This is in contrast to other important complex networks, for which numerous studies have been done such as, the Internet (Calvert et al., 1997; Albert et al., 1999; Cohen et al., 2000) and from the biology domain, yeast protein interaction/co-expression networks (Bu et al., 2003; Jeong et al., 2001; Van Noort et al., 2004). In this work, we analyze the structural properties of the BrainSpan co-expression network, using the k-shell decomposition method, which is a widely used method for finding structurally important nodes in complex networks (Seidman, 1983; Bader and Hogue, 2003; Dorogovtsev et al., 2003; Wuchty and Almaas, 2005; Alvarez-Hamelin et al., 2008; Shao et al., 2009; Kitsak et al., 2010). k-Shell decomposition peels the layers of a network, starting from the least connected (shell) till a dense core (nucleus) with no nodes with less than k edges remain. This way, nodes are classified into groups with different functional roles. This method is preferred over degree-based analyses as it is possible to obtain similar degree distributions with very different network topologies with the latter (Doyle et al., 2005; Carmi et al., 2003). k-Shell decomposition can be computed in polynomial time unlike finding cliques of size k which is another way of detecting densely connected subgraphs. In the context of biology, k-shell decomposition has been used to (1) predict protein function (Altaf-Ul-Amine et al., 2003), (2) analyze cancer mutation rates for cancer in protein domain co-occurance networks (Emerson et al., 2015), and (3) analyze

more). A total of 16 cortical and subcortical brain regions are considered. During the 4-10 PCW, the dataset contains samples from occipital cerebral wall, frontal cerebral wall, parietal cerebral wall, temporal cerebral wall, upper rhombic lip, hippocampal anlage, medial ganglionic eminence, lateral ganglionic eminence, diencephalon, dorsal thalamus, ventral forebrain, caudal ganglionic eminence. From 10 PCW till 82 PY, the dataset contains samples from hippocampus, mediodorsal nucleus of the thalamus, amygdala, striatum, orbital prefrontal cortex, dorsal prefrontal cortex, ventral prefrontal cortex, medial prefrontal cortex, posterior inferior parietal cortex, primary auditory cortex, superior temporal cortex, inferior temporal cortex, primary motor cortex, primary somatosensory cortex, primary visual cortex, cerebellar cortex, (Kang et al., 2011; Willsey et al., 2013).

In this study, we used the microarray data and removed the genes that are not brain expressed. For each gene, expression values were obtained and ordered by brain region and the age when brain sample is obtained. We included all available brain regions and developmental stages. Note that the latter corresponds to the temporal dimension (corresponding to the age of the donor).

2.2. Co-expression Network Construction

Gene pairs are considered co-expressed, if the absolute Pearson's correlation coefficient between their expression patterns, is larger than 0.7. This threshold was also used by Willsey et al., (2013) and Liu et al., (2014). The resulting network is binary and bidirectional.

2.3. k-Shell Decomposition

k-Shell decomposition is an unsupervised method for discovering structurally different layers of a network. The method starts with $k = 1$. Then, removes all the nodes with a

Enriched terms are found using Fisher's exact test. A modified version of the Fisher's exact test which also considers a background distribution based on many random runs also provides an adjusted p value and a corresponding z-score. Finally, a combined score is provided which combines the two p values found. For each enrichment analysis we perform, we report the top 5 GO terms, ordered with respect to the standard Fisher's exact test results.

3. Results

BrainSpan gene co-expression network contains 8007 nodes (genes) and 1562725 edges (co-expressed pairs of genes, self loops removed, threshold = 0.7). Applying the k-shell decomposition onto this network yielded 267 shells (layers). The network is visualized in Figure 2, where shade of each node denotes its shell, darker denoting a deeper shell. Following Carmi et al. (2003), we analyzed (1) crust sizes and (2) sizes of the largest connected component and the second largest connected component in each crust and depicted the results in Figure 3. The number of nodes increases when k is increased, but the rate slows down as k gets close to level of the nucleus. The size of the largest connected component also follows a similar pattern. When the nucleus is added, we see a peak in the size of the deepest crust and the largest connected component in this crust. We observe the percolation transition at k = 8. Percolation transition is the point after which a large connected component is formed and the network is "mostly" connected over long ranges. The transition is similar to that of the autonomous system level Internet (AS) found in Carmi et al. (2003). After k = 8, the largest connected component size doubles and the average distance between nodes peaks as shown in Figure 4. At k = 8 the size of the second largest connected component is also the largest (see Figure 3). Again, similar to the AS, the crust size converges, as it gets deeper. Just before adding

with the KS statistic of the actual data, a p value is obtained. If $p < 0.1$, the test concludes that the data is inconsistent with power law. We obtained a p value of 0.0.

We have investigated the functional meaning of cluster of genes that form the nucleus. GO enrichment analysis shows that the top enriched biological process term is the chromatin modification, followed by covalent chromatin modification and histone modification (Table 1). Disruption of chromatin modification has been implicated as an important player in autism etiology. 9 out of 107 predicted autism risk genes are chromatin modifiers and autism risk genes are found to tightly interact with chromatin modifier genes in a transcription factor regulation network (De Rubeis et al., 2014). We also see "regulation of neuron projection development" term at the 6^{th} ranking. Neuron projection development is also recently implicated as a possible risk source for ASD (Liu et al., 2014).

We also analyzed the connected crust above the nucleus in the same manner and as shown in Table 2. In this region, enriched terms are less specialized than the ones obtained for the nucleus. For instance, we obtained the "behavior" GO term as one of the top terms. "Behavior" is right below the root of the GO biological process term, so it includes many genes and it is very general. Only relevant term obtained in this category is the "synaptic transmission" term.

Finally, analyzing the isolated component, which needs to connect to the rest of the crust via nucleus, yields an interesting result. The top terms obtained are related to "neurotrophin signaling pathway" which plays an important role in the growth and survival of the neurons (Reichardt, 2006). We also observe many immune system related terms like "regulation of inflammatory response" and "neutrophil mitigation". Separation of these functions suggests that the "survival system" of the cell has a

very tight connections to each other each with at least 1089 connections on average. This is another indication that chromatin modification is a complex task and plays an important role in healthy brain development. Finally, the isolated component, which needs the nucleus to connect to the other genes in the crust, is responsible for the survival of the neurons. This is an interesting finding, which indicates that the growth and defense mechanism of the neurons have a different agenda compared to the rest of the crust, and they only interact with the nucleus.

In conclusion, the neurodevelopment is a complex task which involves ~8000 genes which are interacting in a complex network. k-Shell decomposition has helped uncovering the structural components of this complex process in an unsupervised way. We foresee that this initial analysis is going to pave the way towards more detailed analyses. One future direction is going to be focusing on specific time periods and brain regions, which are implied as important for specific diseases. For instance, the prefrontal cortex and primary motor-somatosensory cortex during mid-fetal development has been marked as an important window for autism, as autism genes are clustered there (Willsey et al., 2013; Liu et al., 2014). Understanding the network topology in that region and comparing/contrasting it with other regions has potential to reveal the functions affected autism.

References

Albert R, Jeong H, Barabási AL. Internet: Diameter of the world-wide web (1999). Nature 401(6749):130-1.

Altaf-Ul-Amine M, Nishikata K, Korna T, Miyasato T, Shinbo Y, Arifuzzaman M, Wada C, Maeda M, Oshima T, Mori H, Kanaya S. Prediction of protein functions based

De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Cicek AE, Kou Y, Liu L, Fromer M, Walker S (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. 2014 Nov 13;515(7526):209-15.

Doyle JC, Alderson DL, Lun L, Low S, Roughan M, Shalunov S, Tanaka R, Willinger W. The robust yet fragile" nature of the Internet (2005). Proc Natl Acad Sci USA 102:14497–14502.

Emerson AI, Andrews S, Ahmed I, Azis TK, Malek JA. K-core decomposition of a protein domain co-occurrence network reveals lower cancer mutation rates for interior cores (2015). Journal of clinical bioinformatics. 5(1):1.

Gudenas BL, Wang L. Gene coexpression networks in human brain developmental transcriptomes implicate the association of long noncoding RNAs with intellectual disability (2015). Bioinformatics and biology insights, 9(Suppl 1):21.

Gulsuner S, McClellan JM. De novo mutations in schizophrenia disrupt genes co-expressed in fetal prefrontal cortex (2014). Neuropsychopharmacology 39(1):238-9.

Gulsuner S, Walsh T, Watts AC, Lee MK, Thornton AM, Casadei S, Rippey C, Shahin H, Nimgaonkar VL, Go RC, Savage RM. Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network (2013). Cell 154(3):518-29.

Hormozdiari F, Penn O, Borenstein E, Eichler EE. The discovery of integrated gene networks for autism and related disorders (2015). Genome research 25(1):142-54.

Jeong H, Mason SP, Barabási AL, Oltvai ZN. Lethality and centrality in protein networks (2001). Nature 411(6833):41-2.

Reichardt LF. Neurotrophin-regulated signalling pathways (2006). Philos Trans R Soc Lond B Biol Sci. 361(1473): 1545–1564.

Riazuddin S, Hussain M, Razzaq A, Iqbal Z, Shahzad M, Polla DL, Song Y, van Beusekom E, Khan AA, Tomas-Roca L, Rashid M. Exome sequencing of Pakistani consanguineous families identifies 30 novel candidate genes for recessive intellectual disability (2016). Molecular Psychiatry, 10.1038/mp.2016.109

Seidman SB. Network structure and minimum degree (1983). Social Networks 5 269–87.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks (2003). Genome Research 13(11):2498-504

Shao J, Buldyrev S, Braunstein L, Havlin S, Stanley HE. Structure of shells in complex networks (2009). Phys. Rev. E 80:036105.

Tebbenkamp AT, Willsey AJ, State MW, Šestan N. The developmental transcriptome of the human brain: implications for neurodevelopmental disorders (2014). Current opinion in neurology, 27(2):149.

Van Noort V, Snel B, Huynen MA. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model (2004). EMBO reports 5(3):280-4.

Willsey AJ, Sanders SJ, Li M, Dong S, Tebbenkamp AT, Muhle RA, Reilly SK, Lin L, Fertuzinhos S, Miller JA, Murtha MT. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism (2013). Cell 155(5):997-1007.

Wuchty S, Almaas E. Peeling the yeast protein network (2005). Proteomics 5: 444–9

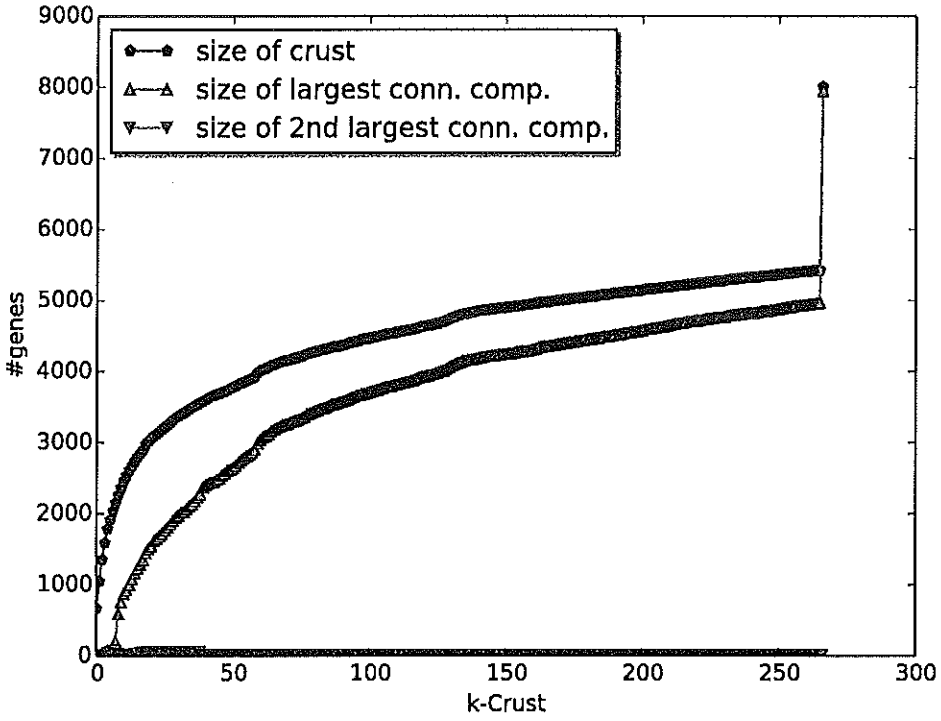

Figure 3. Crust Size Analyses. For each crust, figure shows the number of the nodes in a crust (blue), the size of the largest connected component in that crust (red) and the size of the second largest connected component in that crust (green).

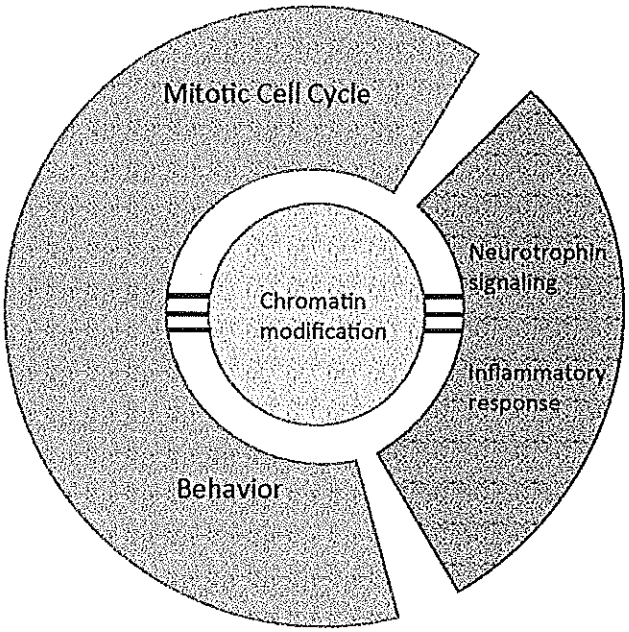

Figure 6. Medusa model components of the decomposed the BrainSpan co-expression network and the enriched GO Biological Process terms. The nucleus: chromatin modification, the connected component on the crust above nucleus: mitotic cell cycle and behavior and the isolated component: neurotrophin signaling pathway.

Table 1. GO Biological Process Enrichment results for the nucleus of BrainSpan gene co-expression network. Only top 5 enriched terms are shown, ordered with respect to p-value. A modified version of the Fisher's exact test is used to calculate adjusted p value based on a background distribution. Combined score combines two scores. Enrichr software is used to obtain this table.

Term	Overlap	P-val	Adjusted P-val	Z-score	Combined Score
chromatin modification	121/475	1.8710	8.83341E	-2.40429	27.9787

		E-06			
organelle fission (GO:0048285)	134/325	8.36204 E-06	0.0065487	-2.33607	11.746928

1

2 Table 3. GO Biological Process Enrichment results for the nucleus of BrainSpan gene
3 co-expression network. Rest of the description is same as Table 1.

Term	Overlap	P-value	Adjusted P-value	Z-score	Combined Score
neurotrophin TRK receptor signaling pathway (GO:0048011)	18/274	0.00010347	0.16891 4534	-2.4328 21159	4.32643 7695
neurotrophin signaling pathway (GO:0038179)	18/278	0.00010 22847	0.16891 4534	-2.4298 97065	4.32123 7596
regulation of inflammatory response (GO:0050727)	15/247	0.00083354	0.36056 9806	-2.4194 03275	2.46795 9981
cellular response to transforming growth factor beta stimulus (GO:0071560)	12/166	0.00067798	0.36056 9806	-2.2812	2.32700 1483
response to transforming growth	12/166	0.0006	0.36056	-	2.32432