BIOINFORMATIC ANALYSIS USING COMPLEX NETWORKS AND CLUSTERING ON PROTEINS LINKED WITH ALZHEIMER'S DISEASE

[Suthinan Rujirapipat, Ken McGarry and David Nelson]

***Abstract*—the detection of protein complexes is an important research problem in bioinformatics, which may help increase our understanding of the biological functions of proteins inside our body. Moreover, new discoveries obtained from identification of protein complexes may be considered important for therapeutic purposes. Several proteins linked with Alzheimer's disease were investigated. By observing the connectivity between proteins using computational methods such as graph theory and clustering, we can uncover previously unknown relationships that are useful for potential knowledge discovery. Furthermore, we demonstrate how Markov Clustering (MCL) and the Molecular Complex Detection (MCODE) algorithm identify interesting patterns from the protein-protein interaction data related to Alzheimer's disease.**

Keywords*—protein network, clustering, styling, insert *(key words)

I. Introduction

The use of various computational techniques to build and analyse networks of protein-protein interactions has begun to rise over the recent years [1, 2]. Using graph based structures commonly practiced in many scientific fields, the protein interactions and their properties can be studied using several algorithms related to the graph theory discipline [3]. Many interesting medical discoveries have been made using protein interactions networks [4, 5, 6]. Furthermore, there is a progressive accumulation of publically available protein interaction data [7].

This raises the popularity and application of network analysis of protein interaction to independent researchers from various scientific areas[23,24]. Following the work of McGarry et al. (2015), the authors conducted extensive research on the application of graph-based model techniques for possible identification of candidate drug re-positioning.

Alzheimer's disease (AD) is the most common form of dementia and is an irreversible, progressive brain disorder (National Institute on Aging, 2011). Alzheimer's disease will slowly destroy person's memory, intelligence, and the ability to complete even the most ordinary tasks (National Institute on Aging, 2011). Dementia is the loss of cognitive functioning, such as thinking, reasoning, and remembering. Scientists are still unsure what causes Alzheimer's disease.

However, major expected causes include plaques, tangles in the brain tissues, and the loss of interconnectedness between nerve cells. Most of the biological processes in our body can be extremely difficult to understand without extensive analysis of vast numbers of interactions and components [19].

University of Sunderland, Faculty of Applied Sciences
City Campus, Dale Building, Wharncliffe St, Sunderland, SR1 3RG, UK
ken.mcgarry@sunderland.ac.uk

II. Graph Theory and Protein Interactions

Graph theory is the study of connectivity patterns, typically describing pairwise relationships between objects[31]. A graph is defined by a set of vertices (nodes) and edges (lines) that connect the vertices together. A mathematical structure used to represent the whole graph is as follows:

$$Graph\ G = (V, E, \mu(V), \mu(E)); E = \{(u, v)\ u, v \in V\}$$

Definition 2.1: Formally, a graph $Graph\ G = (V, E, \mu(V), \mu(E))$; is a mathematical structure consisting of a set V of vertices (also commonly called nodes) and a set E of edges (also commonly called links), where elements of E are unordered pairs {u,v} of distinct vertices u,v $\in$ V. μ (V) is a labelling function that associates a unique label for each node in V, and μ (E) is a labelling function that associates a unique label for each edge in E [20]. In figure 1, a simple protein-protein interaction network is represented. Protein A interacts with B, protein B interacts with proteins A, C, and D, and protein D interacts with proteins B and C.

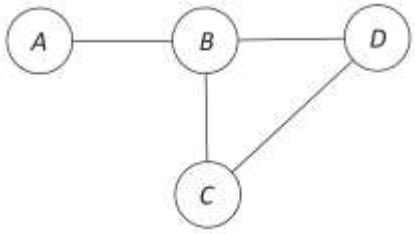


Fig 1 A simple protein-to-protein interaction network between four interacting proteins

There are many applications that can be described using a set of nodes and edges, for example transport networks, political affiliations, financial interactions, scientific collaborations and social networks in particular have received increased attention [21]. Vertices are used to indicate people while edges are used to represent the friendship relation between people. The very same concept can also be used to describe protein-protein interaction networks. Vertices are used to represent proteins while edges illustrate the interactions between proteins.

Interactomics is a discipline at the intersection between bioinformatics and biology. Interactomics focuses on the study and the analysis of interactions and the consequences of those interactions between and amongst proteins. Activities of interactomics include: the study of protein-

protein interaction networks (PINs), the modelling, storage, and retrieval of protein-protein interactions (PPI). Interactomics is an essential key to explaining and interpreting protein interactions, which may involve two or more proteins, founding the protein complexes.

The protein complex is a group of two or more proteins that share the same biological goal [33,34]. Different protein complexes have different protein functions in cell operation (Cannataro et al., 2011). Since this research project will explore how data mining (or graph mining) can be used to find the essential protein complexes, therefore, the computational methods provided by interactomics can be considered as the appropriate approach. Protein-protein interaction (PPI) is the physical interaction established between two or more proteins. PPI is the result of a biochemical event and/or electrostatic forces [2,32]. PPIs are usually stored in specialised databases where each interaction is represented by a pair of interacting proteins (Pi, Pj). PPI can be graphically represented using a specialised network graph, known as a protein-protein interaction network (PIN).

III. MCL and MCODE algorithms

Markov Clustering algorithm (MCL) simulates a flow on the graph by using the successive powers of the associated adjacency matrix [13]. An *inflation* is then applied to enhance the difference between the regions of strong or weak flow in the graph at each iteration. The whole process of MCL converges towards a partition of the graph, with a set of high-flow clusters separated by boundaries with no flow. The value of *inflation* has a direct influence on the number of clusters. However, while the MCL is relatively simple to use and elegant as shown by its popularity in bioinformatics due to its effective and noise tolerant nature of the algorithm, the MCL can be very slow and also prone to output too many clusters. Dongen based this conclusion on the following results; in social network clustering application MCL took 1.2 hours to cluster 76,000 nodes of social network [7], and in protein-protein interaction network also MCL generated 1,416 clusters on 4,741 proteins and 15,148 interactions of protein-protein interaction network of Yeast.

Molecular Complex Detection (MCODE), is used to detect densely connected regions within a graph. First proposed by Bader and Hogue [3], MCODE is one of the first computational methods to predict protein complexes. MCODE assigns a weight to each vertex (node), in conjunction to its local neighbourhood density. Next, it recursively moves outward starting from the top-weighted vertex. The including cluster vertices are controlled by a given threshold. This threshold corresponds to a user-defined percentage of the weight of the top-weighted vertex. MCODE also has optional post-processing options that can filter out non-dense subgraphs and generate overlapping clusters. MCODE can be very beneficial for researchers who are interested in the role of a particular within the cell and its interactions with others proteins [18]. This is considered one of the main advantages of MCODE algorithm. However, MCODE also has a drawback in term of the strictness of MCODE. MCODE tends to miss smaller molecular complexes, especially if the protein interaction data is noisy such as data from the experimental wet lab, such as those generated from mass spectrometer that low-confidence edges in protein-protein interaction network must be discarded before performing MCODE analysis in order to obtain a better result.

IV. Methods

The Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) was used as part of the data collection process. STRING is a database of predicted interactions including protein to protein, protein to DNA, and DNA to DNA [35]. The STRING can be accessed directly using the internet (http://string.embl.de/). Protein interaction data can be obtained by specifying a protein identifier. The interaction unit in STRING is the functional association, a productive functional relationship between two proteins. All the associations are stored with confidence score based on functional associations. The confidence scores are derived from the benchmarking results against a common reference set of trusted protein associated, such as those from KEGG database (Kyoto Encyclopaedia of Genes and Genomes[35]. the confidence score of the interactions from STRING will be strict between 0.999 – 0.900. This is to ensure that the predicted interactions obtained from STRING will be reliable enough.

The flat file containing protein interactions obtained from the STRING database was used to construct protein-protein interaction networks (PINs). The obtained flat file will contain several columns of data related to the pair-wise relationships between two proteins.

Using APP as a starting protein, 20,423 protein interaction pairs were downloaded from the STRING database. The calculated confidence scores given by the STRING database for every predicted protein interaction used in this example are between 0.999 – 0.900.

The R language was used along with the RStudio programming environment on an Intel Xenon CPU, 64-bit with dual processors (3.2GHz) and 128 GB of RAM. The R code was not compiled or optimized. R can be considered as the new de facto standard tool used in statistical research. R is highly versatile and highly expandable; over 5,000 packages have been developed by the highly active R community of researchers and developers. We used the *igraph,* and *ProNet* packages.

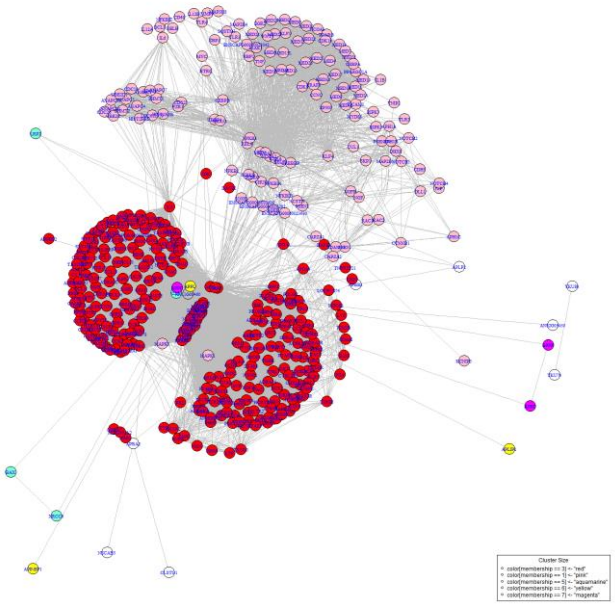
The *igraph* package is one of the many existing extension packages for R used in network sciences. It provides tools to build, import, manipulate, and visualise graphs of the software. Since the software must be able to produce protein-protein interaction network, therefore, the *igraph* is needed as part of the development. The *igraph* package was used in conjunction with the *ProNet* package to find and highlight visual representation of protein complexes. The *ProNet* package provides functions for building, visualisation, and analysis of biological network. *ProNet's*

underlying data structures are based on graphs constructed from the *igraph* package.

v. Results and Discussion

The graphical representation of the protein-protein interaction networks with additional details generated from MCL and MCODE. Fig 2 and Fig 3 show the discovered clusters for the MCL and MCODE algorithms, respectively. For simplicity, only the top five of the largest protein complexes were investigated.

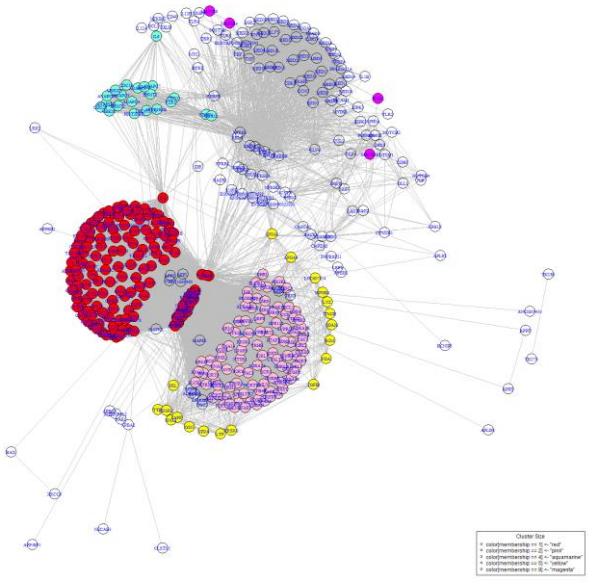
Markov Clustering (MCL), Top 5 largest clusters



The MCL returned nine clusters. The clustering coefficient of MCL is equal to 0.36847. After validating the top five of the largest clusters using GO Term Finder, the most probable cellular process associated with cluster (A) is *G-protein coupled receptor signalling pathway*, involving 142 proteins with *p-value* of 4.87e-174, while 45 proteins of cluster (B) are involved in *positive regulation of macromolecule metabolic process*, with *p-value* of 1.66e-36, whereas cluster (C) has no known association, and no significant terms were found for cluster (D) and cluster (E).

An important aspect of the clustering analysis of protein-protein interaction network is the validation of the clustering results. This is performed in order to investigate whether the returned results are biologically significant or not (Boyle et al., 2004). Such validation can be achieved by the combination of suitable metrics and on-line available tools (Pizzuti et al., 2012).

Molecular Complex Detection Algorithm (MCODE)



The Molecular Complex Detection (MCODE), which returned fourteen clusters. The clustering coefficient of MCODE is equal to 0.99967. The most probable cellular process returned by GO Term Finder for cluster (A) is G-protein coupled receptor signalling pathway, involving 79 proteins, with p-value of 2.75e-102, 59 proteins of cluster (B) are also found to participate in G-protein coupled receptor signalling pathway same as cluster (A), with p-value of 4.95e-75, while 25 proteins of cluster (C) are involved in regulation of transcription from RNA polymerase II promoter, with reported p-value of 3.55e-30, whereas 7 proteins of cluster (D) are found to involve in protein K11-linked ubiquitination, with p-value of 3.22e-19, however, no significant terms were found for proteins of cluster (E).

Complex	Number of protein	Colour	Protein name (omitted)
A	276	Red	{CXCL12,PIK3CA,EDN1,BACE1,CCKBR,CXCR1,CCR7,CXCL10,CXCR2, APP,AGTR2,CCKAR,TACR1,KALRN,TACR2,APBA3,AGTR1, ... BACE2}
B	135	Pink	{PSEN2,PSENEN,IL12B,BTRC,RELA,PSEN1,NCSTN,APH1A,TRAF6,NCO A3,CUL1,NFKBIA,IRAK1,NFKB1,NFKB2,MED1,CHUK, ... KCNIP3}
C	4	Aqua	{CLU,BAX,LRP2,XRCC6}
D	3	Yellow	{APPL,APLIP1,APP-BP1}
E	3	Magenta	{APPC,APPF,APPD}

TABLE I. RESULT FROM MARKOV CLUSTERING ALGORITHM.

Both algorithms agree that the most significant cluster of proteins with a common cellular process is the cluster that participates in G-protein coupled receptor signalling pathway. The validity of such cluster is further supported by the low p-values (MCL:4.87e-174; MCODE: 2.75e-102 & 4.95e-75) and as illustrated by the results from MCL cluster (A) and MCODE cluster (A) and (B). For the other clusters, the variation in the results returned may be associated with varying cellular processes involved in the same proteins of consideration. This suggests that some of the existing proteins may participate not only in a single cluster (as represented in this work) but also in multiple-clusters as well.

TABLE II. RESULT FROM MCODE ALGORITHM.

Complex	Number of protein	Colour	Protein name (omitted)
A	142	Red	{CXCL12,CXCR1,CCR7,CXCL10,CXCR2,APP,AGTR2,CCR9,GNB1,SST,PO MC,CCL5,AGT,PDYN,CCR10,PPBP,CXCL13,NPY1R,BDKRB2, … GALR1}
B	98	Pink	{PIK3CA,EDN1,CCKBR,CCKAR,TACR1,TACR2,AGTR1,EDN2,NMB,TACR 3,HCRT,OXTR,GRPR,TRH,NMBR,EDN3,GNRH1,GCG,OXT, … GRP}
C	51	Aqua	{RELA,NCOA3,NFKB1,MED1,CREBBP,CDK8,MED10,PPARG,PPARGC1A, MED11,MED16,MED18,NCOA2,CDK19,MED24,MED6,MED13, … MED23 }
D	21	Yellow	{CDC27,FZR1,ANAPC10,CDC16,ANAPC7,ANAPC4,ANAPC5,ANAPC1,CDC 26,UBE2C,ANAPC11,UBE2D1,UBBP4,UBA52,EHMT1, … EHMT2}
E	19	Magenta	{ITM2B,APOA1,SNCA,IAPP,TTR,LOC607874,B2M,CRSP-3,CRSP-2,LTF ,FGA,GSN,TGFBI,APOA4,PRL,NPPA,MFGE8,ODAM,LYZ}

An important aspect of the clustering analysis of protein-protein interaction network is the validation of the clustering results. This is performed in order to investigate whether the returned results are biologically significant or not.

One of the most important metric used to validate the clustering results is the clustering coefficient (or transitivity). The clustering coefficient is calculated by considering the nodes within a network and the way nodes linked together[43,44]. The clustering coefficient is used to determine the quality of the clustering results. The definitions for the clustering coefficient of a node and clustering coefficient are given below.

Definition : *(p-value) Given a cluster of size n with m proteins sharing a particular biological annotation, then the probability of observing m or more proteins that are annotated with the same GO term out of those n proteins, according to the Hypergeometric Distribution, is:*

$$p - value = \sum_{i=m}^{n} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Where N is the number of proteins in the database with M of them known to have that same annotation. Thus, the closer the p-value to zero, the more significant the associated GO term.

The first algorithm to be validated is the Markov Clustering (MCL). The MCL returned nine clusters. The clustering coefficient of MCL is equal to 0.36847. After validating the top five of the largest clusters using GO Term Finder, the most probable cellular process associated with cluster (A) is G-protein coupled receptor signalling pathway, involving 142 proteins with p-value of 4.87e-174, while 45 proteins of cluster (B) are involved in positive regulation of macromolecule metabolic process, with p-value of 1.66e-36, whereas cluster (C) has no known association, and no significant terms were found for cluster (D) and cluster (E).

The second algorithm is the Molecular Complex detection (MCODE), which returned fourteen clusters. The clustering coefficient of MCODE is equal to .99967. The most probable cellular process returned by GO Term Finder for cluster (A) is *G-protein coupled receptor signalling pathway*, involving 79 proteins, with *p-value* of 2.75e-102, 59 proteins of cluster (B) are also found to participate in *G-protein coupled receptor signalling pathway* same as cluster (A), with *p-value* of 4.95e-75, while 25 proteins of cluster (C) are involved in *regulation of transcription from RNA polymerase II promoter*, with reported *p-value* of 3.55e-30, whereas 7 proteins of cluster (D) are found to involve in *protein K11-linked ubiquitination*, with p-value of 3.22e-19, however, no significant terms were found for proteins of cluster (E).

The last important aspect of validation is done through biological validation. This analysis is performed in order to verify whether the obtained proteins in a cluster correspond to a biological function or not. This is achieved using the known biological associations from the Gene Ontology Consortium Online Database [5].

The Gene Ontology (GO) database provides three classes of known associations. 1) Molecular function, describing the tasks done by individual gene products (e.g., DNA binding) 2) Cellular component, encompassing subcellular structures, locations, and macromolecular complexes (e.g., nucleus) 3) Biological process, describing broad biological goals (e.g., mitosis) For this example, only the third class (biological process) will be used to exemplify the validation process. Another important metric for clusters validation that GO Term Finder can generate is the hypergeometric p-value. This is a measure of the functional homogeneity of a cluster and is considered useful in enrichment analysis In this example, a protein cluster may be associated with a list of genes, each corresponding to a particular protein in the cluster. The p-value is used to determine the statistical significance of a particular GO term with a group of genes in the list [42]

VI. Conclusions

All the presented algorithms showed that groups of highly connected proteins or protein complexes involved in common cellular processes are presented in protein-protein interaction networks. The computational methods using topological analysis of network (graph mining) can be considered valuable in identifying useful information in protein-protein interaction, such as network components and the connection amongst such components. This paper also illustrates how the developed tool can be used to analyse protein-protein interactions related to Alzheimer's disease, which may lead to better understanding dynamics of the disease. However, all the algorithms used in this example have some parameters that influence the number, the size, the density, and the structure of the clusters produced. Thus, the use of different algorithms in conjunction with different input parameters will yield drastically different results as supported by our work.

A single protein may participate in more than one cellular process. This, in turn, making the considering protein belongs to more than one protein complex, which shares the same cellular process. This implies that in order to achieve an even better understanding of the dynamics of the disease, multiple cluster assignment to proteins must be used. Another implication of research findings is that each result generated by different algorithm with different input parameters can generate a drastically different result from the same data set. A method that can fine-tune the input parameters in relation to the data set is highly encouraged and must be developed in order to yield even better accuracy.

References

[1] Aittokallio, T. (2006) Graph-based methods for analysing networks in cell biology. Briefings in Bioinformatics, 7 (3), pp.243–255.

[2] Alberts, B. (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell, 92 (3), pp.291–294.

[3] Bader, G.D. & Hogue, C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC bioinformatics, 4 (1), p.2.

[4] Barabási, A.-L. & Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. Nature Reviews Genetics, 5 (2), pp.101–113.

[5] Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M. & Sherlock, G. (2004) GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics (Oxford, England), 20 (18), pp.3710–3715.

[6] Brohee, S. & Van Helden, J. (2006) Evaluation of clustering algorithms for protein-protein interaction networks. BMC bioinformatics, 7 (1), p.488.

[7] Chatr-aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L. & Cesareni, G. (2007) MINT: the Molecular INTeraction database. Nucleic Acids Research, 35 (Database issue), pp.D572–574.

[8] Chen, J.Y., Shen, C. & Sivachenko, A.Y. (2006) Mining Alzheimer disease relevant proteins from integrated protein interactome data. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, pp.367–378.

[9] Cho, Y.-R., Hwang, W., Ramanathan, M. & Zhang, A. (2007) Semantic integration to identify overlapping functional modules in protein interaction networks. BMC Bioinformatics, 8, p.265.

[10] Csardi, G. (2015) Network Analysis and Visualization. Available from: <https://cran.rproject. org/web/packages/igraph/igraph.pdf> [Accessed 23 December 2015].

[11] Dhara, M. (2012) Comparative Performance Analysis Of Rnsc And Mcl Algorithms On Power-Law Distribution. Advanced Computing: An International Journal, 3 (5), pp.19–34.

[12] Dhara, M. & Shukla, K.K. (2012) Performance Testing of RNSC and MCL Algorithms on Random Geometric Graphs. International Journal of Computer Applications, 53 (12), pp.5–11.

[13] Feng, J., Jiang, R. & Jiang, T. (2011) A max-flow-based approach to the identification of protein complexes using protein interaction and microarray data. Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 8 (3), pp.621–634.

[14] Goh, K.-I. & Choi, I.-G. (2012) Exploring the human diseasome: the human disease network. Briefings in Functional Genomics, 11 (6), pp.533–542.

[15] Guldener, U. (2006) MPact: the MIPS protein interaction resource on yeast. Nucleic Acids Research, 34 (90001), pp.D436–D441.

[16] IEEE (1990) IEEE standard glossary of software engineering terminology. New York, N.Y, Institute of Electrical and Electronics Engineers.

[17] Kiemer, L. & Cesareni, G. (2007) Comparative interactomics: comparing apples and pears? Trends in Biotechnology, 25 (10), pp.448–454.

[18] King, A.D., Przulj, N. & Jurisica, I. (2004) Protein complex prediction via cost-based clustering. Bioinformatics, 20 (17), pp.3013–3020.

[19] Klapa, M.I., Tsafou, K., Theodoridis, E., Tsakalidis, A. & Moschonas, N.K. (2013) Reconstruction of the experimentally supported human protein interactome: what can we learn? BMC Systems Biology, 7 (1), p.96.

[20] Kolaczyk, E.D. & Csárdi, G. (2014) Statistical Analysis of Network Data with R. New York, NY, Springer

[21] Kumar, M., Agrawal, K.K., Arora, D.D. & Mishra, R. (2011) Implementation and behavioural analysis of graph clustering using restricted neighborhood search algorithm. International Journal of Computer Applications, 22 (5), pp.15–20.

[22] Liao, B., Fu, X., Cai, L. & Chen, H. (2014) Identifying Protein Complexes by Reducing Noise in Interaction Networks. Protein & Peptide Letters, 21 (7), pp.688–695.

[23] Li, X.-L. & Ng, S.-K. eds. (2009) Biological Data Mining in Protein Interaction

[24] Li, X., Wu, M., Kwoh, C.-K. & Ng, S.-K. (2010) Computational approaches for detecting protein complexes from protein interaction networks: a survey. BMC Genomics, 11 (Suppl 1), p.S3.

[25] McGarry, K. & Daniel, U. (2014) Computational techniques for identifying networks of interrelated diseases. In: Computational Intelligence (UKCI), 2014 14th UK Workshop on. IEEE, pp.1–8.

[26] McGarry, K., Slater, N. & Ammaning, A. (2015) Identifying candidate drugs for repositioning by graph based modeling techniques based on drug side-effects. In: The 15th UK Workshop on Computational Intelligence, UKCI-2015,. pp.1–8.

[27] Moschopoulos, C.N., Pavlopoulos, G.A., Iacucci, E., Aerts, J., Likothanassis, S., Schneider, R. & Kossida, S. (2011) Which clustering algorithm is better for predicting protein complexes? BMC research notes, 4 (1), p.549.

[28] Nassa, G., Tarallo, R., Guzzi, P.H., Ferraro, L., Cirillo, F., Ravo, M., Nola, E., Baumann, M., Nyman, T.A., Cannataro, M., Ambrosino, C. & Weisz, A. (2011) Comparative analysis of nuclear estrogen receptor alpha and beta interactomes in breast cancer cells. Mol. BioSyst., 7 (3), pp.667–676.

[29] Ramyachitra, D. & Banupriya, D. (2014) Protein Complex Detection: A Study. International Journal of Computer Science and Information Technology & Security (IJCSITS), 4 (4).

[30] Rao, V.S., Srinivas, K., Kumar, G.S. & Sujin, G.N. (2013) Protein interaction network for Alzheimer's disease using computational approach. Bioinformation, 9 (19), p.968.

[31] Samatova, N.F., Hendrix, W., Jenkins, J., Padmanabhan, K. & Chakraborty, A. (2013) Practical Graph Mining with R. CRC Press.

[32] Sheinerman, F.B., Norel, R. & Honig, B. (2000) Electrostatic aspects of protein–protein interactions. Current opinion in structural biology, 10 (2), pp.153–159.

[33] Spirin, V. & Mirny, L.A. (2003) Protein complexes and functional modules in molecular networks. Proceedings of the National Academy of Sciences, 100 (21), pp.12123–12128.

[34] Sprinzak, E., Sattath, S. & Margalit, H. (2003) How reliable are experimental protein-protein interaction data? Journal of Molecular Biology, 327 (5), pp.919–923.

[35] Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., Kuhn, M., Bork, P., Jensen, L.J. & von Mering, C. (2015)

[36] STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Research, 43 (D1), pp.D447–D452.

[37] Trauger, S.A., Webb, W. & Siuzdak, G. (2002) Peptide and protein analysis with mass spectrometry. Spectroscopy, 16 (1), pp.15–28.

[38] Ulitsky, I. & Shamir, R. (2007) Identification of functional modules using network topology and highthroughput data. BMC Systems Biology, 1 (1), p.8.

[39] Wan, R. & Mamitsuka, H. (2009) Discovering Network Motifs in Protein Interaction Networks. Biological Data Mining in Protein Interaction Networks.

[40] Watts, D.J. (2003) Small Worlds: The Dynamics of Networks between Order and Randomness. Princeton University Press.

[41] Watts, D.J. & Strogatz, S.H. (1998) Collective dynamics of small-world networks. Nature, 393 (6684), pp.440–442.

[42] Wu, M., Li, X.L. & Kwoh, C.-K. (2008) Algorithms for detecting protein complexes in PPI networks: an evaluation study. In: Proceedings of Third IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB 2008). pp.15–17.

[43] Zhang, X.-F., Dai, D.-Q., Ou-Yang, L. & Yan, H. (2014) Detecting overlapping protein complexes based on a generative model with functional and topological properties. BMC Bioinformatics, 15 (1), p.186.