Integrative Network-based multi-OMICs Analysis of Glioblastoma Multiforme

Vasanthi Priyadarshini Gaddi
Institute of Bioinformatics
University Medicine Greifswald
Greifswald, Germany
gaddiv@uni-greifswald.de

Lars Kaderali
Institute of Bioinformatics
University Medicine Greifswald
Greifswald, Germany
lars.kaderali@uni-greifswald.de

I. INTRODUCTION

Glioblastoma multiforme (GBM) is the most aggressive, malignant and deadly intracranial tumour and difficult to treat by surgery, chemo or radiotherapies [1]. Its incidence in Europe is about 3.55 cases per 100,000 persons per year [2]. Intra-tumour heterogeneity of GBM cells varies in different biological aspects such as their morphology, proliferation rate, drug resistance, invasive behaviour and metastatic potential [3-4]. Therefore, molecular characteristics of the tumour as well as treatment response vary between individuals, mandating personalized treatment strategies and thus a patient stratification based on novel biomarkers.

Intracellular molecular components (i.e., DNA, mRNA, miRNA, proteins, metabolites, etc.) within the cell are not independent from each other, but are interconnected, influencing each other and thus jointly affecting biological process [5]. Therefore, disease progression is potentially influenced by any of these molecules. It is therefore crucial to identify and understand the flow of genetic information that lies beneath development and progression of disease. High-throughput experimental technologies provide a rich source of multidimensional molecular data. The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) systematically profile various types of human tumours abundantly at multiple molecular layers such as messenger-Ribonucleic acids (mRNAs), micro-Ribonucleic acids (miRNAs), Methylation patterns, Copy number variations (CNVs), proteomic profiles and genetic information [6]. In addition to experimental measurement data, molecular interaction data have become a valuable resource and play a key role to understand perturbations in biological system [5]. Integrative analysis of different molecular data types along with molecular interaction networks are highly essential, and these data when combined can expand depth, breadth and accuracy to understand the dynamic behaviour of disease progression, to gain comprehensive insights into cell biology and disease etiology, and ultimately to the development of novel, personalised treatment approaches.

Here, we present a methodological pipeline, an integrative network-based approach for disease stratification that utilises multi-OMICs data and different types of network-based approaches. We believe that this methodological pipeline helps to address the complexity of the human disease and may contribute to improve diagnosis and treatment. We show results on Glioblastoma multiforme, where we identify and characterize different molecular subgroup.

II. METHODS AND RESULTS

Multi-OMICs data of GBM cohorts were downloaded from the TCGA data portal (http://tcga-data.nci.nih.gov/tcga/) on January 13, 2015. We used four different types of OMICs data of level 3: mRNA, miRNA, Methylation and Copy Number Variation (CNV). In total, data from 275 patients were available for all these platforms and were retained for further analysis.

A. Network Construction

Molecular interaction data was downloaded from different resources. STRING v10 [7] for protein-protein interaction (PPI) network; TargetScan [8] and microRNA [9] for miRNA-gene interaction network. These data were represented as a graph $G = (V, E, W)$, where V is the set of nodes, E is the set of (undirected) edges and W are the edge weights. We have constructed two different types of molecular networks, first a protein-protein interaction network with W as edge confidence score and miRNA-gene interaction network, where W was set to the unit vector.

B. Identification of Network Modules

Initially, we have used the graph clustering algorithm Clustering with Overlapping Neighbourhood Expansion (ClusterONE) [10] to identify network modules/ submodules on the PPI network. ClusterONE uses a greedy growth process to detect overlapping protein complexes within PPI networks, taking the edge weights into account.

C. Network Propagation

Subsequently, the OMICs data are mapped to the clustered network, and network propagation [11] is used to prioritise genes and gene modules. After projecting the disease associated genes onto the molecular interaction network, the propagation algorithm is run, stimulating

information diffusion through the molecular network by the following function

$$P_{t+1} = \alpha P_{t-1} W + (1-\alpha)P$$

where P is the patient-by-gene matrix containing a specific OMICs profile, W is a degree normalised adjacency matrix of the molecular network, defined by multiplying the laplacian matrix $L = D - A$ (where D is diagonal degree matrix and A the adjacency matrix) with the diagonal matrix (Inverse of the row sums of the adjacency matrix of the diagonal) on both sides to obtain

$$W = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$$

α is a parameter between [0,1] and defines the relative amount of information to be transferred from a node to its neighbours. The algorithm runs iteratively until it convergence, e.g. until the matrix norm of $P_{t+1} - P_t < 1 \times 10^{-6}$. The rows of the resulting propagated matrix P_t was subsequently subjected to quantile normalisation, to make sure every sample within this matrix follows the same distribution.

D. Identification of disease modules

To identify disease relevant gene modules, disease associated genes of every patient after network propagation were projected onto the derived submodules. This was done independently for each data type (gene expression, CNV, etc). We calculated the average over all nodes for every projected submodules and also calculated the mean for unmatched disease genes, which represented as one module for every patient. The resulting matrix is a patient-by-disease modules matrix, containing the module averages per patient in the individual matrix cells.

E. Pairwise similarity

The next step is to calculate the pairwise similarity measure between every pair of samples/patients and construct a patient-by-patient similarity matrix for each data type. This is done using Euclidean distance on the pateint-by-disease modules matrix. We then convert the distance measure d to a similarity matrix S by normalising the distance matrix d to a similarity measure S by normalising the distance matrix d to [0,1] by using $d\_n = \frac{d}{max(d)}$ and calculate the similarity by the formula $S = 1 - d\_n$

F. Network Integration

The resulting matrix of pairwise similarities of multi-OMICs data types were averaged to form one data set. The integrated network was used for disease stratification.

G. Disease stratification

Stratification of patients was then done by clustering using Symmetric nonnegative matrix factorization (SymNMF) [12]. SymNMF is a minimization method and formulated as

$$\min_{H \geq 0} \|A - HH^T\|^2_F$$

Where A is the $n \times n$ pairwise similarity matrix of the graph and n is the number of data points or nodes. H is a nonnegative matrix of size $n \times k$, k is number of clusters requested. $\|.\|F$ denotes the Frobenius norm. H acts as cluster membership indicator; each column of H indicates one cluster and the magnitude of the value in each row indicates the strength of the membership likelihood in the cluster. symNMF is flexible to choose the similarities between data points and it naturally captures the cluster structure embedded in the graph representation.

III. CONCLUSION

We have explained briefly the importance of integrating multi-level OMICs experimental measurement data, interactome data, and different methods of network based approaches. We have introduced the implemented methodological pipeline to identify complex traits and disease subtypes and we believe this methodological pipeline helps to address the complexity of heterogeneous disease and helps to improve better diagnosis.

REFERENCES

1. Ohgaki, H., Dessen, P., Jourde, B., Horstmann, S., Nishikawa, T., Di Patre, P.L., Burkhard, C., Schuler, D., Probst-Hensch, N.M., Maiorka, P.C., Baeza, N., Pisani, P., Yonekawa, Y., Yasargil, M.G., Lutolf, U.M. & Kleihues, P. 2004, "Genetic pathways to glioblastoma: a population-based study", Cancer research, vol. 64, no. 19, pp. 6892-6899.
2. Ohgaki, H. & Kleihues, P. 2005, "Population-based studies on incidence, survival rates, and genetic alterations in astrocytic and oligodendroglial gliomas", Journal of neuropathology and experimental neurology, vol. 64, no. 6, pp. 479-489.
3. Heppner, G.H. 1984, "Tumor heterogeneity", Cancer research, vol. 44, no. 6, pp. 2259-2265.
4. Marusyk, A. & Polyak, K. 2010, "Tumor heterogeneity: causes and consequences", Biochimica et biophysica acta, vol. 1805, no. 1, pp. 105-117.
5. Barabasi, A.L. & Oltvai, Z.N. 2004, "Network biology: understanding the cell's functional organization", Nature reviews.Genetics, vol. 5, no. 2, pp. 101-113.
6. Hofree, M., Shen, J.P., Carter, H., Gross, A. & Ideker, T. 2013, "Network-based stratification of tumor mutations", Nature methods, vol. 10, no. 11, pp. 1108-1115.
7. 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, "STRING v10: protein-protein interaction networks, integrated over the tree of life", Nucleic acids research, vol. 43, no. Database issue, pp. D447-52.
8. Lewis, B.P., Burge, C.B. & Bartel, D.P. 2005, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets", Cell, vol. 120, no. 1, pp. 15-20.
9. Betel, D., Wilson, M., Gabow, A., Marks, D.S. & Sander, C. 2008, "The microRNA.org resource: targets and expression", Nucleic acids research, vol. 36, no. Database issue, pp. D149-53.
10. Nepusz, T., Yu, H. & Paccanaro, A. 2012, "Detecting overlapping protein complexes in protein-protein interaction networks", Nature methods, vol. 9, no. 5, pp. 471-472.
11. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. & Sharan, R. 2010, "Associating genes and protein complexes with disease via network propagation", PLoS computational biology, vol. 6, no. 1, pp. e1000641.
12. Kuang, D., Park, H. & Ding, CHQ. 2012, "Symmetric nonnegative matrix factorization for graph clustering", In: Proceedings of the 12th SIAM international conference on data mining (SDM), Anaheim, CA, pp. 106-117