COMMUNITY DETECTION FROM GENOMIC DATASETS ACROSS HUMAN CANCERS

Nandinee Haq, Z. Jane Wang

Department of Electrical and Computer Engineering
University of British Columbia

ABSTRACT

Cancers originating from different organs can show similar genomic alterations whereas cancers originating from the same organ can vary across patients. Therefore cancer stratification that does not depend on the tissue of origin can play an important role to better understand cancers having similar genomic patterns irrespective of their origins. In this work, we formulated the problem as a weighted graph and communities were found using a modularity maximization based graph clustering method. We classified 3,199 subjects from twelve different cancer types into five clusters. The five communities show significantly different survival rate curves. The distribution of tumor types against communities shows that lung, colon and rectum adenocarcinoma cluster together, whereas breast and ovarian cancers form another cluster.

Index Terms— Community detection, genomic features

1. INTRODUCTION

Cancer, a heterogeneous complex disease, is driven by a combination of genes and these gene combinations can also vary across patients [1]. Over the past decade, researchers have been working on systematic exploration of genetic and epigenetic signatures for different cancer types [2, 3]. Tumor stratification for different tumor types is an active field of research where a population of tumors is divided into biologically meaningful subtypes. However, most of these studies were carried out on tumors originating from the same organ [4, 5].

Analysis of different cancer types has shown that tumors originating from the same organ can significantly vary across subjects [3], whereas similar genomic alteration patterns can be observed across tumors originating from different tissues [6]. The intra-cancer heterogeneity and inter-cancer homogeneity observed across human cancers motivates the design of cancer stratification techniques irrespective of cancer types, which can be helpful for designing genomics-driven personalized medicine [7]. In this work, we tackle the cancer stratification problem from a data-driven clustering perspective by classifying cancers independent of their origins using the community detection approach, and we further investigate whether such data-driven clusters reveal different survival rate patterns.

Community detection methods are graph-based clustering methods that are particularly investigated in social network stratification. Recently Ciriello *et al.* adopted a modularity-maximization based community detection method to classify twelve different tumor types into 31 tumor sub-classes [8]. Inspired by this recent interesting direction, in this work we undertook a similar approach to cluster cancers into subclasses. We adopted a more advanced community detection approach proposed by Blondel *et al.* that was shown to outperform other methods for graphs with well-known structures [9]. This method is also computationally less expensive and more applicable for large graphs. We generated five clusters from a dataset of 3,199 subjects from twelve different cancer types. Furthermore, we investigated the survival characteristics of these clusters and observed significantly different survival rates for different clusters. Unlike the work in [8], where the problem was formulated as bipartite graph, we formulated our problem as a weighted graph to better reflect the commonality between samples. Moreover our adopted method is more advanced and has shown better performance in graphs with well-defind structures. This method is computationally more efficient and suitable for large graph like the graph generated in this work. Furthermore, we studied the survival characteristics of the generated clusters, which was not investigated in [8].

2. MATERIALS AND METHODS

2.1. Dataset and graph generation

For this work we use the cancer genomic dataset for twelve different cancer types from TGCA [3, 2, 10, 6, 11, 12, 13]. In total 3,199 subjects across different tumor types are used and the number of subjects used for each tumor type is shown in Table.1. The genomic and epigenetic changes were reduced to 479 functional alterations as described in [8]. These functional events are comprised of copy number alterations, somatic mutations and gene DNA methylation events. Recurrent regions of copy number change and recurrently mutated genes were determined using the algorithms GISTIC [14], MuSiC [15] and MutSig [16] respectively. DNA hyperme-

thylation was investigated for a selected group of genes as described in [17]. The final set of features consists of 151 copy number losses, 116 copy number gains, 199 recurrently mutated genes and 13 epigenetically silenced genes. The feature set is a binary feature set, where 1 represents that a particular alteration is present in the associated subject.

We formulated the clustering problem as a weighted graph in the subject space, where subjects are considered as nodes and an edge is drawn between the nodes (the subjects) if two subjects have at least one common alteration. The weight of the edge is calculated using the following formulation:

$$W_{ij} = \sum_{p} C_{i,p} * C_{j,p} \tag{1}$$

where W_{ij} means the weight between subject-i and subject-j, $C_{i,p} \epsilon [0, 1]$ where $C_{i,p} = 1$ when feature-p is present in subject-i. Since $C_{i,p}$ is binary, W_{ij} is sum of the number of common alterations between subject-i and j. This weight reflects the commonality between two subjects. If two subjects have no common feature between themselves, that means these two subjects are different in terms of genomic alterations and therefore they do not share any edge. On the other hand, if two subjects have some common genomic alterations, that means they are similar. So they are connected by an edge and W_{ij} is then the number of common alterations they have. The more genomic alterations they have in common, the more similar they are, and therefore the more weight is given to the edge that connects them.

Table 1: Number of cases from different tumor types.

Tumor Type	Number of cases
Bladder urothelial carcinoma (BLCA)	95
Breast invasive carcinoma (BRCA)	466
Colon and rectum adenocarcinoma (COAREAD)	489
Glioblastoma multiformae (GBM)	216
Head and neck squamous cell carcinoma (HNSC)	299
Kidney renal clear-cell carcinoma (KIRC)	378
Acute myeloid leukemia (LAML)	164
Lung adenocarcinoma (LUAD)	224
Lung squamous cell carcinoma (LUSC)	182
Ovarian serous cystadenocarcinoma (OV)	445
Uterine corpus endometrioid carcinoma (UCEC)	241

2.2. Louvain community detection

The resulting graph generated using Eqn. (1) has 3,199 nodes and 1,851,740 edges. In this work, we incorporated the Louvain method [9] to find communities from the weighted graph since this method is suitable for large graph like ours. This method consists of two phases. In the first phase, a different community is assigned to each node of the network. Then for each node-i, the modularity gain is calculated if node-i is placed in the community of each of its neighbouring nodes. Then node-i is finally placed to its neighbouring node-j for which maximum positive modularity gain is achieved. The modularity gain is calculated by the following formula:

$$\Delta\mathcal{M} = \left[\frac{W_j + W_{i,j}}{2 * W_{net}} - \left(\frac{W_j^{tot} + W_i^{tot}}{2 * W_{net}}\right)^2\right] - \left[\frac{W_j}{2 * W_{net}} - \left(\frac{W_j^{tot}}{2 * W_{net}}\right)^2 - \left(\frac{W_i^{tot}}{2 * W_{net}}\right)^2\right] \tag{2}$$

where W_j is the edge-weights in the j^{th} community, $W_{i,j}$ is the sum of the edge-weights from community-i to community-j, W_{net} means the sum of all edge-weights in the network, W_i^{tot} and W_j^{tot} are the sums of the weights of the edges incident to the communities i and j respectively.

This is applied repeatedly until no further improvement can be achieved in terms of modularity. Then the second phase starts where a new network is formed. In this new network each node now represents the communities formed in the first phase. These new nodes share weighted links calculated from the communities of the first phase that they correspond to. The weight of the links between two nodes in the second phase is calculated as the sum of the edge weights between the nodes of the two communities from the first phase that they corresponds to. Then Eqn. 2 is again applied on this new network. At the end of the second phase, new communities are formed by clustering nodes. Note that now each node corresponds to a community from the first phase (hence a group of nodes from the original graph), and therefore connecting two nodes in the second phase to form a community essentially means connecting two groups of nodes from the original graph, and hence a larger cluster is formed. This two-phase process is applied iteratively and in successive iterations larger communities are generated.

3. RESULTS

Table 2: Specifications of the learned communities

Community	Number of subjects (n)	Significant cancers
C-1	638	HNSC, LUSC
C-2	408	GBM
C-3	1016	COAREAD, LAML, UCEC
C-4	802	BRCA, OV
C-5	335	KIRC

With applying the above Louvain algorithm into our dataset, the cancer subjects are divided into five communities automatically. The number of the subjects in each community is reported in Table 2. Each community is comprised of different tumor types. Fig.1 shows the distribution of different tumor types in the five clusters, where black corresponds to

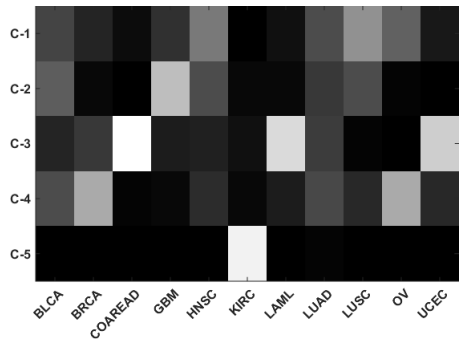


Fig. 1: The normalized distributions of tumor types associated with different clusters, where the white color means 1 and black means 0 in the color bar.

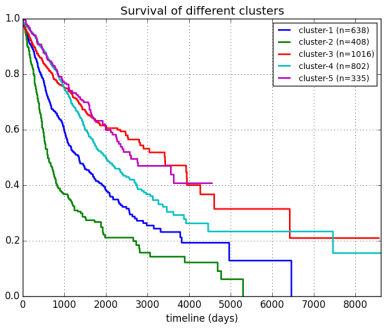
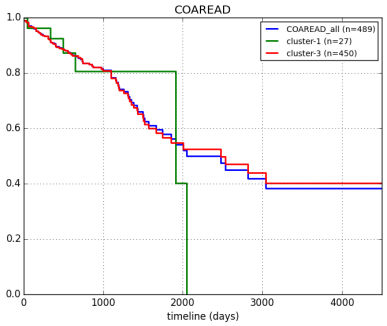


Fig. 2: Survival curves of different clusters.



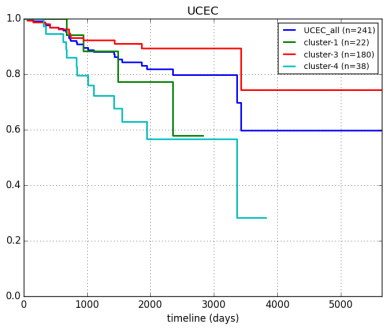


Fig. 3: Survival curves of Colon and rectum adenocarcinoma (COAREAD) and Uterine corpus endometrioid carcinoma (UCEC) tumor types in different clusters.

0% subjects and white corresponds to 100% subjects of the associated cancer type. As we can see, the dominant cancer in cluster-3 is COAREAD and in cluster-5 is KIRC. Approximately 70% subjects of LAML and UCEC tumor types go to cluster-3. Cluster-5 is mainly constituted with subjects of KIRC with a tiny percent of LUAD, LUSC and UCEC. Almost 60% subjects of BRCA and OV go to cluster-4, indicating that these two tumor types share a number of common features. Cluster-2 contains around 60% of GBM subjects. 50% samples of LUSC are in cluster-1. BLCA and LUAD tumors are distributed over the first 4 clusters.

We are particularly interested in studying whether such generated tumor clusters are associated with different survival rate patterns. We investigated the survival rates using the Kaplan-Meier estimator [18] for the generated clusters. The survival curves are shown in Fig. 2. As it can be seen, different clusters have different survival rates. If we take 4000 days as the reference point, the probability of survival is highest in cluster-3 and 5, and is worst in cluster-2. To statistically compare the survival curves, the log rank test is carried out between the clusters. The p-value between cluster-3 and 5 is 0.85. All other combinations are found to be statistically significant ($p < 0.05$) at the 5% significant level. This pattern was also observed when we plotted the survival curves of each tumor type distributed in different clusters. It is worth mentioning that different tumor types also have different survival rates.

Some of the tumor types show interesting patterns in different clusters. Fig. 3 shows the survival plots for two tumor types associated with different clusters. For each of these cases, $\langle tumor\_name \rangle\_all$ denotes the survival curve when all cases of that particular tumor type is taken into account. Survival plots are generated only when the total number of subjects of a particular tumor in a cluster is greater that 10. For COAREAD, we note that the survival curves are significantly different for cluster-1 and cluster-3. After 2100 days, the survival probability of a COAREAD tumor patient is almost 50% if he is in cluster-3, whereas it it 0% if he is in cluster-1. UCEC seems to have a good survival rate in general, and the cases that fall into cluster-3 show even a better rate of survival.

Similar patterns are observed across other tumor types as well, though detail figures are omitted here due to space limit. Cluster-3 also shows better chances of survival for LUAD and HNSC as well. Even after 6000 days, the chance of survival for a HNSC cluster-3 case is almost 38%, which is much higher than that of all other HNSC tumor cases. KIRC tumor cases show good survival probabilities in general, however

the subjects of KIRC in cluster-2 show significantly lower rate of survival. LAML tumors that fall into cluster-3 follow a similar survival pattern as overall LAML tumors. A few LAML cases fall into cluster-1 and these cases show lower survival probabilities. Cluster-2 shows the worst survival rate for OV tumor type as well. Overall, cluster 3 and 5 show better survival probabilities within different tumor types, whereas cluster-2 shows the worst survival rates across tumor types.

4. CONCLUSION

To investigate similar patterns across human cancers, we study a data-driven clustering approach by clustering subjects from twelve cancer types using modularity maximization based community detection technique. The community detection method finds five separate communities, where different communities are dominated by different groups of cancer types. We further explore the survival rates of the generated communities and note that the communities vary significantly in terms of their survival characteristics. The difference of survival rates across communities indicates the potential of cancer stratification that does not depend on cancer origins. In future we will further subdivide the communities and investigate the relation of these communities with biological pathways to validate the learned communities.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

5. REFERENCES

[1] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer genome landscapes," *Science*, vol. 339, no. 6127, pp. 1546–1558, 2013.

[2] Cancer Genome Atlas Network, "Comprehensive molecular characterization of human colon and rectal cancer," *Nature*, vol. 487, no. 7407, pp. 330–337, 2012.

[3] ——, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, 2012.

[4] M. Hofree, J. P. Shen, H. Carter, A. Gross, and T. Ideker, "Network-based stratification of tumor mutations," *Nature methods*, vol. 10, no. 11, pp. 1108–1115, 2013.

[5] J. S. Reis-Filho and L. Pusztai, "Gene expression profiling in breast cancer: classification, prognostication, and prediction," *The Lancet*, vol. 378, no. 9805, pp. 1812–1823, 2011.

[6] Cancer Genome Atlas Research Network, "Integrated genomic analyses of ovarian carcinoma," *Nature*, vol. 474, no. 7353, pp. 609–615, 2011.

[7] L. A. Garraway, "Genomics-driven oncology: framework for an emerging paradigm," *Journal of Clinical Oncology*, vol. 31, no. 15, pp. 1806–1814, 2013.

[8] G. Ciriello, M. L. Miller, B. A. Aksoy, Y. Senbabaoglu *et al.*, "Emerging landscape of oncogenic signatures across human cancers," *Nature genetics*, vol. 45, no. 10, pp. 1127–1133, 2013.

[9] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[10] Cancer Genome Atlas Research Network, "Integrated genomic characterization of endometrial carcinoma," *Nature*, vol. 497, no. 7447, pp. 67–73, 2013.

[11] R. McLendon, A. Friedman, D. Bigner, E. G. Van Meir *et al.*, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, no. 7216, pp. 1061–1068, 2008.

[12] Cancer Genome Atlas Research Network, "Comprehensive genomic characterization of squamous cell lung cancers," *Nature*, vol. 489, no. 7417, pp. 519–525, 2012.

[13] ——, "Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia," *N Engl J Med*, vol. 2013, no. 368, pp. 2059–2074, 2013.

[14] R. Beroukhim, G. Getz, L. Nghiemphu, J. Barretina *et al.*, "Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma," *Proceedings of the National Academy of Sciences*, vol. 104, no. 50, pp. 20 007–20 012, 2007.

[15] N. D. Dees, Q. Zhang, C. Kandoth, M. C. Wendl *et al.*, "MuSiC: identifying mutational significance in cancer genomes," *Genome research*, vol. 22, no. 8, pp. 1589–1598, 2012.

[16] S. Banerji, K. Cibulskis, C. Rangel-Escareno, K. K. Brown *et al.*, "Sequence analysis of mutations and translocations across breast cancer subtypes," *Nature*, vol. 486, no. 7403, pp. 405–409, 2012.

[17] M. Esteller, "Epigenetic gene silencing in cancer: the DNA hypermethylome," *Human molecular genetics*, vol. 16, no. R1, pp. R50–R59, 2007.

[18] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958.