CancerMapper: Explorations of Cancer Study Network

Category: n/a

ABSTRACT

Interactive visualization tools are highly desirable to biologist and cancer researchers to explore the complex structures, detect patterns and find out the relationships among biomolecules responsible for a cancer type. A pathway contains various biomolecules in different layers of the cell which are responsible for specific cancer type. Researchers are highly interested in understanding the relationships among the proteins of different pathways and furthermore want to know how those proteins are interacting in different pathways for various cancer types. We introduce *CancerMapper*, a visual analytics system that helps researchers to explore cancer study interaction network. To fully understand the role of proteins in different cancers, twenty-six cancer studies are merged together. *CancerMapper* also helps biologists to drill down the cancer network based on the common mutated proteins and their frequencies. Proteins which are highly interacted are clustered together. A bubble graph visualize common protein based on its frequency and biological assemblies. Parallel coordinates highlight patterns of patient profiles (obtained from cBioportal by WebAPI services) on different attributes for a specified cancer study.

1 INTRODUCTION

Biologists spend years trying to understand the complex relationships of various cell functions and inter-networks between enormous numbers of components of the cells [4]. It is a very arduous task to explore the cell components and reveal the patterns, anomalies, relations and other signification information just from data which are mostly text and statistics. However, a friendly and efficient visualization of biological data might help to understand that arduous task in a more productive way.

Visualization is the process of representing data visually and is a signification task in network analysis, especially in the biological system where scientists work with the complex structures and relationships of the biomolecules. Hence, there are plethora visualization tools available in these years to utilize the power of visualization to help better understanding of the complex systems such as pathways, proteins network, etc. Despite such available tools and techniques, understanding protein-protein interactions are still ongoing challenges for biologists and researchers for years, even with a small question of how to better understand complicated many-to-many relationships between cancers and mutated genes as long as pointing out which protein plays a central role in multiple cancer studies. Therefore, having a graphical tool that helps biologists can analyze and gain insight interactions of the protein-protein network is highly desirable.

In this paper, we propose *CancerMapper*, a visual analytics tool to explore the relationship among genes (which are essentially proteins or part of proteins) in different cancer studies for providing an interactive way to help the researchers in understanding the complicated many-to-many relationships between cancers and mutated genes.

2 *CancerMapper* VISUALIZATION

We worked closely with a molecular biologist. His expert opinion lead us to identify four important objectives which are difficult to achieve using existing visualization tools. Therefore, the main objectives of this paper are:

- Visualization task **T1**: Allows users to uncover the correlations among cancer studies as well as cancer studies and genes.
- Visualization task **T2**: Filter the cancer network based on the common mutated proteins and their frequencies.
- Visualization task **T3**: Visualize common protein based on its frequency and biological assemblies.
- Visualization task **T4**: Highlight the patterns of patient profiles (obtained from cBioportal by WebAPI services) on different attributes for a specified cancer study.
- Visualization task **T5**: Clustering genes based on the relationship between them.

2.1 Processing input data sets

We demonstrate *CancerMapper* on 26 cancer studies data retrieved from *cBioPortal* [3] as a Comma Separated Value file including proteins name, frequency, cancer studies and study type fields through WebAPI service provided by cBioPortal. For constructing bubble graph to show the Proteins images, we used *Protein Data Bank* (http://www.rcsb.org/) [1] to pull the pictures.

2.2 *CancerMapper* Overview

Figure 1 represents the basic overview of our visualization tool. The control panel in Figure 1(a) provides the basic filtering function of the cancer network based on common gene name(s) or the number of common genes which is set on the slider (visualization task **T2**).

The cancer studies network in Figure 1(b) presents the merged network of 26 cancer studies. The thickness of the link between any two cancer studies indicates the number of common genes, the thicker the link, the more genes found in both studies (visualization task **T1**).

The biological assemblies view highlights mutated proteins in a specific cancer study (visualization task **T3**) as shown in Figure 1(c) . The parallel coordinate view in Figure 1(d) explores the statistical data about the patients of a particular cancer study (visualization task **T4**). It shows the information of patients such as age, cancer type, stage of cancer, location, gender and some more details based on selected cancer study type.

Community detection for protein network In this application, we use Girvan Newman [6] algorithm to detect network community. The algorithm results in a hierarchical structure, named *dendrogram*. We use Modularity [5] as a measure to evaluate the quality of the network.

3 RESULTS

With the support of visualization tool, biologists can quickly notice that Breast Cancer accounts for the biggest proportion of protein found in all cancer studies. AKT3, PIK3R5, CCND1 and NFKB1 are the most important proteins in the network since they interact with many other proteins and stay in the network as a bridge. Drilling down into the Parallel coordinates, users found out that in Germany, the patients who have this particular cancer are ages from around 57-80 years.While in United Kingdom, these years flows between 50-77. If users look at the Australia, they can see that most of the patients of *Pancreatic adenocarcinoma cancer* are from Australia (exactly 216 out of 383) and the ages of the patients are distributed between 30-90. All of those views denotes that *Pancreatic adenocarcinoma cancer* mostly found on adult above 30 years old.

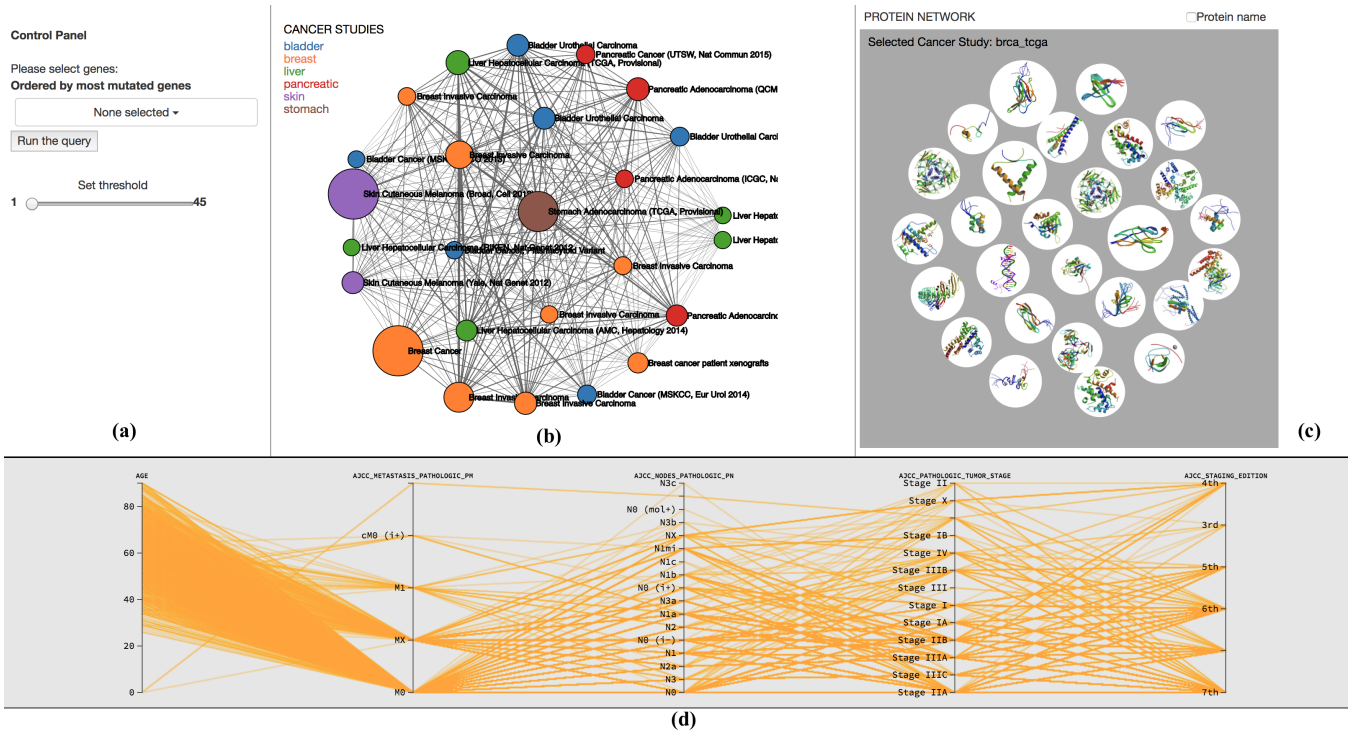

Figure 1: Overview of *CancerMapper*: a) Control panel of *CancerMapper*, b) The merged network of twenty-six cancer studies, c) Bubble chart of proteins within a cancer study. d) The parallel coordinate highlighting patterns of patient profiles.

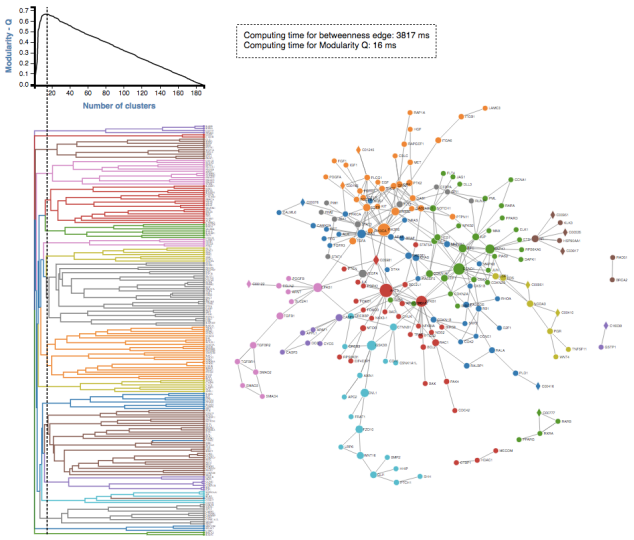

Figure 2: The dendrogram and its corresponding community. The size of the node represents the betweenness centrality value.

4 Conclusions and Future Work

The aim of *CancerMapper* is to assist a researcher in finding patterns between cancer studies and genes. It also helps to explore the patients' data which can help to get a demographic idea about a particular cancer type.

As depicted in this paper, experimental data contain many variables (multidimensional data). Looking into the relationships of different variables in experimental data may lead to interesting discoveries. However, inspecting individual dimension as well as the correlations between dimension in parallel coordinates is a time-consuming process. In the future, we plan to apply visual features [7, 8] to highlight interesting scatterplots (and variables), for examples, scatterplots with clusters [2] representing different classes of cancer patients. Using visual features, we should be able to extract important variables (and genes/proteins) corresponding to the separation of different groups of cancer patients. Highlighting these genes/proteins (using different color encodings) on the biological maps can be significant in analyzing the causality inherent in biological networks and thus very valuable in drug design [4].

References

[1] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[2] T. Dang and L. Wilkinson. Transforming scagnostics to reveal hidden features. In *Proceedings of IEEE Conference on Visual Analytics Science and Technology*, 2014.

[3] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, and N. Schultz. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Science Signaling*, 6(269):pl1–pl1, 2013.

[4] A. Lex, C. Partl, D. Kalkofen, M. Streit, S. Gratzl, A. M. Wassermann, D. Schmalstieg, and H. Pfister. Entourage: Visualizing relationships between biological pathways using contextual subsets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2536–2545, Dec. 2013. doi: 10.1109/TVCG.2013.154

[5] M. E. Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[6] M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[7] J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 4(2):96–113, July 2005. doi: 10.1057/palgrave.ivs.9500091

[8] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *Proceedings of the IEEE Information Visualization 2005*, pp. 157–164. IEEE Computer Society Press, 2005.