Poster: Observing Change in Crowded Data Sets in 3D Space - Visualizing Gene Expression in Human Tissues

Marcin Rogowski [*] Carlo V. Cannistraci [†‡] Gregorio Alanis-Lobato [†‡] Philip Weber [§]
Timothy Ravasi [†‡] Jurgen Schulze [§] Daniel Acevedo-Feliz [¶]

Abstract

We have been confronted with a real-world problem of visualizing and observing change of gene expression between different human tissues. In this paper, we are presenting a universal representation space based on two-dimensional gel electrophoresis as opposed to force-directed layouts encountered most often in similar problems. We are discussing the methods we devised to make observing change more convenient in a 3D virtual reality environment.

Index Terms: I.3.8 [Computer Graphics]—Applications; I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Virtual reality H.5.2 [Information Interfaces and Presentation]: User Interfaces—Interaction styles

1 Problem formulation

The task we were presented with was to observe gene expression changes and find relations in the human genome data. Our data consists of 1292 different genes and their expressions in 23 different human tissues. In addition to that, we were also given data on molecular function (*MF*), biological process (*BP*) and cellular component (*CC*) of the genes which we used to create connections between them and to emphasize their similarity.

Clearly the problem of network visualization and analysis was previously treated by numerous applications. Those commonly used in the field of bioinformatics include Cytoscape [2] and BioLayout Express 3D [5]. Cytoscape offers a number of statistical analysis features as well as robust filters for exploration of network data sets. One of the important features of BioLayout Express 3D is a three-dimensional representation it proposes. Its authors correctly point out that an extra dimension gives possibility to pack the information more efficiently. We recreate some of the functionalities proposed by others in a VR environment and propose some original features. Our contribution is also a representation space based on gene features rather than calculated by an algorithm.

2 Representation space

Historically, the biggest group of algorithms used to find a meaningful layout of a network in 2D and 3D space is force-based. Fruchterman & Reingold algorithm [3] is one of the best known general algorithms and was previously used to represent biological networks, for example in Arena 3D visualization tool [4]. The way we could use a force-based algorithm for our goal of comparison of gene expression between tissues is to determine first two dimensions using an algorithm and then add expression as the third coordinate. The problem is that in this case the first two dimensions do not carry any information, just align the nodes. Another complication is a random element that is present in many force-based algorithms which causes the layout to alter every time.

We decided to go in another direction and rather than aligning the network by an algorithm, use the attributes characteristic to genes themselves to build a unique 3D space. The idea we used comes from two-dimensional gel electrophoresis. In this method, proteins are separated in two dimensions by their isoelectric point (*pI*) and molecular weight (*mW*). We decided to use the same attributes adding the third dimension of gene expression (*E*) to gel electrophoresis therefore creating a universal 3D representation space.

3 Initial visualization - crowding problem

As described in the previous sections, our data consisted of 1292 genes. For each gene its location in two dimensions is fixed by its *pI* and *mW*. There are 23 different *E* values - one for each tissue - and we would like to observe how it changes when we move between tissues. The edges between nodes are determined by the values of *MF*, *BP* and *CC* and for the initial evaluation we only created the edges where all three are maximum. The problem, as expected, was in the number of nodes and edges. Hardly anything can be deduced based on the graph and the situation is even worse if we want to compare the gene expression between tissues.

4 Moving into interactive Virtual Reality space

We used CORNEA - a CAVE-like totally immersive environment offering six-sided stereoscopic display with head tracking. The implementation uses CalVR - a VR framework developed at Calit2 at University of California, San Diego [1]. In this implementation, we gave the user ability to step into the network, manipulate it, change parameters and move between tissues all using a remote controller and easy to use menu system. Moving to VR alone does not solve the problem completely - it can be noticed on Fig. 1 that there is still too much data to be comfortably analyzed in this form. To deal with this complication we implemented a number of functionalities.

The features we implemented to enhance the user interaction can be divided into two groups: the ones that help the users to identify the changes occurring and the ones that reduce the crowding problem. We discuss them in the following sections.

4.1 Visualizing change

4.1.1 Gradual transition between tissues

The first goal of our CalVR implementation was to recreate what we got using traditional tools such as MATLAB in the CAVE environment. That was achieved relatively easily and we quickly started taking advantage of the full interactivity offered. We created a menu of 23 tissues and starting from the initial two-dimensional view as in gel electrophoresis we allow users to pick the tissues from the

[*] e-mail: marcin.rogowski@kaust.edu.sa. Division of Applied Mathematics and Computer Sciences, KAUST, Thuwal, Kingdom of Saudi Arabia

[†] e-mail: {carlo.cannistraci, gregorio.alanislobato, timothy.ravasi}@kaust.edu.sa. Integrative Systems Biology Lab., Division of Biological and Environmental Sciences & Engineering, Division of Applied Mathematics and Computer Sciences, KAUST, Kingdom of Saudi Arabia

[‡] Department of Medicine, Division of Medical Genetics, University of California, San Diego, La Jolla, CA, USA

[§] e-mail: {pweber,jschulze}@ucsd.edu. Calit2, University of California, San Diego, La Jolla, CA, USA

[¶] e-mail: daniel.acevedo@kaust.edu.sa. Visualization Lab., KAUST, Thuwal, Kingdom of Saudi Arabia

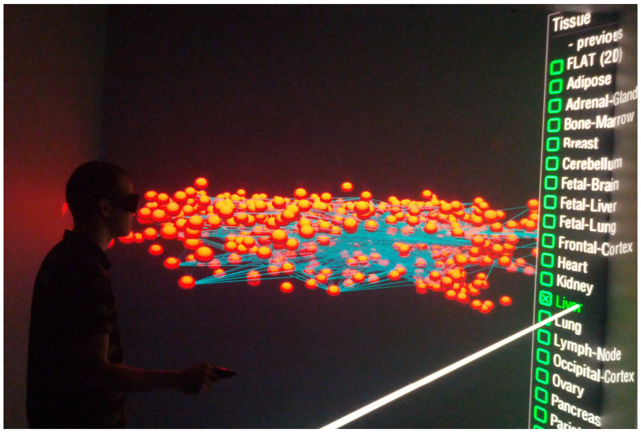

Figure 1: 3D plot recreated in VR with ability to change between tissues.

list. To make it more comfortable to sense the change, transitions are gradual rather than instantaneous.

4.1.2 Ghosts

Another feature we introduced is saving a faint image ("ghost") of the current layout, so changes in the expression between multiple tissues can be compared visually. This is not an ideal solution, as it increases the crowding problem, but it can be useful together with some of the methods described in the section 4.2. In order to avoid escalating the problem too much, we decided not to save the position of edges in a layout - since they are not tissue-specific.

4.1.3 Highlighting groups

We also allowed users to highlight groups of nodes. When the option is used, all the nodes are set to be partially transparent and user can select some of them to be opaque. This way, an expert user with knowledge in the field can focus on the genes that are relevant to him without losing the sense of the big picture. The selection is normally done for regions of the *pI*-*mW* space.

4.1.4 Heat map coloring

One of the most observable features we introduced to track changes of gene expression between tissues is node coloring. Since the colors of the nodes initially do not have any meaning related to the genes themselves, we were free to manipulate colors to highlight changes occurring between tissues. We used an entropy-like measure between the expression in the tissue currently viewed and the one that was being shown previously. We color the nodes whose expression changes the most with red and gradually transition to white for the nodes that did not change significantly. For edges, the color is based on the two nodes it connects and is based on the one that is closer to pure red. As can be observed in Fig. 2, most of the time majority of the nodes stay white while only for a few the expression changes significantly. Another feature we added is hiding the genes whose expression did not change substantially, implemented as a slider bar as described for other functionalities in the next section.

4.2 Reducing crowding

4.2.1 Manipulating parameters

We decided to take advantage of the interactivity not only to compare between the gene expression in different tissues with set parameters, but we also allowed users to change those parameters. We created a set of slider bars that allow the user to modify the parameters of *MF*, *BP*, *CC* used to create edges as well as another one determining a threshold for edge lengths visible at any point. For all the parameters, both minimum and maximum values can be set, which gives an opportunity for thresholding and hence viewing the whole network in subsets.

Thresholding can also be based on the entropy-like measure of similarity of gene expression between tissues. Using the same values as calculated for heat map coloring, effectively it allows users to hide nodes of particular colors. It is most often used to focus on the nodes that change the most while hiding the ones where only insignificant changes occur.

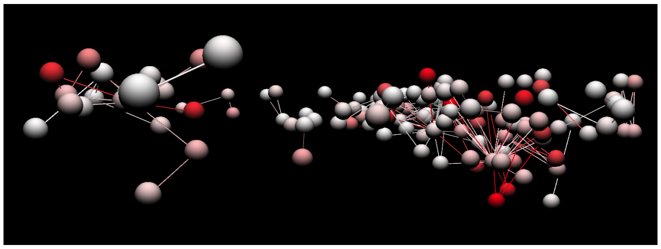

Figure 2: Threshold of 10% of the shortest edges visible with coloring reflecting expression difference between fetal lung and lung, isolated nodes hidden.

4.2.2 Transparency

Transparencies we used are tightly coupled to the heat map coloring we discussed in the previous section. Unfortunately, even with only a few nodes colored red and the rest very close to white, the crowding problem still exists. Some of the nodes whose expression changed considerably might be enclosed by nodes whose expression stayed the same or might simply not be visible from a particular point of view. The solution to the problem is adding transparency - using the same measure as for coloring from white to red - the node whose expression changed the most is fully opaque while the one that changed the least is fully transparent. Everything else is scaled accordingly. Thanks to this method, we are able to see through the nodes representing genes whose expression did not change and focus on the ones where significant changes occur.

4.2.3 Hiding isolated nodes

Since the parameters of a network can be manipulated by the user interactively, the conditions which an edge has to satisfy may be very strict or fairly loose, hence the number of the edges varies. To reduce the crowding problem, we added a switch in the menu that allows the user to remove all the nodes that are isolated. This way, groups of genes with similar properties may be more easily compared without distractions from numerous isolated nodes.

5 Conclusions

In this paper we proposed a three-dimensional universal representation space that can be used to visualize gene expression. We also showed some of the advantages of analyzing complex, crowded networks in VR environment. A subset of the methods that we devised to perform visual comparison of gene expression between human tissues was described. As shown on Fig. 2, our techniques can be used to greatly reduce the complexity of a given network and analyze it in parts. Heat map coloring and gradual transitions of both positions and colors vastly help to notice changes occurring in the network. We are now working to identify significant relationships in the data set analyzed.

References

[1] CalVR. http://ivl.calit2.net/wiki/index.php/calvr, accessed 06/12/2012.

[2] Cytoscape. http://www.cytoscape.org, accessed 08/12/2012.

[3] T. M. Fruchterman and E. M. Rheingold. Graph drawing by force directed placement. *Softw. Exp. Pract.*, pages 1129–1164, 1991.

[4] G. A. Pavlopoulos, S. I. O'Donoghue, V. P. Satagopam, T. G. Soldatos, E. Pafilis, and R. Schneider. Arena3D: visualization of biological networks in 3D. *BMC Systems Biology*, 2(1):104, 2008.

[5] A. Theocharidis, S. van Dongen, A. J. Enright, and T. C. Freeman. Network visualization and analysis of gene expression data using BioLayout Express3D. *Nature protocols*, 4(10):1535–1550, 2009.