Accelerating Translational Insights through Visual Analytics

Suresh K. Bhavnani, PhD[1] , Bryant Dang[1], BS, Rohit D. Divekar, MD PhD[2]

[1]Inst. for Translational Sciences, [2]Div. of Allergy, Dept. of Med, Univ. of Texas Medical Branch, Galveston, TX.

Abstract

The exponential growth of biomedical data related to complex diseases such as asthma and Alzheimer's far exceeds our cognitive abilities to comprehend it for tasks such as biomarker discovery, pathway identification, and molecular-based phenotyping. Here we begin by discussing the theoretical foundations for the emerging field of visual analytics, with a focus on the cognitive and task-based motivations to use methods from this field to analyze complex biomedical data. Next, we present the state of the practice for one such approach called network visualization and analysis, and demonstrate through a concrete example how networks are particularly useful for deriving translational insights from complex molecular and phenotype information. This exposition helps to identify the strengths and limitations of network analysis that are critical for its practical application. The presentation is targeted towards members of interdisciplinary translational teams consisting of translational bioinformaticians, biologists, and clinicians, who wish to comprehend the interaction of molecular and phenotype information, leading to translational insights. The educational goals include acquiring a theoretical understanding of visual analytics, and the practical knowledge to begin the analysis of biomedical data using methods from visual analytics.

Introduction

The explosion of molecular information generated by multidimensional measurements of proteins, genes, and metabolites, coupled with digital access to patient clinical records has created unprecedented opportunities for a more comprehensive understanding of complex diseases such as asthma and Alzheimer's disease. However, this explosion of information has also created a challenge for researchers, especially those in multidisciplinary translational science teams, to comprehend and integrate such disparate and large amounts of data.

One approach to integrate and comprehend such complex information is through methods being developed in the new field of visual analytics. We begin by presenting an overview of the evolving theoretical foundations for visual analytics, and the cognitive and task-based motivations to use methods from this field to analyze complex biomedical data. Next, we focus on one form of visual analytics called networks which are particularly useful for analyzing complex molecular and clinical data, with the goal of identifying sub-phenotypes in the disease, and to infer the molecular pathways involved in those phenotypes. These analyses reveal the strengths and limitations of the method, which are critical for its practical use to analyze ever increasing and complex biomedical data.

Visual Analytics: Theoretical Foundations

Visual analytics is defined as the science of analytical reasoning, facilitated by interactive visual interfaces[1]. The primary goal of visual analytics is to augment cognitive reasoning by translating symbolic data (e.g., numbers in a spreadsheet) into visualizations (e.g., a scatter plot), which can be manipulated through interaction (e.g., highlight only some data points in the scatter plot). As discussed below, **visualizations**, and **interaction** with those visualizations, are powerful for helping analysts comprehend complex relationships in biomedical data because of the nature of human cognition, and the nature of tasks performed by analysts.

Motivation for Visualization. Visualizations are powerful because they leverage the massively parallel architecture of the human visual system consisting of the eye and the visual cortex of the brain.[2] This parallel cognitive architecture enables the rapid comprehension of multiple graphical relationships simultaneously, which often leads to insights about relationships in complex data such as similarities, trends, and anomalies.[1] For example, the detection of an outlier in a scatter plot is fast because the graphical relationships between the outlier and the rest of the points can be processed in parallel by the visual cortex. Such parallel processing is independent of the number of non-outlying points and therefore scales up well to large amounts of data. In contrast, finding an outlier in a spreadsheet of numbers involves numerical comparisons to identify the outlier, which is dependent on the much slower symbolic processing areas of the human brain. Such symbolic processing is serial in nature, and therefore highly dependent on the number of data points, which when large can quickly overwhelm an analyst. Data visualizations therefore help to shift processing from the slower symbolic processing areas of the human brain, to the faster graphical parallel processing of the visual cortex enabling comprehension of large and complex data sets such as those currently available for complex diseases such as asthma and Alzheimer's disease.

However, not all data visualizations are effective in augmenting cognition. For example, an organizational chart of employee names and their locations laid out in a hierarchy based on seniority is not very useful if the task is to determine patterns related to the geographical distribution of the employees. Therefore visualizations need to be aligned with tasks[3], data, and mental representations of the user[4], before they can be effective for augmenting cognition.

Motivation for Interactivity. While static visualizations of data can be powerful if they are aligned with tasks, data, and mental representations, they are often not sufficient for comprehending complex data. This is because data analysis typically requires many different tasks performed on the same data such as discovery, inspection, confirmation, and explanation[5], each requiring different views of the data. Furthermore, when analysis is done in teams consisting of different disciplines, each member often requires a different representation of the same data. For example, in a translational team, a molecular biologist might be interested in which cytokines are co-expressed across patients, whereas a clinician might be interested in the clinical characteristics of patients with similar cytokine profiles, and later how they integrate with the molecular information. To address these changes in task and mental representation, visualizations require interactivity or the ability to transform parts, or the entire visual representation.

Theories Related to Visual Analytics. Although the field of visual analytics has drawn on theories and heuristics from different disciplines such as cognitive psychology, computer science, and graphic design, the development of theories and taxonomies for visual analytics are still in early stages of development[1]. For example, there are a number of attempts to classify visual analytical representations[6,7], and interaction methods at different levels of granularities and tasks[8]. One such classification attempt categorizes visual analytical representations into (1) time series (e.g., line graphs showing how the expression of different cytokine change over time), (2) statistical distributions (e.g., box-and-whisker plots), (3) maps (e.g., pie charts showing percentages of different races at different city locations on the US map), (4) hierarchies (e.g., top-down tree showing the management structure of an organization), and networks (e.g., a social network of how friends connect to other friends such as on Facebook). Once these visualizations are generated, they are considered visual analytical if they enable interaction directly or indirectly with part, or all of the information being represented. Examples for such interactivity include transforming a top-down tree into a circular tree, coloring nodes in the tree based on specific properties such as gender, or dragging a node in the tree to swap its location with another sibling node. Each of these visual analytical methods can show trends through animation, or represent "big data" through different granularities of information.

It is important to note that visual analytics has considerable overlap with the fields of scientific visualization (focused on modeling real-world geometric structures such as earthquakes), and information visualization (focused on modeling abstract data structures such as relationships). However, visual analytics places a large emphasis on approaches that facilitate reasoning and making sense of complex information individually and in groups[1], which makes this approach particularly pertinent for tasks such as inferring biological pathways from molecular and clinical information in translational teams.

Visual Analytics: Application to Translational Science

As described above, there are numerous visual analytical representations that have been proposed and used. However, networks[9] are one of the most advanced forms of visual analytics because they enable not only an interactive visualization of complex associations, but because they are based on a graph representation, also enable the quantitative analysis and validation of the patterns that become salient through the visualization. This visual and quantitative coupling enables comprehension *and* significance, both of which are critical in translational research.

Network Visualization and Analysis. Networks have been used to analyze a wide range of molecular measurements related to gene regulation[10], disease-gene associations[11], and disease-protein associations.[12] A network (also called a graph) consists of a set of nodes, connected in pairs by edges; nodes represent one or more types of entities (e.g., patients or cytokines). Edges between nodes represent a specific relationship between the entities (e.g., a patient has a particular cytokine expression value). Figure 1 shows a bipartite network where edges exist only between different types of entities[9], in this case between asthma patients and cytokines.

Network analysis of biomedical data typically consists of three steps: (1) **exploratory visual analysis** to identify emergent bipartite relationships such as between patients and cytokines; (2) **quantitative analysis** through the use of methods suggested by the emergent visual patterns; (3) **inference of the biological mechanisms** such as across different emergent phenotypes. This three-step method used across several studies[5,13,14] has revealed complex but comprehensible visual patterns, each prompting the use of quantitative methods that make appropriate assumptions about the underlying data, which in turn have led to inferences about the biomarkers and underlying mechanisms involved.

For example, Figure 1 shows the results of using the above method to analyze asthma patients and cytokine profiles.[14] The

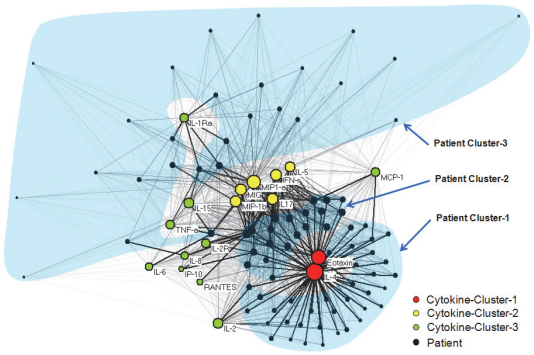

Fig. 1. **Bipartite network analysis of asthma patients and cytokines.** Three clusters of patients (encircled with blue shapes) have a complex but understandable relationship with three clusters of cytokines (colored nodes). The results led to the identification of molecular-based sub-phenotypes, and their inferred pathways.[14]

figure shows a bipartite network where black nodes represent asthma patients, colored nodes represent candidate cytokines, and the edges connecting the nodes represent normalized cytokine expression. Global patterns between patients and cytokines in the network were visualized and analyzed using the *Kamada-Kawai* layout algorithm[9]. This algorithm pulls together nodes that are strongly connected, and pushes apart nodes that are not. The result is that nodes with a similar pattern of connections are placed close to each other, and those that are dissimilar are pushed apart. Because the network analysis suggested the existence of strong clustering, the visual results were quantitatively verified using hierarchical clustering to identify the boundaries for the patient and cytokine clusters, and validated by comparing them to 1000 random networks of the same size and node distribution. The results revealed 3 clusters of patients (encircled by the blue shapes) that had a complex but understandable relationship with three clusters of cytokines (colored nodes).

These results led to the identification of three sub-phenotypes of asthma, and their inferred biological pathways. For example, Patient Cluster-1 (at the bottom of the figure) had a strong association with the cytokine cluster containing Eotaxin and IL-4. This strong co-occurrence of cytokines within Patient Cluster-1 suggested a sub-phenotype that has a T-helper-2 (Th_2) lymphocyte-skewed immune response. Such a response is known to result in the secretion of IL-4, which in turn triggers Eotaxin production by non-immune cells such as bronchial epithelial cells, fibroblasts, and smooth muscle cells, that precipitate downstream actions including the activation and recruitment of tissue-resident eosinophils, an important marker of early stage asthma. The network visualization and analysis therefore helped to identify sub-phenotypes of asthma and their inferred biological pathways, which led to translational insights for therapeutics that are targeted to specific biological processes.

Strengths and Limitations of Network Analysis. The strengths of networks include a unified representation for both sides of a bipartite relationship (e.g., patients and cytokines), common in data analyzed by interdisciplinary translational teams, in addition to the nature of each relationship through the edge weights. As demonstrated above, this unified representation enables the rapid comprehension of underlying biological mechanisms. Furthermore, it guides the use of appropriate measures to verify and quantitatively analyze the observed patterns, and therefore requires no *a priori* assumptions about the relationships. The limitations include a constraint on the number of variables that can be simultaneously represented through graphical properties such as color, size and shape, which often requires alternate or multiple visual analytical representations to conduct a comprehensive analysis of all the variables.

Target Audience and Educational Goals

This presentation is targeted towards members of interdisciplinary translational teams consisting of translational bioinformaticians, biologists, and clinicians, who wish to comprehend the complex interaction of molecular and phenotype information, leading to translational insights. The educational goals include (1) a theoretical understanding of visual analytics, (2) the practical knowledge to design the analysis of biomedical data using visual analytics, (3) motivation and concepts to use network analysis, and (4) the strengths and limitations of the method, which guides the appropriate use of this emerging methodology.

Acknowledgements. Supported by NIH 1U54RR02614 UTMB CTSA (ARB), and CDC/NIOSH #R21OH009441-01A2.

References

1. Thomas JJ, Cook KA. Illuminating the Path: The R&D Agenda for Visual Analytics, National Visualization and Analytics Center. (2005).
2. Card S, Mackinlay JD, Shneiderman B. Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann Publishers, San Francisco (1999).
3. Norman, D. Things that make us smart. New York: Doubleday/Currency (1993).
4. Tversky B, et al. Animation: can it facilitate? International Journal of Human-Computer Studies 57 (2002):247-262.
5. Bhavnani SK, Bellala G, Victor S et al. The Role of Complementary Bipartite Visual Analytical Representations in the Analysis of SNPs: A Case Study in Ancestral Informative Markers. J Am Med Inform Assoc.19 (2012):e5-e12.
6. Heer J, Bostock M, Ogievetsky V. A Tour through the Visualization Zoo. Communications of the ACM 53 (2010):59-67.
7. Shneiderman B. The Eyes Have It. A Task by Data Type Taxonomy for Information Visualization. Visual Languages (1996):336–343.
8. Yi JS, Kang YA, Stasko J, et al. Toward a Deeper Understanding of the Role of Interaction in Information Visualization. IEEE Transactions on Visualization and Computer Graphics 13 (2007).
9. Newman MEJ. Networks: An Introduction. Oxford University Press (2010).
10. Albert RK. Boolean Modeling of Genetic Regulatory Networks. Complex Networks (2004):459-481.
11. Goh K, Cusick M, Valle D, et al. The human disease network. Proc Natl Acad Sci U S A 104 (2007):8685.
12. Ideker T, Sharan R. Protein networks in disease. Genome Research 18 (2008):644.
13. Bhavnani SK, Bellala G, Ganesan A et al. The Nested Structure of Cancer Symptoms: Implications for Analyzing Co-occurrence and Managing Symptoms. Methods of Information in Medicine 49 (2010):581-591.
14. Bhavnani SK, et al. How Cytokines Co-occur across Asthma Patients: From Bipartite Network Analysis to a Molecular-Based Classification. Journal of Biomedical Informatics, 44 (2011) S24-S30.