Chapter 18
The Role of Visual Analytics in Asthma Phenotyping and Biomarker Discovery

Suresh K. Bhavnani, Justin Drake, and Rohit Divekar

Abstract The exponential growth of biomedical data related to diseases such as asthma far exceeds our cognitive abilities to comprehend it for tasks such as biomarker discovery, pathway identification, and molecular-based phenotyping. This chapter discusses the cognitive and task-based reasons for why methods from visual analytics can help in analyzing such large and complex asthma data, and demonstrates how one such approach called network visualization and analysis can be used to reveal important translational insights related to asthma. The demonstration of the method helps to identify the strengths and limitations of network analysis, in addition to areas for future research that can enhance the use of networks to analyze vast and complex biomedical datasets related to diseases such as asthma.

Keywords Asthma • Phenotypes • Visual analytics • Network analysis • Visualization • Bipartite networks • Multivariate analysis • Exploratory visual analysis • Quantitative verification • Emergent clusters • Inference of biological pathways • Molecular-based classification • Phenotyping • Biomarker discovery

S.K. Bhavnani, Ph.D. (✉) • J. Drake, B.S.
Institute for Translational Sciences, University of Texas Medical Branch,
6.168 Research Building 6, 301 University Blvd, Galveston, TX, USA
e-mail: skbhavnani@gmail.com; jad952@gmail.com

R. Divekar, M.D. Ph.D.
Division of Allergy and Immunology, University of Texas Medical Branch,
301, University Boulevard, Galveston, TX 77555, USA
e-mail: rddiveka@utmb.edu

A.R. Brasier (ed.), *Heterogeneity in Asthma*, Advances in Experimental Medicine and Biology 795, DOI 10.1007/978-1-4614-8603-9_18,

18.1 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 asthma. 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 information. For example, the identification of molecular pathways involved in different asthma phenotypes requires an interdisciplinary understanding of (1) biomarkers that are co-expressed across different groups of patients, (2) clinical characteristics of the patient groups across the biomarkers that are co-expressed, and (3) known and novel molecular pathways suggested by the patterns related to molecular and clinical patient profiles.

One approach to integrate and comprehend such complex information is through methods being developed in the new field of visual analytics. In this chapter we begin by presenting an overview of the evolving theoretical foundations for visual analytics, and the motivations to use methods from this field to analyze asthma data. Next, we focus on one form of visual analytics called networks which are particularly useful for analyzing complex molecular and clinical data. In contrast to the supervised learning methods (discussed in the last chapter) that use a priori information (e.g., cases and controls) to build predictive models, networks are considered to be an exemplar of unsupervised learning methods which do not use a priori information (e.g., cases and controls). We will demonstrate how this approach can be been used to identify asthma phenotypes and infer the molecular pathways involved in those phenotypes. These analyses reveal the strengths and limitations of the method, which are used to define a research agenda for advanced methods to enable in the future, comprehension of complex relationships in ever-increasing and complex asthma data.

18.2 Visual Analytics: Definition, Motivation, and Theoretical Foundations

Visual analytics is defined as the science of analytical reasoning, facilitated by interactive visual interfaces (Thomas and Cook 2005). 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 asthma data because of the nature of human cognition, and the nature of tasks performed by analysts.

18.2.1 Motivation for Visualizations

Visualizations of data are often powerful because they leverage the massively parallel architecture of the human visual system consisting of the eye and the visual cortex of the brain (Card et al. 1999). 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 (Thomas and Cook 2005). 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 processing of large and complex datasets such as those currently available for asthma.

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. Similarly, if a chart lacks a legend or axes labels, the visualization is difficult to comprehend because it cannot be mapped to concepts in the data. Finally, a road map pointing south is not very useful to a driver who is facing north because it requires a mental rotation of the map before it can be useful for navigation. Therefore visualizations need to be aligned with tasks (Norman 1993), data, and mental representations of the user (Tversky et al. 2002), before those visualizations can be effective in augmenting cognition.

18.2.2 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 (Bhavnani et al. 2012), 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, 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.

18.3 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 (Thomas and Cook 2005). For example, there are a number of attempts to define heuristics for the design of effective visualizations (e.g., Tufte 1983), and to classify visual analytical representations (e.g., Heer et al. 2010; Shneiderman 1996), and interaction methods at different levels of granularities and tasks (Yi et al. 2007). 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.

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 (Thomas and Cook 2005), which makes this approach particularly pertinent for tasks such as inferring biological pathways from molecular and clinical information in translational teams.

18.4 Network Visualization and Analysis: Making Sense of Asthma Molecular and Phenotype Information

Networks (Newman 2010) 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

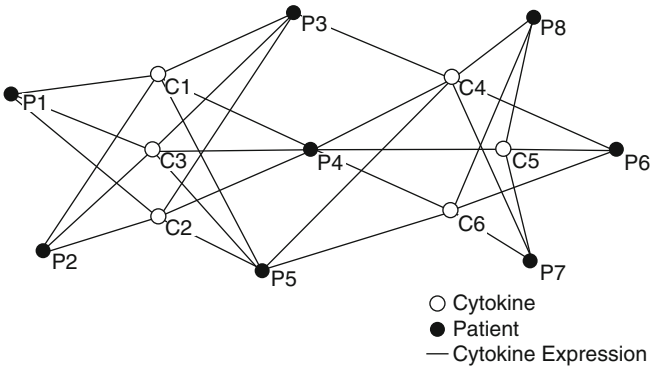

Fig. 18.1 A sample bipartite network where edges exist only between two different types of nodes. In this case, nodes represent either patients (*black*) or cytokines (*white*), and edges connecting the two represent cytokine expression

visualization. Networks are increasingly being used to analyze a wide range of molecular measurements related to gene regulation (Albert 2004), disease–gene associations (Goh et al. 2007), and disease–protein associations (Ideker and Sharan 2008). 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 18.1 shows a sample bipartite network where edges exist only between different types of entities (Newman 2010), in this case between patients and cytokines.[1]

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 involved across different emergent phenotypes. This three-step method used across our earlier studies (Bhavnani et al. 2007, 2010a, b) have revealed complex but comprehensible visual patterns, each prompting the use of quantitative methods that make the appropriate assumptions about the underlying data, which in turn led to inferences about the biomarkers and underlying mechanisms involved. Below we describe the methods used in each step, and their application to analyze a dataset of asthma patients and their cytokine expressions.

[1] Researchers have explored a wide range of network types including unipartite, directed, dynamic, and networks laid out in three dimensions to analyze complex data. As this wide range is beyond the scope of this chapter, we suggest other excellent sources (Newman 2010) for such information.

18.4.1 Exploratory Visual Analysis

Network analysis typically begins by transforming symbolic data into graphical elements in a network. To achieve this, the analyst needs to decide which *entities* in the data represent the nodes in the network, in addition to how other useful information can be mapped onto the node's shape, color, and size. Similarly, the analyst needs to decide which *relationships* between the entities in the data are represented by the edges in the network, in addition to how to map other useful information to the edge's thickness, color, and style. These selections are made based on an understanding of the kinds of relationships that are needed to be explored, and is often an iterative process based on an understanding of the domain and the nature of the data.

Once the symbolic data has been mapped to graphical elements, the resulting network is laid out so the nodes and edges can be visualized. The layout of nodes in a network can be done where either the distances between nodes has no meaning (e.g., nodes laid out randomly or along a geometric shape such as a line or circle) or where the distance between nodes represents a relationship such as similarity (e.g., similar cytokine expression profiles). Layouts where distance has meaning are typically generated through force-directed layout algorithms (Newman 2010). For example, the application of the *Kamada–Kawai* layout algorithm (well suited for small- to medium-sized networks in the range of 50–1,000 nodes) (Kamada and Kawai 1989; Nooy et al. 2005) to a network results in nodes with a similar pattern of connecting edge weights to be pulled together and those with different patterns to be pushed apart.

Figures 18.2–18.6 show the steps that were used to generate a bipartite network of 83 asthma patients, and 18 cytokines. Figure 18.2 shows how asthma patients, were represented as black nodes, and cytokines (molecules involved in intercellular signaling) were represented as white nodes. Furthermore, normalized cytokine expression values were represented as edges connecting each patient to each cytokine. These nodes were laid out equidistantly around a circle. Figure 18.3 shows the same network but where the edge thicknesses are proportional to the normalized cytokine expression values. Therefore, thick edges represent higher cytokine expression values compared to thin edges. Furthermore, the size of the node was made proportional to the total expression value of the connecting edges. Therefore, large patient nodes have overall higher aggregate cytokine expression values compared to smaller patient nodes.

Although the patients, cytokines, and the cytokine expression have been visually represented, the distances between the nodes have no meaning. To better comprehend the data, the patients who have higher cytokine expression value for a particular cytokine should be spatially closer to that cytokine compared to those who have lower cytokine expression value. This approach of using short distances between entities to show similarity, and long distances between entities to show dissimilarity is typical across clustering algorithms. As shown in Fig. 18.4 and reported in (Bhavnani et al. 2011a), application of the force-directed algorithm Kamada–Kawai to the circular layout results in nodes that have a similar pattern of cytokine

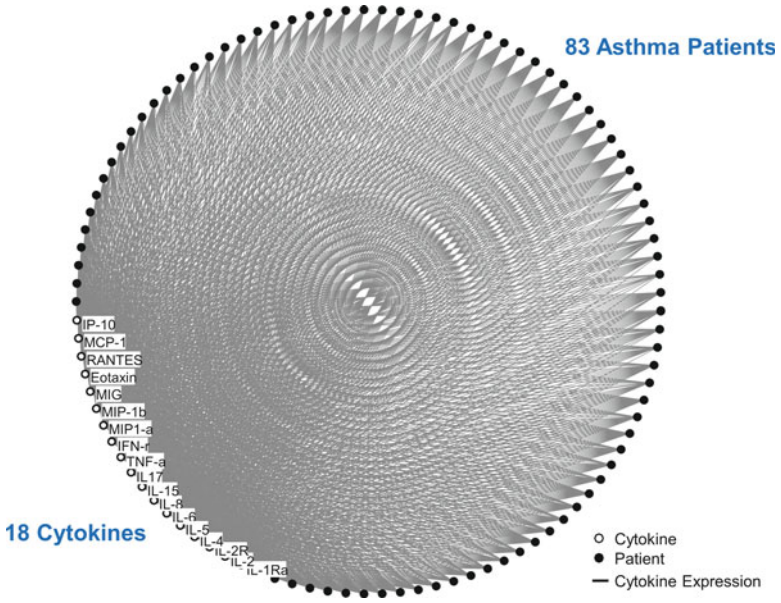

Fig. 18.2 A bipartite network showing patient nodes (*black*) and cytokine nodes (*white*) connected in pairs by edges which represent normalized cytokine expression. Patient and cytokine nodes were separately grouped and randomly laid out equidistantly around a *circle*

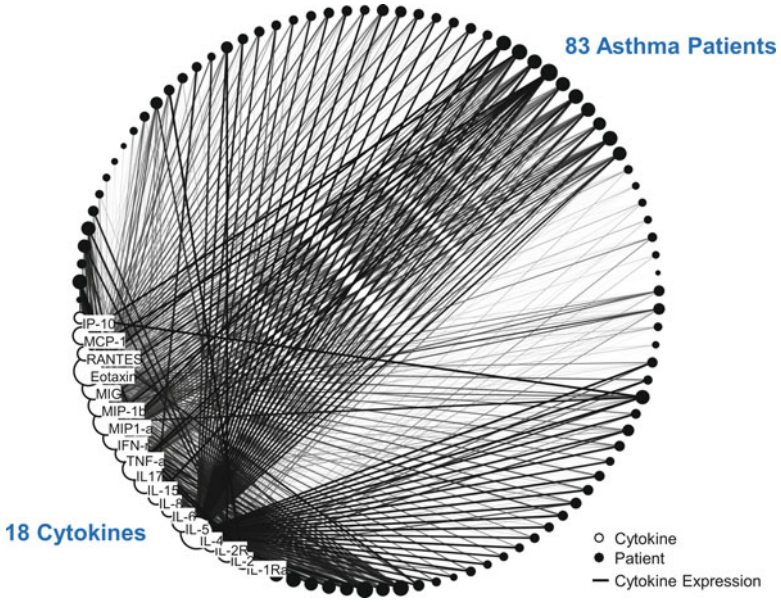

Fig. 18.3 The same network as in Fig. 18.2 but where edge thickness is proportional to the normalized cytokine expression value and the size of each node is proportional to the total expression values of the connecting edges. *Thick edges* represent higher cytokine expression values compared to *thin edges*. Similarly, larger patient nodes have higher aggregate cytokine expression values compared to smaller patient nodes

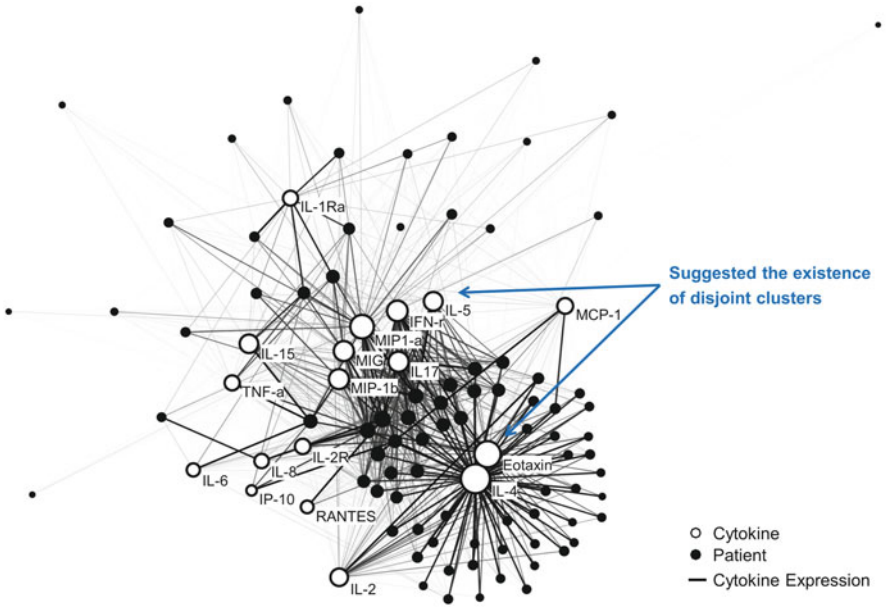

Fig. 18.4 Application of Kamada–Kawai, a force-directed algorithm, to the circular layout. The algorithm pulls nodes with similar cytokine expression patterns closer together, while pushing apart those with dissimilar expression patterns. The layout of the network suggested the existence of disjoint patient and cytokine clusters, and revealed intercluster relationships such as how the patient clusters express particular cytokine clusters. However, quantitative methods must be used to identify cluster boundaries

expression to be pulled together, and those that are not similar to be pushed apart. The resulting layout suggests that there exist distinct clusters of patients and cytokines. Furthermore, the layout also reveals the intercluster relationships such as which patient clusters are most closely related with which cytokine clusters.

While the network layout suggests the existence of distinct clusters, it is not designed to reveal the members of each cluster. We therefore need to use quantitative methods that are explicitly designed to identify the boundaries of clusters based on a multivariate analysis of the data.

18.4.2 Quantitative Verification and Validation

There exist a wide range of quantitative methods to verify and validate patterns discovered through network visualization methods. While in principle any statistical method can be used to quantitatively analyze a pattern observed in a network, many patterns are often analyzed using graph-based methods (Newman 2010) that

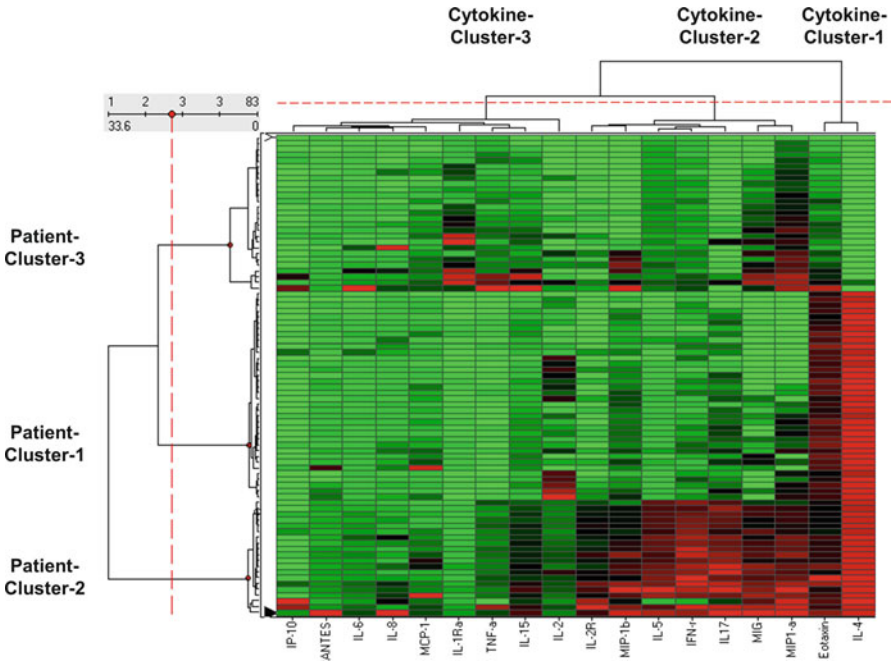

Fig. 18.5 A heat map where the *rows* represent patients, the *columns* represent cytokines, and the *colors* represent normalized cytokine values (*green* = 0, *red* = 1). The *rows* and *columns* are ordered based on the results of agglomerative hierarchical clustering. The patient and cytokine dendrograms are shown on the *vertical* and *horizontal axes*, respectively. Each dendrogram shows a natural break at three clusters indicated by the *red lines*

specialize in analyzing complex relationships. For example, *degree assortativity* measures whether one type of nodes in a network which have high weighted degree (e.g., patients who have large nodes in Fig. 18.3) are preferentially connected to another type of nodes that have high degree (e.g., cytokines that have large nodes in Fig. 18.1), or vice versa.

Another approach that can be used to verify patterns in a network is hierarchical clustering (Johnson and Wichern 1998). This unsupervised learning method attempts to identify the number and boundary of clusters in the data. For example, hierarchical clustering can be used to identify clusters of patients based on their relationship to cytokines, or clusters of cytokines based on their relationship to patients. The method begins by putting each node in a separate cluster, and then progressively joins nodes that are most similar based on their relationship to connected nodes. This progressive grouping generates a tree structure called a *dendrogram*, where distances between subsequent layers of the tree represent the strength of dissimilarity between the respective clusters; the larger the distance between two subsequent layers, the stronger the clustering. Analysts therefore determine the

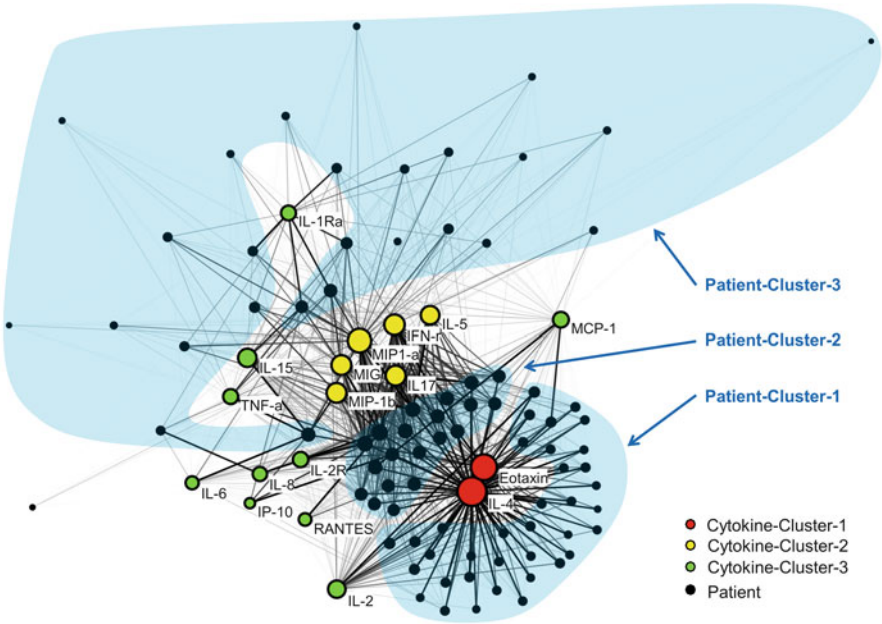

Fig. 18.6 Results of the agglomerative hierarchical clustering from Fig. 18.5 superimposed onto the *network* in Fig. 18.4 using *colors* to denote the three cytokine clusters and translucent *shapes* to denote the three patient clusters. The *network* shows that Patient-Cluster-1 highly expresses Cytokine-Cluster-1, while Patient-Cluster-3 highly expresses Cytokine-Cluster-3. The *network* also shows that Patient-Cluster-2 primarily expresses Cytokine-Cluster-2 and Cytokine-Cluster-1

number and membership of the clusters by identifying relatively large breaks between the layers in the dendrogram.

Given the wide range of quantitative methods available, the patterns in the network are used to guide the selection of the appropriate method. For example, if distinct clusters do not exist in a network, then it is not appropriate to apply a clustering algorithm to the network. This approach of selecting methods based on the inspection of the data is similar to how statisticians determine whether to use parametric or non-parametric inferential methods based on the underlying distribution of the data.

Because the network in Fig. 18.4 suggested the existence of disjoint clusters, hierarchical clustering was used to identify the boundary and members of the clusters. As shown in Fig. 18.5, the horizontal dendrogram represents the cytokine clusters, the vertical dendrogram represents the patient clusters, and the colored cells represent normalized cytokine expression ranging from green (0) to red (1). Each dendrogram shows a clear break at three clusters for cytokines, and for patients (as shown by the corresponding red dotted lines across each dendrogram).

While there may be clear breaks in the dendrograms, the overall pattern could have occurred by random chance. Patterns discovered in networks, and subsequently

the dendrograms, are therefore validated by determining their significance. One approach to do this is to compare the patterns in the data to random permutations of the network.

To test whether there were significant breaks in the dendrogram (denoting the existence of disjoint clusters), the variance, skewness, and kurtosis of the dissimilarities (generated by the hierarchical clustering algorithm) in the asthma network were compared to 1,000 permutations of the asthma network. For each network permutation, the number of nodes and the number of edges connected to each node, in addition to the edge weight distribution of patients were preserved when analyzing the cytokine dendrogram, and vice versa. Significant breaks in the asthma patient or cytokine dendrograms would result in a significantly larger variance, skewness, and kurtosis of the dissimilarity measures, compared to the same measures generated from the random networks.

As reported in (Bhavnani et al. 2011a) the results showed the clusteredness of the patients in the asthma network was significant as measured by the variance of the dissimilarities (Asthma=64.95, Random Mean=20.08, $p<0.001$ two-tailed test), skewness of the distribution of dissimilarities (Asthma=4.9, Random Mean=2.81, $p<0.001$ two-tailed test), and kurtosis of the distribution of dissimilarities (Asthma=30.24, Random Mean=14.78, $p<0.001$ two-tailed test). Furthermore, the results also showed that the clusteredness of the cytokine clusters was significant as measured by the variance of the dissimilarities (Asthma=837.62, Random Mean=46.69, $p<0.001$ two-tailed test), the skewness of the distribution of dissimilarities (Asthma=2.18, Random mean=0.49, $p<0.001$ two-tailed test), and kurtosis of the distribution of dissimilarities (Asthma=7.25, Random mean=2.49, $p<0.001$ two-tailed test).

To understand why the patients or cytokines were clustered, and how they related to each other, the cluster memberships were superimposed onto the network. As shown in Fig. 18.6, the cytokine nodes were colored to denote their membership in three separate clusters. In contrast, the patient clusters were denoted by closed translucent shapes to enable visual discrimination between patient and cytokine[2] clusters. As shown, Patient-Cluster-1 and Patient-Cluster-3 are enriched with Cytokine-Cluster-1 and Cytokine-Cluster-3, respectively. However, Patient-Cluster-2 is enriched with Cytokine-Cluster-1 and Cytokine-Cluster-2. The results of the quantitative analysis superimposed over the network visualization therefore helped to identify the intercluster relationships in the data.

[2] Such visual design decisions are currently loosely based on graphic design heuristics (Johnson and Wichern 1998) such as limiting the number of colors in the visualization to reduce visual overload. However, successful visualizations are often based on the graphic design expertise of the analyst who explores many variations of a visualization, and uses judgment to determine which one is most effective for the data, task, and mental representations of the domain experts who will be interpreting the results.

Table 18.1 Comparison of six independent pulmonary functions across the three patient clusters identified by the network analysis

Pulmonary function	*p* value with FDR correction
Max FVCpp/MPVLung	0.006*
Max FEV_1pp/MPVLung	0.0375*
Baseline FEV_1pp	0.0375*
Baseline FEV_1/FVC	0.1944
Max FEV_1 reversal	0.583
PC_{20} methacholine	0.0375*

Significant differences between the groups are indicated by asterisks based on a one-way, two-tailed Kruskal–Wallis test with an FDR correction. (*FVC* forced vital capacity, *FEV_1* forced expiratory volume in 1 s, *PC_{20} methacholine* dose of methacholine that produces 20 % fall in FEV_1, *FEV_1 albuterol reversal* percent change in FEV_1 in response to albuterol inhalation, *MPV* maximal postbronchodilator value, *pp* percent predicted). Permission pending

18.4.3 Inference of Biological Mechanisms and Asthma Phenotypes

While the visual and quantitative analysis helped to reveal patterns in the data, the ultimate goal of the network analysis is to infer the biological mechanisms involved, and the emergent sub-phenotypes in the data. This inferential step requires an integrated understanding of the molecular and clinical variables. One approach is to analyze how the patients in each emergent cluster (based on molecular profiles) differ in their clinical variables. This can be done with well-known statistical tests such as Kruskal–Wallis, a nonparametric test used to determine if the median of a variable is significantly different across many groups such as the clusters.

The Kruskal–Wallis test revealed patterns of pulmonary function across the three patient clusters (Bhavnani et al. 2011a). As shown in Table 18.1, four out of six pulmonary function measures were significantly different across the three clusters. In addition, we conducted a pairwise intercluster analysis, which revealed that Patient-Cluster-3 had three lung functions (Max FEV1pp/MPVLung, Baseline FEV1pp, and PC_{20} Methacholine) that were significantly higher than Patient-Cluster-1, and one lung function (Max FVCpp/MPVLung) that was significantly higher than Patient-Cluster-2. In contrast, Patient-Cluster-1, had only one lung function (Max FVCpp/MPVLung) that was significantly higher than Patient-Cluster-2. Patient-Cluster-3 therefore had less baseline airway obstruction (both FEV1 values were significantly higher), less hyper-reactivity to methacholine challenge (significantly higher PC20 Methacholine), and preserved pulmonary capacity (significantly higher FVC values) compared to the other two patient clusters.

The molecular and clinical profiles of the patients therefore helped to identify hypotheses for the mechanisms involved in asthma. As discussed in (Bhavnani et al.

2011a) the co-occurrence of Eotaxin and IL-4 (Cytokine-Cluster-1) is well aligned with a known sequence of molecular changes in asthma patients who often have a T-helper-2 (Th_2) lymphocyte-skewed immune response. This response results in the secretion of IL-4, which in turn triggers Eotaxin production by bronchial epithelial cells (Fujisawa et al. 2001). The resulting downstream actions include the activation and recruitment of tissue-resident eosinophils, a hallmark of early-stage asthma. The presence of Eotaxin and IL-4 in lung fluids therefore appears to indicate key substages of a complex molecular pathway in asthma, which explains their high co-occurrence in the network.

To comprehend the biological significance of cytokines in Cytokine-Cluster-2 (IL-5, IFN-γ, MIP1a, MIG, IL-17, and MIP-1β), they were entered into the Ingenuity Pathway Analysis (IPA) application. The results suggest that the frequent co-occurrence of these cytokines is regulated by the innate inflammatory nuclear factor-κB pathway (NF-κB). NF-κB is a potent pro-inflammatory transcription factor that activates expression of cytokine networks. In addition, persistent NF-κB activation has been linked to uncontrolled/acute exacerbations of asthma (Gagliardo et al. 2003). The frequent co-occurrence of this set of cytokines therefore implies the presence of a distinctly different pro-inflammatory state, when compared to the IL-4–Eotaxin process discussed above.

The above cytokine clusters combined with the pulmonary functions of the patients, provide a biological explanation for the patient clusters. The strong relationship of Patient-Cluster-1 to Cytokine-Cluster-1 suggests that patients in this cluster have disease driven primarily by Th_2 inflammation. In contrast, Patient-Cluster-2 has a strong relationship to both Cytokine-Clusters-1 and -2. This result implies that patients in Patient-Cluster-2 have a component of activated innate inflammatory pathways. Additional evidence for this inference of state-based clusters is evidenced by differences in pulmonary function across the clusters discussed earlier. Patient-Cluster-3 which has the lowest cytokine values for both of the above cytokine clusters also has the largest number of significant differences in obstructive airway disease parameters in pulmonary function testing, and lowest airway reactivity response to methacholine compared to Patient-Clusters-1 and -2. This result implies that Patient-Cluster-3 represents a subgroup of asthmatics with preserved pulmonary function and greatest response to albuterol without active inflammation.

Informed by these underlying molecular processes, the network analysis of patients and cytokines therefore implies a state-based classification of asthma patients. The results also provide evidence for the growing consensus (Bousquet et al. 2010) that asthma is a dynamic disease where the same patient could enter different asthmatic states based on environmental and/or other triggers. Future studies that include such information could lead to a better understanding of the relationship between triggers and resulting asthmatic states, which could translate into more effective personalized treatment and prevention approaches for each patient.

18.5 Strengths and Limitations of Network Analysis

Network analysis has several strengths and limitations, whose understanding can lead to informed uses of the method, appropriate interpretation of the results, and insights for future enhancements and complementary methods.

18.5.1 Strengths

Network visualization and analysis provide four distinct strengths for enabling rapid discovery of patterns in complex biomedical data.

1. Networks (that are based on graph theory) provide a tight integration between visual and quantitative analysis. For example, as shown in the Fig. 18.6, networks enable the simultaneous visualization of multiple raw values (e.g., patient–cytokine associations, cytokine values, patient attributes), aggregated values (e.g., sum of cytokine values), and emergent global patterns (e.g., clusters) in a uniform representation. This uniform visual representation leverages the parallel processing power of the visual cortex enabling the comprehension of complex multivariate, quantitative relationships.
2. Networks do not require a priori assumptions about the relationship of nodes within the data, in contrast to hierarchical clustering or k-means which assume the data is hierarchically organized or contain disjoint clusters, respectively. Instead, by using a simple pairwise representation of nodes and edges, network layouts enable the identification of multiple structures (e.g., hierarchical, disjoint, overlapping, nested) in a single representation (Nooy et al. 2005). Therefore, while layout algorithms such as Kamada–Kawai depend on the force-directed assumption and its implementation, such algorithms are viewed as less biased for data exploration because they do not impose a particular cluster structure on the data, often leading to the identification of more complex structures in the data (Bhavnani et al. 2010a). The overall approach therefore enables a more informed selection of quantitative methods to verify the patterns in the data.
3. Networks preserve highly correlated variables (such as cytokines) and display them through clustering. Furthermore, the bipartite network representation enables the comprehension of intercluster relationships such as between variable (e.g., cytokines) clusters and patient clusters. These features provide important clues to domain experts about the pathways that involve those variables. This is in contrast to many supervised learning methods which drop highly correlated variables in an attempt to identify a small number of variables that together can explain the maximum amount of variance in the data. While this approach is powerful for developing predictive models, the reduction in variables could limit the inference of biological pathways involved in the disease.
4. Networks enable high interactivity enabling the rapid modification of the visual representation to match the changing task and representation needs of analysts during the analysis process. For example, nodes that represent patients in a network can be

interactively colored or reshaped to represent different variables such as gender and race, enabling the discovery of how they relate to the rest of the network.

18.5.2 Limitations

Networks have three important limitations that need to be understood for their proper use.

1. While node shape, color, and size can represent different variables, there is a limit on the number of variables that can be simultaneously represented. Furthermore, a visual representation can get overloaded with too many colors and shapes, which can mask rather than reveal important patterns in the data. Therefore, while networks can reveal complex multivariate patterns in the data based on a few variables, they often require complimentary visual analytical representations such as Circos ideograms (Krzywinski et al. 2009; Bhavnani et al. 2011b) to explore data that is high dimensional (e.g., large number of attributes related to entities such as patients in the network).
2. While networks provide a rich vocabulary of graphical elements to represent data, their design and use requires iterative refinement based on an understanding of the domain, knowledge of graphic design and cognitive heuristics, and the use of complex interfaces that are designed for those facile in computation. This combination of knowledge required to conduct network analyses makes domain experts dependent on network analysts to generate and refine the representations, which can limit the rapid exploration and interpretation of complex data.
3. While network layout algorithms are designed to reveal complex and unbiased patterns in multivariate data, they often fail to show any patterns in the data resulting in what is colloquially called a "hairball." In such cases, the nodes appear to be randomly laid out providing little guidance for how to proceed with the analysis. While network applications offer many interactive methods to filter data such as by dropping edges and nodes based on different thresholds, many of these methods are arbitrary and therefore unjustifiable to use when searching for patterns especially in important domains such as biomedicine. There is therefore a need to develop more systematic and defensible methods to find hidden patterns in network hairballs.

18.6 Future Directions in Network Analysis Related to the Analysis of Biomedical Data

The limitations of networks discussed above motivate three important future research directions to make network analysis more effective for the analysis of biomedical data such as those related to asthma: (1) As nodes can only represent a

limited number of variables simultaneously, there is a need to use complementary visual analytical representations. This motivates the development of a framework designed to guide the selection and use of multiple visual analytical representations based on the nature of the tasks and of data. (2) Because network analyses currently require many iterations to design the representation through the use of complex interfaces, there is a need for systems that are streamlined for specific tasks such as biomarker discovery. (3) Given that many network layouts show no structure, future algorithms should attempt to integrate methods from supervised learning to enable the discovery of hidden patterns. These research directions could enable the rapid discovery of patterns in the age of big data and translational medicine.

References

Albert RK (2004) Boolean modeling of genetic regulatory networks. Complex Networks; 459–481

Bhavnani SK, Abraham A, Demeniuk C et al (2007) Network analysis of toxic chemicals and symptoms: implications for designing first-responder systems. Proc of AMIA 2007:51–55

Bhavnani SK, Bellala G, Ganesan A et al (2010a) The nested structure of cancer symptoms: implications for analyzing co-occurrence and managing symptoms. Methods Inf Med 49:581–591

Bhavnani SK, Carini S, Ross J et al (2010b) Network analysis of clinical trials on depression: implications for comparative effectiveness research. Proc of AMIA 2010:51–5

Bhavnani SK, Victor S, Calhoun WJ et al (2011a) How cytokines co-occur across asthma patients: from bipartite network analysis to a molecular-based classification. J Biomed Inform 44:S24–S30

Bhavnani SK, Pillai R, Calhoun WJ, et al. (2011b) How circos ideograms complement networks: a case study in asthma. Proc of AMIA Summit on Translational Bioinformatics

Bhavnani SK, Bellala G, Victor S et al (2012) 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:e5–e12

Bousquet J, Mantzouranis E, Cruz AA et al (2010) Uniform definition of asthma severity, control, and exacerbations: document presented for the World Health Organization Consultation on Severe Asthma. J Allergy Clin Immunol 126(5):926–38

Card S, Mackinlay JD, Shneiderman B (1999) Readings in information visualization: using vision to think. Morgan Kaufmann, San Francisco, CA

Fujisawa T, Kato Y et al (2001) Chemokine production by the BEAS-2B human bronchial epithelial cells: differential regulation of eotaxin, IL-8, and RANTES by T_H2- and T_H1-derived cytokines. J Allergy Clin Immunol 105:126–133

Gagliardo R, Chanez P, Mathieu M et al (2003) Persistent activation of nuclear factor–κB signaling pathway in severe uncontrolled asthma. Am J Respir Crit Care Med 168:1190–1198

Goh K, Cusick M, Valle D et al (2007) The human disease network. Proc Natl Acad Sci USA 104:8685

Heer J, Bostock M, Ogievetsky V (2010) A tour through the visualization zoo. Commun ACM 53:59–67

Ideker T, Sharan R (2008) Protein networks in disease. Genome Res 18:644

Johnson RA, Wichern DW (1998) Applied mutlivariate statistical analysis. Prentice-Hall, NJ

Kamada T, Kawai S (1989) An algorithm for drawing general undirected graphs. Inform Process Lett 31:7–15

Krzywinski M, Schein J, Birol I et al (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19:1639–1645

Newman MEJ (2010) Networks: an introduction. Oxford University Press, Oxford
Nooy W, Mrvar A, Batagelj V (2005) Exploratory social network analysis with Pajek. Cambridge University Press, Cambridge
Norman D (1993) Things that make us smart. Doubleday/Currency, New York
Shneiderman B (1996) The eyes have it: a task by data type taxonomy for information visualization. Visual Languages; 336–343
Thomas JJ, Cook KA (2005) Illuminating the path: the R&D agenda for visual analytics. National Visualization and Analytics Center
Tufte ER (1983) The visual display of quantitative information. Graphics Press, Chesire, CT
Tversky B, Morrison JB, Betrancourt M (2002) Animation: can it facilitate? Int J Hum Comput Stud 57:247–262
Yi JS, Kang YA, Stasko J et al. (2007) Toward a deeper understanding of the role of interaction in information visualization. IEEE Transact Visualiz Comput Graph 13