HHS Public Access

Author manuscript

Proteomics. Author manuscript; available in PMC 2015 June 18.

Published in final edited form as:

Proteomics. 2015 April ; 15(8): 1405–1418. doi:10.1002/pmic.201400451.

Unlocking Proteomic Heterogeneity in Complex Diseases through Visual Analytics

Suresh K. Bhavnani, Ph.D.[1,2,3,§], **Bryant Dang, BS.**[1], **Gowtham Bellala, Ph.D.**[4], **Rohit Divekar, M.B.,B.S., Ph.D.**[5], **Shyam Visweswaran, M.D., Ph.D.**[6,7], **Allan Brasier, M.D.**[1,3], and **Alex Kurosky, Ph.D.**[8]

[1]Institute for Translational Sciences, University of Texas Medical Branch, Galveston, TX, USA

[2]Institute for Human Infections and Immunity, University of Texas Medical Branch, Galveston, TX, USA

[3]Sealy Center for Molecular Medicine, University of Texas Medical Branch, Galveston, TX, USA

[4]Hewlett Packard Laboratories, Palo Alto, CA, USA

[5]Division of Allergic Diseases, Mayo Clinic, Rochester, MN, USA

[6]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA

[7]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA

[8]Department of Biochemistry & Molecular Biology, University of Texas Medical Branch, Galveston, TX, USA

Abstract

Despite years of preclinical development, biological interventions designed to treat complex diseases like asthma often fail in phase III clinical trials. These failures suggest that current methods to analyze biomedical data might be missing critical aspects of biological complexity such as the assumption that cases and controls come from homogeneous distributions. Here we discuss why and how methods from the rapidly evolving field of visual analytics can help translational teams (consisting of biologists, clinicians, and bioinformaticians) to address the challenge of modeling and inferring heterogeneity in the proteomic and phenotypic profiles of patients with complex diseases. Because a primary goal of visual analytics is to amplify the cognitive capacities of humans for detecting patterns in complex data, we begin with an overview of the cognitive foundations for the field of visual analytics. Next, we organize the primary ways in which a specific form of visual analytics called networks have been used to model and infer biological mechanisms, which help to identify the properties of networks that are particularly useful for the discovery and analysis of proteomic heterogeneity in complex diseases. We describe one such approach called subject-protein networks, and demonstrate its application on two proteomic datasets. This demonstration provides insights to help translational teams overcome

[§]Corresponding author. Suresh K. Bhavnani, Institute for Translational Sciences, Institute for Human Infections and Immunity, University of Texas Medical Branch, 301 University Blvd, Galveston, TX, USA, skbhavnani@gmail.com.

CONTRIBUTIONS

S.K.B. conceived and supervised the project, S.K.B., A.R.B., S.V., R.D., A.K., and B.D. refined the method and interpreted the results, and S.K.B., S.V., R.D., and B.D. wrote and refined the manuscript.

theoretical, practical, and pedagogical hurdles for the widespread use of subject-protein networks for analyzing molecular heterogeneities, with the translational goal of designing biomarker-based clinical trials, and accelerating the development of personalized approaches to medicine.

Keywords

Visual Analytics; Proteomic Heterogeneity; Network Analysis; Subject-Protein Networks; Molecular and Clinical Profiles; Personalized Medicine

INTRODUCTION

Although vast resources have been spent in developing new therapies for complex diseases such as asthma, many drugs designed to target specific proteins have failed in clinical trials for reasons ranging from drug ineffectiveness to toxic side effects [1, 2]. For example, although several *in-vitro* studies strongly suggested that blocking IL-5 (critical in Th2 inflammation and allergic response) would be effective in asthma treatment [3, 4], clinical trials using mepolizumab (a monoclonal antibody to IL-5) failed to show a statistically significant improvement in key clinical parameters [5]. Subsequent studies found that only a subgroup of asthma patients might benefit from mepolizumab treatment [6, 7], suggesting that there existed considerable heterogeneity in molecular etiologies among asthma patients.

Such realizations have led to a growing consensus that current methods used for identifying proteomic targets in complex diseases (defined as having multifactorial etiologies) are not designed to reveal *proteomic heterogeneities* (defined as differences in the proteomic profiles of patients), resulting in missed opportunities for the design of therapies that are targeted to specific patient subgroups. For example, most methods used to analyze molecular data assume that cases and controls can each be characterized by a single mean and variance, and identify variables that are univariately (e.g., chi-square) or multivariately (e.g., regression) significant across the two distributions. This focus on identifying variables that explain the difference between cases and controls potentially conceals patient subgroups, whose identification could lead to more targeted therapeutics, a necessary component of personalized medicine [8].

One approach to help multidisciplinary translational teams [9] (typically consisting of biologists such as proteomic researchers, clinicians, and bioinformaticians) integrate and comprehend such complex proteomic data is through methods from the evolving field of visual analytics [10]. Because a primary goal of visual analytics is to help humans amplify their cognitive capabilities for detecting complex patterns in data, we begin by presenting an overview of the theoretical foundations for visual analytics, and the motivations to use methods from this field to analyze proteomic data. Next, we organize the major ways in which a specific form of visual analytics called networks have been used to model and infer biological mechanisms such as genetic regulatory pathways. This organization helps to identify the properties of networks that are especially effective for the analysis of molecular heterogeneities and their respective mechanisms. We demonstrate the use of an approach that uses these network properties to help identify proteomic heterogeneity and their respective pathways across two proteomic datasets. These demonstrations reveal the

strengths and limitations of the method leading to insights for the development of future advanced approaches that can accelerate the discovery of molecular heterogeneities through the integrated analysis of *multi-omics* data.

VISUAL ANALYTICS: THEORETICAL FOUNDATIONS

Visual analytics is defined as "the science of analytical reasoning facilitated by interactive visual interfaces" [10]. Visual analytical methods are designed to augment cognitive reasoning by transforming symbolic and numeric data (e.g., numbers in a spreadsheet) into *visualizations* (e.g., a scatter plot), which can be manipulated through *interaction* (e.g., highlight outliers in the scatter plot). As described below, visualizations and interactions with those visualizations can be powerful for making important discoveries in proteomic data because of the nature of cognition and the tasks that translational teams typically perform.

The Role of Visualizations in Analytical Reasoning

Data visualization can be powerful for analyzing biomedical data because it leverages the parallel architecture of the human visual system consisting of the eye and the visual cortex. This parallel cognitive architecture enables the rapid comprehension of multiple complex relationships simultaneously such as similarities, anomalies, and trends, which can lead to insights about relationships in the data [10, 11]. For example, Figure 1A shows a spreadsheet that contains normalized cytokine expression levels in patients before and after taking a drug. Determining which of the two conditions have more patients with cytokine level >= 0.8 is tedious and error prone because the analyst needs to compare the entry in each cell with >= 0.8, remember the result of each comparison, and then count the number of patients with cytokine levels >= 0.8 in each column to make the final comparison. Because such symbolic processing is done serially, the cognitive load increases with an increasing number of entries, and therefore a large number of data points can easily overwhelm the cognitive capacities of an analyst.

As shown in Figure 1B, when cells with values >= 0.8 are highlighted in red, the resulting visual representation allows the analyst to determine more rapidly that the left column has more red cells compared to the right column. This occurs because the representation leverages the power of the human visual system to process in parallel the red cells in each column. Moreover, the visualization in Figure 1B shifts information from an internal to an external representation, which speeds up the task of counting the number of individuals with high cytokine levels in each column [12].

Nevertheless, not all data visualizations are helpful in enhancing cognition. For example, a road map that is oriented to the south is not helpful for a driver who is facing north since she will have to rotate the map mentally before identifying the route. Similarly, an organizational chart that shows employee names and locations laid out in a hierarchy based on rank is not helpful if the task is to identify patterns based on the geographic location of the employees. Furthermore, if a chart has a missing legend, it is difficult to map domain concepts onto the visualization. Visualizations therefore need to be aligned with tasks [13],

mental representations of the user [14], and the data before they can be helpful in enhancing cognition.

The Role of Interactivity in Analytical Reasoning

Although data visualizations can be useful if they are aligned with tasks, mental representations, and data, they often can be inadequate for analyzing complex data. This is because analysis typically entails multiple subtasks such as discovery, inspection, confirmation, and explanation [15], each of which requires a different representation of the data. For example, if the task in Figure 1 is to comprehend the relationship of gender to condition, then it is useful to sort the data based on gender. As shown in Figure 1C, interaction with the data through such sorting reveals that the drug has no effect on females (low values remain low, and high values remain high), but consistently lowers cytokine expression in males (all high values become low). Thus, interaction with data visualizations can reveal relationships that are often not apparent in a single static visualization of the data.

Furthermore, interactivity is particularly critical when an interdisciplinary team is involved in the analysis because each member of the team typically needs a different representation of the same data. For example, a molecular biologist might be interested in which genes are co-expressed across individuals of a specific phenotype, while a clinician might be interested in the response to a therapy in patients with similar gene expression profiles. To be able to handle several tasks and mental representations, interactivity is critical to transform parts, or the entire visual representation to generate new representations.

Theories Related to Visual Analytics

Visual analytics draws on existing theories and taxonomies from cognitive psychology, computer science, and graphic design. However, integrative theories and principles underlying visual analytics are still in early stages of development [10]. For example, researchers have developed several classifications of visual representations [16, 17], and articulated the goals of interaction at different levels of granularity [18, 19].

One such classification [16] of visual representations categorizes them into (1) **time series** (e.g., line plots showing how the expression of cytokines change over time), (2) **statistical distributions** (e.g., box-and-whisker plots of cytokine expression across patients), (3) **maps** (e.g., heatmaps of proteomic expression across patients), (4) **hierarchies** (e.g., dendrograms resulting from hierarchical clustering showing how cytokines cluster based on differences in expression across patients), and (5) **networks** (e.g., protein-protein interaction networks). Each of these visual representations are constructed of basic elements referred to as *marks* (e.g., points, lines, and areas), which can have graphical attributes referred to as *channels* (e.g., position, size, value, texture, color, orientation, and shape) [20, 21]. A mark and associated channels are together referred to as a *visual encoding* of symbolic information such as protein expression.

Once a visual representation of the data is generated, interactivity allows it to be transformed in part or in whole into a new visualization. For example, a top-down tree may be transformed into a circular tree, and nodes in a tree may be colored based on specific

properties such as gender. Several attempts have been made to categorize such interactions with visualizations at different levels of granularity. For example, low-level interaction intents have been classified [19] as: filter, retrieve value, compute derived value, sort, determine range, find extremum, characterize distribution, find anomalies, cluster, and correlate. In contrast, higher level interaction intents have been classified as: select, explore, reconfigure, encode, abstract/elaborate, filter and connect [18].

While classifications help to organize different types of visualizations and interactions, several principles and heuristics have been proposed to guide the design of effective visualizations and interaction methods. These include the *principle of closure* (tendency of an incomplete shape like a circle drawn with a missing segments to appear like a closed form) from the Gestalt Laws of Grouping [22], and the *pop-out effect* (tendency of objects with bright colors, movement, and large size to stand out compared to other objects in the visual field, and conditions when this phenomenon fails leading to effects like visual clutter) from the Biased Competition Theory [23]. Furthermore, several design heuristics have been proposed for creating effective visualizations such as (1) to increase the "data-ink" ratio [24] when designing quantitative displays of information (stripping unnecessary decoration and other "chartjunk" that interferes with comprehension), (2) how to select visual channels for different data types (e.g., position in a Euclidean plane is the most effective channel to represent nominal, ordinal, and continuous data) [25] and (3) the strategy of *overview first, zoom and filter, then details-on-demand* to help interaction designers create interfaces that help humans explore and comprehend large and complex visual displays of information [17]. Many of these principles interact in complex ways, and therefore require skill for their application (often requiring trial and error) in order to generate an effective visualization that is useful for specific tasks and users.

In addition to the above classifications of visual representations and interactions, there have been early attempts at developing theoretical frameworks that explain how key elements of visual analytics enable analytical reasoning. For example, Liu and Stasko [26] proposed a framework that integrates the three major components of visual analytics: visual representation, interaction, and analytical reasoning. In this framework, internal and external representations are coupled to enable three related goals: (1) *External anchoring* or the process of associating conceptual structures (e.g., cytokine expression >= 0.8) to elements of the visualization (red colored cells), (2) *Information foraging* or the process of exploring the external visual representation through extraction (e.g., counting red cells related to cytokine expression) or through transformation (e.g., sorting according to gender) of the representation, and (3) *Cognitive offloading* or the process of transferring a conceptual structure onto the visual representation to reduce cognitive load (e.g., encircling or annotating in Figure 1C all female patients who have cytokine expression >= 0.8 before and after taking the drug). Though such frameworks are still in early stages of development, they provide a first step in developing a theoretical understanding of how visual analytics enables analytical reasoning,

Finally, it is pertinent to note that the field of visual analytics has considerable overlap with the fields of *scientific visualization* (concerned with the visualization of real-world three-dimensional phenomena such as earthquakes) and *information visualization* (focused on

visual representations of abstract data such as relationships). However, while visual analytics obviously shares visualization with both fields, it differs from them because it is also focused on developing interactive methods that facilitate analytical reasoning and making sense of complex information by analysts individually, or in groups [10].

In summary, although many heuristics, principles, and frameworks from cognitive psychology, computer science and graphic design have been proposed to inform visual analytics, integrated theories for this field have yet to emerge. However, despite the lack of such theories, one form of visual analytics, namely networks, have been widely used to model and infer biological mechanisms. The next section organizes these attempts, with the goal of identifying the properties of networks that make them particularly suitable for modeling and inferring proteomic heterogeneity.

APPLICATION OF VISUAL ANALYTICS TO MODEL AND INFER BIOLOGICAL MECHANISMS

In recent years there has been a growing realization that most biological phenomena emerge from complex relationships among numerous components of a cell such as DNA, RNA, proteins, and metabolites. This realization has motivated a shift in the analysis of biological phenomena from a *reductionist* approach of analyzing individual molecules and their immediate neighbors, to a *holistic* approach of modeling and inferring relationships among all molecules related to a biological system [27, 28]. Understanding biological processes using this holistic approach by integrating the individual molecular associations has become a central goal of systems biology [29].

Because the systems biology approach embraces complex relationships among numerous molecular components, network analysis has gained primacy as a fundamental approach for modeling and inferring biological phenomena [30, 31]. This is because networks enable (1) the visual representation of associations between pairs of molecules, in addition to how all the pair-wise associations in the network result in a biological system, (2) the quantitative analysis and validation of local and global patterns because the representation is a graph and therefore has mathematical properties, and (3) interaction with the visualization which helps translational teams to explore different aspects of the data with the goal of comprehending the overall biological system.

A network consists of a set of nodes, which are connected in pairs by edges [32]. Nodes can represent one or more types of entities (e.g., subjects or cytokines), and edges between nodes represent a relationship between the entities (e.g., a case has a particular cytokine expression value). *Unipartite* networks have nodes that are of one type of entity (e.g., proteins), and the edges represent associations between them (e.g., protein-protein interaction). In contrast, Figure 2 shows an example of a *bipartite* network where nodes represent two types of entities, and edges exist only between different types of entities [32] (e.g., between subjects and cytokines representing cytokine expression).

Nodes and edges (marks represented as points and lines in a network) can have several graphical attributes (channels) to represent different aspects of the data. Nodes can represent

different aspects of the data through shape and color (e.g., white and red circles can represent cases and controls), and edges connecting nodes can either be *undirected* (as shown in Figure 2), or *directed* such as an arrow representing directionality (e.g., a transcription factor regulates a gene in a biological pathway). Furthermore, edges can be *weighted* to represent continuous values (e.g., thickness of the edge is proportional to the normalized amount of cytokine expression as shown in Figure 2), or *unweighted* to represent a binary relationship (e.g., a drug is related to a target). Additionally, edges can be colored or have style to represent a type of relationship (e.g., red and blue dashed lines representing up and down regulation), or be arced, tapered, or bundled to improve comprehension [33, 34]. Networks can also be dynamic representing temporal changes (e.g., spread of a virus through a social network [35], and laid out in three dimensions to analyze complex data which are described in more detail elsewhere [36]. Furthermore, there is a wide range of network analytical measures (e.g., modularity and degree centrality), whose description is beyond the scope of this paper, but which have been extensively covered in recent reviews and books [32].

The above graphical properties of nodes and edges designed to represent different aspects of the data have been combined to generate different network types to help model or infer a wide range of biological mechanisms. As shown in Table 1, these networks can be organized based on four major types of biological relationships:

1. **Process networks**. This class of networks is designed to directly model biological mechanisms typically using a bipartite or multipartite (where nodes can represent many types of entities) network. For example, in a *gene regulatory network* the nodes represent genes or transcription factors, and a directed edge either shows which gene generates which transcription factor, or which transcription factor regulates which gene. This approach has been used to help biologists comprehend a biological system as a whole, and to identify regulation phenomena such as NfkB signaling [37] or cytokine signaling [38]. As shown in Table 1, signal transduction networks [39] and metabolic networks [40] also model biological processes, but the nodes and edges have different semantic meaning compared to gene regulatory networks. Process networks have been modeled using tools such as Cytoscape [41] which can layout the nodes from left to right to reflect the directionality of the overall sequence of regulation (see [42], and [43] reviews of network analysis tools).

2. **Interaction networks**. This class of networks is designed to model molecules that interact with each other. Because such interactions have no directionality, they are typically modeled using a unipartite network with undirected edges. For example, in a protein-protein interaction network, the nodes represent proteins and the undirected and unweighted edges represent the binary relationship that two proteins can interact with each other. Protein-protein interaction networks [44] have been used to identify network properties of individual proteins (e.g., hub proteins that have many edges because they can interact with many other proteins and are evolutionarily important), in addition to global network properties (e.g., densely connected clusters representing protein complexes related to a specific function).

As shown in Table 1, interaction networks have also been used to model how genes interact (via proteins) with each other [45]. Typical tools for such analyses include Cytoscape, and STRING [44].

3. **Similarity networks**. This class of networks is designed to model how entities such as molecules or subjects are similar to each other using a statistical measure to represent similarity, and a weighted unipartite network to represent the pair-wise similarity. For example, in a *gene co-expression network* the nodes represent genes, and the weighted edges represent some statistical measure of similarity between gene pairs, such as the Pearson's correlation of two genes co-occurring across subjects (see [46] for an analysis of key similarity measures). Gene co-expression networks are therefore not designed to directly model biological mechanisms, but rather are used to infer mechanisms based on how genes cluster. For example, a gene-gene co-expression network was used to infer the function of five genes with cellular processes of cell proliferation and cell cycle that were previously uncharacterized [47]. As shown in Table 1, other examples of similarity networks include patient-patient similarity networks which aim to reveal how patients are similar or dissimilar based on molecular or clinical variables [48], in addition to metabolite-metabolite correlation networks [49]. Typical tools to construct such networks include Cytoscape and Pajek [50].

4. **Affiliation networks**. This class of networks is designed to model how one kind of entity is affiliated to another kind of entity. This is typically done by explicitly representing both entities using a bipartite network with weighted or unweighted edges. For example, in a *drug-target network*, the nodes represent drugs or targets, and unweighted edges represent which drug is affiliated with which target. Drug-target networks [51] have been used to infer new purposes for known drugs. For example, researchers have (1) modeled addictive drugs and their targets as a bipartite network, (2) included in the network non-addictive drugs that shared at least one target with the addictive drugs, and (3) analyzed how the non-addictive drugs clustered with the addictive drugs suggesting a new purpose for the non-addictive drugs. Other types of networks in this class include disease-gene networks [52], and species-microbiome networks [53]. Finally, subject-protein networks contain nodes which represent subjects or proteins, and weighted edges represent protein expression. Such networks have been useful in identifying proteomic heterogeneity within subjects (based on how the subjects are clustered), and their respective pathways (based on which proteins are enriched in subject clusters) [54]. Typical tools to construct such bipartite networks include Pajek.

The above classification of networks that have been used to model and infer biological phenomena suggests that subject-protein networks are most useful in analyzing proteomic heterogeneity because they explicitly model both subjects and proteins in the same representation. This duality can therefore reveal not only how subject clusters are similar or different based on their proteomic profile, but also how those subject clusters are related to protein clusters, and therefore the possible mechanisms activated or absent in those subject clusters.

Subject-protein networks therefore differ from similarity networks which are unipartite, and created by aggregating one side of the bipartite relationship in the data such as patient clusters, based on an aggregated similarity score of proteins, or vice versa.

While similarity networks can reveal either subject clusters, or protein clusters, they cannot reveal how subjects are *related* to protein clusters. This relationship is fundamental for inferring the mechanisms that are unique or shared among the subject clusters and are therefore critical for comprehending proteomic heterogeneity.

Furthermore, as described in the next section, because of their dual-node representation, subject-protein networks also enable an integrated visualization and analysis of not only the above discussed proteomic profiles, but also how those profiles are associated with subject variables (e.g., clinical, demographic, and environmental), and with the protein variables (e.g., function and pathways to which they belong). Furthermore, an important property of subject-protein networks is the modeling of experimental data, in comparison to modeling existing knowledge culled from databases of molecules and their function and interactions.

Despite the power of the above subject-protein bipartite representation, to the best of our knowledge there appear to be few attempts to use them for analyzing proteomic heterogeneity. The next section therefore aims to address this missed opportunity by describing the methodology for modeling and analyzing subject-protein networks, and the subsequent section demonstrates how that method has been used to reveal different forms of heterogeneity in two proteomic datasets.

METHOD FOR SUBJECT-PROTEIN NETWORK ANALYSIS: DISCOVERING PROTEOMIC HETEROGENEITY

Because subject-protein networks have key properties that help translational teams to infer proteomic heterogeneity and the respective mechanisms, we have found it useful to use its bipartite representation throughout the modeling and analytical phases. This approach is different from the commonly-used approach of starting with a bipartite network representation but then converting it into a unipartite network [52] of only subjects or only molecules. As discussed earlier, this approach cannot reveal how biomarkers and subjects co-cluster. Below we describe a three stage approach of analyzing proteomic data using subject-protein networks throughout the analytical process.

1. Exploratory Visual Analysis

The first stage is to transform the symbolic relationships (protein expression) between subjects and proteins in the data into a visual representation for analysis of heterogeneities. As illustrated in Figure 3.1A, to maintain a strong intuition about the relationships across the variables, we normalized each variable to range from 0–1 using the min-max range normalizing method [54, 55]. This enabled a straight-forward interpretation of the edges in the network such as being able to compare the maximum value in one variable, to the maximum in another variable (see Appendix A for more details and rationale for the min-max range normalization method). Next, we identified outliers in each variable using the Grubb's test [56], and discussed with the domain expert (typically a biologist with extensive

experience in the domain of the data) whether the identified outliers were biologically feasible (and therefore important to preserve in the data), or an error in measurement (in which case it was removed, and the min-max normalization repeated).

Next, as shown in Figure 3.1B, the normalized data were transformed into a network representation. Similar to Figure 2, nodes represented subjects or molecules, and edges represented normalized molecular measurements. Additionally, the size of the nodes was made proportional to the sum of the edges that connected to them, which provided a visual cue about the variance of the molecular expressions. For example, large nodes represented either subjects who had high overall cytokine expressions, or cytokines that were highly expressed across the subjects.

Because Euclidean inter-node distance is an effective channel to represent similarity [25], as shown in Figure 2.1C, we used a force-directed algorithm called *Kamada Kawai* [57] in the network analysis tool called *Pajek* [50] to lay out the nodes. This algorithm results in pushing together nodes with similar edge weight profiles, and pushing apart those with dissimilar profiles. Layouts generated through force-directed algorithms are approximate and designed to reveal overall topologies, rather than to show exact distances between nodes.

The overall network topology (Figure 3.1D) was then inspected by a domain expert to identify the nature of the heterogeneities in the data. Examples of topologies with heterogeneities include: (1) *distinct clustering* where subject clusters are associated with one or more molecule clusters, and (2) a *core-periphery* topology where there exists a network core consisting of subjects with high expression of some or all variables, and a network periphery consisting of subjects with low expression of the same variables.

2. Quantitative Verification and Validation

As shown in Figure 3.2A, the topology identified in the network was used to guide the selection of appropriate quantitative methods for verification and validation. For example, if there were distinct clusters in the network, we used modularity [32] to identify the number and boundaries of the clusters. However, if there were more complex topologies in the network such as a core-periphery, then we used hierarchical clustering [55] which we have found to be more successful in distinguishing the core from the periphery [58, 59]. The topologies were then compared to 1000 random permutations of the data to test whether the patterns could have occurred by chance [54]. Once the subgroups of patients and molecules were identified and validated, the cluster boundaries were superimposed onto the bipartite network either by making the nodes in a cluster the same color, or by drawing outlines around the node clusters to denote the boundaries (Figure 2.2B).

Because the subject clusters were determined based on their molecular profiles, we integrated the clinical variables by analyzing which of them was significantly different across the clusters. This can be done using univariate statistical methods such as Kruskal-Wallis [55], or a multivariate analysis using regression to determine for example which combination of clinical variables best distinguish one patient cluster from the others. To further explore the association of clinical and molecular variables, significant categorical

(e.g., gender) and continuous variables (e.g., systolic blood pressure) were superimposed onto the original network using color and node size respectively (Figure 3.2C–D).

3. Inference of Heterogeneity and Biological Mechanisms

The ultimate goal of our analytical method was to generate data-driven hypotheses about the molecular and clinical heterogeneities. Towards that goal, we used databases such as Ingenuity Pathway Analysis [60] and STRING [44] to identify pathways that included some or all of the molecules that co-occurred in the identified clusters. This analysis helped the translational team to recognize pathways that were either already known to be activated in the disease being analyzed, or those known to be activated in another disease, and therefore novel for the current disease. If no known pathway was found, then the biologist proposed a new pathway that was activated for the patient subgroup identified (Figure 3.3A).

The translational team then integrated the inferred pathways that were significantly associated with specific patient subgroups, with the significant clinical variables across the subgroups to define a hypothesis for the heterogeneities (Figure 3.3B). The resulting heterogeneities and pathways pinpointed hypotheses which could be tested through future laboratory experiments, or in other datasets of the same disease.

APPLICATIONS OF THE SUBJECT-PROTEIN NETWORK ANALYSIS METHOD

We have used the above general methodology on several biomedical datasets [54, 58, 59], of which two are briefly described here because (1) the respective bipartite networks had distinctly different topologies demonstrating the power of the methodology to identify important associations in the data, and (2) the hypotheses of heterogeneity and molecular pathways generated from these topologies were considered by domain experts to be novel contributions worthy of publication. While both of these analyses have been published before [54, 58], we briefly present the results here to highlight and compare the kinds of topologies and inferences that can be made using subject-protein networks, with the goal of identifying their strengths for analyzing proteomic heterogeneity.

(a) Asthma—The asthma network consisted of 83 asthma patients, 18 candidate cytokines, and 9 lung function variables (see Figure 3 for the intermediate steps of the analysis, and Supplementary Material A for details of the data and steps of the method). The analysis (Figure 4) revealed three distinct patient clusters that had a complex but comprehensible association with three distinct cytokine clusters: Patient-Cluster-1 and Patient-Cluster-3 were associated with two separate cytokine clusters and different levels of expression, but Patient-Cluster-2 had high expression of two cytokine clusters, resulting in a complex but comprehensible inter-cluster relationship among patients and cytokines.

Analysis of the clinical variables revealed that the three patient clusters were significantly different based on six lung function variables (e.g., FEV_1, a measure of the lung capacity). A biologist and pulmonologist integrated the molecular-based clustering with the clinical variables and inferred three separate subgroups of patients with activation of different biological mechanisms [54]. For example, Patient-Cluster-3 had high co-expression of eotaxin and IL-4, significantly lower co-expression of the other cytokines, and significantly

lower lung function. The domain experts therefore inferred that these patients with significantly lower lung function have a Th2 lymphocyte skewed immune response resulting in the secretion of IL-4, which induces eotaxin production by bronchial epithelial cells. This in turn results in downstream actions including the activation and recruitment of tissue-resident eosinophils, a marker of early stage asthma, suggesting a different approach to their treatment compared to patients in other clusters.

(b) Rickettsial Infections—The rickettsia network consisted of 49 Mediterranean Spotted Fever (MSF) patients, 36 Dermacentor spp.-borne necrosis-erythema lymphadenopathy (DEBONEL) patients, and 26 candidate cytokines (see Supplementary Material B for details of the data and steps of the method). The DEBONEL infection is considered milder compared to the MSF infection, and the goal was to analyze how the candidate cytokines were expressed across both phenotypes.

The analysis (Figure 5) revealed a core-periphery network topology where there were 12 MSF patients with high overall cytokine expression of 5 cytokines in the network core, and the remaining patients of both phenotypes with low overall cytokine expression were distributed in the network periphery [58]. Furthermore, 7 of the 12 patients in the core had evidence of thrombocytopenia, and the 5 cytokines in the core were implicated in pro-inflammatory pathways. A pulmonologist from the translation team integrated the molecular-based clustering with the clinical variables and inferred that the patients in the core had an amplification of inflammatory responses, resulting in diffused endothelial injury and vascular leakage, and therefore at highest risk of severe disease [58].

Discussion—The above two applications of subject-protein network analysis revealed two substantially different types of heterogeneity in proteomic data. The asthma network revealed patients characterized by three distinct multivariate combinations of cytokine expression. These profiles resulted in three patient clusters with a complex but comprehensible relationship with three cytokine clusters. Furthermore, by integrating this network topology with the clinical variables, the translation team inferred three distinct proteomic heterogeneities, each with their respective mechanisms [54]. In contrast, the rickettsia network had one patient group that had high expression of 5 cytokines and low expression of the remaining 21 cytokines, and another patient group that had low expression of majority of the 26 cytokines. In other words, both patient groups had similarly low expression for most cytokines, but one patient group had substantially higher co-expression of just 5 cytokines. These two profiles resulted in the core-periphery topology reflecting the high overlap between the two groups. This result enabled the translational team to integrate four types of information: (1) associations revealed by the bipartite network topology as described above, (2) relationship of the topology to the clinical variables of the patients, (3) prior domain knowledge that only a small percentage of patients with rickettsial infections have severe reactions, and (4) the mechanisms implied by the 5 strongly-expressed cytokines. This led the team to infer that only the patients in the core had an amplification of inflammatory responses (as evidenced by the 5 highly expressed cytokines also in the core) in a phenomenon referred to as a *cytokine storm* [58]) resulting in a severe form of the disease. This phenomenon was largely absent in the periphery patients.

In summary, the bipartite network layout in both projects revealed a topology consisting of all subjects, proteins, and their relationships together in the *same* external representation. This externalization of key elements and relationships in the data into a unified visualization was designed to leverage the parallel processing power of the visual cortex to detect and comprehend complex patterns among the represented elements and relationships. Practically, it enabled the translational team to derive a holistic understanding of how the subjects were similar or different based on their proteomic profiles, leading to a cogent understanding of the heterogeneity and mechanisms involved.

As we have argued elsewhere [15], such complex differences in molecular profiles across subjects are difficult to discover and comprehend from the heatmap representation commonly used to analyze bipartite molecular relationships. This is because heatmaps have only one degree of freedom on each of the vertical and horizontal axes, allowing a row or column (e.g., representing subjects and variables respectively) to be adjacent to a maximum of two other columns or rows. This severely restricts the kinds of inter-cluster relationships that can be easily discovered. In contrast, networks have two degrees of freedom where nodes can move in the x-axis and y-axis simultaneously enabling the discovery of complex inter-cluster relationships that are comprehensible as demonstrated by the asthma and rickettsia networks.

CONCLUSIONS AND FUTURE RESEARCH

In response to a growing realization that current methods to analyze proteomic data might be missing critical aspects of biological complexity such as molecular and phenotypic heterogeneity, we explored why and how methods from visual analytics could help translational teams overcome this hurdle. A review of the theoretical foundations of visual analytics suggests that although there exist many heuristics, principles, and frameworks from cognitive psychology, computer science and graphic design that inform visual analytics, integrated theories for this field have yet to emerge. Furthermore, a review of how network visualization and analysis have been used to model and infer biological phenomena helped to identify the properties of networks exemplified by subject-protein networks that are particularly useful for the analysis of proteomic heterogeneity. Given the growing realization that both target and patient selection play a critical role in the success of clinical trials [61], we believe that subject-protein networks could be used to identify subgroups of patients (e.g., those with and without activation of the IL-5 signaling pathway) as inclusionary criteria in clinical trials that target a specific biological pathway.

Reflecting on our experience in using subject-protein networks to identify heterogeneities in complex diseases [54, 58, 59], we have come to appreciate two factors that are critical for the successful application of this method. First, we believe that the bipartite representation itself should be used consistently to layer different types of information during the exploratory, verification, and inferential stages. This representational consistency enables translational teams to comprehend the complex associations between the molecular and clinical information. While this fact can be derived from cognitive theories related to external representations [10, 12], most projects either transform bipartite networks into unipartite networks often for the convenience of analysis, or use the bipartite representation

as a way to present results of analyses conducted without the use of the representation. Second, the use of bipartite networks (and for that matter the use of many other visual representations) is dependent on the involvement of a domain expert who is willing to complement the dominant paradigm of hypothesis testing (which tends to focus on micro phenomena such as single molecules or pathways) with a willingness to explore the macro phenomena about a disease, of which heterogeneity is a prime example.

However, bipartite networks currently have several theoretical and practical limitations. Theoretically, while subject nodes can simultaneously represent a few variables such as gender, blood pressure, and phenotype using color, size, and shape respectively, there is an upper limit on the number of variables that can be simultaneously visualized. We have explored alternate representations such as Circos [62–64] which overcomes this limitation, but which have important trade-offs such as being unable to show patient or protein clusters through positioning in the Euclidean plane, an important channel provided by network layouts. Therefore, there is a need for integrative frameworks which could for example help to determine which combination of visual representations are best suited for different tasks such as discovering heterogeneities. Furthermore, force-directed layout algorithms 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, and which are often arbitrarily removed in the search for network structure. Therefore we need more systematic, defensible, and transparent methods to discover hidden structures in network hairballs.

Another limitation of the method in its current form is that it has been used to model only one type of molecular data, namely protein levels in the same network. However, there are increasing opportunities and need to conduct multi-omics analysis such as the integrated analysis of protein levels, gene expression, and metabolite concentrations across a cohort of subjects. Our current research is therefore exploring two natural extensions for analyzing such multi-omics data: (a) the bipartite network could represent subject clusters based on the primary molecular type such as proteins, and a regression analysis could be used to determine which combination of the other omics variables are significantly expressed across the subject clusters; (b) the bipartite network could represent subjects and *all* the molecular types using a normalization method that enables each molecular type to have the same *interpretive range* (e.g., 0=lowest value, 0.5=middle value, 1=highest value, for a specific molecule across the subjects) to enable comparison across the different omics types. If such approaches are successful, the concept of subject-protein networks could be generalized to *subject-molecule* or even *subject-variable* networks to model a wide range of variables ranging from genes to co-morbidities across subjects.

Practically, visual analytical tools tend to be designed for analysts, often requiring substantial programming and knowledge to generate appropriate visualizations, and therefore limiting the use of the methods by biologists and clinicians. This limitation motivates the need for tools that enable biologists and clinicians to explore data on their own so that they can better leverage their domain knowledge in interpreting the patterns in the data. Furthermore, as visual analytics progressively becomes a necessary part of data-driven hypotheses generation, there is a need to include the skills of generating and interpreting integrated visual analytics in biomedical informatics curricula. Such theoretical, practical,

and pedagogical advances have the potential for accelerating the identification of molecular and phenotypic heterogeneities in complex diseases, which is an important step towards the design of biomarker-based clinical trials, and for achieving the goals of personalized medicine.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This research was supported in part by NIH 1UL1TR000071 UTMB CTSA (ARB), the Institute for Human Infections and Immunity at UTMB, the Rising Star Award from University of Texas Systems (SKB), CDC/NIOSH #R21OH009441-01A2 (SKB), and NHLB: Contract # HHSN268201000037C-0-0-1 (AK) for the NIH/NHLBI Proteomics Center for Airway Inflammation.

REFERENCES

1. Arrowsmith J. Trial watch: Phase II failures: 2008–2010. Nat Rev Drug Discov. 2011; 10:328–329. [PubMed: 21532551]
2. Arrowsmith J. Trial watch: phase III and submission failures: 2007–2010. Nat Rev Drug Discov. 2011; 10:87. [PubMed: 21283095]
3. Mauser PJ, Pitman AM, Fernandez X, Foran SK, et al. Effects of an antibody to interleukin-5 in a monkey model of asthma. Am. J. Respir. Crit. Care Med. 1995; 152:467–472. [PubMed: 7633694]
4. Shardonofsky FR, Venzor J 3rd, Barrios R, Leong KP, Huston DP. Therapeutic efficacy of an anti-IL-5 monoclonal antibody delivered into the respiratory tract in a murine model of asthma. J. Allergy Clin. Immunol. 1999; 104:215–221. [PubMed: 10400864]
5. Flood-Page P, Swenson C, Faiferman I, Matthews J, et al. A study to evaluate safety and efficacy of mepolizumab in patients with moderate persistent asthma. Am. J. Respir. Crit. Care Med. 2007; 176:1062–1071. [PubMed: 17872493]
6. Nair P, Pizzichini MMM, Kjarsgaard M, Inman MD, et al. Mepolizumab for Prednisone-Dependent Asthma with Sputum Eosinophilia. N. Engl. J. Med. 2009; 360:985–993. [PubMed: 19264687]
7. Ortega HG, Liu MC, Pavord ID, Brusselle GG, et al. Mepolizumab Treatment in Patients with Severe Eosinophilic Asthma. The New England Journal of Medicine. 2014
8. Waldman SA, Terzic A. Therapeutic targeting: a crucible for individualized medicine. Clin. Pharmacol. Ther. 2008; 83:651–654. [PubMed: 18425084]
9. Calhoun WJ, Wooten K, Bhavnani S, Anderson KE, et al. The CTSA as an exemplar framework for developing multidisciplinary translational teams. Clin. Transl. Sci. 2013; 6:60–71. [PubMed: 23399092]
10. Thomas JJ, Cook KA. Illuminating the path: the R&D agenda for visual analytics. 2005
11. Stuart, KC.; Jock, DM.; Ben, S. Readings in information visualization: using vision to think. Morgan Kaufmann Publishers Inc.; 1999. p. 686
12. Zhang J, Norman DA. Representations in Distributed Cognitive Tasks. Cognitive Science. 1994:87–122.
13. Norman, D. Things that make us smart. New York: Doubleday/Currency; 1993.
14. Norman, DA. The Design of Everyday Things. Basic Books, Inc.; 2002.
15. Bhavnani SK, Bellala G, Victor S, Bassler KE, Visweswaran S. 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. 2012; 19:e5–e12. [PubMed: 22718038]
16. Heer J, Bostock M, Ogievetsky V. A Tour through the Visualization Zoo. Communications of the ACM. 53:59–67.
17. Shneiderman B. The Eyes Have It : A Task by Data Type Taxonomy for Information Visualization. Visual Languages. 1996:336–343.

18. Yi JS, Kang Ya, Stasko JT, Jacko JA. Toward a Deeper Understanding of the Role of Interaction in Information Visualization. IEEE Transactions on Visualization and Computer Graphics. 2007; 13
19. Amar, R.; Eagan, J.; Stasko, J. Low-Level Components of Analytic Activity in Information Visualization; Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization; 2005. p. 15
20. Bertin, J. Semiology of graphics : diagrams, networks, maps / Jacques Bertin; translated by William J. Berg. Madison, Wis: University of Wisconsin Press; 1983.
21. Munzer, T. Visualization Analysis and Design. A K Peters/CRC Press; 2014.
22. Smith, B. Foundations of Gestalt Theory. Philosophia: 1988.
23. Desimone R, Duncan J. Neural mechanisms of selective visual attention. Annu. Rev. Neurosci. 1995; 18:193–222. [PubMed: 7605061]
24. Tufte, ER. The visual display of quantitative information. Graphics Press; 1986.
25. Mackinlay J. Automating the design of graphical presentations of relational information. ACM Trans. Graph. 1986; 5:110–141.
26. Liu Z, Stasko J. Theories in Information Visualization: What, Why and How. Workshop on the Role of Theory in Information Visualization. 2010
27. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004; 5:101–113. [PubMed: 14735121]
28. Fernie AR, Trethewey RN, Krotzky AJ, Willmitzer L. Metabolite profiling: from diagnostics to systems biology. Nat. Rev. Mol. Cell Biol. 2004; 5:763–769. [PubMed: 15340383]
29. Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. Nat Rev Genet. 2010; 11:476–486. [PubMed: 20531367]
30. Junker, BH.; Schreiber, F. Analysis of Biological Networks (Wiley Series in Bioinformatics). Wiley-Interscience; 2008.
31. Ma'ayan A. Introduction to Network Analysis in Systems Biology. Science signaling. 2011; 4:tr5–tr5. [PubMed: 21917719]
32. Newman, MEJ. Networks: An Introduction. Oxford University Press; 2010.
33. Holten D, Isenberg P, Fekete J-D, Van Wijk JJ. Performance Evaluation of Tapered, Curved, and Animated Directed-Edge Representations in Node-Link Graphs. 2010:10.
34. Holten D. Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data. Visualization and Computer Graphics, IEEE Transactions on. 2006; 12:741–748.
35. Christakis NA, Fowler JH. Social Network Sensors for Early Detection of Contagious Outbreaks. PLoS One. 2010; 5:e12948. [PubMed: 20856792]
36. Bhavnani SK, Ganesan A, Hall T, Maslowski E, et al. Discovering hidden relationships between renal diseases and regulated genes through 3D network visualizations. BMC Res. Notes. 2010; 3:296. [PubMed: 21070623]
37. Brasier AR. The nuclear factor-kappaB-interleukin-6 signalling pathway mediating vascular inflammation. Cardiovasc. Res. 2010; 86:211–218. [PubMed: 20202975]
38. Barnes PJ. The cytokine network in asthma and chronic obstructive pulmonary disease. The Journal of Clinical Investigation. 2008; 118:3546–3556. [PubMed: 18982161]
39. Finkel T. Signal transduction by reactive oxygen species. J. Cell Biol. 2011; 194:7–15. [PubMed: 21746850]
40. Okuda S, Yamada T, Hamajima M, Itoh M, et al. KEGG Atlas mapping for global analysis of metabolic pathways. Nucleic Acids Res. 2008; 36:W423–W426. [PubMed: 18477636]
41. Shannon P, Markiel A, Ozier O, Baliga NS, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13:2498–2504. [PubMed: 14597658]
42. Pavlopoulos GA, Wegener AL, Schneider R. A survey of visualization tools for biological network analysis. BioData mining. 2008; 1:12. [PubMed: 19040716]
43. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, et al. Visualization of omics data for systems biology. Nature methods. 2010; 7:S56–S68. [PubMed: 20195258]

44. Jensen LJ, Kuhn M, Stark M, Chaffron S, et al. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res. 2009; 37:D412–D416. [PubMed: 18940858]
45. Tong AH, Evangelista M, Parsons AB, Xu H, et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science. 2001; 294:2364–2368. [PubMed: 11743205]
46. Bass JIF, Diallo A, Nelson J, Soto JM, et al. Using networks to measure similarity between genes: association index selection. Nat Meth. 2013; 10:1169–1176.
47. Stuart JM, Segal E, Koller D, Kim SK. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science. 2003; 302:249–255. [PubMed: 12934013]
48. Bauer-Mehren A, Lependu P, Iyer SV, Harpaz R, et al. Network analysis of unstructured EHR data for clinical research. AMIA Jt Summits Transl Sci Proc. 2013; 2013:14–18. [PubMed: 24303229]
49. Weckwerth W, Loureiro ME, Wenzel K, Fiehn O. Differential metabolic networks unravel the effects of silent plant phenotypes. Proc. Natl. Acad. Sci. U.S.A. 2004; 101:7809–7814. [PubMed: 15136733]
50. Nooy, W.; Mrvar, A.; Batagelj, V. Exploratory Social Network Analysis with Pajek. Cambridge University Press; 2005.
51. Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M. Drug-target network. Nat. Biotechnol. 2007; 25:1119–1126. [PubMed: 17921997]
52. Goh K-I, Cusick ME, Valle D, Childs B, et al. The human disease network. Proc. Natl. Acad. Sci. U.S.A. 2007; 104:8685. [PubMed: 17502601]
53. Muegge BD, Kuczynski J, Knights D, Clemente JC, et al. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science. 2011; 332:970–974. [PubMed: 21596990]
54. Bhavnani SK, Victor S, Calhoun WJ, Busse WW, et al. How Cytokines Co-occur across Asthma Patients: From Bipartite Network Analysis to a Molecular-Based Classification. Journal of Biomedical Informatics. 2011; 44:S24–S30. [PubMed: 21986291]
55. Johnson, RA.; Wichern, DW. Applied Multivariate Statistical Analysis. Prentice-Hall, NJ: 1998.
56. Grubbs FE. Sample Criteria for Testing Outlying Observations. 1950:27–58.
57. Kamada T, Kawai S. An algorithm for drawing general undirected graphs. Information Processing Letters. 1989; 31:7–15.
58. Bhavnani SK, Drake J, Gowtham B, Dang B, et al. How Cytokines Co-occur across Rickettsioses Patients: From Bipartite Visual Analytics to Mechanistic Inferences of a Cytokine Storm. Proceedings of AMIA Summit on Translational Bioinformatics. 2013
59. Bhavnani SK, Dang B, Caro M, Bellala G, et al. Heterogeneity within and across Pediatric Pulmonary Infections: From Bipartite Networks to At-Risk Subphenotypes. Proceedings of AMIA Summit on Translational Bioinformatics. 2014
60. Ingenuity. http://www.ingenuity.com/products/ipa.
61. Cook D, Brown D, Alexander R, March R, et al. Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. Nat Rev Drug Discov. 2014; 13:419–431. [PubMed: 24833294]
62. Krzywinski M, Schein J, Birol I, Connors J, et al. Circos: an Information Aesthetic for Comparative Genomics. Genome Res. 2009; 19:1639–1645. [PubMed: 19541911]
63. Bhavnani SK, Pillai R, Calhoun WJ, Brasier AR. How Circos Ideograms Complement Networks: A Case Study in Asthma. Proceedings of AMIA Summit on Translational Bioinformatics. 2011
64. Bhavnani SK, Abbas M, McMicken V, Oezguen N, Tupa JA. iCircos: Visual Analytics for Translational Bioinformatics. Proceedings of ACM International Health Informatics Symposium. 2012
65. Brasier AR, Victor S, Boetticher G, Ju H, et al. Molecular phenotyping of severe asthma using pattern recognition of bronchoalveolar lavage-derived cytokines. J. Allergy Clin. Immunol. 2008; 121:30–37. e36. [PubMed: 18206505]
66. Proceedings of the ATS workshop on refractory asthma: current understanding, recommendations, and unanswered questions. American Thoracic Society. Am. J. Respir. Crit. Care Med. 2000; 162:2341–2351. [PubMed: 11112161]

67. Kotera M, Goto S, Kanehisa M. Predictive genomic and metabolomic analysis for the standardization of enzyme data. Perspectives in Science. 2014

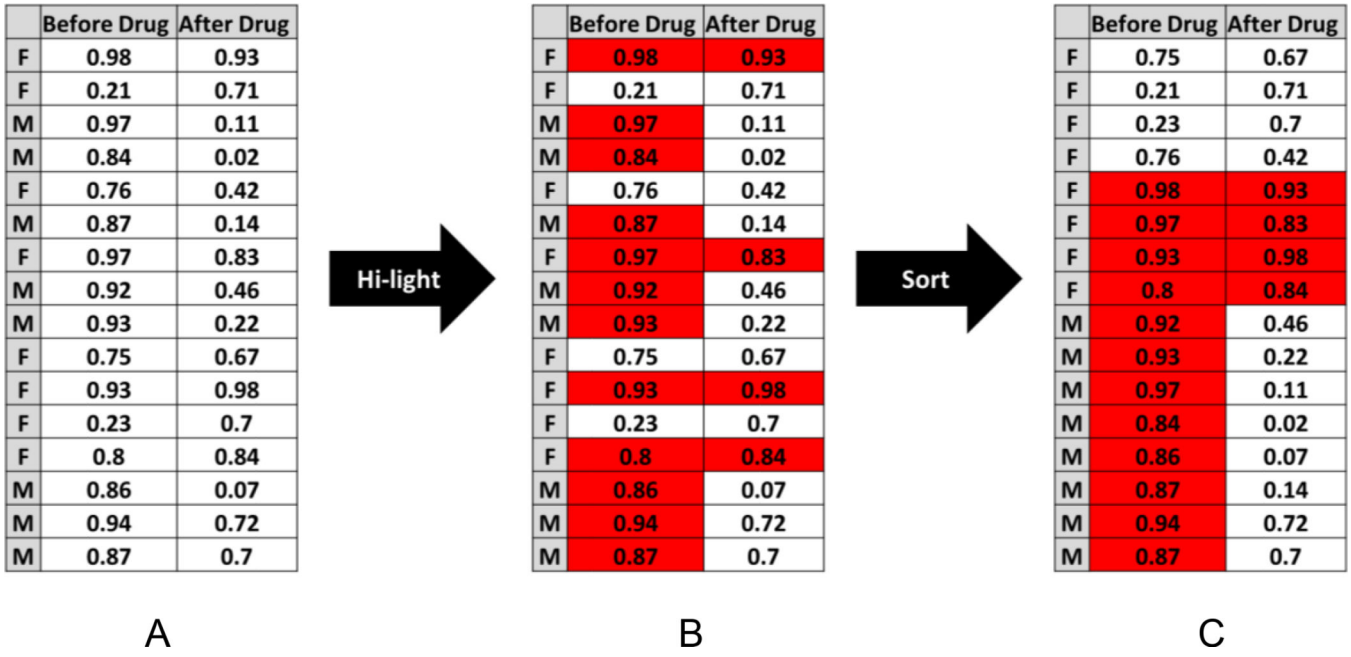


Figure 1.

An example of how symbolic data in a spreadsheet (A) when converted into a visual representation (B) leverages the parallel processing abilities of the visual cortex which enables faster comprehension of patterns in the data. Because visual processing is parallel in nature, it scales to handle large amounts of data. When the same data is sorted by gender (C), the visual representation reveals yet another pattern demonstrating how interaction with the data is a critical aspect of visual analytics, and can guide the verification of the patterns using the appropriate quantitative measures.

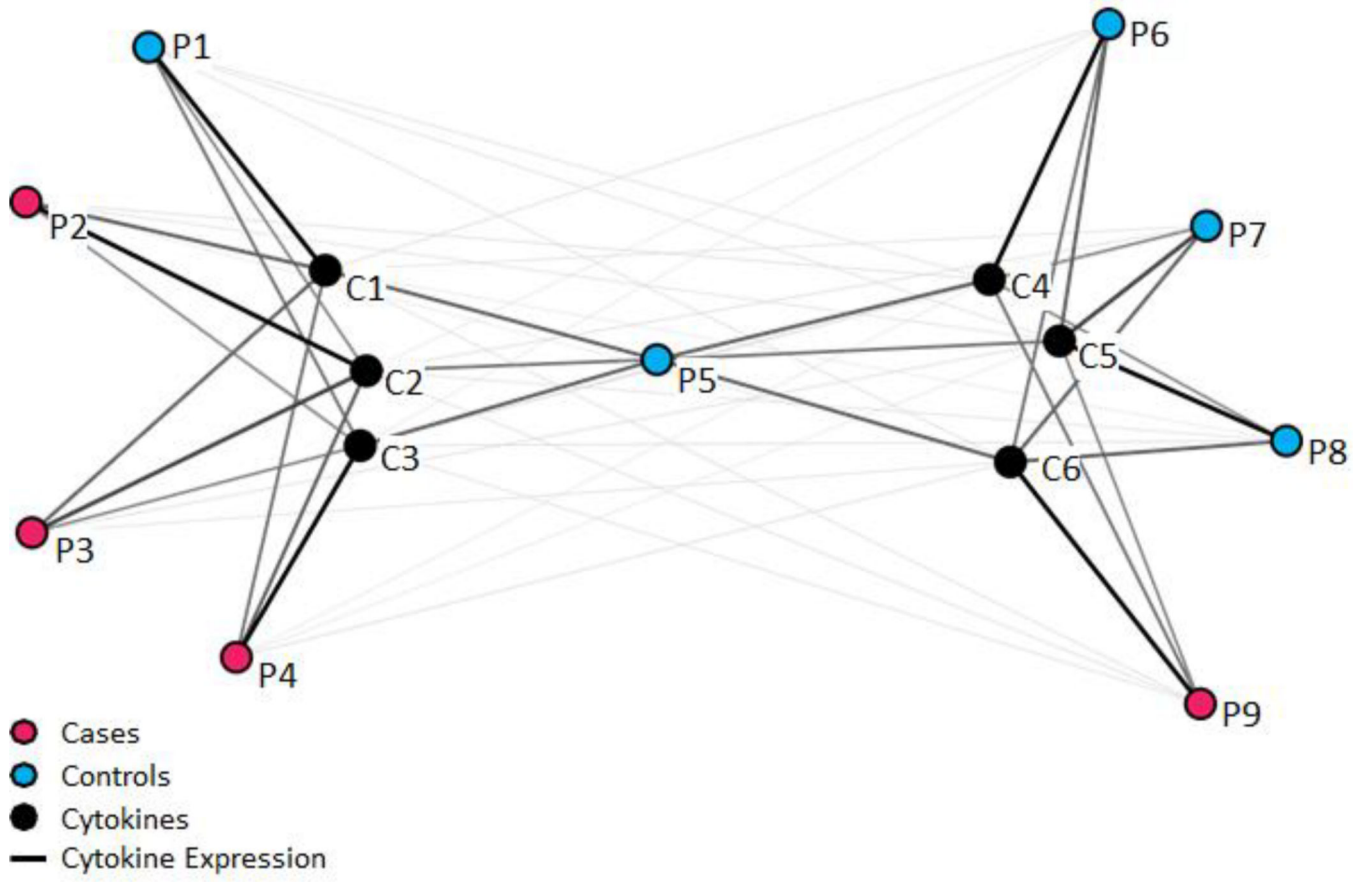
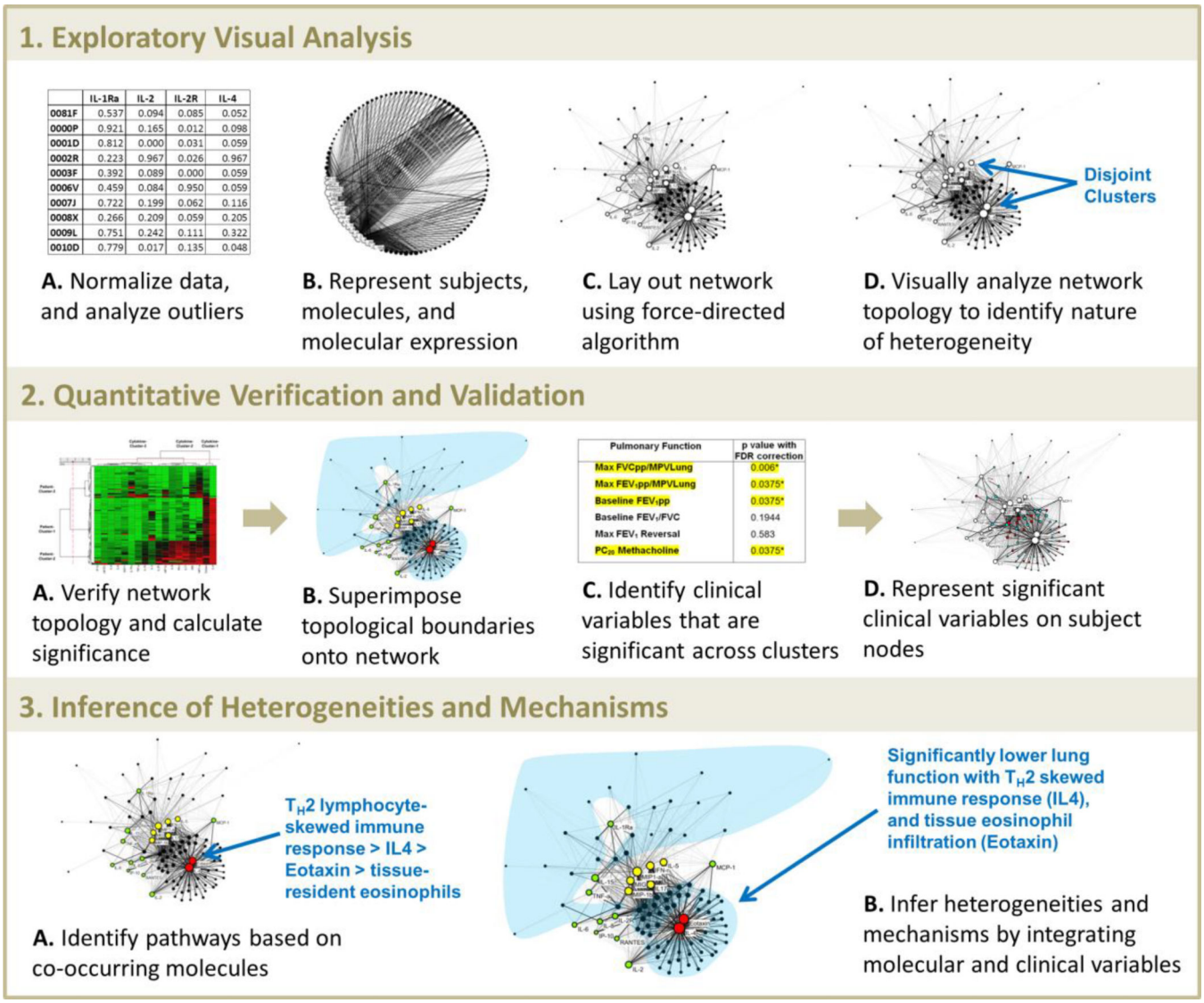


Figure 2.

An example of a bipartite network where edges exist only between two different types of nodes. Here nodes represent either subjects (cases = pink, controls = blue, or cytokines (black), and the undirected weighted edges connecting the two represent gene expression.



	IL-1Ra	IL-2	IL-2R	IL-4
0081F	0.537	0.094	0.085	0.052
0000P	0.921	0.165	0.012	0.098
0001D	0.812	0.000	0.031	0.059
0002R	0.223	0.967	0.026	0.967
0003F	0.392	0.089	0.000	0.059
0006V	0.459	0.084	0.950	0.059
0007J	0.722	0.199	0.062	0.116
0008X	0.266	0.209	0.059	0.205
0009L	0.751	0.242	0.111	0.322
0010D	0.779	0.017	0.135	0.048

Figure 3.

The three analytical stages of subject-protein network analysis for identifying proteomic heterogeneities in complex diseases. These stages are often iterative such as when the inference stage triggers new hypotheses about subject-molecular relationships, which in turn require quantitative verification and validation.

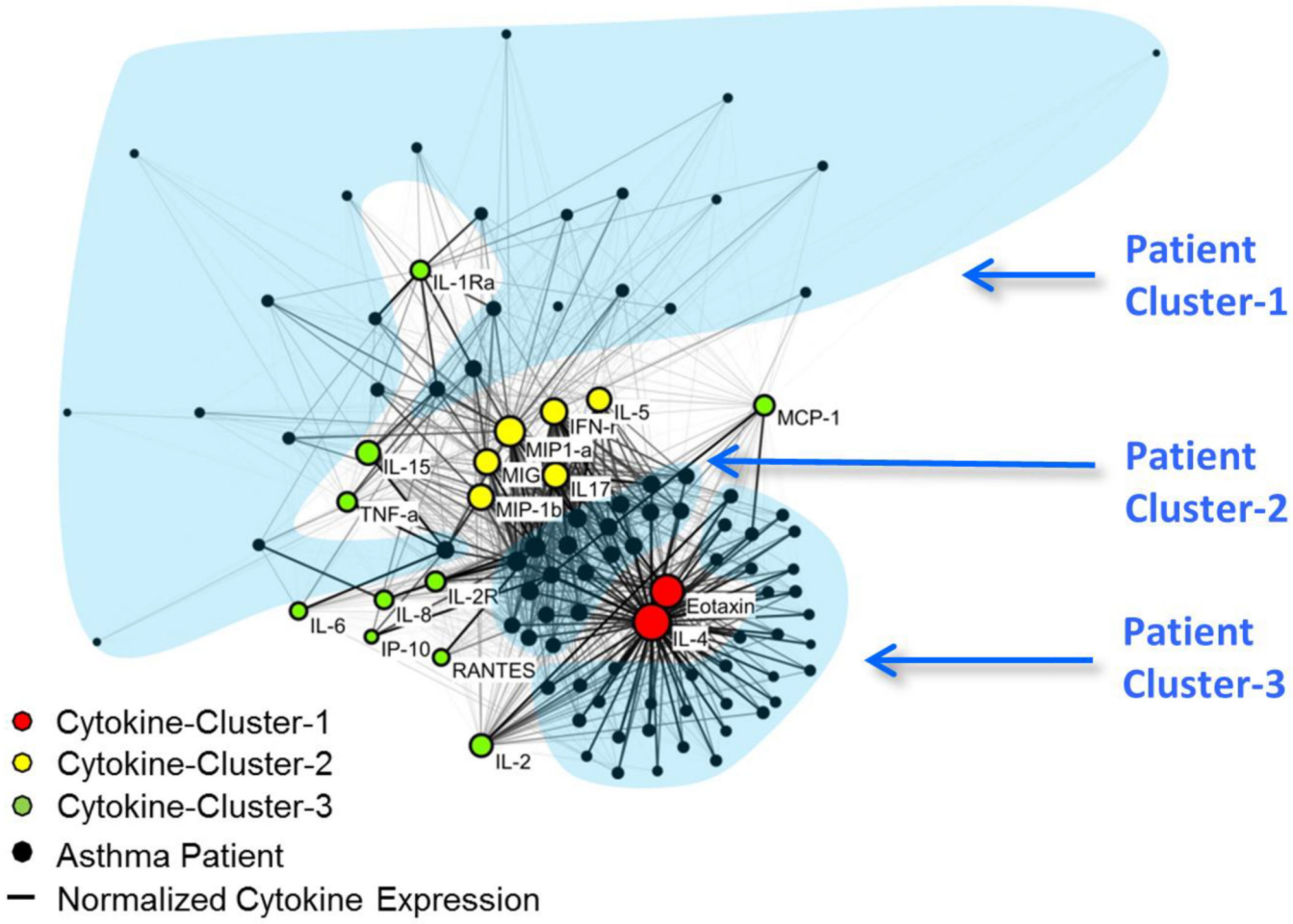


Figure 4.

Results of the subject-protein network analyses showing a complex but understandable clustering of patients and their associations with cytokine clusters. The cytokine-based clusters were integrated with clinical variables to help infer the proteomic heterogeneities and their biological mechanisms.

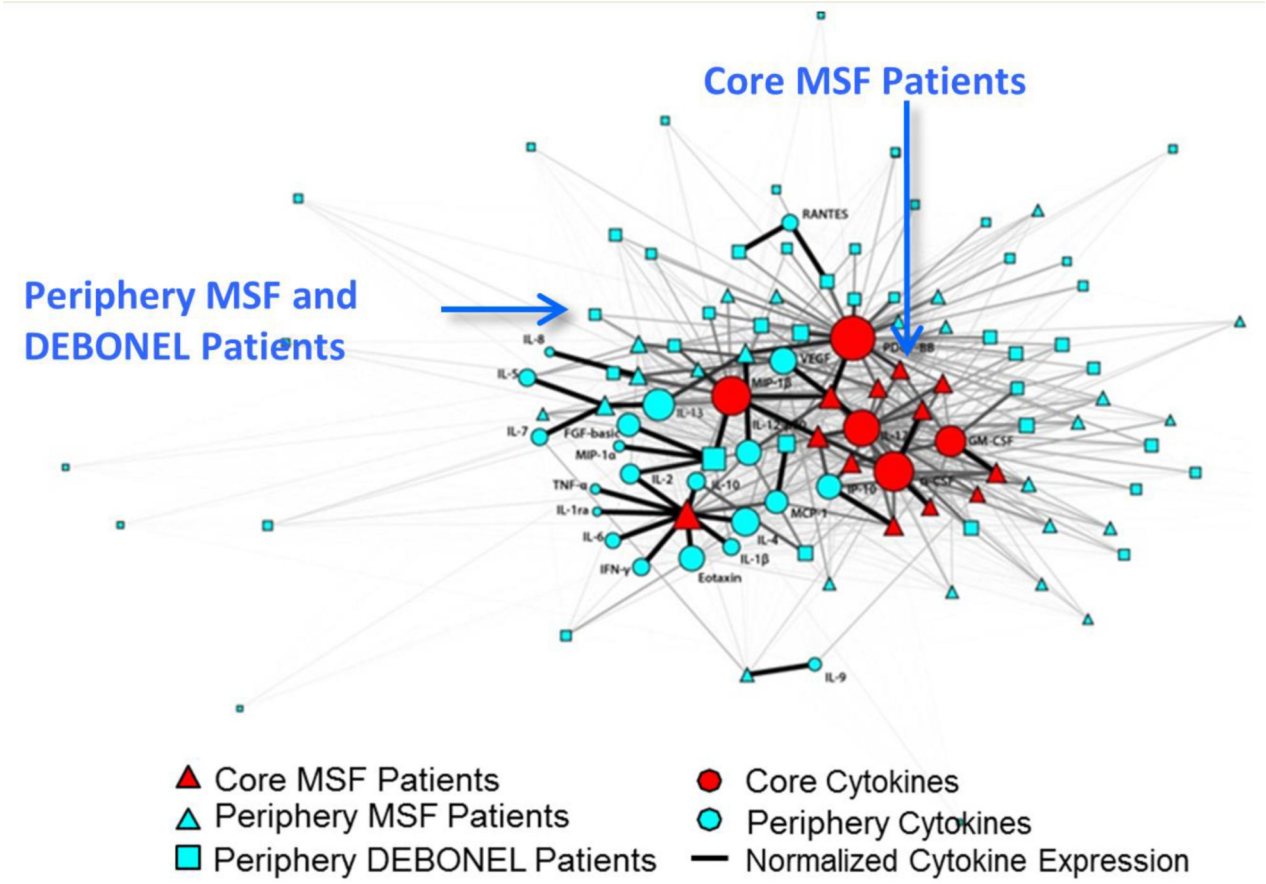


Figure 5.

Results of the subject-protein network analyses showing a core-periphery topology of patients with rickettsial infections, and cytokines clusters. The cytokine-based clusters were integrated with clinical variables to help infer the patient heterogeneities and their biological mechanisms.

Table 1

Four major types of networks used to model biological phenomena with examples of their motivation, typical network representation used, example data sources, and example tools used to conduct the analysis.

Network Application	Motivation	Network Representation			Example Data Sources	Example Tools
		Type	Nodes	Edges		
1. Process Networks						
Gene Regulation (e.g., Brasier, 2010)	Represent gene regulatory pathway	Directed, unweighted, with multiple entities	Genes, transcription factors	Regulation	Experimental, Published literature	Pathvisio, Ingenuity pathway analysis
Signal Transduction (e.g., Finkel, 2011)	Represent and comprehend cell signaling pathway	Directed, unweighted, with multiple entities	Proteins, metabolites, lipids, etc.	Signaling	KEGG	Pathvisio, Ingenuity pathway analysis
Metabolic Reactions (e.g., Kotera et al., 2014)	Represent and analyze metabolic reactions	Directed, unweighted, bipartite, or undirected unipartite	Metabolites and reactions/enzymes	Reaction	EcoCyc, KEGG	Pathvisio, Ingenuity pathway analysis
2. Interaction Networks						
Protein-Protein Interaction (e.g., Jensen et al., 2009)	Infer protein complexes and types of protein interaction	Undirected, unweighted, unipartite	Proteins	Mutual binding	HPRD, BIND, STRING db	STRING
Gene-Gene Interaction (e.g., Tong et al., 2001)	Identify hub genes	Undirected, unweighted, unipartite	Genes	Interaction	BIND, Experimental results	Pajek
3. Similarity Networks						
Gene-Gene Co-Expression (e.g., Stuart et al., 2003)	Identify gene modules	Undirected, weighted, unipartite	Genes	Degree of statistical similarity	Experimental data	VxInsight
Metabolite-Metabolite Correlation (e.g., Weckwerth, et al., 2004).	Analyze network topology for differences between test & control	Undirected, weighted, unipartite	Metabolites (sugars etc)	Correlation	Experimental data	Cytoscape, Pajek
Patient-Patient Similarity (Bauer-Mehren et al., 2013)	Identify patient modules	Undirected, weighted, unipartite	Patients	Degree of statistical similarity	Electronic medical record	Cytoscape, Pajek
4. Affiliation Networks						

Network Application	Motivation	Network Representation			Example Data Sources	Example Tools
		Type	**Nodes**	**Edges**		
Disease-Gene Affiliation (e.g., Goh et al., 2007)	Identify how genes are shared within and across diseases	Undirected, unweighted, bipartite	Diseases and genes	Gene affiliation to diseases	OMIM	Pajek
Species-Microbiome Affiliation (e.g., Muegge et al., 2011)	Analyze how diet and microbiome associate across species	Undirected, weighted, bipartite	Species and microbes	Bacterial 16s RNA level	Experimental data	Pajek
Drug-Target Protein Affiliation (e.g., Yıldırım et al., 2007)	Infer new purposes for existing drugs	Undirected, unweighted, bipartite	Drugs and target proteins	Protein affiliation to drugs	DrugBank	Pajek
Subject-Protein Affiliation (e.g., Bhavnani et al., 2011).	Infer molecular heterogeneity and respective pathways	Undirected, weighted, bipartite	Subjects and proteins	Degree of protein expression in subjects	Experimental data	Pajek