Graduate School ETD Form 9
(Revised 12/07)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Qian You

Entitled

Iterative Visual Analytics and its Applications in Bioinformatics

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Shiaofen Fang
Chair

Luo Si

Mihran Tuceryan

Elisha Sacks

To the best of my knowledge and as understood by the student in the *Research Integrity and Copyright Disclaimer (Graduate School Form 20)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): Shiaofen Fang

Approved by: Sunil Prabhakar / William J. Gorman 11/10/2010
Head of the Graduate Program Date

Graduate School Form 20
(Revised 1/10)

PURDUE UNIVERSITY
GRADUATE SCHOOL

Research Integrity and Copyright Disclaimer

Title of Thesis/Dissertation:
Iterative Visual Analytics and its Applications in Bioinformatics

For the degree of Doctor of Philosophy

I certify that in the preparation of this thesis, I have observed the provisions of *Purdue University Teaching, Research, and Outreach Policy on Research Misconduct (VIII.3.1)*, October 1, 2008.*

Further, I certify that this work is free of plagiarism and all materials appearing in this thesis/dissertation have been properly quoted and attributed.

I certify that all copyrighted material incorporated into this thesis/dissertation is in compliance with the United States' copyright law and that I have received written permission from the copyright owners for my use of their work, which is beyond the scope of the law. I agree to indemnify and save harmless Purdue University from any and all claims that may be asserted or that may arise from any copyright violation.

Qian You
Printed Name and Signature of Candidate

09/21/2010
Date (month/day/year)

*Located at http://www.purdue.edu/policies/pages/teach_res_outreach/viii_3_1.html

ITERATIVE VISUAL ANALYTICS AND ITS APPLICATIONS IN BIOINFORMATICS

A Dissertation
Submitted to the Faculty
of
Purdue University
by
Qian You

In Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

December 2010
Purdue University
Indianapolis, Indiana

To my parents

ACKNOWLEDGMENTS

I am heartily thankful to my advisor Dr. Shiaofen Fang, whose encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of the subject. I also owed my deepest gratitude to Dr. Jake Chen. He has tremendously supported me in a number of ways, including providing the high quality data sets, spending tremendous effort on manuscript revisions and offering many inspiring discussions and encouragement. I am also grateful to Dr. Luo Si, Dr. Mihran Tuceryan and Dr. Elisha Sacks for their warm support and many instructive comments during the development of my research topic and the dissertation.

Also, this dissertation would not have been possible unless my parents showed their greatest love and support from the other end of the Pacific Ocean. I am indebted to my co-workers who have ever worked with me or helped me as well. Finally I would like to show my gratitude to many friends, because they have always believed in me and encouraged me to do my best.

TABLE OF CONTENTS

Page

Page

.

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABSTRACT

You, Qian. Ph.D., Purdue University, December, 2010. Iterative Visual Analytics and its Applications in Bioinformatics. Major Professors: Shiaofen Fang and Luo Si.

Visual Analytics is a new and developing field that addresses the challenges of knowledge discoveries from the massive amount of available data. It facilitates humans' reasoning capabilities with interactive visual interfaces for exploratory data analysis tasks, where automatic data mining methods fall short due to the lack of the pre-defined objective functions. Analyzing the large volume of data sets for biological discoveries raises similar challenges. The domain knowledge of biologists and bioinformaticians is critical in the hypothesis-driven discovery tasks. Yet developing visual analytics frameworks for bioinformatic applications is still in its infancy.

In this dissertation, we propose a general visual analytics framework – **Iterative Visual Analytics (IVA)** – to address some of the challenges in the current research. The framework consists of three progressive steps to explore data sets with the increased complexity: **Terrain Surface Multi-dimensional Data Visualization**, a new multi-dimensional technique that highlights the global patterns from the profile of a large scale network. It can lead users' attention to characteristic regions for discovering otherwise hidden knowledge; **Correlative Multi-level Terrain Surface Visualization**, a new visual platform that provides the overview and boosts the major signals of the numeric correlations among

nodes in interconnected networks of different contexts. It enables users to gain critical insights and perform data analytical tasks in the context of multiple correlated networks; and the **Iterative Visual Refinement Model**, an innovative process that treats users' perceptions as the objective functions, and guides the users to form the optimal hypothesis by improving the desired visual patterns. It is a formalized model for interactive explorations to converge to optimal solutions. We also showcase our approach with bio-molecular data sets and demonstrate its effectiveness in several biomarker discovery applications.

CHAPTER 1 INTRODUCTION

1.1 Objectives

Over the past decades, the development of computing technologies has largely been driven by the tremendous amount of data. Those data are from numerous domains and applications, including structured or unstructured text from web pages, emails, documents and blogs; medical, biological, climate, commercial transactions, internet activities, geographical and sensor data. Not only due to the amount, but also due to the heterogeneity and uncertainty of the data, there is an urgent need to advance the data processing capabilities of current computing technologies. The primary reason of processing these data is to discover hidden knowledge for better decision making or problem solving. It becomes an essential means for benefitting both the human users and the automatic computations. Human have superior pattern recognition, comprehension and reasoning capability that have not fully been understood. However, in terms of storage, processing speed, computers are much more advantageous. Motivated by the complementary advantages human beings and computers have in information processing, Visual Analytics (VA) is a newly developing discipline, a “science of analytical reasoning facilitated by interactive visual interfaces” [1].

VA comes to play when massive amounts of data does not only overwhelm the analysts, but also makes the traditional data analysis and mining techniques fall short. Automatic data analysis or mining models essentially searches for optimal solutions after objectives of the computing tasks are defined. However, for the

majority of today's data sets, the meaningful patterns and hidden knowledge are not known beforehand, hence it is hard to formulate the goals of discovery at the first place. VA is advantageous over automatic data mining primarily because it leverages human perception, intelligence and reasoning capability, and cooperates with the automatic computing in solving complex real-world problems.

Earlier research in VA and its relevant applications set the stepping stones [2-4]: the interactive visualization needs to be an integral part of the cycles where human make decisions and form insights. In the iterative process, users use visual interfaces to explore the data set, to observe phenomena, to see alternative solutions and making hypotheses, and to reflect on what they would be interested in. Their preference can be a short cut to reduce complexity. After they have made their decisions, they input their feedback. Then the new intermediate visual results are presented and a new cycle will start. The process stops once the tasks at hand are accomplished or users have developed sufficient insights on the data sets. However, to substantiate such an iterative cycle, there are challenges and ongoing research in at least the following three aspects [5-7]:

- High-dimensional or non-visual data sets need to go through a series of properly designed transformations into user comprehensible forms,
- Right tools, methods and models need to be developed, along with interactive visual representations, to scaffold users' knowledge construction and insight provenance during the visual analytical process,
- Formal models need to be studied and established on how, in complex data analysis applications, to take advantage of both human cognition and computers: when and which part of the tasks are dispatched to one party or the other, and how the changes to the data set made by one party can be understood and handled by the other.

Considering the first challenge, the information visualization community over the past decades extensively studied and developed numerous interactive visual representations for high-dimensional data sets [8-14]. But the primary focus of the visual representation designs in information visualization is not assisting users to track the development of the insights and the knowledge. The interactions are not fully designed for the purpose of feedback users' intentions to drive the underlying data analysis model. To tightly couple interactive visualization with users' reasoning process remains an early research topic. Because not only to VA, but also to psychology and different behavioral sciences, human's higher recognition remains a "black box". For the second and third challenges, the research is still in its early stage [15-17].

Bioinformatics research is an area that has benefitted from information visualization, and also poses challenges on existed visualization techniques. For example, graph and network visualization techniques are used extensively to help biologists understand and communicate the biological data sets [18, 19], including biological networks with multi-category nodes and semantically differing sub-networks [20]. The exposed visual patterns and clues [21-23] becomes extremely helpful when biologists and bio-informaticians analyze the rapidly growing "omics" data, from numerous public databases [24, 25] and high throughput experiments [26]. Holistic investigations of the differing but related biology networks can lead to the discovery of the newer biology functional properties [27]. However, with the existing visualization techniques, biologists can be overwhelmed by the dense nodes, clusters of links, colors etc. Moreover, how their observed visual patterns can relate to functional hypotheses remains at a descriptive level.

Visual Anlaytics addresses the need of analyzing the increased volume of biological data by integrating the power of visualization and the domain knowledge of biologists. Visualization has the capability of presenting the large volume of data in a succinct and comprehensible form. And the biologists reason with the visual phenomenon and their domain knowledge for forming new insights and hypotheses. With the visualization, they also piece together the evidence for the verifications of their assumptions. So developing visual analytical models for bioinformatics applications has the following two critical requirements: first, to create clear, meaningful visualizations without overwhelming the biologists by the intrinsic complexity of data; second, to create simple and effective visual interface and process for biologists to carry out their analytical tasks, form and improve their hypothesis, and eventually arrive at optimal solutions.

In this work we propose a general visual framework – **the iterative visual analytic (IVA)** – to address the challenges and requirements in the current visual analytics research and its applications in bioinformatics. Our framework consists of three progressive steps: **Terrain Surface Multi-dimensional Data Visualization**, **Correlative Multi-level Terrain Surface Visualization**, and **Iterative Visual Refinement Model**. The three steps deal with increasing complexity in the underlying data sets, and enable domain users to perform more and more sophisticated visual exploratory tasks. Therefore the discoveries from each step are less and less straightforward for automatic analysis methods. We showcase our approach with bio-molecular data sets and demonstrate its effectiveness in biomarker discovery applications that are critically important for, drug design, clinical diagnosis and treatment development. **Terrain Surface Multi-dimensional Data visualization** renders a surface profile over a large scale bio-molecular interaction network, using a newly proposed graph drawing algorithm and the Scatter Data Interpolation. We have applied this method to

Alzheimer's Disease protein interaction subnetwork and microarray expression samples, and are able to identify diagnostic, prognostic, and stage markers that are consistent with previous studies. Then we develop the **Correlative Multi-level Terrain Surface Visualization**, to visualize the profiles of multiple correlated biological networks. This method uses the terrain surface visualization to render a profile of each network by interpolating the correlation numeric values as a surface over each the networks. The correlative terrains visually highlight the patterns hidden in the correlations among nodes, while preserving their locality and neighborhood in the networks. When applying this method to a pair of correlated bio-molecular interaction network and disease association network, we are able to use the visual patterns to identify molecular biomarkers and compare their performance in terms of sensitivity and specificity measures. Finally the **Iterative Visual Refinement Model** is a formal four-step approach which enables users to iteratively improve biomarkers' performance according to visual assessment on the changing terrain profiles. We have applied this model to the correlated cancer biomarker protein interaction network and the cancer association network. As a result we are able to discover a new group of biomarkers that achieves optimal specificity for lymphoma cancer. We also validate the newly found biomarker panel by classifying the third party microarray expressions. As a result, this panel outperforms 90% of the benchmark biomarkers. In summary, the three steps of IVA have the following major contributions:

- Terrain Surface Visualization we developed is a new high-dimensional data visualization technique, where the relationships among data can be appropriately described as a graph or a network. The technique exposes the globally changing patterns over large scale network. The base network of the terrain surface is laid out by a new graph layout model that captures the inherent structural properties of the original network. The data interpolation and surface rendering avoids the scalability problem and represents features derived from the data set as prominent geographic

landmarks. Interacting with regions prioritized as prominent landmark features, with interactive visualizations, can lead to new hypotheses based on domain knowledge.

- Correlative Multi-level Terrain Surface Visualization is a new visual analytical platform to study correlations among nodes in interconnected subnetworks of different contexts. It visually highlights the major signals in the correlation as well as preserves the major topology of the subnetworks, regardless of the noise inherent in the networks. The visual patterns of the correlative multi-level terrain enables users to perform visual analytical tasks on correlations in the context of more than one networks, thus enable them to gain critical insights and form hypotheses from the complex data set.
- Iterative Visual Refinement Model is a novel visual analytical process. The model treats users' perceptions as the objective function, and guides the users to the final formation of the optimal hypothesis by improving the desired visual patterns. The changing visual patterns observed from the terrain surfaces represent intermediate hypotheses formed, and the ultimate satisfactory visual patterns mark the final optimal discoveries. So the patterns serve as a form of reasoning artifacts which can record users' temporary findings as well as enable visual comparison among findings. To ensure that the interactive exploratory process will reach to the optimal solutions, the model consists of four steps that assist users in implementing the elimination heuristics using the visualization components.
- We also identified a new biomarker panel of four protein biomarkers for lymphoma cancer, using the iterative visual refinement model. The four used as a panel has not yet reported, but has surprisingly high sensitivity (both type I errors and type II errors are at the <1% level) and high specificity against leukemia (at the >99% level) on a separately prospective microarray data set. After the good performance is further

validated by thorough perspective validations, the panel can possibly be translated into markers for clinical diagnosis and drug design.

The IVA can be used to develop visual analytic toolkits for bioinformatics applications, including disease-wide visual biomarker discovery, personalized microarray biomarker development and potentially drug discovery. IVA can also be extended to a visual analytical platform on semantically complex networks other than biology subnetworks. Particularly, the iterative refinement model presents a few guidelines for visual analytical models. First the visual interface and the process represent the domain experts' hypotheses as visual patterns. This enables users to assess the quality of their hypotheses in the iterations which update the solutions. The formation of desired knowledge is clearly marked, that is, the development of the shape of the patterns. Additionally, IVA supports domain experts to follow their problem-solving heuristics when refining their hypotheses. It is valuable to discuss and research about developing visual analytical models that would explicitly support various types of human problem solving heuristics.

1.2 Organization

This dissertation covers all three steps of IVA and has six chapters. The next chapter comprehensively surveys related high-dimensional data visualization techniques, the important aspects and models for visual analytical science, and the visualizations used for biomolecular networks and biomarker discovery applications. Chapter 3 elaborates the motivation, methods and applications of **Terrain Surface Multi-dimensional Data visualization,** followed by **Correlative Multi-level Terrain Surface Visualization** in Chapter 4. The **Iterative Visual Refinement Model** and its applications are elaborated in Chapter 5. I also present the data sets, the statistical tests and results for validating our newly identified panel biomarker. The last chapter discusses the advantages, limitations and possible alternatives of our framework. It also concludes the dissertation with

future work, including further validating the discovered panel and using statistical and machine learning methods to leverage the iterative visual analytics framework.

CHAPTER 2 RELATED WORK

2.1 Visual Analytics Techniques and Models

In light of the data deluge from numerous real world applications, the need to analyze the data raises a fundamental problem: how users' reasoning and analysis capabilities of the data set can be facilitated by interactive visual interfaces. The 2005 book *illuminating the path*: *The R&D Agenda for Visual Analytics* [1] marked the birth of Visual Analytics (VA) and posed a general paradigm for solving this problem. Visual Analytics has a unique data-driven origin and the interdisciplinary characteristics. Therefore, since early five university-led Regional Visualization Centers (http://nvac.pnl.gov/centers.stm) were established, and people from academia, governments and industries are forming a diverse and interdisciplinary team. They have actively engaged in this new research [28], and have developed successful visual analytics system and applications in very diverse domains: real-time situation assessments and decision making [29, 30], spatial-temporal relationships in traffic control/epidemic disease management [31-34], internet activity and cyber security [35-38], large scale social networks [39-42], multi-media understanding and explorations [43-45], documents and on line text analysis [16, 46-49] [50], financial transaction management and fraud detections [51, 52], the latest bioinformatics applications [53-56] etc.

For establishing a science for VA, a number of challenges and theoretical issues are in on-going discussions. One of the major issues is how existed information

visualization techniques can be leveraged to better cope with the increasing scale and heterogeneity of the available data sets. The improvements on the techniques also require the focus on assisting users reasoning and analytical tasks on the data sets. The second major issue is that how VA can provide interactive framework that scaffolds the human knowledge construction process, with the right tools and methods to support the accumulation of evidence and observations. The third issue is, how VA could harness the complimentary advantages of both computers and human beings, and closes the problem-solving and reasoning cycles [4] in which users and computers take turn to accomplish parts of the tasks.

In the rest of section 2, we first survey some of the existed techniques in information visualizations, particular visual representation for non-linear high-dimensional data. Among the techniques, graph/network visualizations are the most relevant techniques to our framework. So we focus on large scale graph/network visualization in section 2.1.1, then we briefly introduce other representative techniques in section 2.1.2. For understanding how current research addresses the last two challenges, in section 2.1.3 we discuss representative works of scaffolding the knowledge construction process, and of integrating reasoning capability of human and computers.

2.1.1 Graph and Network Visualization Techniques

Graph or networks have long been used to characterize non-linear high dimensional relationships among attributes. To characterize such relationships, typical concerns of graph drawing algorithms are separation of vertices and edges so they can be distinguished visually, and preservation of properties such as symmetry and distance. Many graph drawing algorithms attempt to achieve an optimized graph lay out by minimizing a pre-defined system energy function. The

energy functions derived from the spring model (force-direct or energy-based model) [57], and its variant [58] are the most popular and the easiest to implement. Other proposed models are Linlog energy model [59]. The energy function varies among different algorithms, but in general it is a function of the distance between nodes and the weights of edges among them. A number of multi-dimensional minimization methods, such as Downhill Simplex Method, Powell's Method and Conjugated Gradient Methods, are common options to implement the minimization [60]. Graph drawing problems have also been studied in the context of Multi-dimensional Scaling (MDS) [9]. MDS aims to map a data set in higher dimensions to lower dimensions by non-linear projections, so that the distance between data points in lower dimensions best preserves the similarities or dissimilarities in the original distance matrix [61]. The cost function or stress function of this non-linear embedding is in fact a generalization of the energy function in a force-based graph drawing model. Therefore, Stress Majorization [62] used in MDS can also be applied to graph drawing. The major advantage of Stress Majorization over the energy function minimization is that Stress Majorization ensures that stress monotonically decreases during the optimization; thus, Stress Majorization effectively avoids the energy value oscillation in optimization and shows improved robustness over local minima [63]. MDS implementations are available in both commercial [64] and open source [65] packages.

Scalability and avoiding visual clutters remains an important issue in graph and network visualization, because the scale of graph for representing real-world applications keeps increasing. Simple graph drawing algorithms are not usually scaling well. So in many cases the nodes in graph are first clustered to create a hierarchy for overview navigations, and then can be interactively explored [66]. Existed agglomerative and divisive hierarchical clustering [67], can merge nodes into subgroups [68] or "communities" [69] based on the connectivity of nodes. In

addition, other graph features, for example, semantics [70], topological [71] and geometric features [72] of the networks are studied and extracted by statistical analysis methods to highlight relevant network structure. In this way the presentations of large graphs could be simplified and the persevered features [21] are highlighted. The clusters of nodes can be laid out afterwards with space filling visualizations, in order to achieve even better screen space utilizations and better preservations on the semantics conveyed in the networks. For instances, Itoh et al. [73, 74] and Muelder et al. [75] hierarchically cluster a graph then spreads out nodes using a treemap-like space-filling layout techniques. Also Muelder et al. [76] in a later paper proposes a large graph layout, built on top of the hierarchy, using space-filling curves. It also extensively compares existed layouts models, including the common force-direct models, the fast layout models for large graphs, and the treemap space-filling layouts. Unlike space-filling model which relies on the hierarchy of nodes, Hierarchical Edge Bundles distinguishes adjacent edges and hierarchical edges , draws edge bundles accordingly [77], in order to reduce the visual clutters caused by dense edges. Another way that assists users to read the large graph is that coping with their constantly changing intentions in the analysis process. Numerous interaction models, such as overview+detail [78, 79] or iterative explorations [80], are also developed to support users' changes in their mental context, in their analytical models and their focus of trust in various regions of data.

An alternative approach to ease the congestion problem of large scale graph is to use adjacency matrix for presenting graphs. Previous studies [81, 82] show that adjacency matrices are better than node-link for displaying dense or large scale networks. A non-zero entry in the matrix represents an edge between two vertices that the row and column entries represent in a graph. Therefore matrices have the advantages that each node has the position in a confined cell in the screen. Interactive multi-scale visualization has also been incorporated into

matrix-based network to assist users' exploration when the size of the graphs becomes large. For example, Frank Van Ham [83] developed a multi-level matrix visualization for call graphs among the subsystem of very large software projects, according to the uniform visual representation and recursive structure of matrices. Using the same property, MGV is a system for visualizing large multidigraph [84]. A disadvantage with adjacency matrix is that a path in the graph can be mapped to any loose pattern in the matrix. It needs extra mental mapping steps for users to interpret the patterns. Visualizing the properties associated with nodes or its surrounding neighborhoods can raise the same problem. When the properties of nodes and their proximities in a large scale graph are of primary interest, mapping properties of a node to different color gradient can better preserve an informative overview and demonstrate meaningful patterns. Research in information visualization community have demonstrated human perceptive advantages on spatial phenomena, such as landscape (surface) spatialization [85], over points arrangements. Taking advantage of these findings, there are graph visualization methods which render continuous fields over the underlying graph layout, by interpolating numeric values of nodes over every point of the 2D plane the graph resides. Among these methods, ThemeScape [86] and VxInsights [87] are the first to use elevation as the interpolated value to indicate the strength of certain themes in a given region in document visualization. The overall 3D surface (landscape) visualization is claimed to be effective in providing both a overview and the inter-relationships among the documents and their themes. The formal model of rendering another scalar field over a graph layout is presented in GraphSplatting [88], which assumes that significant structural information can be provided from the density of vertices. In this work, a 2D kernel or basis function plus a noisy factor is placed at the center of a vertex's 2D position to create a continuous 'splatting' signature around the vertex.

2.1.2 Other Data Visualization Techniques

Besides graph visualization and analytics, other frequently studied techniques for visualizing high dimensional data sets are parallel coordinates (PC) [8, 89, 90], RadViz [11, 91], Stacked Graph [14, 92] and so on. Among these techniques, PC, is the most relevant to the terrain surface high-dimensional data visualization technique. Dimsdale and Inselberg [8] first proposed PC where each dimension is drawn as a vertical (or horizontal) line, and each multi-dimensional point is visualized as a polyline that crosses each axis at the appropriate position to reflect the position as in a N dimension space. PC has the advantages that it visualizes the data item as well as the high-dimensional geometry in 2D. There are two major problems with PC. The first is the line crossings and overlappings caused by the polylines of large data sets. Too much clutter result in incomprehensible rendering and little insights. To alleviate this problem, different clustering methods are used to create initial clusters within the data sets: Johansson et al. uses K-means for initial clustering [93]; Fua et al. [94] propose a multi-resolution view of the data via hierarchical clustering. The clusters can be represented by rendering a representative item within each cluster, e.g. the centroid, as a solid line. The data items in clusters are then represented with faded regions or differing colors for each data item to show their cluster membership. A few more work has proposed sophisticated rendering techniques, such as high-precision texture [93], edge-bundling through B-splines and “branched” clusters [90]. Focus+context techniques, for example, Sampling Lens [95], are proposed to reduce clutters and allow users to gain insights from extremely large data sets. Another way to tackle this problem of cluttered parallel coordinates display are via line density plots [96, 97]. The second problem is that the linear arrangement of the dimension vertical bars, will, to some extent, lose the original geometry of the data distribution of the high dimensions. Although methods have been proposed to reorder the dimensions [98], there is no guarantee that there are linear arrangements of dimensions can reveal all significant patterns in high dimensions.

2.1.3 "User-in-the-loop" Interactions Models in Visual Analytics

Merely developing novel visual metaphors is rarely sufficient to trigger insight from users. These visual displays must be embedded in an interactive framework that scaffolds the human knowledge construction process, with the right tools and methods to support the accumulation of evidence and observations into theories and beliefs. Understanding human's reasoning process for developing insights, therefore, is the first step for designing such tools. There have long been three established humans inquiry phases that form the process of knowledge construction- abduction, deduction and induction [99]. Recently Pike et al. [6] have elaborated how analysts use the three steps to form a cycle and are used iteratively to form hypotheses and get answers. Then for scientific data visualizations, Upson et al. [2] have proposed an analysis cycle where the rendered visualization is then used by the user to provide feedback into the previous steps, restarting the cycle. Card et al. [3] describes a similar cycle of visual transforms with users interactions.

The researchers have realized some interactions with the information might take place within the context of a software tool, but much of them occur internally in one's mind. Insights can be generated and tested wherever the mind is – not whenever the data and the tool happen to be. Therefore, the effectiveness of "User-in-the-loop" interaction models is firstly affected by the fact whether the interactions design can reflect the users' inquiry and intentions coherently and consistently, and whether the interactions capabilities are at the user's disposal whenever and wherever he or she is thinking about a problem space. To further study users' interactions and to externalize their mental reasoning activities, lower level interactions are extensively recorded, analyzed and categorized. For examples, for lower-level interactions, Amar et al. [100] defines a set of primitive analysis task, including retrieving values, filtering, calculating values, sorting, clustering, etc. Yet understanding the users' intentions requires mapping from

low-level manipulations on data to high-level user goals. Yi et al. [101] defines a taxonomy of interactions intent – select, explore, reconfigure, encode, abstract/elaborate, filter and connect – that can be components to constitute the knowledge discovery process. In order to reuse, share or even learn from the occurred interactions, there are a few meta-visualization models or history-preserving tools being developed to capture, analyze, present or parameterize the interactions of exploration processes in VA applications. CzSaw [102] uses a script-language to record and program the sequences of analysis steps in investigative document collection analysis. It also builds visual history views showing progress and alternative paths, and presents dependency graphs among primary data objects to characterize the current state of analysis process. Its major advantages are that it explicitly presents the analytical process for users to gain insights, and that it enables reusing of the existed interaction and analytical flows onto new or dynamic data set. VisTrail [103] system manages final visualization products, e.g. an image, as well as the vistrail data flow specifications that generate the products. Using XML, VisTrail can represent, query, share and publish the vistrail specifications. Furthermore, the steps in specifications can be used as templates, and the concrete actions hence are parameterization of the templates. Therefore users interactions are not only presented as data flows but also are translated into a parameterized space. This is an interesting feature of the system. (Several earlier novel visualization user interfaces assume visualization exploration is equivalent to navigating a multi-dimensional parameter space [104].) P-Set (subset of parameters) [17] method fully explores the idea of parameterzing and formalizing the visualization process: users exploratory interactions are translated into parameter sets which then applied to visualization transform and renders the result; users feedback are translated as modifying the parameters repeatedly until the results of interest are generated. The exploration sessions are then documented in the form of a derivational model by XML. The generation of final parameter sets is heuristic exploration of parameter space resulted from users' intentions. So with P-Set and

their derivations, the framework has high potential in understanding the how users arrive at the satisfactory visualization. Yet which portion of the information for the sessions to be extracted and how they could be studied and generalized for optimal visualization generation remains open. HARVEST [15] is a visual analytic system designed and augmented by a high-level semantic model which tracks an insight's provenance to record how and from where each insight was obtained. The model first characterizes user analytic behavior at multiple levels of granularity based on the semantic richness of the activity. Then it is able to locate an action level as a set of generic but semantically meaningful behaviors that can constitute to the semantic building blocks for insight provenance.

The effectiveness of the "user-in-the-loop" interaction models is secondly required to harness the advantages of human intelligences and the power of computing technology and seamless integrate them to boost the problem solving capabilities. The models, thus, have to deal with two loops: one loop happens in users mind where decisions are made and leads to feedback actions; the other loop is data foraging loop which takes users input and visualizes the intermediate results for better sense making and insight development. Green et al. [105] have studied and explicitly addressed the complementary cognitive advantages of human and computers, and present a few design guidelines for visual analytics design. According to them, human has the superb adaption of relating unfamiliar or new phenomenon to something in the existed knowledge schema. And human beings master a compendium of reasoning and problem solving heuristics, e.g. eliminating pertinent information with prior knowledge. Meanwhile, computer has superior working memory and is lack of inherent biases. Therefore Green with others proposes a scheme describing how human analysts and computer can collaborate and complete the reasoning loop in the knowledge discovery process: user create knowledge by relating two previously irrelevant patterns and make this understood by the computer; the computer then learn from what users are

interested and recommends the semantically related information. The created knowledge is not only a set of declarative facts, but also the sequential steps and semantic inferential process in which users give facts, patterns and relationships. Using the same two loops, RESIN [106] approaches the predicative analytics tasks by combining a AI blackboard reasoning module with the interactive visual analytical tools. An underlying Markov Decision Process (MDP) captures the essence of sequential processes and is used to compute the optimal policies that identify, track, and plan to resolve confidence values associated with blackboard objects.. Users, assisted by the interactive visualization interface, can revise the confidence value of the partial solutions presented in the blackboard. The feedback adjusts the final confidence score which is constituted by a linear combination of difference confidence values and weights on sources during the predictive process.

Lately a few machine learning models are coupled with interactive visualizations for better integrating the strengths of both human reasoning and computers. In the work proposed by Xiao et al. [107], users' discovered interesting visual patterns of network traffic can be constructed by a declarative pattern language derived from the first-order logic. The patterns can then be saved and built into a knowledge base for further use. It is an iterative process that users identify, evaluate and refine interesting patterns via the visualizations, and then the system searches and recommends candidate predicates and their possible combinations to describe the patterns. It is significant for this work that the discoveries are driven by users' pattern recognition capabilities and their domain knowledge, and that users' input and preferences are described by a formal and computable logic model. This way, in the problem solving process, user' intentions and discovered knowledge can be captured, understood and used by the system, and the system can provide better recommendations based on accumulated users knowledge. While the system can recommend predicates, it

is still up to the users to construct clauses for describing the model. Therefore, the generalization of this model is not only limited by the expressiveness of Boolean logic, but also limited by users' capability of constructing complex predicate logic clauses. Starting with similar ideas, Garg et al. [108] proposes a model with the following two advantages: first it enables automatic learning of the rules using inductive logic programming with annotated positive and negative examples; second it has a full-fledged visual interface, N-D projection visualization, for users to interactively define projecting plane in d interesting dimensions out of N high dimensions. Therefore, it allows the users to gain much freedom to construct and refine models for arbitrary relationships in the complex data set. A major concern with the Logic Programming based VA models, is that it is only suitable for the domains and applications where the pattern discovery tasks can be characterized by predicates and clauses. VA can also be used to accomplish the general exploratory data analysis tasks, e.g. clustering, where the interpretation of results are largely dependent on users' subjectivity and application context. Schreck et al. [109] propose a visual clustering model for trajectory data by augmenting Self-Organization Map (SOM), a popular black-box neural network unsupervising learning model, with users' preferences, expectations or application context. Users' preferences are first input as template patterns and their positions are for initializing the SOM. The clustering is essentially iterative and can be paused to get users input, who can edit the patterns and adjust the learning parameters and layout. Therefore the clustering would converge based on minimizing quantization error and at the same time reflect desired application-dependent patterns and layout criteria.

2.2 Visual Analytics in Bioinformatics Applications

2.2.1 Visualizations of Biomolecular Networks

Graph and network visualization tools are becoming essential for biologists and biochemists to store and communicate bio-molecular interaction networks, including protein interaction networks [110], gene regulatory networks [111], and metabolic networks [112]. General large graph drawing techniques and toolkits, such as Pajek [113] and Tulip [114] are transferred into biology domains. At the same time, more and more biomolecular interaction databases[115] [25, 116] drive graph/network visualization toolkit developed for users to visualize, to annotate, and to query biomolecular interaction networks. Several popular biomolecular network visualization software packages are Cytoscape [22], NAViGaTOR [117], Osprey [118], Proteolens [119]. These software tools use graph metaphor and show biological macromolecules such as proteins and genes as nodes and their interacting relationships as edges; annotations of the graph are represented as nodes or edges of different colors, sizes, and distances. A comprehensive survey of visualization tools for biomolecular networks can be found at [120].

Biomolecular networks have the same scalability issues as the size of networks increases. Especially, the intensive investigations into biological systems result in increased volume of complex, interconnected data in recent years. For example, the development of a wide range of high-throughput experiments and public databases produce tremendous amount of interconnected biomolecular subnetworks, including metabolic networks [112], gene regulatory networks [121] and protein interaction network [122]. Therefore the rich semantics contained by those biomolecular networks can hardly be communicated clearly and effectively by a single planar graph with numerous annotations and legends. Visualizing

multi-category graphs remains a complicated problem, and there are very few general graph visualization techniques to solve it. This is because that connectivity, edge and node categories can all play a role in the final layout, and the optimal design largely depends on the requirements of specific domain contexts and applications. Itoh et al. [20] is one of the very few works in graph drawing community that proposes a formal framework for visualizing graphs consisting of nodes belonging to more than one categories. It first clusters categorized nodes together and then spreads out nodes using a force-direct model where the edges among clusters are quantified as constraints. Then the following space-filling step uses the result of the layout as the template for to adjust the position of the clusters of nodes. The framework also enables interactive layout modifications to bring clusters of the same categories close together. As a result, the framework provides an uncluttered and brief graph representation for displaying the clusters of categorized nodes, the clusters of uncategorized ones and the relationships among them. It is also applies the framework to address the complexity in gene/protein interaction networks and successfully discovered meaningful relations among protein complexes. The relations are otherwise hard to find using computational methods when no objective functions are defined around them.

In addition to general graph visualization framework, various biomolecular network visualization tools [123-125] have also been developed for displaying and analyzing complex information in interconnected biological subnetworks. In most tools, the integration of rich information is incrementally built on the previously simpler representations, and supports the interactive integration where users decide when and what to add in the existed visual representations. GenApp [123] can view, analyze and filter the gene expression data built on the context of biological pathways and can support users to modify and design their own pathway networks ; BiologicalNetworks [124] enables systematic integration,

retrieval, construction and visualization of complex biological networks, including genome-scale integrated works of protein-protein, protein-DNA and genetic interactions; the VisANT project [125] dose not only support simultaneously visualizing and overlaying multiple types of biomolecular networks, but also provides tools analyzing topological and statistics features. Unlike other tools, VisANT also introduces an interesting function — enabling comparisons between experimental interactions gathered from different data sets: it allows scientists to visualize each stage sequentially, by updating node colors to reflect values for a selected data set. This leads to preliminary yet promising investigations of how biomolecular network visualizations can demonstrate the dynamics of properties due to different data resources or experiment conditions. An alternative strategy for viewing the changing patterns over the network is to arrange changes on the nodes properties as changing color, and then to arrange networks at different time spot in a grid. Cerebral [126] is a well-designed suit that supports such strategy to analyze microarray experimental data in the context of a biomolecular interaction graph. The changing patterns over networks become more prominent when nodes properties being mapped to 3D landscape spatialization, as demonstrated in GraphSplatting and other user study works (refer to section 2.1.2). Following the same idea, Gene Maps [127] uses co-expression profiles of genes and builds clustered coexpression data on a 2D surface, and further incorporates the density of gene clusters as the latitude of high-density clustered areas-mountains' of a 3D visualization map. However, accurate gene co-expression similarity profiles usually require dozens, if not hundreds or thousands, of expression experiments; therefore, as more data become available, the topology and relative positioning of genes to each other in a gene map may dramatically differ from one another. Therefore only visualizing the density of underlying clusters dose not scale well. The complexity of biological networks remains a valuable challenge for network visualization, and hopefully will spin off a new research direction when more interdisciplinary work is taking place.

2.2.2 Visualization in Biomarker Discovery Applications

Molecular biomarkers refer to a group of biological molecules that can be assayed from human samples to help medical decision making, ranging from disease diagnosis, disease subtyping, disease prognosis classification, drug toxicity testing, to targeted therapeutics [128]. One primary way to identify molecular biomarkers is to study differentially expressed genes from microarrays [26] — a widely used high-throughput and large scale assaying technology which enables simultaneous genome-wide measurement of gene expression level for humans — across control (healthy) samples and positive (disease) samples. To extract only a small subset of relevant features and to achieve good performance on classifying samples, statistical analysis, dimensional reduction and machine learning methods, such as t-test [13], Principle Component Analysis [129], Support Vector Machine [130] or K-Nearest Neighbor classifier [131] etc are researched and applied. The key challenge in microarray analysis for biomarker discovery is that the features are usually noisy and the number of features is much larger than the number of samples. Therefore the data analysis method tends to yield unstable results and the found candidate biomarkers are subject to "the curse of dimensionality" [132].

Information visualization techniques have played central roles in helping exposing change patterns microarray samples. 2D Heatmap is used widely to help identify patterns of gene expression values. In a heatmap, each cell represents the expression value of a gene as the row entry in the corresponding observation as the column entry. The quantitative value is usually color-coded and the color patterns in heatmaps can lead to insights of the highly complicated and noisy microarrays. For example, Golub et al. applied 2-D heatmap visualizations to identify two distinct clusters of differentially expressed genes [133]. Eisen et al. applies pair wise average-linkage hierarchical clustering methods to cluster genes with similar expression values among observations

[134] in a gene expression analysis to distinguish two subclasses of leukemia. Heatmap visualization has the advantage to enable biologists to assimilate and explore in a naturally intuitive manner. However, clustering algorithms applied to the same data set will typically not generate the same sets of clusters. This is especially true to microarray data sets which are subject to data changes due to data normalizations and experimental noises. To address the uninterruptable clusters when clustering genes as row entries of the heatmap, first order matrix approximation is used and then the resulting patterns are filtered by human [135] to produce meaningful clusters. Sharko et al. [136] proposes a formal heat-map based method to visually assess both the stability of clustering results as well as the overall quality of the data set . They use heatmap to visualize the cluster stability matrix, which reflects the extent to which one gene tends to be in the same cluster any other gene across the entire set of clusters. As a result, the darkness and distribution of color patterns can be visually evaluated to assess the stability of the clustering algorithms on microarrays, to investigate the correlations among clusters of genes and to assess the qualities of the microarray data sets.

Identifying biomarkers, which are molecules such as genes and proteins, from microarray data sets need to identify relevant subset of genes whose expression changes are consistent with class annotation, instead of being pure noise. This task is essentially finding the several dimensions, projecting samples on which can achieve a reasonable separation. Other high dimension data visualization techniques are used for biomarker discovery as well. For example, to explore relationships between gene expression patterns and sample subgroups, M. Sultan*,* et al*.* [137], use self-organizing maps (SOM) [138] to create a signature from gene expressions for each sample, the collection of which are then classified and arranged onto a binary tree. VizRank [139] uses Radviz [140], a technique similar to star coordinates, to project microarray sample data as points

inside a 2D circle where a selective subset of gene features are anchors. Given the large number of possible combination of gene features, VizRank adopts a heuristic search to rank then sample the combinations of genes, and scores each of projections according to the degree of separation of data points with class labels. VizRank can further uncover outliers which reveal intrinsic properties of the data sets, and can perform classification for newer samples. SpRay [53] is a Parallel Coordinates based visual analytical suit for gene expression data. It conjoins original data with its statistical derived measure, and enables interactive selection on the range of statistical measure for highlighting a desired portion of the data. Its rendering techniques are designed to assist the recognition of uncovered traits in the large data sets and the qualitative relations between data dimensions. When being applied to a number of expression data sets, this suit can facilitate biologists in common expression analysis tasks, including detecting periodic variation patterns, studying differing p-value correction methods and detecting outliers.

Recent bioinformatics research has expanded to study “omics” data which includes genome, proteome, metabolome, expressome, and their interactions. Systematically integrating microarray profiling, other types of “omic” data sets, biological networks and knowledge resources can lead to biomarker discovery breakthroughs [141]. For example, Chuang et al. [142], have hypothesized that a more effective means of marker identification may be to combine gene expression measurements over groups of genes that fall within common pathways. And molecular biomarkers, which could only be discovered and evaluated in a single disease context in previous approaches, can now be investigated and tested in a disease-wide environment. Although there are many visualization tools (see section 2.2.1) has the functionality that maps expressions level to biomolecular networks e.g. pathway networks, not much has been reported on how biomarker discovery applications can be benefitted from such

mappings and integration. The information diversity and complexity certainly is one of the major challenges on the existed molecular biomarker discovery methodologies and tools. Integrated data sets, without appropriate representations and tools for supporting users exploration, can only overwhelm users rather than intriguing any insights or hypotheses. Therefore there is an urgent demand for visual analytical platforms and tools that can harness the advantage of computational models that deal with the vast amount of data, as well as the knowledge and reasoning capabilities of experts [141]. Yet relatively few established works of interactive visual interfaces for biomarker discoveries have been reported. The design and evaluation of such visualizations becomes an active research topic.

CHAPTER 3 TERRAIN SURFACE HIGH-DIMENSIONAL VISUALIZATION

3.1 Problems with the Node-Link Diagram Graph Visualization

Conventional node-link diagram graph and network visualization is a widely used approach to reveal non-linear relationships among data items in high-dimensions. It is useful in describing large number of words and phrases and how they are related in literatures. It can also present the biomolecular networks — interactions among thousands molecules (e.g. genes, proteins), of a biological context. However, large scale graph is prone to the visual clutter caused by dense edge crossings. It becomes difficult to identify and interpret any pattern from the graph. In addition, it is hard for users to perceive patterns reflecting the changes of nodes and their neighborhoods. For instance, visualizing biomolecular networks can help biologists to understand the high-level protein categorical interplays in a network. However, a large and cluttered biomolecular networks is inadequate when the focus of biological questions is on the patterns of functional changes of genes, proteins, and metabolites with biological significance such as the following:

- What are the significant functional changes in a given biological condition such as human disease?
- Where are such changes in the biomolecular network context?
- Can we focus attention on biologically significant changes in gene/protein expression measurements, while allowing for inherent data noise introduced by imperfect data collection instruments?

These questions are of central concern in post-genome molecular diagnostics research, particularly, biomarker discovery. The reason conventional node-link

diagram network visualization methods are insufficient to address these biomarker discovery related questions is simple: the graph-based network visualization cannot allow users to shift the focus of visual analysis to the “nodes” (encoding proteins, genes, and metabolites) away from “edges” (encoding relationships between pairs of proteins, genes, and metabolites). In general, when the primary interest is reading the overview of the properties of regions in the graph and identifying patterns on the change on the properties, the graph visualization needs to be augment and improved.

In this chapter, we present a novel framework to address the above challenges using scattered data interpolation, surface rendering and interactive visual exploration. Figure 3.1 shows the pipeline of this visualization method. The method offers the following advantages:

1) A modified force-direct graph layout model captures the inherent structural properties of original network which is both edge-weighted and node-weighted.

2) Scattered data interpolation and surface rendering avoid the scalability problem and represent features derived from data set as prominent geographic landmarks.

3) Interaction with areas and heights supports multi-scale visualization and visual exploration, which can lead to the discovery of new hypotheses based on domain knowledge.

The application of this method to bio-molecular interaction networks of Alzheimer’s Disease (AD) enables the enhancement and detection of regions of bio-markers of disease progression, which addresses a major concern of the biology community.

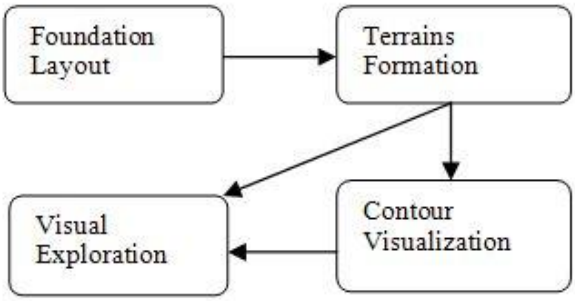

Figure.3.1 Framework of GeneTerrain visualization.

The rest of the chapter is organized as follows: In the section 3.2, we then discuss the methods of **Terrain Surface Visualization**, which has two critical components: the foundation layout algorithm for **a base network** and the scattered data interpolation algorithm of **a response variable**. In the same section, we then discuss a terrain and contour rendering technique for the generation of smooth surface geometry. In section 3.3 Applications and Results, we apply this general graph/network visualization pipeline to Alzheimer's disease (AD) protein-protein interaction networks. Gene terrains are rendered with gene expression values as the height dimension. Next in the same section, we demonstrate the interactive exploration and multi-scale visualization techniques. Additionally, bio-discoveries related to the progression of disease are presented. In the final section of this chapter, we conclude the paper with additional discussions.

3.2 Foundation Layout of the Base Network

3.2.1 Initial Layout

We define a general node-weighted edge-weighted undirected graph as G = (V, E, *f*, *g,* O, C), where V is the node set $\{v_i\}_{i=1}^{n}$, E denotes the edge set $\{e_{ij} \mid (v_i, v_j) \in V\}$, f assigns a weight value to each node, $f : V \to R$, g assigns a score to each edge $g : E \to R$, O is the center position of the planar graph in world coordinates, and C is the scale of graph. The grid scale for base map of terrain rendering can be defined based on C.

In the foundation layout, nodes in the original networks are laid out in two steps: initial layout and optimization. Though the layout algorithm gives priority to nodes with larger weights, it also keeps them as compact as possible. This is because drastically different distances among pairs of nodes can cause the resolution of later defined grids to be arbitrarily small, which in turn leads to aliasing problems in rendering. Intuitively, nodes with larger weights push others aside while edges are pulling end nodes closer. The final position of each node is the accumulated effect of the constraints imposed on it. *f* and *g* are used to quantify the constraints. Formal definitions of the constraints are introduced and explained later. The improved layout of the graph is achieved by optimizing this constraints-based system.

In the initial layout, the graph can be configured manually to approximate the global minimum before the optimization, to avoid local minima in the process of optimization. Nodes are arranged in 2D and are always kept planar during the optimization. Each node v_i, with $f(v_i)$ larger than threshold θ_f is radially laid out

around point O. The radius is proportional to log ($f(v_i)$) which reflects the idea that nodes with larger weight push each other asides. A logarithmic scale is used here and later in the model, because it can reduce any significant difference of distances among pairs of nodes. Starting from one of those nodes, an extended version of Breadth First Search (BFS) is carried out to determine the position of other nodes. The node is radially laid out around its parent when it is first visited. The position is adjusted each time it is revisited by other nodes. The algorithm can be outlined in the following pseudo-code:

```
Proc BFSdraw:
if (queue is empty) return;
v_c = queue.head();
n = number of v_c's neighbors
step = 360/n;
for each neighbor v_i of v_c
if (v_i is not visited ) {
   radius = cal_radius(v_i);
   angle = angle+step;
   cal_position (v_c, n, angle, radius);
   set v_i as visited;
   queue.add(v_i);
 } else {
   adj_position(v_c,v_i);
 }
 BFSdraw();
```

where *cal_radius()* calculated the radius of v_C for the radial layout around v_c depending on $g(v_i, v_c)$, $f(v_i)$, and $f(v_c)$, *cal_position()* calculates the actual position for v_i and *adj_position()* adjusts v_i's position depending on $g(v_i, v_c)$, $f(v_i)$, and $f(v_c)$. The actual algorithms of *cal_position()* and *adj_position()* are designed similarly as the energy minimization model discussed in next section.

3.2.2 Energy Minimization

To optimize the constraints-based system, the spring embedder (force-direct) model is applied. The classical spring model is:

$$E = \frac{1}{2}\sum_{i \neq j} k_{ij}(|\, p(v_i) - p(v_j)\,| - l_{ij})^2$$

Where $p(v_i)$ is the position of node v_i; l_{ij} is the ideal *spring length* for node v_i and v_j, which is usually a predefined path between the two nodes; k_{ij} is Hook coefficient. This model can be generalized as a MDS model, where $|\, p(v_i) - p(v_j)\,|$ is the original distance of the two nodes in *d* dimension and l_{ij} is the distance in projected d' dimension (d >= d'). In our model, however, each of the terms in the general model is redefined based on constraints. Node that weight f and interaction strength of an edge g are two important factors. So there are two types of constraints for placing the node pairs (v_i, v_j):

- Node constraint. Nodes are kept together to keep the layout compact. We introduce the concept of *area of influence* for each node, which is a circular area with the node at the center. When the pair of nodes does not have edges between them, nodes tend to push other nodes out of their *area of influence*. In other words, two *areas of influence* cannot overlap under this circumstance. The radius of the *area of influence* is determined by $f(v_i)$ and $f(v_j)$.
- Edge constraint. The edge between two nodes tends to pull them closer. The *area of influence* can somewhat overlap. However, the distance between the centers of the two *areas of influence* is still preserved by $g(v_i, v_j)$.

Both nodes and edge constraints influence the final position of node pair (v_i, v_j). Pairs of nodes having no edges between them are subject to node constraints, whereas pairs of nodes having edges between them are subject to edge constraints. Therefore the formal definition of our force-direct model is:

$$E = \frac{1}{2}\Big(\sum_{(v_i,v_j)\notin E}(|\,p(v_i)-p(v_j)\,|-\log(f(v_i)+f(v_j)))^2 + \sum_{(v_i,v_j)\in E}(|\,p(v_i)-p(v_j)\,|-g(v_i,v_j))^2\Big)$$

where $\log(f(v_i)+f(v_j))$ is the ideal projected distance for v_i and v_j when they do not have edges and $g(v_i, v_j)$ is the ideal projected distance when they share an edge.

Nonlinear system minimization techniques can be applied to minimize the energy of this model. As Newton's methods suffer from the complexity of computing the Hessian matrix, we use conjugate gradient to estimate the descent direction in *N* dimensions. After the optimization, we obtained the positions of each node in the base network.

3.3 Terrain Formation and Contour Visualization

3.3.1 Definition of the Grids

In previous definitions, *O* is the center and *C* is the scale of the graph. The optimized base network is scaled to fit into a bounding square that centers at *O* and has edge length *C*. The grids are defined to be as the same size as the bounding square that centers at *O* as well. If the shortest distance between any pair of nodes is βC after minimization, where β<1, the resolution of the grids is defined to be smaller than βC. So no cell of the grid has more than one node.

3.3.2 Scattered Data Interpolation of the Response Variable

Now the grids containing the optimized base network is ready for surface rendering. Suppose *s* is a discrete scalar function defined over nodes $s: V \to R$. The result of s mapping can represent a numeric attribute the node

has, in other words, s is the response variable. The complete scalar field over grid points should be interpolated from the available response variable of nodes. The scalar field is rendered as elevations to generate a height field from the 2D plane where the nodes reside.

Shepard's method is one of the simplest interpolation techniques, which was originally proposed in 1968 [143]. It takes the distance weighted average of the interpolation points as the interpolation function. An improved Shepard's method was proposed later [144], which interpolates the displacements of the points. In our scattered data interpolation, a scalar value is used as "displacement". Therefore the unknown scalar value for each grid point can be computed by:

$$s(p) = \sum_{i=1}^{n} \frac{s(v_i)}{d_i^r(p)} / \sum_{i=1}^{n} \frac{1}{d_i^r(p)}$$

Where p is the grid point with unknown scalar value, $s(v_i)$ is the scalar value of node v_i , $d_i^r(p)$ is the distance from node v_i to p and r is the exponent parameter to weigh the factor of distance. Since *area of influence* is introduced, nodes with different weights $f(v_i)$ cannot be interpolated as they are symmetric points in interpolation. The scalar value of nodes with larger weight should have more influence on the scalar value of grids than nodes with smaller weight. Thus, the modified Shepard's method is as follows:

$$s(p) = \sum_{i=1}^{n} \frac{s(v_i) * f(v_i)}{d_i^r(p)} / \sum_{i=1}^{n} \frac{* f(v_i)}{d_i^r(p)}$$

where the $f(v_i)$ is the weight factor in interpolation.

3.3.3 Elevation and Surface Rendering

The scalar value of each grid point is rendered as an elevation from the 2D plane of the foundation layout. The position of the elevated point q of grid point $p(x,y)$ is

$(x, y, \alpha * s(q))$, where α is a uniform scale factor. The height field can then be rendered as a surface, given the scalar values of the grids points are available. We use the Visualization Tool Kit (VTK) to generate the terrain surface and contours based on constant height values. The color scheme is adopted to denote different height. Let $\alpha * s(v_i)$ be $H(v_i)$. If $H(v_i)$ is larger than certain value S_i, v_i in 2D plane of contour rendering will be enclosed by the contour of value S_i.

3.4 Visualization of Gene Terrains

Software such as ProteoLens [33] can be used to model and visualize biomolecular interaction networks of AD. The network visualization of the software tries to distinguish nodes and edges using different colors, sizes, and distances. As the scale of the network increases, more proteins become involved and the edge-crossings and overlapping nodes obscure the network, thus, making exploration difficult. In the gene terrain visualization, edges are replaced by geographical neighborhoods implied by the terrain appearance. The neighborhood size is also proportional to the significance of proteins implied by the weight of the nodes.

3.4.1 Experimental Data Sets

In the original network [145], 20 highly significant proteins were based on a method developed by Chen et al. [146], as shown in Table 3.1. Given that each node is a protein, each edge is indicating interaction between two protein nodes in a biological process, the scalar for each node is defined to be the gene expression value of each gene/protein. The gene expression microarray correlation analyses prove to be successful in indicating major transcriptional response [147]. The gene expression value for each protein can be derived from probe set, each of which is indentified by its AFF_ID [148] and contains a single gene expression value. Each probe set gene expression value is then mapped to

a gene expression value identified by UNIPROT_ID, which were made compatible with protein node identifiers in the foundation layout. Algebraic average is used to represent the aggregated expression value if multiple probe set can be mapped to a unique protein identified by its UNIPROT_ID. After this aggregation, 218 out of 625 protein nodes and 19 out of top 20 significant protein nodes remained. The AD gene expression data were collected from a prior AD study consisting of 31 individuals [148].

Table 3.1 Top 20 significant proteins UNIPROID and weights.

PROTEIN ID	RANK	AD RELEVANCE WEIGHT
A4_HUMAN	1	35.5
LRP1_HUMAN	2	34.0
PSN1_HUMAN	3	27.1
PIN1_HUMAN	4	18.6
FHL2_HUMAN	5	16.4
PSN2_HUMAN	6	13.6
NP1L1_HUMAN	7	11.3
S100B_HUMAN	8	11.2
CDK5_HUMAN	9	11.1
NOG1_HUMAN	10	11.0
CLUS_HUMAN	11	10.9
NCOA6_HUMAN	12	9.7
CATB_HUMAN	13	9.3
ARLY_HUMAN	14	8.6
FLNB_HUMAN	15	8.3
CTND2_HUMAN	16	7.8
APBA1_HUMAN	17	7.3
C1TC_HUMAN	18	5.9
ODO2_HUMAN	19	5.1
MK10_HUMAN	20	5.0

3.4.2 Gene Terrain and Contours Rendering

Figure 3.2a is the foundation layout of the data set before optimization, and Figure 3.2b is the foundation layout after optimization. 5% of the nodes with largest weight are colored red, the rest of the nodes are colored blue. After

minimization, the most significant nodes are spread out. Black circles indicate the regions of interests, which contain at least one highly significant AD protein.

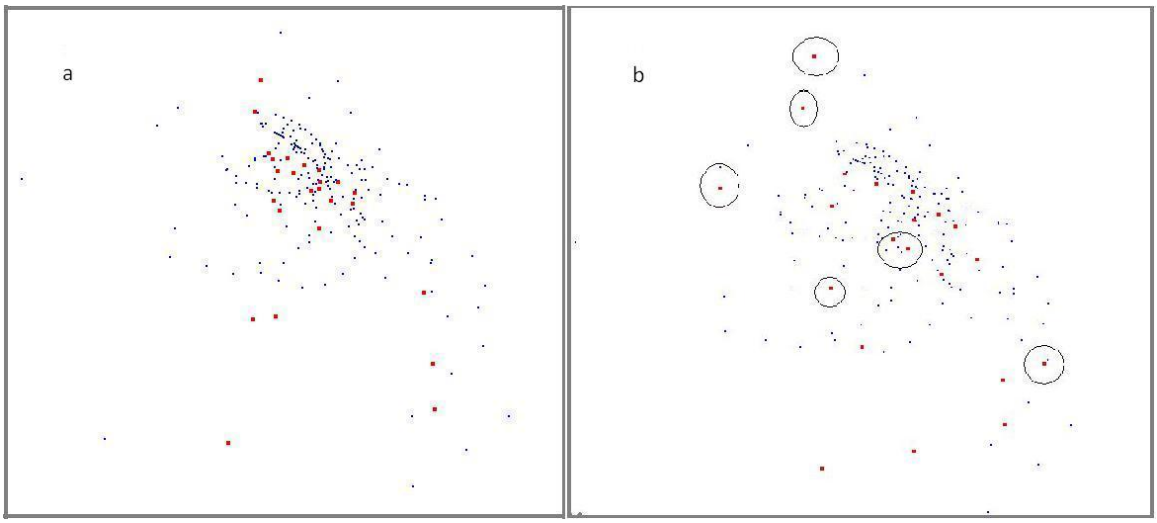

Figure 3.2 Foundation layout before optimization (a) and after optimization (b). The nodes with high weights are circled in the right panel.

Gene/protein expression values [148] are used to render gene expression profile visualization. During the transcription process from DNA to mRNA, the number of mRNA produced is taken as the value of expression of this gene/protein. (In the case study of Alzheimer's disease, the mapping between mRNA and proteins is generally 1-to-1; therefore, we use the gene expression value as the protein expression value). This rendering is based on the foundation layout and interpolation method described earlier. Figure 3.3 is an example showing gene terrain surface and its contour for the normal control group. Note that height value in Z direction is adjusted to a proper scale of gene expression suitable for display and exploration. The scale of Z direction is different from the scale of grids used in the X-Y plane.

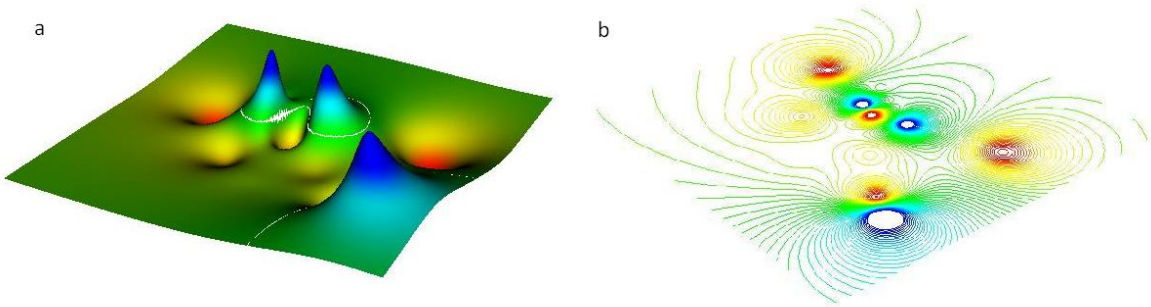

Figure 3.3 GeneTerrain visualization for averaged absolute gene expression profile of a group of samples (size=9) from normal individuals. (a) is a GeneTerrain surface map. (b) is a GeneTerrain Contour map.

3.5 Interactive and Multi-scale Visualization on Gene Terrains

User interaction is provided for visual exploration. All labels can be toggled on to support overview of the distribution of protein nodes. The label of individual protein can be toggled on by querying the name of the protein. To enable multi-scale visualization, a threshold T (T>0) can be set and only proteins whose height values are larger than T will be displayed. In this way, multi-scale visualization can organize hundreds of proteins and gradually narrow down the search space by increasing the value of T. Meanwhile, it can group proteins by different threshold and may yield biologically meaningful clusters. Figure 3.4a is the terrain with proteins threshed by T=3; b is the contour visualization. Since the annotations are not readable, multi-scale schemes must be developed, which is shown in Figure 3.5. Details of local region can be obtained by zooming in. To support more advanced visual explorations, proteins names in regions of interest could be shown by clicking the area. Note that only proteins whose height value are above the current threshold T and whose coordinates are within a circle centered at clicking point with predefined radius ε are shown. Figure 3.5 a shows all protein names in a peak area in contour visualization; b is a zoomed-in view to identify each protein's name.

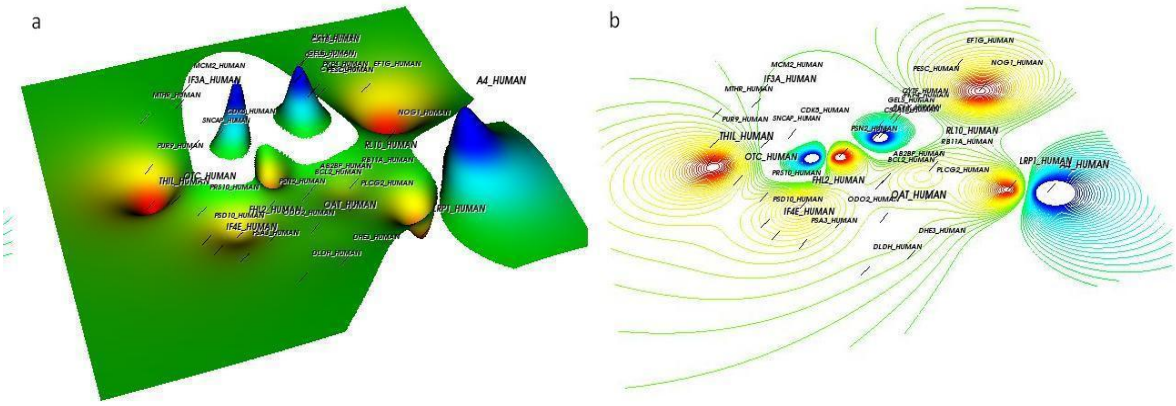

Figure 3.4 (a) GeneTerrain surface map with labels on when threshold T=3 (b) Contour map for (a)

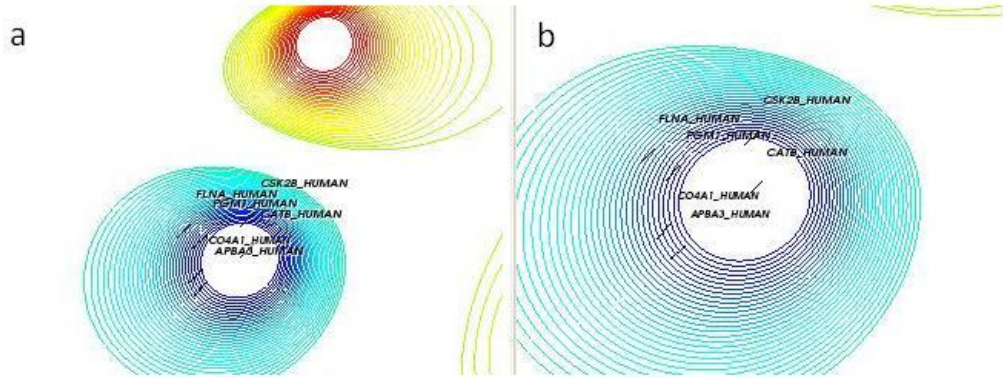

Figure 3.5 (a) Proteins with names in one peak area. (b) Proteins in the same peak area can be identified by zooming in. They are “FLNA_HUMAN” “PGM1_HUMAN” “CSK2B_HUMAN” “CATB_HUMAN” “APBA3_HUMAN” “CO4A1_HUMAN”.

3.6 Visual Exploration on Differential Gene Expression Profiles

The protein expression value varies in different individuals and according to the circumstances of diseases, such as AD, where certain transcriptions are up-regulated or down-regulated abnormally. In our data set, 9 individuals are normal controls, 7 are incipient AD patients, 8 are moderate AD patients, and 7 are severe AD patients. We verified that the gene expression data sets obtained from the publication was already previously normalized. In each of the four groups, we

averaged the absolute gene expression value across all grouped individuals to the mean value. We also paired the AD patient groups (incipient, moderate, and severe) with the normal control group to derive relative gene expression. Relative (differential) gene expressions are rendered as a new type of gene terrain sharing the same foundation layout of the gene terrains for absolute gene expressions. Relative gene expression values are calculated here according to standard gene expression analysis conventions as the following:

$$\mathrm{Re}\,Exp(pro\_id) = \begin{cases} \dfrac{Exp2(pro\_id)}{Exp1(pro\_id)}, Exp2(pro\_id) >= Exp1(pro\_id) \\ -\dfrac{Exp1(pro\_id)}{Exp2(pro\_id)}, Exp2(pro\_id) < Exp1(pro\_id) \end{cases}$$

where *ReExp(pro_id)* represents the differential gene expression ratio for the diseased stage vs. normal control condition for a given protein with *pro_id* as the identifier, *Exp1(pro_id)* is the absolute gene expression value for the same protein under condition 1, and *Exp2(pro_id)* is the absolute gene expression value for the same protein under condition 2.Differential gene expression values is therefore either larger than or equal to 1 or smaller than -1. To filter differential gene expression values due to possible noise, we only consider changes beyond 5% of normal controls, or >= 1.05 and <-1.05 cases, setting those changes between 0.95 and 1.05 to 1.00. In Figure 6 and 7, we show a series of differential expression surfaces and contours for control vs. incipient, control vs. mild, and control vs severe conditions. We set threshold of height values for the surface. The part of surface whose height value is out of the range is set to be transparent and no contour will be displayed either. Peak and valley areas are colored separately. (Though usually red represents over-expressed value, here we use red to represents areas with comparatively lower height value as the surfaces are control vs. condition)

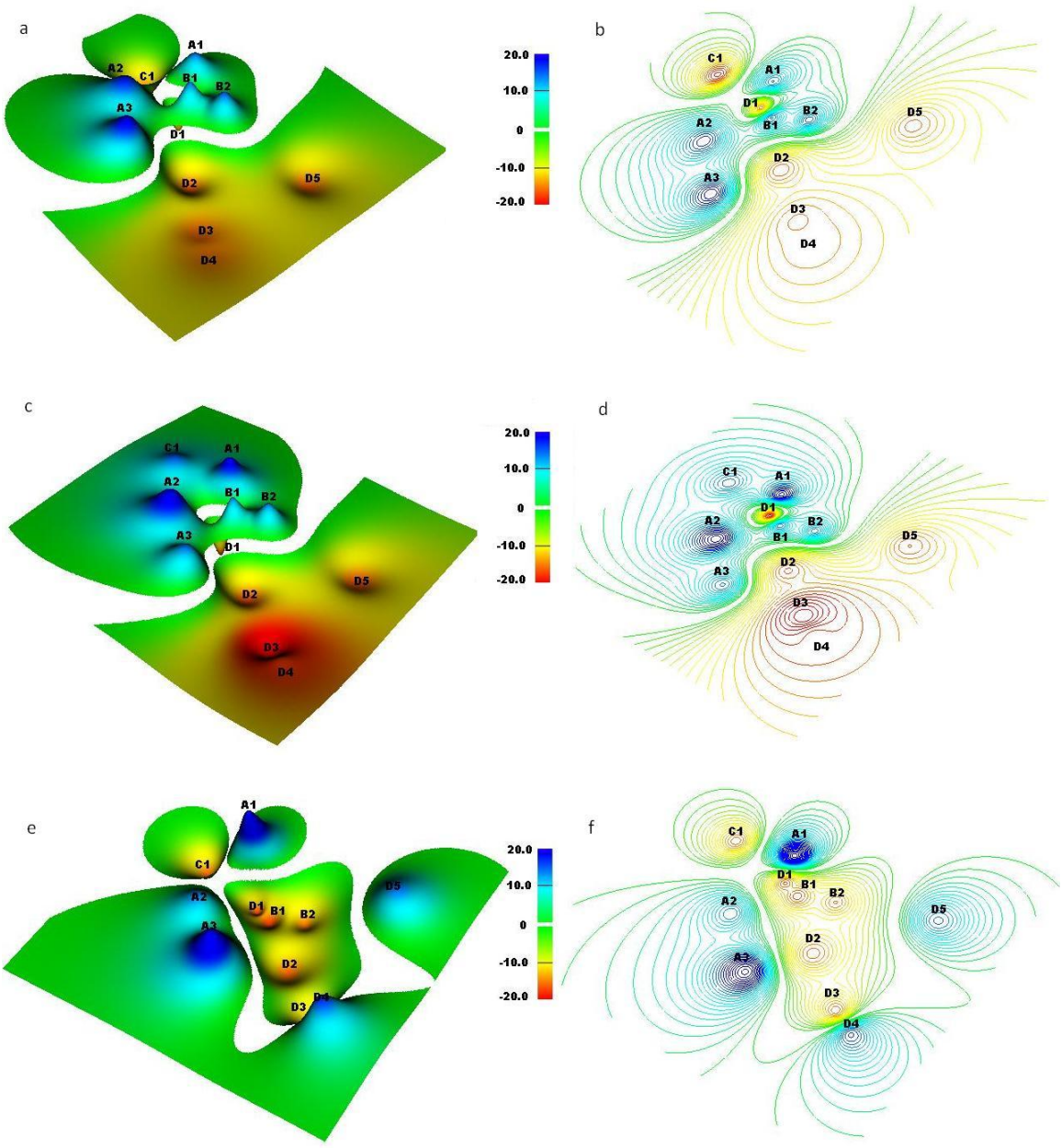

Figure 3.6 GeneTerrain surface maps (a) (c) (e) and contour visualization (b) (d) (f) for averaged AD differential gene expression profiles. Among them, (a) is the differential expression profile of control versus incipient, and (b) is the corresponding contour visualization; (c) (d) are for control versus moderate; (e) (f) are for control versus severe.

From these figures, we observe peaks and valleys in the Gene Terrain surface maps and rings of concentric circles in the Gene Terrain contour maps. These distinct visual features serve as “visual cues”, allowing a biologist to quickly comprehend the results of AD differential gene expressions in their biological context. In Figure 3.6 a, c, e , peaks are clearly identifiable with colors ranging from yellow, green, and blue. We labeled major peaks and valleys for easy comparisons between different panels in Figure 3.6 a-f. Area of base with height value within certain range is set into transparency to separate features. With these visual representations, several observations can be made readily. First, we observe that peaks A1, A2 and A3 are present in all panels, indicating that relative to controls, the AD conditions lack the expressions for these genes. The proteins in these peak areas, especially those determined to have significant links to AD (protein nodes with high weight scores from previous studies), are candidate AD *diagnostic biomarkers*. Similarly, valley D1 and D2 can be *diagnostic biomarkers* too. Second, we observe that the height of peak A1 increases as AD progressed in stages. Therefore, proteins in this peak can be considered candidate *prognostic biomarkers*. Third, we observe that peaks B1 and B2 disappear in severe form of AD and vally D3 appears in severe form of AD. This makes the up-regulation of proteins within peaks B1 and B2 as well as down-regulation of proteins within peaks D3 candidate *staging biomarkers*. Fourth, we observe that the small peak C1 appears in moderate AD vs. control normal whereas it transformed to a valley in incipient or severe differential AD gene expression profiles. The inconsistent behavior of the protein in the area of C1 certainly poses an interesting question.

To identify proteins of interests within peaks/valleys in the gene terrain, we may click the regions of interest and toggle the label on. Figure 3.7a displays the name of proteins in the peak or valleys whose height value is above threshold in control vs. incipient terrain. So those nodes are of primary interest. Figure 3.7b is

the corresponding contour visualization. Using the interactive functionality introduced in previous section, more protein names will appear in the region of interest by decreasing the threshold value. By examining all relative gene terrain, we identified that the *prognostic biomarker* in peak A1 is mainly explained by protein “CDK5_HUMAN” in the top 20 significant proteins listed in Table 3.1. In separate independent studies reported in biomedical literature, the link between cdk5 and Alzheimer’s Disease were well documented [149].

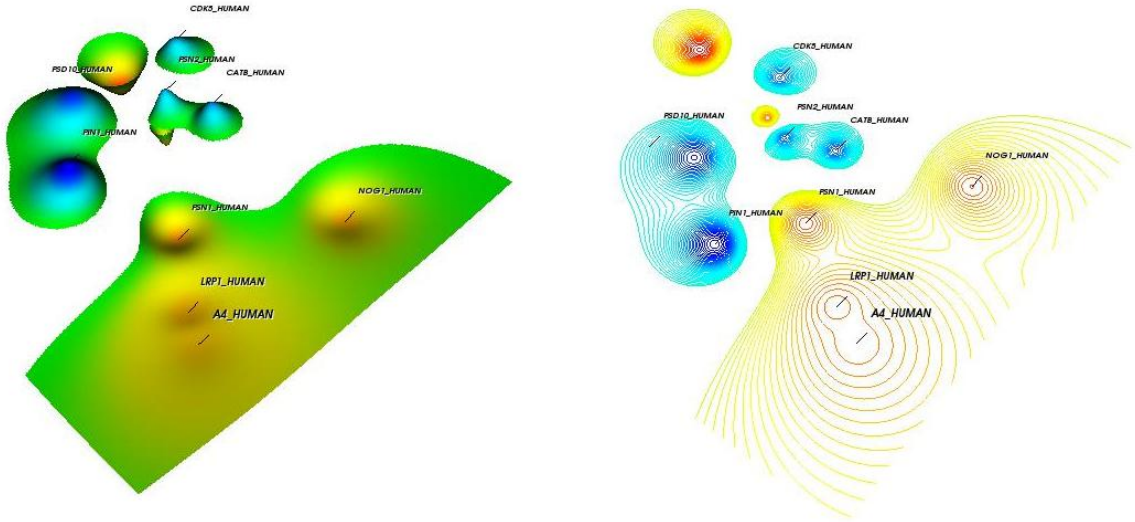

Figure 3.7 (a) Control vs incipient GeneTerrain surface map with labels in regions of interest, height value threshold = 17. (b) Contour map for (a).

3.7 The Advantages of the Terrain Surface Visualization

In this chapter, we introduced a new graph visualization technique, the terrain surface visualization, which offers an alternative overview of networks for high dimensional data sets. The continuous surface takes advantage of human perception and prioritizes local regions of nodes as landscape patterns. The differential surface, which is obtained from the changing response variable over the same base network, exposes the change or dynamics of the region properties as visual patterns as well. We applied the new framework for

biomolecular interaction network visualization and exploration. Differential expression surfaces and contours on gene terrains could give much visual information to help biologists with biomarker discovery tasks and clinicians with molecular diagnostics tasks. The visual features of gene terrain were successful primarily because of several innovations. First, we developed enhanced visual representation of network node features as peaks and network edge features as neighborhoods. Second, we developed a gene terrain foundation layout model, in which nodes with high biological relevancy weight scores are well separated from each other spatially with inter-node distances as a function of the weight scores. Third, we encoded differential gene expression information as terrain surface heights. Fourth, interactions with areas and heights facilitate the visual exploration.

Our visualization method is scalable, because 3D surfaces are rendered and contours are used to differentiate groups of proteins at different height levels, avoiding the intricate navigation in the conventional graph-based network. Our visualization yielded biologically significant results — several types of candidate biomarkers, useful for hypothesis generation and derivation of biomarker panels for future clinical use. On the one hand, we believe that the principle and framework of our work can be generalized for biomarker discovery data explorations beyond the case study of Alzheimer's Disease. We are in the process of refining this technique, developing user-friendly software tools, and applying it to other disease biology studies. We believe our graph-based terrain rendering and interactive visual exploration can serve well in other application domain of graph and network visualization.

CHAPTER 4 CORRELATIVE MULTI-LEVEL TERRAIN SURFACE VISUALIZATION

4.1 Challenges of Visualizing the Complex Networks

In the last chapter, we have introduced the terrain surface visualization. It addresses users' difficulties in reading and understanding one large scale network, because there are visual clutters induced by condensed edges and nodes. The technique visualizes the profile of an attribute of nodes over a base network, and successfully shifts users' attentions to the network nodes and their surroundings prioritized as the landscape features. However, in many large networks that model real world relationships, the data entities belong to different categories, and the edges represent heterogeneous interactions. To decide which aspect of information should be encoded for showing in the limited screen space are usually up to the application context. Recent advances in the biology community raise a similar problem: how visualization can assist the biologists to study complex networks with rich semantics. The complex networks, for example, consist of a bio-molecular interaction network that represents interacting molecules and a disease association network that represents the genetic commonalities among each other. At the same time, one molecule and one disease are correlated by measurements on the disease capability of the molecule. The appropriate correlations can lead to new functional hypotheses yet are hard to be identified among the rich and complex signals presented in the correlative networks. Such problems require appropriate visualization techniques for studying the correlative multiple networks, but few have been reported.

In this chapter, we propose the correlative multi-level terrain surfaces for visual investigation of the correlations among networks. It is based on the terrain surface visualization we introduced in Chapter 3. We study the correlative biology networks to showcase the construction, properties and the applications of the approach. The correlative multi-level terrain surfaces are built from the following component: **molecular network terrain surfaces**, the base network of which is a biomolecular interaction network; **phenotypic network terrain surfaces**, the base network of which is a phenotypic association network; the response variable of both terrains — numerical correlations derived from the molecule-phenotype measurements. The approach has the following advantages:

1) The approach provides an intuitive overview for the correlations between nodes in the correlative multiple networks,

2) It visually highlights the major signals for users to gain critical insights from the networks, while depressing the irrelevant information,

3) It enables users to perform visual analytical tasks on the correlations among networks based on visual patterns, which lead to hypotheses generation as well as visual evaluation,

4) It scales well and is robust when being exposed to the noise inherent in the large networks. Therefore the discoveries are stable.

In the next section, we first show the construction and properties of the correlative multi-level terrain surface, using a small data set of correlative core gene term association network and core cancer term association network. We derive their correlations using translational mining model on biomedical literature collections. After constructing the correlative multi-level terrain surfaces, we arrange those surfaces according to the node position in the original network. We show that not only the correlative multi-level terrain surface model is correct, but also it visually captures the major signals from the correlations, result of which is comparable to the automatic algorithms. Then in section 4.3, we apply our

approach to another pair of high-quality correlative biological networks, a candidate biomarker protein interaction network and a cancer disease association network. We demonstrate the correlative multi-level terrain surfaces are effective in performing visual analytical tasks: molecular biomarker discovery and performance assessment. We also do perturbance study to show the robustness of our approach. Finally we conclude this chapter in section 4.4 with additional discussions.

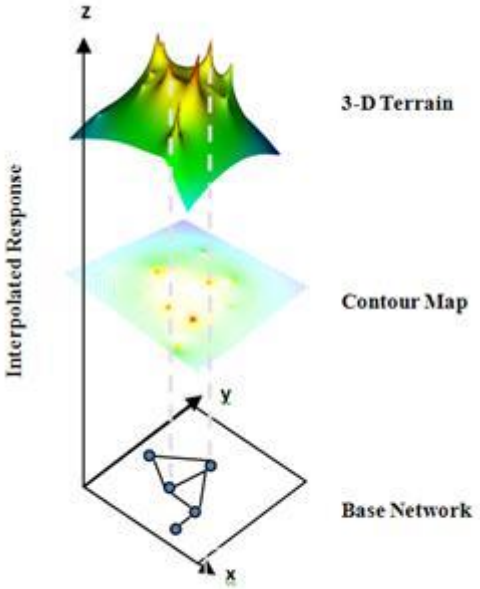

Figure 4.1 The Terrain Surface Visualization concept.

4.2 Terrain Surface Visualization

We have introduced the terrain surface visualization in Chapter 3. In summary it interpolates a responsible variable value of nodes into a smooth surface over the foundation layout (see Figure 4.1). Here we introduce the *consensus terrain*, a terrain surface where the response variable can consist of the functional mapping from multiple response variables. For this study, we use a linear equal-weighted function to combine the response variables. For each point p in consensus terrain, its vertical elevation is calculated as the *average* elevation of individual response variables. Figure 4.2 a is the consensus terrain of c-e.

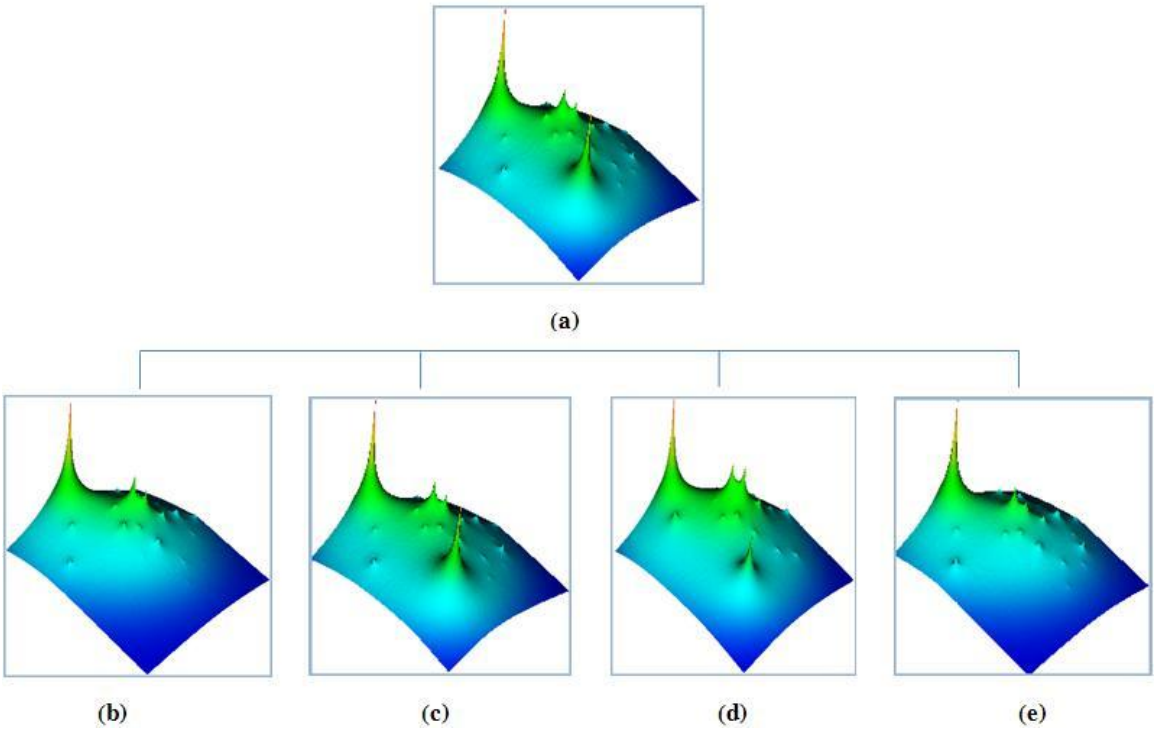

Figure 4.2 The terrain surface in (a) is the consensus terrain of (b) (c) (d) (e).

4.3 Construction of Correlative Multi-level Terrain Surface Visualization

Figure 4.3 shows the framework of correlative multi-level terrain surfaces construction. Figure 4.3a is the process of molecular network terrain construction: a biomolecular interaction subnetwork (BAN) is first constructed from a comprehensive, literature-curated list of molecues for a specific phenotypic context, with physical molecular interactions or functional associations. The physical interactions can be obtained from a variety of sources, including: gene co-expression network, protein-protein interaction network, microRNA and RNA target measurements. Then a molecular network terrain is interpolated by treating BAN as the base network, and the molecular measurements as the response variable. Figure 4.3b is the process of phenotypic network terrain construction: a phenotypic association subnetwork (PAN) is first constructed from a set of similar phenotypic conditions as the base network. The associations among the phenotypes can be derived from a variety

of sources, including “omics” data, genome-wide association study, literature mining. Then a phenotypic network terrain is interpolated by treating PAN as the base network and treating the phenotypes’ measurement as the response variable. The measurements for the response variable for both types of terrain can be the normalized phenotype-molecule associations’ score shown in Figure 4.3c. A number of literature mining algorithms can be used to derive the association score for every pair of a phenotype and a molecule.

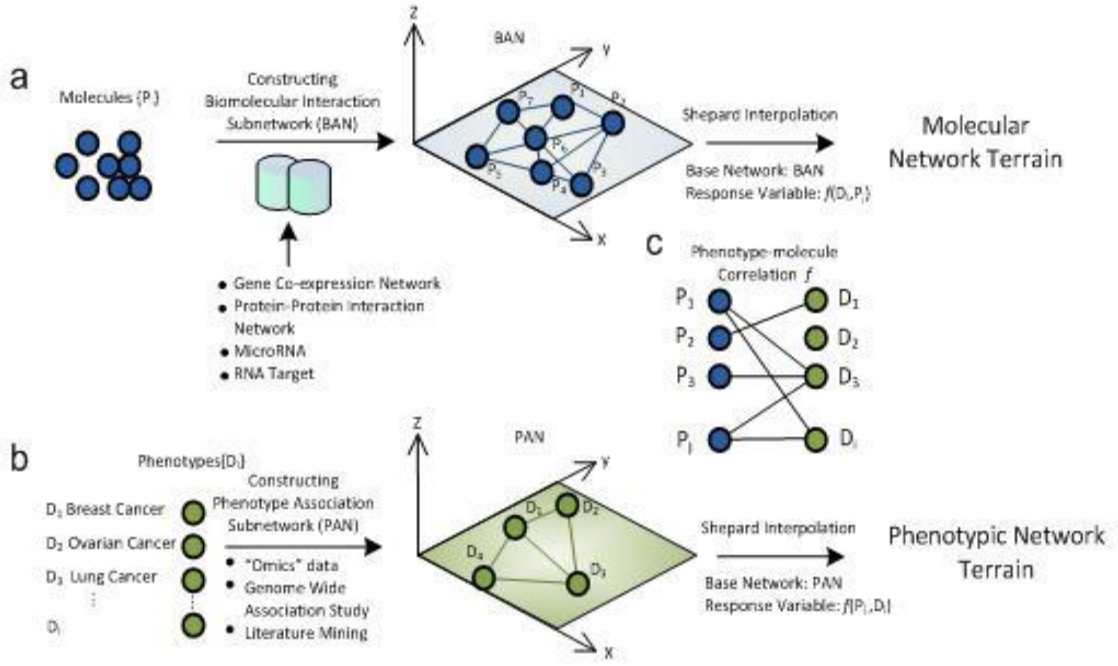

Figure 4.3 Correlative Multi-level Terrain Surfaces construction: (a) Molecular Network Terrain construction, (b) Phenotypic Network Terrain construction, (c) Phenotype - Molecule correlation.

4.4 A Pilot Study of the Correlative Multi-level Terrain Surface

To show the construction and properties of the approach, we create a small data set of a cancer term network and a gene term network.

4.4.1 Retrieving the Biological Entity Terms

We select unique cancer terms from MeSH [150]. The gene terms are then retrieved by using the cancer terms to query the PubMed [125] abstracts collection. For every query pass, only a constant number of returned genes terms are kept (in this study, the constant number is 20), and subsequently , unique gene terms are kept. We use the Uniprot [151] naming convention to label each gene throughout the paper. Also, during the querying process, top 20% of all article abstracts returned are kept for mining the association among the terms. As a result, we have identified the 25 cancer terms representing top killing cancers, and have chosen the connected sub-network of 25 terms as the core cancer network. Second we have chosen a connected sub-network of 20 genes which present in the core cancer genes.

4.4.2 Mining the Term Correlations

The associations between any two terms a_p and a_q are calculated by the method proposed in [152] for trans-associations mining ,which factors in both co-occurrences in the abstracts collection and the indirect associations inferred by transitive closures. We now summarize the method as the following steps:

Step1.Calculate the weight of term a_k in one document i, W_{ik}, using tf-idf algorithm [153].

Step 2.Identify the score of co-occurrences between any two terms a_k and a_l , by summing up their weight in each document *i*.

$$associations[k][l] = \sum_{i=1}^{N} W_{ik} * W_{il}, k = 1,2 \dots m, l = 1,2, \dots . m \quad (4.1)$$

Step 3.Identify the indirect association between any two terms, assuming that a transitive relation R could apply onto the terms associations:

$\forall a_p a_r a_q, (R(a_p, a_r), R(a_r, a_q)) \rightarrow R(a_p, a_q)$ where a_p, a_r, a_q are terms. We first obtain a binary matrix A for the co-ocurrences of all such pair of terms in *association*. Then a transitive closure A^* of the binary matrix is computed. Then

$TA = A^* - A$, where each non zero $TA(i,j)$ indicates the existence of an indirect association between the two terms.

Step 4.Score the associations between two terms. In each non zero cell $TA(i,j)$, identify the segments of the paths, and looking up the score of each segments in *associations* calculated before. The score of such a path is the summations of segment scores. The score of association between terms are the minimum among scores of all paths.

4.4.3 Building the Terrain Surfaces

After calculating the associations between any pair of the terms, the gene term association network are laid out, as a type of biomolecular interaction network (BAN). We lay out both networks using Multi-dimensional Scaling (MDS) with the optimal distance between any nodes proportional to the association values. Other graph drawing algorithms may apply as well. Then for each gene term node in the gene term association network, we build its phenotypic network terrain, i.e. a disease terrain based on the cancer term association network. We replace the gene term node with its disease terrain surface. Figure 4.4a shows a schematic arrangement of the disease terrain surfaces, on top of a gene term network, and (b) shows the formation of one disease terrain surface in (a), with a cancer term network as the base network.

As the scale of the network in Figure 4.4a increases, only limited space is available. So in the arrangement, based on the resolution, we could cluster entity nodes to render their consensus terrain surface, and put the consensus terrain surface in the centroid of the cluster as the summary. Similarly, we replace each cancer term with its molecular network terrain, i.e. a gene terrain surface based on the gene term association network.

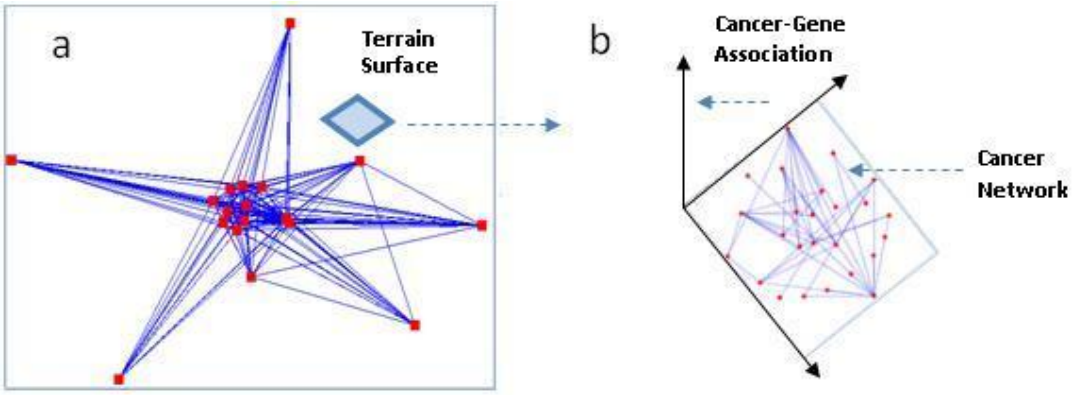

Figure 4.4 The arrangement of terrain surfaces: (a) a terrain surface on top of a node in a gene network; (b) the formation of the terrain surface in (a).

4.4.4 Properties of the Correlative Multi-level Terrain Surfaces

The final arrangement of the terrain surfaces is shown in Figure 4.5: Panel A shows the disease terrain surfaces for gene terms in the gene association network; Panel D shows the gene terrain surface for cancer terms in the cancer association network. In Panel A, the disease terrain labeled ‘RBM4 cluster’ (also see L-shaped Panel C), is a consensus terrain generated from neighboring genes ‘RBM4’ ‘SHBG’ ‘LHCGR’ together. We do not show individual disease terrain surface in Panel A, because the three genes are cluttered with each other in the gene core network. However, we show each of the disease terrain surfaces in Panel C. In Panel B, we show the disease terrain surfaces of four well separated genes in the gene association network.

From observing the terrains of RBM4 cluster in Panel C, we can see that genes that are close together in the network have similar terrain shapes. From observing the terrain surfaces in Panel B, we can see that the four well separated genes have distinctively different terrain surface shapes. So the layout of the gene network is consistent with the shape variations appeared in the disease terrains of gene nodes. The further away two genes are, the more differing shapes they have. Similar observations could be made by the changing trend on terrain

surfaces in Panel D. The results validate the method we used to build up the correlative multi-level terrain surfaces of biology networks, as the terrain surface shape variations are consistent with the nodes positions in the network.In the disease terrains in Panel A, each peak represents a strong correlation between this gene and one of the diseases in the base network. We identify the major peaks in all disease terrains in the gene network, and mark the disease in the disease-gene association matrix where each row represents a cancer and each column represents a gene. The colors represent different scales of the peaks. Then we do a two-way clustering on the heat map we generated from observing the disease terrains and marking the corresponding cells. In the clustering results of genes, we have observed the four genes (Panel B, “BCL2” “HDAC1” “ERBB2” “EGFR”) that are far away to each other and have differing terrain shapes belong to four well separated clusters. In the clustering results of cancer terms, we have observed four cancers, namely ‘adenoma’ ‘melanoma’ ‘non-hodgkin lymphoma’ and ‘radiation-induced leukemia’, belong to four well separated clusters. After referring the four cancers in the core cancer network, we have found the corresponding nodes are spatially far away. So we have concluded from the results present in Figure 4.5, the major peaks in terrains, as represented as dark cells in the heatmap are well preserved features that could indicate how the nodes should be positioned among others.

The results show the correctness of the correlative multi-level terrain surfaces construction. The terrain surface shapes of the nodes in the networks are consistent with the proximities among the nodes. The insignificant peaks in the terrain surfaces are filtered out by human’s perception. Nevertheless, using the correlations which are preserved as the major landmark features, clustering both the cancers and genes simultaneously yields clusters that are consistent with the proximities of nodes in both networks. The results show that the correlative multi-

level terrain surfaces preserve the major signals in the correlations and the networks.

4.5 Correlative Multi-Level Terrain for Biomarker Discovery

In this section, we use a much larger high quality data set for constructing the correlative multi-level terrain surface. And we later show how it can assist the visual analytical tasks of biomarker discovery and performance analysis.

4.5.1 Protein Terrain for Candidate Biomarker Protein-Protein Interactions Network

The base networks of all molecular network terrains are constructed from candidate cancer biomarker protein-protein interaction networks. We refer to this type of molecular network terrain as *cancer biomarker protein terrain.* In this study, we take candidate cancer biomarker proteins from a literature-curated collection of 1,049 cancer candidate biomarkers [154], which primarily consist of differentially expressed proteins or genes in cancer. The sources of human protein-protein interaction data were collected from the Human Annotated and Predicted Protein Interaction database (HAPPI) described in [155], which is a comprehensive compilation of experimental and computationally-predicted human protein interactions primarily from the STRING [156] and OPHID [157] databases. The reliability of protein-protein interaction information in HAPPI is quantified using *H score*s ranging from 0 to 1 or a quality star rank grade of 1, 2, 3, 4, or 5. Increased protein interaction grades from 1 to 5 have been shown to be associated with improved quality of physical interacting proteins and a decreased amount of non-physical interactions found primarily in text mining or gene co-expression studies [155].Protein interactions in the HAPPI database with a star grade of 3 are comparable to the overall quality of HPRD [158] and consist of mostly physical protein interactions. We use HAPPI instead of HPRD because

of its coverage of more than 280,000 human protein interactions with star grade of 3 and above, compared favorably with a count of <40,000for HPRD. Of the 1,049 cancer candidate biomarkers, 762 can be matched with UniProt accession numbers in the HAPPI database. We refer to the HAPPI-*n* base network as one generated by building a protein-protein interaction network involving only these candidate biomarker proteins that are connected to HAPPI protein interactions by a quality grade of *n* and above.

4.5.2 Disease Terrain for Major Cancer Disease Associations and Base Network Constructions

We built two classes of base networks for the phenotypic network terrains. The first class, *CNG*, was built from disease-gene associations reported in the OMIM database. Therefore we used *cancer association terrain* to refer to this specialized type of phenotypic association network terrain. The CNG network was built by connecting a pair of cancer types if they shared at least one gene reported by the OMIM, similar to the method reported in [159]. In CNG, we kept only 98 different cancer subclasses from all 1,284 disease subclasses defined in the work of Goh et al. [159], and we narrowed it down to 60 major cancer categories for this study. We further classified CNG into CNG-I and CNG-II, based on the minimum number of shared cancer genes in the OMIM for the CNG. Therefore, CNG-I is the same as the original CNG, sharing minimally one gene in common between any two cancers, whereas CNG-II is a more stringent version of CNG, sharing at least two genes in common with any two cancers. CNG-I contains 39 major cancer nodes in its largest connected sub-network, whereas CNG-II contains 16 major cancer nodes in its largest connected sub-network.

The second class of base network, CNL, was built from disease-gene term co-occurrence reported in the literature. The edge score between the two terms,was

calculated according to [160] shown below:

$$f(v_a, v_b) = \ln\left(df_{v_a,v_b} * N + \lambda\right) - \ln(df_{v_a} * df_{v_b} + \lambda) \quad (4.2)$$

where df_{v_a} or df_{v_b} , is the number of documents in which term v_a or term v_b occurred, and df_{v_a,v_b} is the number of documents in which v_a and v_b co-occurred in the same document. *N* is the number of documents in all PubMed abstracts. *λ* is a small constant (λ=1 here) introduced to avoid out-of-bound errors. There is no edge for v_a and v_b if the edge score is not considered, which means any of df_{v_a}, df_{v_b}, or df_{v_a,v_b}values is 0. The resulting is positive when the co-occurrences of the pair of terms are over-represented and negative when under-represented. In this method, each cancer-cancer association edge in CNL also carries a normalized positive score, $conf$, to indicate the strength of the disease association relationships. Similar to the classification of CNG, we also classified CNL into CNL-I and CNL-II, to indicate their different qualities. CNL-I contains CNL sharing two diseases with a minimal strength $conf$ of 1.0, whereas CNL-II contains CNL sharing two diseases with a minimal strength of 2.0. Of the 60 major cancers, 56 are preserved in both CNL-I and CNL-II. In both types of base networks, CNG and CNL, we define node weight function to measure the node's connectivity based on the scores $conf$ of its edges [161].

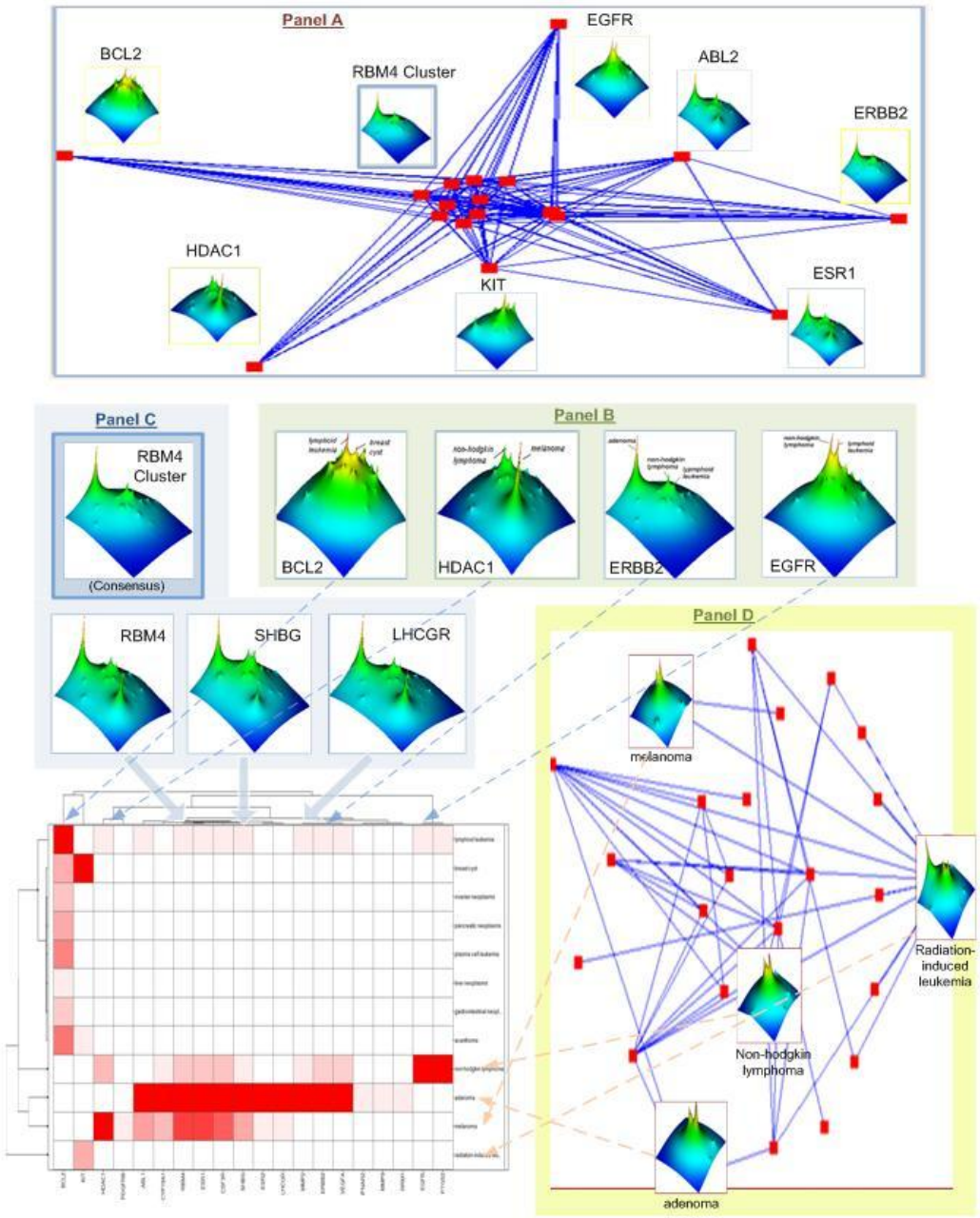

Figure 4.5 Panel A are gene terrains arranged on a core gene network; Panel B are detailed view of thumbnails in Panel A; Panel C are enlarged local regions of panel A. Panel D are terrains of major cancer terms which are identified by observing gene terrains in Panel A.

4.5.3 Correlative Protein Terrain and Disease Terrain

The response variable of both molecular and phenotypical network terrains in this experiment is either a protein-to-disease association strength score or a disease-to-protein association strength score. We used the reported functions between genes and diseases in the GeneRif database [162] to generate the disease-gene association matrix. A strength score was recorded in the association matrix between two associated terms—a disease was represented using its MeSH term and a gene (with all gene or protein synonyms)—regardless of the direction of the associations identified. The proteins were taken from the 762 HAPPI-overlapped cancer candidate biomarkers, whereas the diseases were taken from the 56 major cancers in CNL. For each cancer-protein association, we calculated its association strength using the same equation in 1.1. We normalized all the association strength scores between a pair of cancer and candidate protein biomarkers by dividing the original association strength score with the average of all association scores for the cancer involved in the normalization. This ensured a fair comparison of response values across both popular and rare cancer types for our study.

There are two major criteria of candidate biomarkers: *disease sensitivity* is how strong the protein(s) are correlated to certain diseases; and *disease specificity* is how specific the protein(s) are related to certain diseases. In the following sections, Protein Terrain Surface enables the visual analytical task of sensitivity evaluation, and Disease Terrain Surface enables the visual analytical task of specificity evaluation.

4.5.4 Candidate Biomarker Sensitivity Evaluation with Protein Terrain Surface

In Figure 4.6, we show 3x4 protein terrains developed for breast cancer, ovarian cancer, and lung cancer and varied among four types of protein interaction base

networks. We can make four interesting observations from the protein terrains shown.

First, we can identify well known genetic markers for these cancers, by following any column (fixed protein interaction base network quality), e.g., for “HAPPI-5” base network, and relate major peaks to regions of gene cluster regions highly associated to any of the three cancers. Here, the heights of major peaks suggest the sensitivity performance of a candidate biomarker: the higher the peak rises above the surface, the more sensitive the candidate protein biomarker is. For breast cancer, BRCA1_HUMAN (Breast cancer 1), BRCA2_HUMAN (Breast cancer 2), ESR1_HUMAN (estrogen receptor 1), and ERBB2_HUMAN (Human Epidermal growth factor receptor 2, HER2) are four major characteristic peaks. For ovarian cancer, the same set of four proteins still dominates the protein terrain landscape. For lung cancer, EGFR_HUMAN (Epidermal growth factor receptor 1), RASK_HUMAN (KRas proto-oncogene protein), GSTM1_HUMAN (Glutathione S-transferase Mu 1) are four characteristic peaks. Abundant literature studies can be found to confirm that BRCA1, BRCA2, HER2, and ESR1, among other genes, are major genetic markers and risk factors for breast cancer and ovarian cancer [163-165] [166]. Defects in EGFR, RASK, and GSTM1 are also strongly associated with lung cancer [167-170].

Second, major landscapes and peaks from these dominant genetic cancer markers do not appear to be affected by different base network layouts, developed from protein interaction data of varying qualities. This can be confirmed by comparing gene terrains across different columns for the same cancer type in Figure 4.6. However, subtle patterns of landscape differences on smaller peaks do exist. This could be attributed to the fact that the base network layout for higher quality cancer biomarker protein interactions contains fewer

proteins (727 for HAPPI-2, 717 for HAPPI-3, 679 for HAPPI-4, and 562 for HAPPI-5) and protein interaction clusters on the protein terrain. During the surface interpolation step to generate protein terrains, regions filled with proteins with higher node weights (due to higher degree of interaction connections) could lead to higher peaks. Therefore, more details of small peaks can be observed for the breast cancer protein terrain series generated with lower interaction data qualities, while higher peak levels can be observed for the ovarian cancer protein terrain series generated with lower interaction qualities as well.

Third, the relative distances and topological relationships of major peaks also seem to be stable, resistant to variations of interaction data quality of the base networks. For example, the BRCA1_HUMAN and BRCA2_HUMAN peaks are consistently closely clustered together. This type of clustering is not found for any of the other protein peaks, including ESR1_HUMAN or ERBB2_HUMAN, in breast and ovarian cancers.

Fourth, diseases that are similar to each other share more similar protein terrain landscapes than diseases that are different. Compare the protein terrains between two female cancers, breast and ovarian cancers, and a female cancer against lung cancer within the same column. It is apparent that protein terrains for breast and ovarian cancers not only share similar genetic markers but also similar protein terrain landscapes. It was not the case for breast cancer and lung cancer. Although our observations suggest that the choice of base network does not significantly impact the finding of major biomarker peaks, we still decide to use HAPPI-3 base network for the remainder of the work to strike good balance between high protein coverage and reliable protein-protein interactions (recall that HAPPI-3 consists of protein interactions of 3-stars and above, which are

high-quality physical interactions comparable to the overall quality of the HPRD manually curated database of human protein interactions).

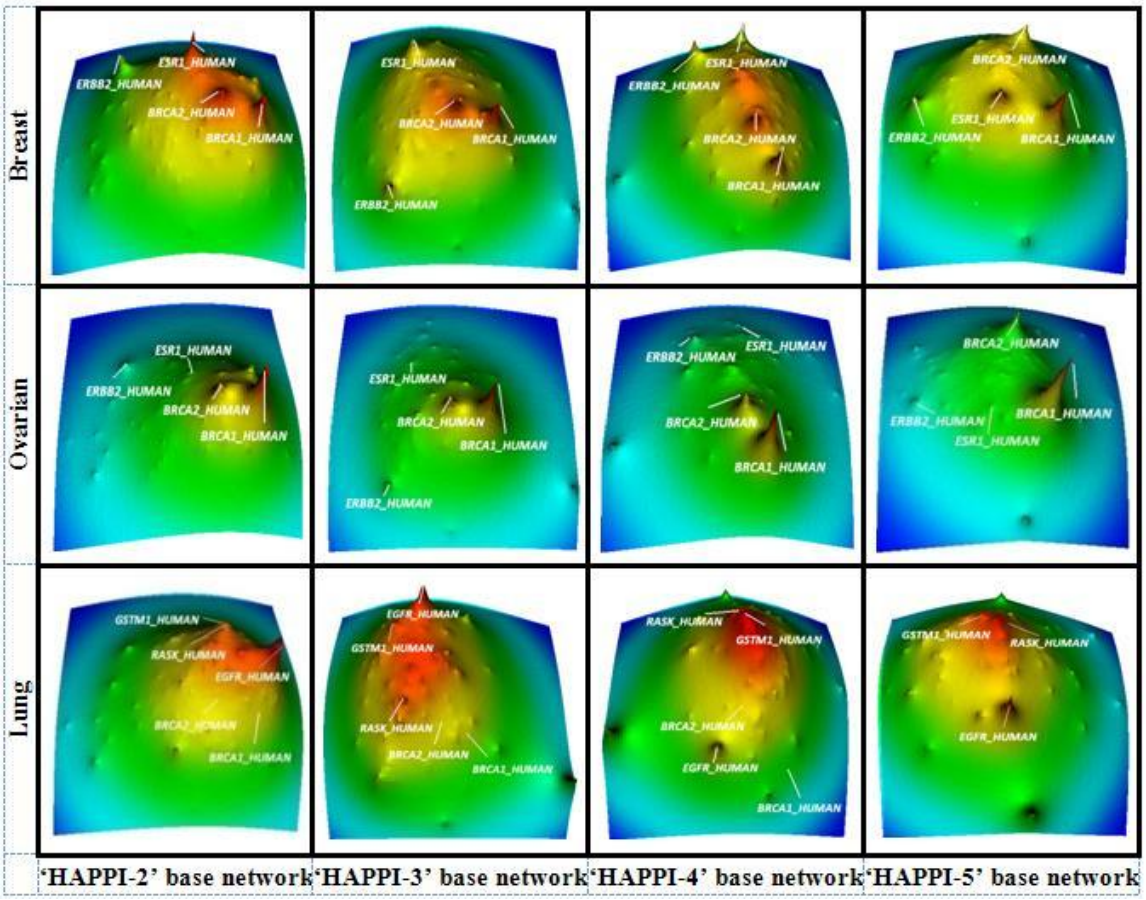

Figure 4.6 Major peaks on the 3x4 molecular network terrains are consistently identified as known sensitive cancer genetic markers.

4.5.5 Candidate Biomarker Specificity Evaluations with Disease Terrain Surface Visualization

In Figure 4.7, we show 4 disease terrains developed for four cancer biomarkers well-documented in the literature to examine their disease biomarker specificity. All these disease terrains have the same base network, the cancer disease association network (type CNL II), which is derived from a method described in

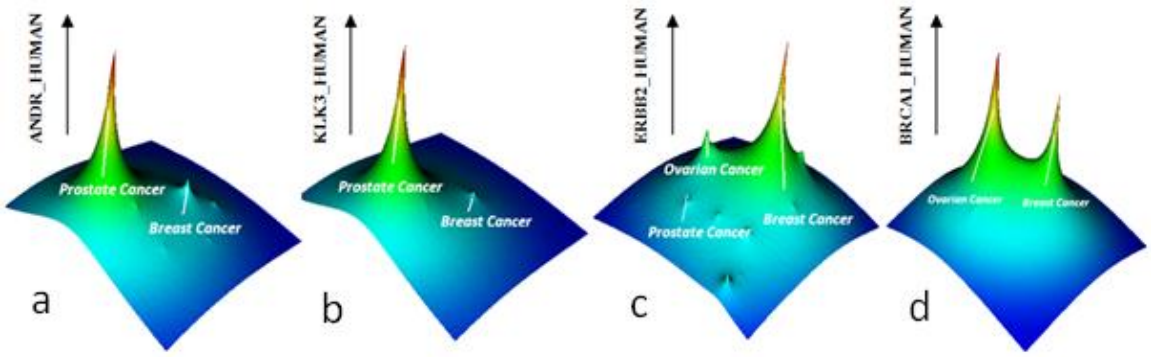

Figure 4.7 Major peaks on 4 phenotypic network terrains show different cancer disease specificity for each of the four tested candidate biomarker proteins.

the Method section. Note that we have made similar experimentations as we have done for protein terrains by altering disease base networks to make the choice of an overall good CNL II base network. (Results not shown due to space limits)

By comparing 5.4a and 5.4b, we can make three observations. First, both ANDR_HUMAN (Androgen Receptor) and KLK3_HUMAN (Prostate specific antigen, PSA) are “decent” candidate biomarkers for prostate cancer, because the peaks for prostate cancer in the two disease terrains—suggesting the sensitivity performance of these two protein biomarkers for prostate cancers—are both much higher than other peaks (e.g., breast cancer as the second most visible peak). Second, since the disease terrain surface for candidate biomarker PSA is cleaner than ANDR and the second most visible peak for breast cancer is much smaller, we can hypothesize that PSA is perhaps a better single biomarker for prostate cancer. Third, since the disease terrains between PSA and ANDR are similar, a panel biomarker by simple aggregating these two proteins in a same assay perhaps would not be a good idea. Many literature reports have covered the performance of these two biomarkers in prostate cancer [171].

Similarly, we compared candidate biomarker proteins ERBB2_HUMAN (HER2) and BRCA1_HUMAN for the specific detection of ovarian cancer in Figure 4.7c and 3d. The results are consistent with literature knowledge that HER2 is broadly associated with many types of cancers while BRCA1 is strongly associated with female cancers more specifically. Neither of the two proteins, therefore, should be used for general-purpose cancer subtyping applications. With better specificity than HER2, however, BRCA1 could be arguably developed for distinguishing female cancers from other cancer types.

4.6 Conclusions

In this chapter, we present the correlative multi-level terrain surfaces approach, to visualize multiple correlative networks. And we explain the approach with biology networks. There are three critical components in this approach: Molecular Network Terrain, built with a molecular interaction network as the base network; Phenotypic Network Terrain, built with a phenotype association network as the base network; the response variable of both terrains are the numeric correlation derived from literature mining or other measurements. Using a small pilot data set, we visually show that the design is correct and consistent with the network topology. We have also used automatic clustering methods to verify that the visual approach preserves the major signals in the correlations and in the networks. Then we use a much larger and high quality data set to show how the approach can assist users to perform visual analytical tasks of biomarker discovery and biomarker performance assessment The significances of this approach are many folds: first it visually encodes and prioritizes the correlations among nodes in correlative multiple networks. Using a pair of correlative terrain surfaces, it offers an intuitive overview of the patterns hidden in the network and in their correlations. Second, prominent visual patterns boost the major signals of correlations therefore the most relevant are captured by users' perceptions. Third, after we apply the approach to a pair of correlative cancer candidate biomarker

interaction network and cancer association network, we are able to identify stable biomarkers given the noise inherent in the biology networks. Based on the visual patterns on the terrain surfaces, we are able to visually evaluate and compare the performance of identified biomarkers. Through quick investigations on the patterns over the terrain surface, we can develop more insightful hypotheses for the phenotype-molecule correlations.

The correlative networks to be studied are not limited to biomolecular networks and phenotypic networks. They can be other types of biology networks, or networks in other domain. The specific context of the networks is depending on the applications. The functional hypotheses generated will also change accordingly.

We believe with appropriate adjustments, this approach can be generalized to to study networks and their correlations in many other domains.

CHAPTER 5 ITERATIVE VISUAL REFINEMENT MODEL

5.1 How to Improve the Hypotheses from the Complex Networks

As we can see in section 4.4.4, using correlative multi-level terrain surfaces, biologists can investigate the rich information of the hundreds of molecule-phenotype correlations. Based on their observations of the landscape features, biologists can form disease-protein functional hypotheses. However, being able to observe the phenomenon, biologists usually ask more complex questions, which can only be solved by interactively manipulating the data sets and changing the visual profiles. For example, in Figure 4.7, we have learnt that a highly sensitive biomarker for one cancer does not necessarily have high specificity. In fact, the lack of specificity for many disease biomarkers is the ultimate challenge for biomarker development today. Biologists are generally interested in the question: can panel biomarker improve the performance, especially for achieving disease specificity, at all?

In this chapter, we have invented the four step iterative visual refinement model, a visual analytical approach that helps biologists improve their disease-protein functional hypothesis. In the model, users can interactively reason and manage their findings with the visual patterns on correlative multi-level terrain surfaces and with their prior knowledge. The model supports users to manipulate the graphics in a process that iterates through three steps, namely *construction step*, *filtering step* and *evaluation step*, and stops at the *rendering step*. This approach is unique in its iterative nature and is innovative in the following aspects:

1) The iterative refinement model treats users' perceptions as the objective function, and guides the users to the final formation of the optimal hypothesis by visual patterns. The process is intuitive but also set a clear benchmark for users to estimate the progress of reasoning cycles.
2) The changing visual patterns observed from the terrain surfaces represent intermediately formed hypotheses. So the patterns serve as a form of reasoning artifacts which can record users' temporary findings and can enable visual comparison among findings. The ultimately satisfactory visual pattern can be delivered as the representation of the optimal hypothesis.
3) The iterative refinement model ensures the otherwise dynamic users interactions leading to the final discoveries. The model maps the elimination heuristics human use for problem solving into the concrete four steps and uses the correlative terrain surfaces to support such a process.

After applying the iterative visual refinement model to the correlative multi-level terrain we build from the biology networks in Chapter 4, we achieved a biomarker panel for lymphoma cancer with surprisingly high sensitivity and specificity. The panel's performance is validated using microarray samples from separate studies. The discoveries are significant in biology domain because the panel has not been reported. The forward sections are organized as follows. The next section first summarizes the workflow of using iterative visual refinement model for discovering functional knowledge between phenotypes and molecules. Then in section 5.3, we use a specific cancer, lymphoma, as an example and follow the iterations in the model to derive a highly performed biomarker panel. Section 5.4 we compare the classification performances on microarray samples, using our discovered panel with bench biomarkers. The results validate the competitiveness of the panel we derived from the iterative model.

5.2 Iterative Visual Refinement Model Workflow

Figure 5.1a shows the **iterative refinement model** loops in three steps, namely *construction step*, *filtering step* and *evaluation step*, and stops at the *rendering step*. Construction step enables users to construct protein terrain for selected disease, and the filtering step preserves proteins of major peaks and other interesting regions then removes others. In the evaluation step, the consensus disease terrain of the preserved proteins are rendered and visually evaluated. When there are high and obvious “noisy” peaks, the refinement process thus needs to go back to the filtering step, to compare the protein terrain surfaces of the targeted disease and the disease represented by the ‘noisy’ peak and further removes proteins that are sensitive to both diseases. The improved group of biomarkers therefore would result in the refined distinctive peak in evaluation steps. The process continues to the point that the number of biomarker is manageable and the peaking pattern is considered optimal. Figure 5.1b is an optional step for biomarker performance variance checking, the color intensity here maps the variance of the association scores between biomarkers in the panel and phenotypes. Figure 5.1c shows the the achieved candidate panel with satisfactory performance: high sensitivity indicated by the visual pattern of the molecular network terrain, and high specificity indicated by the phenotypic network terrain. Section 5.3 shows a working example of the process.

5.3 Iterative Visual Refinement for Biomarker Discovery

Here, we use *lymphoma* as a case study, as our visual analytic analysis of several known single protein markers for lymphoma on disease terrain shows major peaks on *leukemia* and *lymphoma*. This discovery is consistent with the fact that several subtypes of late-stage lymphoma are known to be clinically co-occurring with *leukemia*. Now we use the iterative refinement process to improve the biomarkers’ specificity for lymphoma by identifying a group of proteins that collectively contribute to a high specificity.

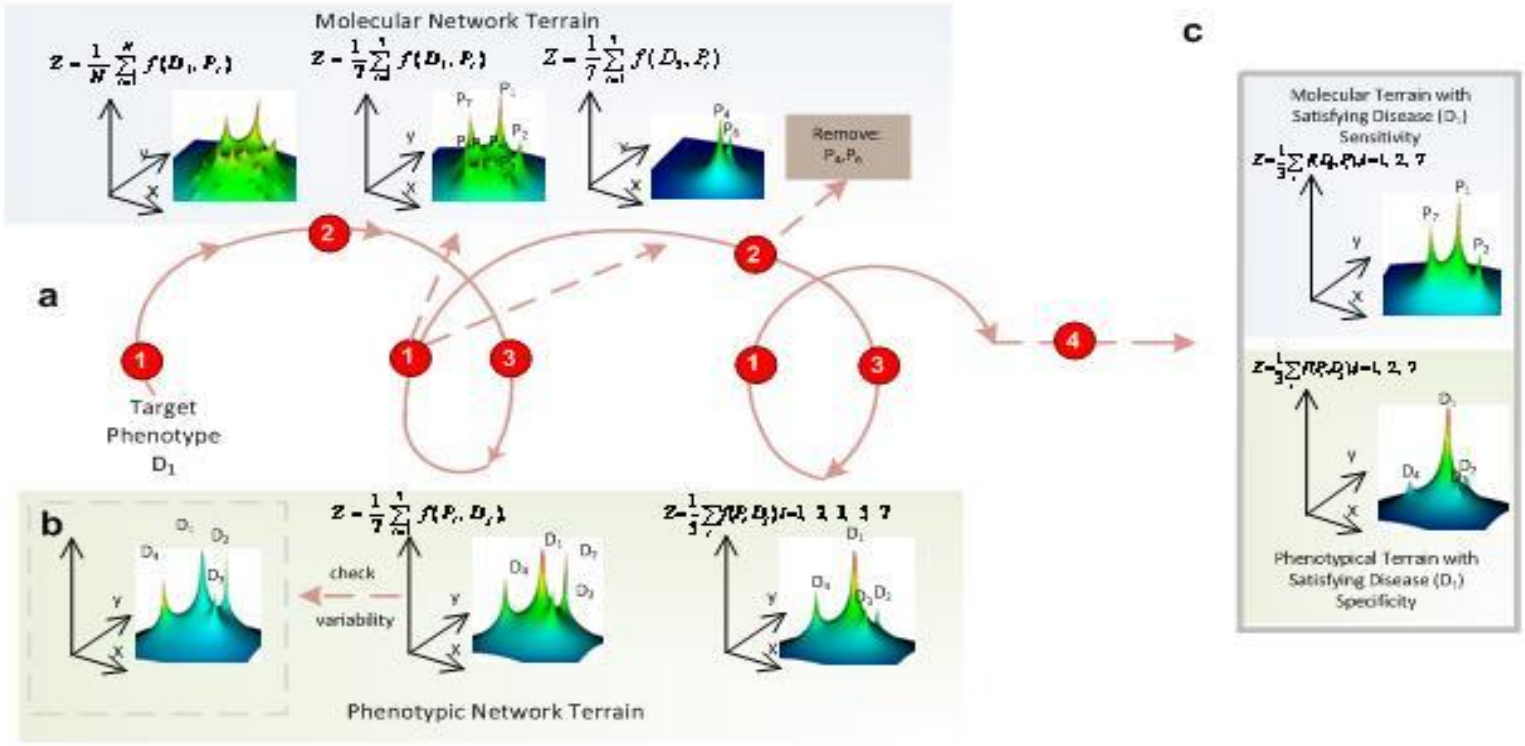

Figure 5.1 The four-step iterative refinement process of biomarker panel development using terrain visualization panels: for phenotype D_1, achieve a high-quality molecular biomarker panel with satisfying disease sensitivity and specificity using: (a) the four-step process: 1. constructing, 2. filtering, 3. evaluating, 4. rendering; (b) an optional variability check step of the current molecular biomarker panel; (c) the achieved candidate panel with satisfactory performance an optional variability check step of the current molecular biomarker panel; (d) the achieved candidate panel with satisfactory performance.

The process begins with the initial *construction step*. We build a lymphoma protein terrain by choosing the HAPPI-3 base network (see Figure 5.2a). Among all the candidate cancer biomarkers used for this study, 169 curated lymphoma candidate biomarkers are covered.

In the second *filtering step*, we zoom in to two regions, A and B, as labeled in Figure 5.2a. Region A contains major clustered peaks characteristic of the entire lymphoma protein terrain, while Region B is a peripheral area of Region A with extended surface slopes and small "buds". Altogether, regions A and B contain 31 of 169 curated lymphoma candidate biomarkers. In this study, we choose to focus on candidate protein markers within these two regions only and use them to build the initial panel (shown in the list of proteins preceding Figure 5.2b).

In the third *evaluation step***,** we evaluate the lymphoma disease specificity of an identified cluster of proteins filtered from the previous step. The difference practice here compared to evaluating single protein biomarker is that we must render a *consensus* disease terrain for all filtered proteins in a panel. In the consensus disease terrain shown in Figure 5.2b, we used the same base disease association network (type CNL II). This consensus disease terrain contains two dominating peaks, one for lymphoma and the other for leukemia.

As we intend to filter out more proteins to improve the uniqueness of the lymphoma peak pattern, it is usually necessary to go back to earlier steps to pick other "noisy" regions and remove proteins in those regions iteratively. We show contours of the two protein terrains for clear natural regions of proteins, one for lymphoma (Figure 5.2d) and the other for leukemia (Figure 5.2e), during iterative refinements. In both contours, we identify a common peak region, Region C,

prioritizing a common group proteins (compare example annotations in Region C in Figure 5.2d and 5.2e), thus infer those proteins in Region C would not be distinguishable between the two cancer types. As a result, we further filtered the identified 20 out of 31 curated candidate proteins as located in Region C. In Region D, we keep proteins TNR8_HUMAN (Lymphocyte activation antigen, CD30) and BCL6_HUMAN (B-Cell Lymphoma 6, BCL6) because they show peaks only in lymphoma protein terrain contour but not in leukemia protein terrain contour. Additional evaluations of what other proteins to keep in Region D are done manually by continuing the iteration of evaluating the quality of distinctive lymphoma peak in protein's disease terrain. Two more proteins, PIM1_HUMAN (Proto-oncogene serine/threonine-protein kinase, PIM-1) and FSCN1_HUMAN (Fascin, p55), are found to be able to improve the quality of lymphoma peaks and added to the biomarker panel for lymphoma. (Note that corresponding Gene Symbols and description of proteins are in the parentheses).

In the fourth and final *rendering step*, we build a consensus disease terrain for the completed biomarker panel of the 4 proteins (see Figure 5.2c). Comparing Figure 5.2b (before further filtering) and 5.2c (after filtering), we show that dramatically improved lymphoma disease specificity has been accomplished by us. This new biomarker panel consists of manageable number of proteins, with both high sensitivity (high peak) and high specificity (unique peak).

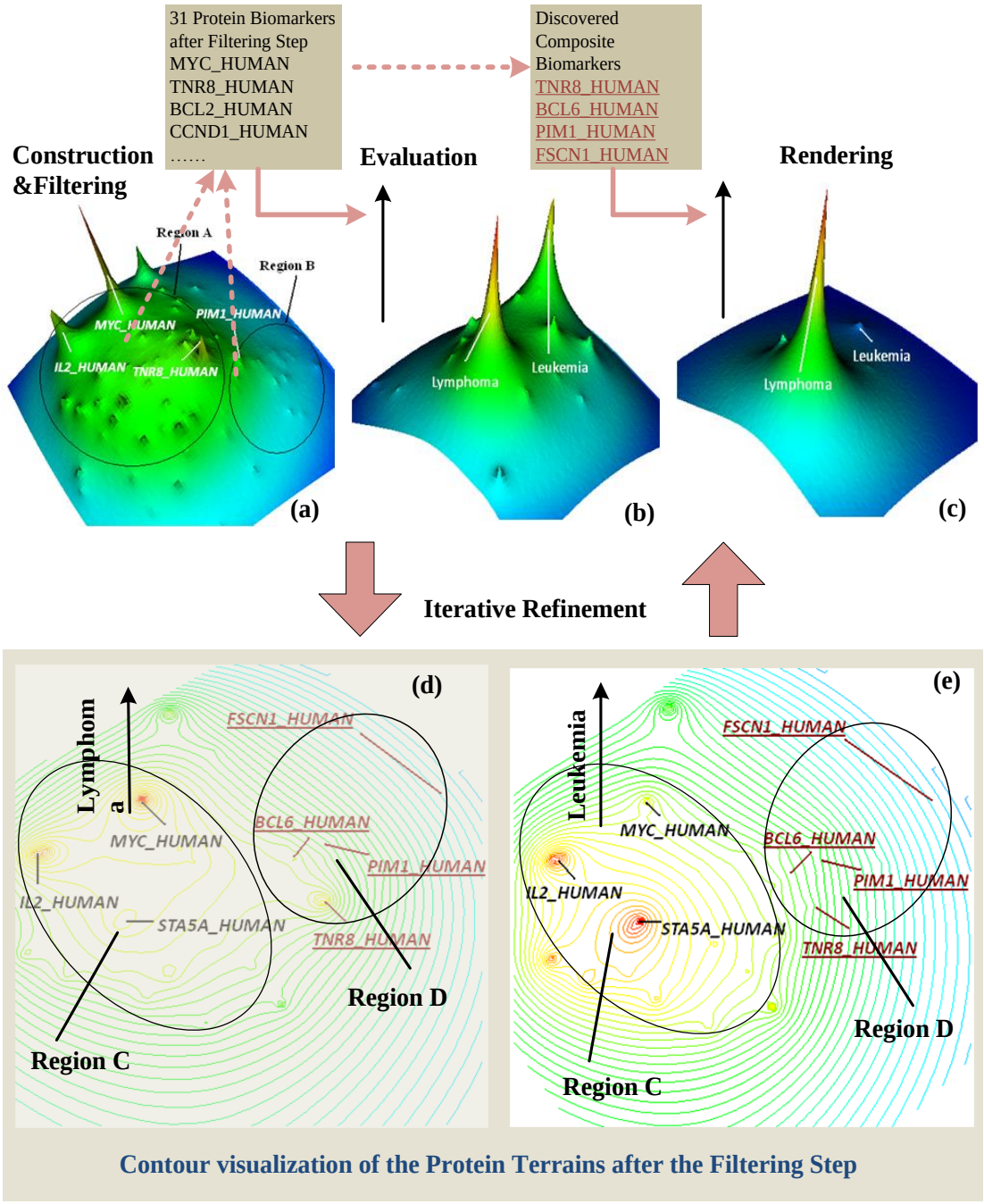

Figure 5.2 Development of the biomarker panel for diagnosing lymphoma to achieve high sensitivity and specificity.

5.4 Validation of the Lymphoma Biomarker Panel

We validated the performance of the newly found biomarker panels by measuring their disease sensitivity and disease specificity. The disease sensitivity was defined by the bi-classification results on microarray expression samples, where the case was the lymphoma samples and the control was the normal samples; the disease specificity was defined by the results of the bi-class classification, where the case was the leukemia samples and the control was the lymphoma samples. In the following sections we first introduce the microarray expression data sets and the normalization we used in the validation process, and then we describe the model, the parameters, and the results of the bi-class classification.

5.4.1 Microarray Expression Data Sets

All of the microarray expression samples in our validation are from a recent, high-quality, comprehensive study [172]. From the study, 25 normal samples, 29 lymphoblastoid lymphoma cell line tissue samples, and 34 B-Cell chronicle lymphocytic leukemia cell lines were normalized and then used for classification. Normal samples and lymphoblastoid lymphoma cell line tissue samples were used for assessing the biomarker disease sensitivity; the lymphoma cell line tissue samples and the B-Cell chronicle lymphocytic leukemia cell lines were used for assessing the biomarker disease specificity.

5.4.2 Microarray Expression Normalization

We aligned the total of 88 samples, of which each had 12,533 probes, and we normalized them with the expressions of the identified “housekeeping” probes. There are two steps in performing “housekeeping” probe normalization:

Step 1. Quantile Normalization Check

First we checked to see if the data set needed any routine normalization, e.g. quantile normalization. For each of the 88 samples, we excluded the top 5 percentile and the bottom 5 percentile of the expressed probes, and we

calculated the mean of the expressions for the remaining probes. Then we found that the standard deviation of the mean values from all samples was small enough (8.22 in this study). We repeated the normalization checks by temporarily removing the top and bottom 10 percentile, and then removing the top and bottom 25 percentile. In each check the standard deviation among the mean values was acceptably small, so we considered this data set already quantile normalized.

Step 2. “Housekeeping” Probes-Based Normalization

In this step, we first identified “housekeeping” probes, which are probes with relatively more stable expressions across all the samples. We distinguished between “housekeeping” probes and probes that barely function because the expressions of the barely-functioning probes are low and not reliable due to the unavoidable artifacts introduced by the chips. To identify the housekeeping probes, we examined the P/M/A calls for probe expressions in all samples, and used the maximum expression marked with an absence call as the minimal threshold, T (T is 41.4 in this study), for presence and absence. We then temporarily removed probes that had intensity values dropping below the threshold, a minimum of 5% of all the samples used (5% assumes that some samples may be outliers). In this study, 4,912 out of 12,533 probes remained. For the remaining probes, the bottom 100 probes with least variance across all samples were the “housekeeping” probes. We denoted the average of the housekeeping probe expressions as IO (in this study, IO is 98.23) for the baseline.We then used the baseline from the house-keeping probes as the “internal standard” to normalize each expression: each new expression value was the relative fold change with regard to the standard, i.e. the normalized expression IX’ is max (0, (IX-T)/(IO-T)). Note that expressions lower than the baseline were set to zero.

5.4.3 Bi-class Classification Model for Validating Biomarker Performance

We consolidated the probe expressions into gene expressions and then mapped those genes to obtain the expressions for the 169 known lymphoma biomarkers. Those biomarkers were from the 762 candidate biomarkers used for constructing the candidate biomarker protein interaction network. When multiple gene symbols are mapped to the same protein Uniprot ID, a simple linear average is used to calculate the expression of the protein. As a result, 156 out of the 169 lymphoma biomarkers survived. Among them, the four protein biomarkers in our newly detected panel, i.e. TNR8_HUMAN, BCL2_HUMAN, PIM1_HUMAN, FSCN1_HUMAN, survived as well.

We used the four marker expressions as a four-dimensional feature vector for each sample for classification. We also used the surviving 156 single biomarkers as bench markers. Then we used hierarchical clustering to cluster the feature vectors of the samples, in order to approximate the best possible bi-class classification results. In the hierarchical clustering, we used the “euclidean” default distance measure and the “mean” default linkage method. The results were compared to the known annotations, and the errors defined two performance criteria: disease sensitivity and specificity.

Disease sensitivity is characterized by two types of errors: a Type I error is the ratio between the lymphoma samples (in this study, lymphoblastoid lymphoma cell line tissue samples) classified as normal and the total number of lymphoma samples; a Type II error is the ratio between the number of normal samples misclassified as lymphoma and the total number of normal samples. The disease specificity is defined as the ratio between the lymphoma samples in the lymphoma-dominated class and the total number of samples in that class.

To compare the performance between the newly detected four markers and the other 156 bench markers, we marked the Type I and II errors and the disease specificity of the new panel on the empirical cumulative density function (CDF) from the bench markers. In CDF, the x value is the performance, e.g. a Type I error, and the y value is the portion of bench markers whose performance is less than x. In the case of both Type I and II errors, the lower the y value in our panel was, the more accurately classified the normal and lymphoma samples were; in the case of disease specificity, the higher the y value in our panel was, the more specific they were in distinguishing lymphoma conditions from leukemia conditions.

As a result, Figure 5.3a shows the Type I error of the panel's sensitivity is 0.0069, lower than 79% of the benchmark population; Figure 5.3b shows the Type II error of the panel's is 0.01, lower than 90% of the bench marker population. The panel's specificity against leukemia is surprisingly high, 0.9914, higher than 97% of the benchmark biomarkers.

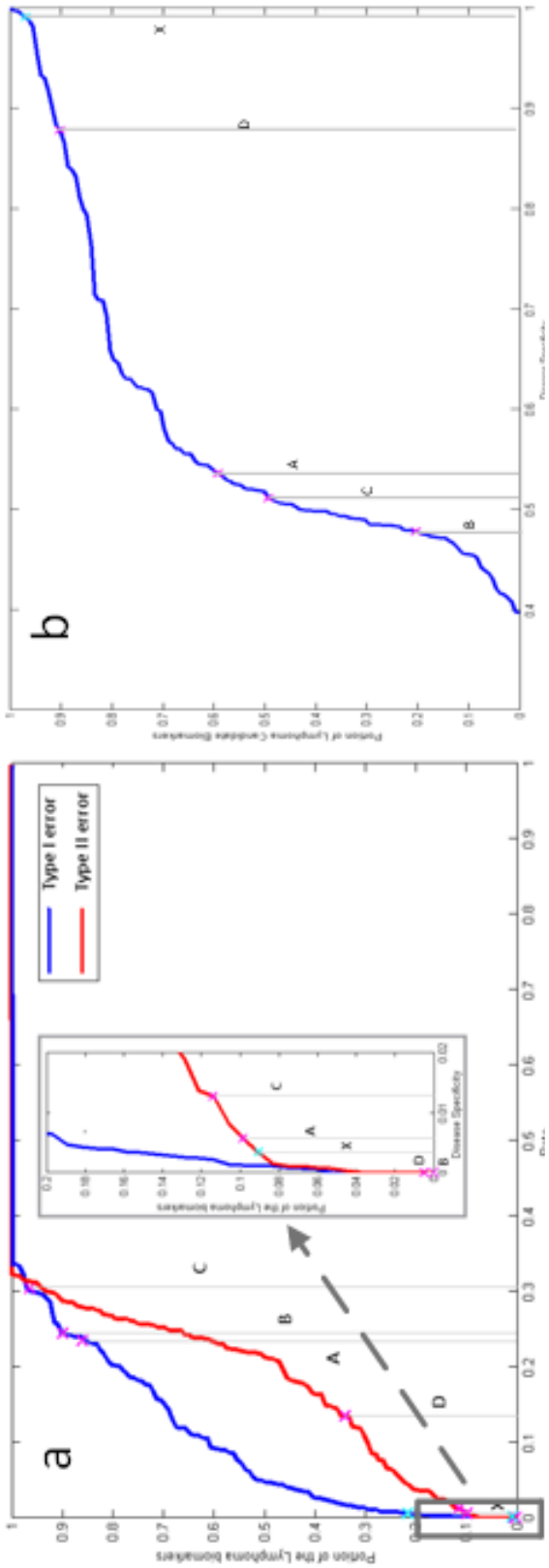

Figure 5.3 The prospective evaluation results of the new biomarkers panel's performance: (a) cumulative distribution plots (CDF) of Type I (blue) and Type II (red) error rate of disease sensitivity; (b) cumulative distribution plots (CDF) of disease specificity.

5.5 The Importance of the Interactive Iterative Visualization

Using the correlative multi-level terrain surface visualization in Chapter 4, users are able to observe the pre-attentive landmark features in the visualized correlations between multiple correlated biological networks. The visual features can reflect some of the most obvious characteristics of the correlations. However, given the initial insights, users would like to have more complex questions answered. It hence requires interactive visualization to support users to further carry out their reasoning cycle: form the hypotheses, manipulate the graphics for options that improves the hypothesis, and then restarts the cycle with the newly improved hypothesis. Therefore in this chapter, we propose a novel iterative refinement model for externalizing users' reasoning process. The model is innovative because of the following reasons: first users can use the observed changing visual patterns as a guide for improving hypothesis, and as an estimator for the progress. Second, the changing patterns are an external visualization presentation for the hypothesis and can be used for assessment and for further delivery and disseminations. Third, the four steps in the model essentially follows the principle of humans' elimination heuristics, which is known as the most frequently used and the most effective approach when human being are facing a complex problem. Therefore it can come to a reasonably good solution for a complex problem with relatively short time. We applied the approach to the correlative candidate biomarker interaction network and the cancer association network. Using automatic computing, obtaining the biomarkers for lymphoma cancer with the best sensitivity and specificity, is NP hard. However using the visual analytical approach we proposed, a new biomarker panel is achieved from heuristics by integrating the power of the visualization design and users reasoning. The discovered 4-protein biomarker panel has not yet reported, but has surprisingly good performance. It validates that our iterative refinement model can significantly benefit the visual analytics community as well as the biomarker development study of systems biology.

CHAPTER 6 DISCUSSIONS AND CONCLUSIONS

6.1 Design Effective Graph Visualization for Bioinformatics Applications

High-volume complex data drives the development of Visual Analytics. Although in the past decades, various information visualization techniques have been designed to visually represent the high-dimensional data or information, the increasing data volume and complexity poses a challenge to identify meaningful patterns from the data, a “needle in a haystack” scenario. Biology is one of the areas that have the urgent need to discover hidden knowledge from the vast amount of high-throughput experimental data or literature. Biologists have long been using the graph/network visualization to communicate the many types of different relationships, because “a picture is worth a thousand words”. However, when large collections of diverse relationships are generated from high-throughput experiments or from the biological systems, it is hard to make decisions on which aspects of information to be presented in the network. Had there been certain holistic visual representations, the users would still be likely to be overwhelmed by the richness of the visual information. And there would be no guarantee that the exposed patterns can help users develop their insights.

Our proposed iterative visual analytics (IVA) framework is in fact developed around the theme of exposing visual patterns that both are revealing interesting properties of the data set and are pre-attentive to users’ perceptions. And the most innovative part of IVA is that it substantiates a graphics-aided process for users to reason and distill their satisfactory visual patterns, when the initial visual

designs themselves are not sufficient for answering the question, i.e. biomarkers with both high sensitivity and specificity. Among many design decisions which are made during the development of the IVA framework, there are two most influential factors for the visual pattern formations: the base network layout algorithms, and the surface interpolation and color encoding scheme.

6.2 Design Decisions of the Base Network Layout

We design the base network layout with the criteria that it will contribute to preserving the proximity imposed by data item similarities, as well as to isolating the significant nodes from the clutters. Therefore we have both used the multi-dimensional scaling (MDS) and the proposed node-weight edge-weight force-directed model. Both of them in principle preserve the similarities of data item in high-dimensions. MDS is a generalized standard force-direct model which is a common choice of layout, when little semantics is known about the underlying network. As the scale of network increases (>500 nodes), the results of standard force-direct layout models become difficult to be interpreted. Therefore the node-weight edge-weight force direct model we proposed factors in the domain knowledge to isolate known hub nodes from other. We ensure this property by defining the two-phase layout and defining the “area of influence”. However, the final layout largely still depends on the initial positions, and the optimization process can result in varied final positions. Hence we suggest interactive adjustments to tune up the initial layout or automatic layout results. This way the layout is likely to yield more biological relevance when being interpolated as a surface.

6.3 Design Decisions of the Surface Visualization

The surface visualization we designed and used throughout the paper belongs to the widely used spatialization information visualization. There have been many discussions and sometime controversies around the 3D spatialization visualization [87, 88]. The first controversy is whether it is more advantageous

than the corresponding 2D (image) approach. Our design and applications of terrain surface visualization are for detecting the global visual pattern and localizing interesting regions. The visualization serves as a screening step for the visual reasoning process. For subsequent detailed information, we have developed multi-scale interactive visualization to provide on-demand details. A more in-depth user study, though, is necessary in the future development of our approach. The 2D approach needs to be compared with the 3D one based on different requirements from specific visual analytical tasks. The second issue of the debate issue is whether rainbow or grayscale is the optimal coloring scheme for the responsible variable. In our framework, we used rainbow color coding because there is evidence showing that it is better for quickly directing user to the regions of interest (e.g. identifying highly sensitive biomarkers) and for detecting the dynamics of visual patterns in a series of terrain surfaces (e.g. identifying progressive biomarkers and assessing the constant refining visual patterns). Rainbow color coding is also consistent with the red-blue variable encoding schemes extensively used in biology. Another issue is how different interpolation methods will affect the shapes of terrain surfaces and their visual patterns. Essentially all interpolation methods change a discrete distribution of network nodes into a continuous field. The reason we used interpolation in surface rendering is to expose the global patterns of the network and to prioritize certain regions. Not all interpolated values are meaningful with respect the network context. We have used the Shepard Interpolation methods, while other methods, e.g. radial basis function interpolation can be used as well. We expect different interpolation would result in slightly different surface profiles. However, further investigation and user studies need to be carried out to determine whether or not the differences will affect users' interpretations on the shapes and patterns.

6.4 Design Decisions for the Scalability

The terrain surface visualization for the properties of the base network is inherently scalable. It exposes the general patterns of the network property

regardless of dense edge crossings caused by the large scale. Besides this, we have addressed the scalability of our methods in two other places in our framework. First, in designing the correlative multi-level terrain surfaces, we break the complex large network, which consists of different types of nodes, into two smaller networks. It helps users to tackle the problem in the larger networks by visually analyzing and correlating information in the smaller ones. Second, based on the robustness of the terrain surface visualization, we show that the major visual patterns of the large network can be preserved, even when part of the weak or spurious edges are included and cause noise. In the visual assessment of biomarkers we presented in Chapter 4, we have compared the landscape features generated from differing underlying base networks. Each of the base networks is from the same protein-protein interaction data set, but includes edges with different confidence thresholds. The stableness of visual patterns on the terrain surface indicates the additions of weak interactions hence the scale of the large scale network does not affect users' interpretations in the visual analytic tasks.

6.5 Future Directions

Developing web-based tools or software suits for our IVA framework is beyond the scope of this work. However, it is nice to have them to maximize the immediate impact of this work. For example, the Gene Terrain application can either be expanded into biomarker discovery software tool or be a plug-in for existing biology network visualization software; the intractability of the Iterative Refinement Model can be improved. When being continued to use in the bioinformatics application, IVA needs more successful applications as the biomarker panel for lymphoma cancer. The discovered biomarker panel for lymphoma, needs further validation. Comparing with biomarkers selected by existing statistical test, such as t-test, is a necessary next step and is now undergoing.

The most important future improvement for the framework itself is to improve the base network layout. Now the final configuration of the network is largely depends on the initial layout. If grid layout and random layout is used in the initial layout, the layout after optimization would significantly different. Also sparse networks can cause unstable behavior of the layout. This is because the formula of the base network has one term quantifying the constraints on nodes, and the other term on edges. Unbalanced number of nodes and edges can cause the final layout be determined by solely one term. Therefore the model needs more investigation: first, adjustable parameters stabilizer terms can be added as well; second, the model can be more flexible depending on the semantics of the context.

The iterative refinement model needs to be improved to become a fully-fledged visual analytical framework. Statistical significance of the assessment on the correlation and the convergence of the discovery need to be generated from the exploratory process. It then can be further indicated by visual encoding to reveal the uncertainty of the results. However, we have to be cautious that the visual stimuli encoding doesn't pose additional perceptual complexity. The current model is essentially a process that externalizes a human problem solving heuristics. Other iterative schemes can be developed to prune the search space for the optimal solution. The effectiveness largely depends on what content to be presented to the users, what type of decisions they make and when. A further step is to investigate how the underlying data transformations can learn from users' preferences and interactions, in order to provide further suggestions in the discovery process.

We would also like to apply our framework for knowledge discovery tasks in other domains. For example, we have applied our framework in textual analysis where human knowledge plays a critical role in identifying concepts from the

unstructured text. The terrain surface visualization provides an overview of the term association network, and also highlights regions of interest as visual cues. The iterative refinement model can then enable users to select desired clusters of terms, in order to learn from their preferences. The success of the learning requires the visualization to expose patterns of textual features that are semantically meaningful. It also requires investigation of a learning model that can take users interaction as part of the input. Using our framework in unstructured text analysis is a bigger challenge because semantics in the text is usually hard to be described and distilled.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] J. Thomas and K. Cook, *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, IEEE Computer Society, 2005.

[2] C. Upson, T. Faulhaber Jr, D. Kamins, D. Laidlaw, D. Schlegel, J. Vroom, R. Gurwitz, and A. Van Dam, "The application visualization system: A computational environmentfor scientific visualization," *IEEE Computer Graphics and Applications*, vol. 9, pp. 30-42, 1989.

[3] S. Card, J. Mackinlay, and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann, 1999.

[4] P. Pirolli and S. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in *Proceedings of the International Conference on Intelligence Analysis*, 2005, pp. 2-4.

[5] D. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler, "Visual analytics: Scope and challenges," *Lecture Notes in Computer Science*, vol. 4404, pp. 76-90, Springer, 2008.

[6] W. A. Pike, "The science of interaction," *Information Visualization*, vol. 8, pp. 263-274, 2009.

[7] W. Ribarsky, B. Fisher, and W. Pottenger, "Science of analytical reasoning," *Information Visualization*, vol. 8, pp. 254-262, 2009.

[8] A. Inselberg, "The plane with parallel coordinates," *The Visual Computer*, vol. 1, pp. 69-91, 1985.

[9] I. Borg and P. Groenen, "Modern multidimensional scaling: Theory and applications," *Journal of Educational Measurement*, vol. 40, pp. 277-280, 2003.

[10] E. Kandogan, "Visualizing multi-dimensional clusters, trends, and outliers using star coordinates," in *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2001, pp. 107-116.

[11] P. Hoffman, G. Grinstein, and D. Pinkney, "Dimensional anchors: A graphic primitive for multidimensional multivariate information visualizations," in *Proceedings of Workshop on New Paradigms in Information Visualization and Manipulation*, 1999, pp. 9-16.

[12] B. Shneiderman, "Tree visualization with tree-maps: 2-d space-filling approach," *ACM Transactions on Graphics*, vol. 11, pp. 92-99, 1992.

[13] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, and M. Caligiuri, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531-537, 1999.

[14] S. Havre, E. Hetzler, P. Whitney, L. Nowell, B. P. N. Div, and W. A. Richland, "ThemeRiver: Visualizing thematic changes in large documentcollections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, pp. 9-20, 2002.

[15] D. Gotz and M. Zhou, "Characterizing users' visual analytic activity for insight provenance," in *Proceedings of theIEEE Symposium on Visual Analytics Science and Technology*, 2008, pp. 123-130.

[16] J. Stasko, C. Görg, and Z. Liu, "Jigsaw: Supporting investigative analysis through interactive visualization," *Information Visualization*, vol. 7, pp. 118-132, 2008.

[17] T. Jankun-Kelly, K. Ma, and M. Gertz, "A model and framework for visualization exploration," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, pp. 357-369, 2007.

[18] D. Holten, "Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data," *IEEE Transactions on Visualization and Computer Graphics*, pp. 741-748, 2006.

[19] Z. Shen, K. Ma, and T. Eliassi-Rad, "Visual analysis of large heterogeneous social networks by semantic and structural abstraction," *IEEE Transactions on Visualization and Computer Graphics* , pp. 1427-1439, 2006.

[20] T. Itoh, C. Muelder, K. Ma, and J. Sese, "A hybrid space-filling and force-directed layout method for visualizing multiple-category graphs," in *Proceedings of the IEEE Pacific Visualization Symposium*, 2009, pp. 121-128.

[21] Y. Jia, J. Hoberock, and M. Garland, "On the visualization of social and other scale-free Networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, pp. 1285-1292, 2008.

[22] P. Shannon, A. Markiel, O. Ozier, N. Baliga, J. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: A software environment for integrated models of biomolecular interaction networks," *Cold Spring Harbor Lab*, vol. 13, pp. 2498-2504, 2003.

[23] T. Huan, A. Sivachenko, S. Harrison, and J. Y. Chen, "ProteoLens: A visual analytic tool for multi-scale database-driven biological network data mining," *BMC Bioinformatics*, vol. 9, pp. 1-13, 2008.

[24] J. Y. Chen, S. Mamidipalli, and T. Huan, " HAPPI: An online database of comprehensive human annotated and predicted protein interactions," *BMC Genomics*, vol. 10, pp. S16, 2009.

[25] G. D. Bader, I. Donaldson, C. Wolting, B. F. F. Ouellette, T. Pawson, and C. W. V. Hogue, "BIND — The biomolecular interaction network database," *Nucleic Acids Research*, vol. 29, pp. 242-245, 2001.

[26] M. Schena, D. Shalon, R. Davis, and P. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, vol. 270, p. 467, 1995.

[27] T. Ideker, T. Galitski, and L. Hood, "A new approach to decoding life: Systems biology," *Annual Reviews in Genomics and Human Genetics*, vol. 2, pp. 343-372, 2001.

[28] P. Wong and J. Thomas, "Visual analytics: Building a vibrant and resilient national science," *Information Visualization*, vol. 8, pp. 302-308, 2009.

[29] C. Aragon, S. Poon, G. Aldering, R. Thomas, and R. Quimby, "Using visual analytics to maintain situation awareness in astrophysics," in *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, 2008, pp. 27-34.

[30] D. Tesone and J. Goodall, "Balancing interactive data management of massive data with situational awareness through smart aggregation," in *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, 2007, pp. 67-74.

[31] G. Andrienko and N. Andrienko, "Spatio-temporal aggregation for visual analysis of movements," in *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, 2008, pp. 51-58.

[32] F. Janoos, S. Singh, O. Irfanoglu, R. Machiraju, and R. Parent, "Activity analysis using spatio-temporal trajectory volumes in surveillance applications," in *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, 2007, pp. 3-10.

[33] R. Maciejewski, S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani, W. Cleveland, S. Grannis, M. Wade, and D. Ebert, "Understanding syndromic hotspots-a visual analytics approach," in *Proceedings of theIEEE Symposium on Visual Analytics Science and Technology*, 2008, pp. 35-42.

[34] P. Proulx, S. Tandon, A. Bodnar, D. Schroh, R. Harper, and W. Wright, "Avian flu case study with nSpace and Geotime," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2006, pp. 27-34.

[35] W. Wang and A. Lu, "Interactive wormhole detection in large scale wireless networks," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2006, pp. 99-106.

[36] D. Keim, F. Mansmann, J. Schneidewind, and T. Schreck, "Monitoring network traffic with radial traffic analyzer," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2006, pp. 123-128.

[37] C. Muelder, K. Ma, and T. Bartoletti, "A visualization methodology for characterization of network scans," in *Proceedings* of the *IEEE Workshop of Visualization for Computer Security*, 2005, pp. 29-38.

[38] G. Fink, P. Muessig, and C. North, "Visual correlation of host processes and network traffic," in *Proceedings of the IEEE Workshop of Visualization for Computer Security*, 2005, pp. 11-19.

[39] M. Bilgic, L. Licamele, L. Getoor, and B. Shneiderman, "D-dupe: An interactive tool for entity resolution in social networks," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2006, pp. 43-50.

[40] A. Pattath, B. Bue, Y. Jang, D. Ebert, X. Zhong, A. Ault, and E. Coyle, "Interactive visualization and analysis of network and sensor data on mobile devices," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2006, pp. 83-90.

[41] P. Wong, G. Chin, H. Foote, P. Mackey, and J. Thomas, "Have green: A visual analytics framework for large semantic graphs," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2006, pp. 67-74.

[42] T. von Landesberger, M. Görner, and T. Schreck, "Visual analysis of graphs with multiple connected components," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2009, pp. 152-168.

[43] M. Ghoniem, D. Luo, J. Yang, and W. Ribarsky, "NewsLab: Exploratory broadcast news video analysis," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2007, pp. 123-130.

[44] J. Yang, J. Fan, D. Hubball, Y. Gao, H. Luo, W. Ribarsky, and M. Ward, "Semantic image browser: Bridging information visualization with automated intelligent image analysis," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2006, pp. 191-198.

[45] C. Yu, Y. Zhong, T. Smith, I. Park, and W. Huang, "Visual mining of multimedia data for social and behavioral studies," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2008, pp. 155-162.

[46] P. J. Crossno, D. M. Dunlavy, and T. M. Shead, "LSAView: A tool for visual exploration of latent semantic modeling," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2009, pp. 83-90.

[47] D. Fisher, A. Hoff, G. Robertson, and M. Hurst, "Narratives: A visualization to track narrative events as they develop," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2008, pp. 115-122.

[48] D. Keim and D. Oelke, "Literature fingerprinting: A new method for visual literary analysis," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2007, pp. 115-122.

[49] D. Oelke, M. Hao, C. Rohrdantz, D. Keim, U. Dayal, L. Haug, and H. Janetzko, "Visual opinion analysis of customer feedback data," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2009, pp. 187-194.

[50] C. Collins, F. B. Viegas, M. Wattenberg, "Parallel tag clouds to explore and analyze faceted text corpora," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2009, pp. 91-98.

[51] R. Chang, M. Ghoniem, R. Kosara, W. Ribarsky, J. Yang, E. Suma, C. Ziemkiewicz, D. Kern, and A. Sudjianto, "WireVis: Visualization of categorical, time-varying data from financial transactions," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2007, pp. 155-162.

[52] S. Rudolph, A. Savikhin, and D. Ebert, "FinVis: Applied visual analytics for personal financial planning," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2009, pp. 195-202.

[53] J. Dietzsch, J. Heinrich, K. Nieselt, and D. Bartz, "SpRay: A visual analytics approach for gene expression data," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2009, pp. 179-186.

[54] M. Meyer, T. Munzner, and H. Pfister, "MizBee: A multiscale synteny browser," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2009, pp. 897-904.

[55] C. B. Nielsen, S. D. Jackman, I. Birol, and S. J. M. Jones, "ABySS-explorer: Visualizing genome sequence assemblies," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2009, pp. 881-888.

[56] M. Valle and A. R. Oganov, "Crystal structures classifier for an evolutionary algorithm structure prediction," in *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, 2008, pp. 11-18.

[57] T. Fruchterman and E. Reingold, "Graph drawing by force-directed placement," *Software Practice and Experience*, vol. 21, pp. 1129-1164, 1991.

[58] T. Kamada and S. Kawai, "An algorithm for drawing general undirected graphs," *Information Processing Letters*, vol. 31, pp. 7-15, 1989.

[59] A. Noack, "Energy-based clustering of graphs with nonuniform degrees," *Lecture Notes of Computer Science*, vol. 3843, pp. 309-320, 2006.

[60] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C*, Cambridge University Press, 1992.

[61] H. Abdi, D. Valentin, A. J. O'Toole, and B. Edelman, "Distatis: The analysis of multiple distance matrices," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2005, pp. 42-47.

[62] E. Gansner, Y. Koren, and S. North, "Graph drawing by stress majorization," in *Proceedings of the 12th International Symposium of Graph Drawing*, *2004*, *Lecture Notes in Computer Science*, vol. 3383, pp. 239-250.

[63] J. Kruskal and M. Wish, *Multidimensional Scaling*, Sage Publications, Inc, 1978.

[64] I. MathWorks, *MATLAB: The Language of Technical Computing*, *Desktop Tools and Development Environment*, *Version 7*, Mathworks, 2005.

[65] R. Ihaka and R. Gentleman, "R: A language for data analysis and graphics," *Journal of Computational and Graphical Statistics*, vol. 5, pp. 299-314, 1996.

[66] E. Gansner, Y. Koren, and S. North, "Topological fisheye views for visualizing large graphs," *IEEE Transactions on Visualization and Computer Graphics*, vol. 11, pp. 457-468, 2005.

[67] Y. Tian, R. Hankins, and J. Patel, "Efficient aggregation for graph summarization," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2008, pp. 567-580.

[68] A. Clauset, M. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, p. 66111, 2004.

[69] B. Shneiderman and A. Perer, "Balancing systematic and flexible exploration of social networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, pp. 693-700, 2006.

[70] M. Girvan and M. Newman, "Community structure in social and biological networks," in *Proceedings of the National Academy of Sciences*, vol. 99, p. 7821, 2002.

[71] Z. Shen, K. L. Ma, and T. Eliassi-Rad, "Visual analysis of large heterogeneous social networks by semantic and structural abstraction," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, pp. 1427-1439, 2006.

[72] P.C. Wong, F. Hoot, G. Chin Jr., P. Mackey, K. Perrine, "Graph signatures for visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, pp. 1399-1413, 2006.

[73] D. Archambault, T. Munzner, and D. Auber, "TopoLayout: Multilevel graph layout by topological features," *IEEE Transactions on Visualization and Computer Graphics*, pp. 305-317, 2007.

[74] W. Cui, H. Zhou, H. Qu, P.C. Wong, and X. Li, "Geometry-based edge clustering for graph visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, pp. 1277-1284, 2008.

[75] T. Itoh, Y. Yamaguchi, Y. Ikehata, and Y. Kajinaga, "Hierarchical data visualization using a fast rectangle-packing algorithm," *IEEE Transactions on Visualization and Computer Graphics*, pp. 302-313, 2004.

[76] T. Itoh, H. Takakura, A. Sawada, and K. Koyamada, "Hierarchical visualization of network intrusion detection data," *IEEE Computer Graphics and Applications*, vol. 26, pp. 40-47, 2006.

[77] C. Muelder and K.-L. Ma., "A treemap based method for rapid layout of large graphs," in *Proceedings of the IEEE Pacific Visualization Symposium*, 2008, pp. 231-238.

[78] C. Muelder and K. Ma, "Rapid graph layout using space filling curves," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, pp. 1301-1308, 2008.

[79] D. Holten, "Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data," *IEEE Transactions on Visualization and Computer Graphics*, vol 12, pp. 741-748, 2006.

[80] T. Jankun-Kelly and K. Ma, "MoireGraphs: Radial focus+ context visualization and interaction for graphs with visual nodes," in *Proceedings of the IEEE Symposium on Visualization*, 2003, pp. 59-66.

[81] T. Dwyer, K. Marrior, F. Schreiber, P. J.Stuckey, M. Woodward, and M. Wybrow, "Exploration of networks using overview+detail with constraint-based cooperative layout," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, pp. 1293-1300, 2008.

[82] H. Kang, C. Plaisant, B. Lee, and B. Bederson, "NetLens: Iterative exploration of content-actor network data," *Information Visualization*, vol. 6, pp. 18-31, 2007.

[83] M. Ghoniem, J. Fekete, and P. Castagliola, "On the readability of graphs using node-link and matrix-based representations: A controlled experiment and statistical analysis," *Information Visualization*, vol. 4, pp. 114-135, 2005.

[84] R. Keller, C. Eckert, and P. Clarkson, "Matrices or node-link diagrams: Which visual representation is better for visualising connectivity models?" *Information Visualization*, vol. 5, pp. 62-76, 2006.

[85] F. van Ham, "Using multilevel call matrices in large software projects," in *Proceedings of the IEEE Symposium on Information Visualization*, 2003, pp. 227-232.

[86] J. Abello and J. Korn, "MGV: A system for visualizing massive multidigraphs," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, pp. 21-38, 2002.

[87] S. Fabrikant, "Spatial metaphors for browsing large data archives," *PhD Thesis*, University of Colorado at Boulder, Boulder, CO, USA, 2000.

[88] M. Tory, D. Sprague, F. Wu, W. So, and T. Munzner, "Spatialization design: Comparing points and landscapes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, pp. 1262-1269, 2007.

[89] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow, "Visualizing the non-visual: Spatial analysis and interaction with information from text documents," in *Proceedings of the IEEE Symposium on Information Visualization*,1995, pp. 51-58.

[90] G. Davidson, B. Hendrickson, D. Johnson, C. Meyers, and B. Wylie, "Knowledge mining with VxInsight: Discovery through interaction," *Journal of Intelligent Information Systems*, vol. 11, pp. 259-285, 1998.

[91] R. van Liere and W. de Leeuw, "GraphSplatting: Visualizing graphs as continuous fields," *IEEE Transactions on Visualization and Computer Graphics*, vol. 9, pp. 206-212, 2003.

[92] G. Ellis and A. Dix, "Enabling automatic clutter reduction in parallel coordinate plots," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, pp. 717-724, 2006.

[93] K. McDonnell and K. Mueller, "Illustrative parallel coordinates," *Computer Graphics Forum*, vol. 27, pp. 1031-1038, 2008.

[94] J. Sharko, G. Grinstein, and K. Marx, "Vectorized radviz and its application to multiple cluster datasets," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, pp. 1444-1427, 2008.

[95] L. Byron and M. Wattenberg, "Stacked graphs-geometry & aesthetics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, pp. 1245-1252, 2008.

[96] J. Johansson, P. Ljung, M. Jern, and M. Cooper, "Revealing structure within clustered parallel coordinates displays," in *Proceedings of the IEEE Symposium on Information Visualization*, 2005, pp. 125-132.

[97] Y. Fua, M. Ward, and E. Rundensteiner, "Hierarchical parallel coordinates for exploration of large datasets," in *IEEE Visualization* 1999, pp. 43-50.

[98] G. Ellis, E. Bertini, and A. Dix, "The sampling lens: Making sense of saturated visualisations," in *CHI'05 Extended Abstracts on Human Factors in Computing Systems*, 2005, pp. 1351-1354.

[99] J. Miller and E. Wegman, "Construction of line densities for parallel coordinate plots," *Computing and Graphics in Statistics*, vol. 36, pp. 107-123, 1991.

[100] E. Wegman and Q. Luo, "High dimensional clustering using parallel coordinates and the grand tour," *Computing Science and Statistics*, vol. 28, pp. 352–360, July 1996.

[101] W. Peng, M. Ward, and E. Rundensteiner, "Clutter reduction in multi-dimensional data visualization using dimension reordering," in *Proceedings of the IEEE Symposium on Information Visualization*, 2004, pp. 89-96.

[102] C. Peirce, "The fixation of belief," *Popular Science Monthly*, vol. 12, pp. 15-16, 1877.

[103] R. Amar, J. Eagan, and J. Stasko, "Low-level components of analytic activity in information visualization," in *Proceedings of the IEEE Symposium of Information Visualization*, 2005, pp. 111-117.

[104] J. Yi, Y. ah Kang, J. Stasko, and J. Jacko, "Toward a deeper understanding of the role of interaction in information visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, pp. 1224-1231, 2007.

[105] N. Kadivar, V. Chen, D. Dunsmuir, E. Lee, C. Qian, J. Dill, C. Shaw, and R. Woodbury, "Capturing and supporting the analysis process," in *Proceedings of the IEEE Symposium of Visual Analytics and Technology*, 2009, pp. 131-138.

[106] L. Bavoil, S. Callahan, P. Crossno, J. Freire, C. Scheidegger, C. Silva, and H. Vo, "Vistrails: Enabling interactive multiple-view visualizations," in *Proceedings of the IEEE Visualization*, 2005, pp. 135-142.

[107] K. Ma, "Image graphs — a novel approach to visual data exploration," in *Proceedings of the IEEE Visualization*, 1999, p. 88.

[108] T. Green, W. Ribarsky, and B. Fisher, "Visual analytics for complex concepts using a human cognition model," in *Proceedings of the IEEE Symposium on Visual Analytics and Technology*, 2008, pp. 21-23.

[109] J. Yue, A. Raja, D. Liu, X. Wang, and W. Ribarsky, "A blackboard-based approach towards predictive analytics," in *Proceedings of the AAAI Spring Symposium on Technosocial Predicitive Analytics*, 2009, pp. 154-161.

[110] L. Xiao, J. Gerth, and P. Hanrahan, "Enhancing visual analysis of network traffic using a knowledge representation," in *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, 2006, pp. 107-114.

[111] S. Garg, J. Nam, I. Ramakrishnan, and K. Mueller, "Model-driven visual analytics," in *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, 2008, pp. 19-26.

[112] T. Schreck, J. Bernard, T. Von Landesberger, and J. Kohlhammer, "Visual cluster analysis of trajectory data with interactive Kohonen maps," *Information Visualization*, vol. 8, pp. 14-29, 2009.

[113] A. Walhout and M. Vidal, "Protein interaction maps for model organisms," *Nature Reviews Molecular Cell Biology*, vol. 2, pp. 55-62, 2001.

[114] N. Luscombe, M. Babu, H. Yu, M. Snyder, S. Teichmann, and M. Gerstein, "Genomic analysis of regulatory network dynamics reveals large topological changes," *Nature*, vol. 431, pp. 308-312, 2004.

[115] D. Kell, "Metabolomics and systems biology: Making sense of the soup," *Current Opinion in Microbiology*, vol. 7, pp. 296-307, 2004.

[116] V. Batagelj and A. Mrvar, "Pajek — Analysis and visualization of Large Networks," *Lecture Notes of Computer Science*, vol. 21, pp. 477-478 2002.

[117] D. Auber, "Tulip — A huge graph visualization framework," *Mathematics and Visualization*, pp. 80-102, Springer, 2003.

[118] I. Xenarios, "DIP, the database of interacting proteins: A research tool for studying cellular networks of protein interactions," *Nucleic Acids Research*, vol. 30, pp. 303-305, 2002.

[119] I. Vastrik, P. D'Eustachio, E. Schmidt, and L. Stein, "Reactome: A knowledge base of biologic pathways and processes," *Genome Biology*, vol. 8, p. R39, 2007.

[120] K. Brown, D. Otasek, M. Ali, M. McGuffin, W. Xie, B. Devani, I. Toch, and I. Jurisica, "NAViGaTOR: Network analysis, visualization and graphing toronto," *Bioinformatics*, vol. 25, p. 3327, 2009.

[121] B. Breitkreutz, C. Stark, and M. Tyers, "Osprey: A network visualization system," *Genome Biology*, vol. 4, p. R22, 2003.

[122] T. Huan, A. Sivachenko, S. Harrison, and J. Chen, "ProteoLens: A visual analytic tool for multi-scale database-driven biological network data mining," *BMC bioinformatics*, vol. 9, p. S5, 2008.

[123] N. Gehlenborg, S. O'Donoghue, N. Baliga, A. Goesmann, M. Hibbs, H. Kitano, O. Kohlbacher, H. Neuweger, R. Schneider, and D. Tenenbaum, "Visualization of omics data for systems biology," *Nature Methods*, vol. 7, p.S56, 2010.

[124] N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein, "Genomic analysis of regulatory network dynamics reveals large topological changes," *Nature*, vol. 431, pp. 308-312, 2004.

[125] A. J. Walhout and M. Vidal, "Protein interaction maps for model organisms," *Nature Review Molecular Cell Biology*, vol. 2, pp. 55-62, 2001.

[126] K. Dahlquist, N. Salomonis, K. Vranizan, S. Lawlor, and B. Conklin, "GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways," *Nature Genetics*, vol. 31, pp. 19-20, 2002.

[127] M. Baitaluk, M. Sedova, A. Ray, and A. Gupta, "BiologicalNetworks: Visualization and analysis tool for systems biology," *Nucleic Acids Research*, vol. 34, p. W466, 2006.

[128] Z. Hu, J. Mellor, J. Wu, T. Yamada, D. Holloway, and C. DeLisi, "VisANT: Data-integrating visual framework for biological networks and modules," *Nucleic Acids Research*, vol. 33, p. W352, 2005.

[129] A. Barsky, T. Munzner, J. Gardy, and R. Kincaid, "Cerebral: Visualizing multiple experimental conditions on a graph with biological context," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, pp. 1253-1260, 2008.

[130] S. Kim, J. Lund, M. Kiraly, K. Duke, M. Jiang, J. Stuart, A. Eizinger, B. Wylie, and G. Davidson, "A gene expression map for Caenorhabditis elegans," *Science*, vol. 293, pp. 2087-2092, 2001.

[131] Y. He, "Genomic approach to biomarker identification and its recent applications," *Cancer Biomarkers*, vol. 2, pp. 103-133, 2006.

[132] K. Yeung and W. Ruzzo, "Principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, pp. 763-774, 2001.

[133] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389-422, 2002.

[134] S. Dudoit, J. Fridlyand, and T. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, pp. 77-87, 2002.

[135] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, 2000.

[136] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531-7, Oct 15 1999.

[137] M. Eisen, P. Spellman, P. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," in *Proceedings of the National Academy of Sciences*, vol. 95, pp. 14863-14868, 1998.

[138] W. Li, Y. Peng, H. Huang, and Y. Liu, "Efficient generalized matrix approximations for biomarker discovery and visualization in gene expression data," in *Proceedings of Computational Systems Biology*, pp. 523-529, 2006.

[139] J. Sharko, G. Grinstein, K. Marx, J. Zhou, C. Cheng, S. Odelberg, and H. Simon, "Heat map visualizations allow comparison of multiple clustering results and evaluation of dataset quality: Application to microarray data," in *Proceedings of the11th International Conference Information Visualization*, 2007, pp. 521-526.

[140] M. Sultan, D. Wigle, C. Cumbaa, M. Maziarz, J. Glasgow, M. Tsao, and I. Jurisica, "Binary tree-structured vector quantization approach to clustering and visualizing microarray data," *Bioinformatics*, vol. 18, p. S111, 2002.

[141] T. Kohonen and P. Somervuo, "Self-organizing maps of symbol strings," *Neurocomputing*, vol. 21, pp. 19-30, 1998.

[142] M. Mramor, G. Leban, J. Demsar, and B. Zupan, "Visualization-based cancer microarray data classification analysis," *Bioinformatics*, vol. 23, pp. 2147-2154, 2007.

[143] P. Hoffman, G. Grinstein, K. Marx, I. Grosse, and E. Stanley, "DNA visual and analytic data mining," in *Proceedings of the IEEE Visualization*, 1997, pp. 437-441.

[144] F. Azuaje, Y. Devaux, and D. Wagner, "Computational biology for cardiovascular biomarker discovery," *Briefings in Bioinformatics*, vol.10, pp. 367-377, 2009.

[145] H. Chuang, E. Lee, Y. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular Systems Biology*, vol. 3, pp. 140, 2007.

[146] D. Shepard, "A two-dimensional interpolation function for irregularly-spaced data," *Proceedings of the 23rd ACM national conference*, pp. 517-524, 1968.

[147] D. Ruprecht and H. Muller, "Image warping with scattered data interpolation," *IEEE Computer Graphics and Applications*, vol. 15, pp. 37-43, 1995.

[148] A. Sivachenko, J. Chen, and C. Martin, "ProteoLens: A visual data mining platform for exploring biological networks," *BMC Bioinformatics*, vol. 9, p. S5, 2008.

[149] J. Y. Chen, C. Shen, and A. Y. Sivachenko, "Mining Alzheimer disease relevant proteins from integrated Protein Interactome Data," in *Proceedings of Pacific Symposium on Biocomputing*, vol. 11, pp. 367-368, 2006.

[150] E. M. Blalock, J. W. Geddes, K. C. Chen, N. M. Porter, W. R. Markesbery, and P. W. Landfield, "Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses," in *Proceedings of the National Academy of Sciences*, vol. 101, pp. 2173-2178, 2004.

[151] C. Wu and D. W. Nebert, "Update on genome completion and annotations: Protein Information Resource," *Human Genomics*, vol. 1, pp. 229-233, 2004.

[152] J. C. Cruz and L. H. Tsai, "Cdk5 deregulation in the pathogenesis of Alzheimer's disease," *Trends in Molecular Medicine*, vol. 10, pp. 452-458, 2004.

[153] C. Lipscomb, "Medical Subject Headings (MeSH)," *Bulletin of the Medical Library Association*, vol. 88, pp. 265-266, 2000.

[154] R. Apweiler, A. Bairoch, C. Wu, W. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, and M. Magrane, "UniProt: The universal protein knowledgebase," *Nucleic Acids Research*, vol. 32, pp. D115, 2004.

[155] M. Palakal, M. Stephens, S. Mukhopadhyay, R. Raje, and S. Rhodes, "Identification of biological relationships from text documents using efficient computational methods," *Journal of Bioinformatics and Computational Biology*, vol. 1, pp. 307-342, 2003.

[156] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, pp. 11-21, 1972.

[157] M. Polanski and N. Anderson, "A list of candidate cancer biomarkers for targeted proteomics," *Biomarker Insights*, vol. 2, pp. 1-48, 2006.

[158] J. Y. Chen, S. Mamidipalli, and T. Huang, "HAPPI: An online database of comprehensive human annotated and predicted protein interactions," *BMC Genomics*, vol. 10, p. S16, 2009.

[159] C. von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, and P. Bork, "STRING: Known and predicted protein-protein associations, integrated and transferred across organisms," *Nucleic Acids Research*, vol. 33, pp. D433-7, 2005.

[160] K. R. Brown and I. Jurisica, "Online predicted human interaction database," *Bioinformatics*, vol. 21, pp. 2076-2082, 2005.

[161] T. S. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, and A. Venugopal, "Human protein reference database--2009 update," *Nucleic Acids Research*, vol. 37, pp. D767-D772, 2008.

[162] K. Goh, M. Cusick, D. Valle, B. Childs, M. Vidal, and A. Barabasi, "The human disease network," in *Proceedings of the National Academy of Sciences*, vol. 104, pp. 8685-8690, 2007.

[163] J. Li, X. Zhu, and J. Chen, "Mining disease-specific molecular association profiles from biomedical literature: A case study," in *Proceedings of the ACM Symposium on Applied Computing*, 2008, pp. 1287-1291.

[164] J. Chen, C. Shen, and A. Sivachenko, "Mining Alzheimer disease relevant proteins from integrated protein interactome data," in *Proceedings of the Pacific Symposium on Biocomputing*, 2006, pp. 367-378.

[165] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez Gene: Gene-centered information at NCBI," *Nucleic Acids Research*, vol. 35, p. D26, 2007.

[166] R. M. Wenham, J. M. Lancaster, and A. Berchuck, "Molecular aspects of ovarian cancer," *Best Practice & Research Clinical Obstetrics & Gynaecology*, vol. 16, pp. 483-497, Aug 2002.

[167] J. Palacios, M. J. Robles-Frias, M. A. Castilla, M. A. Lopez-Garcia, and J. Benitez, "The molecular pathology of hereditary breast cancer," *Pathobiology*, vol. 75, pp. 85-94, 2008.

[168] O. Imamov, G. J. Shim, M. Warner, and J. A. Gustafsson, "Estrogen receptor beta in health and disease," *Biology of Reproduction*, vol. 73, pp. 866-871, Nov 2005.

[169] R. R. Tubbs and D. G. Hicks, "HER-2 testing in breast cancer," *Journal of the American Medical Association*, vol. 292, pp. 1817-1818, Oct 2004.

[170] D. W. Bell, B. W. Brannigan, K. Matsuo, D. M. Finkelstein, R. Sordella, J. Settleman, T. Mitsudomi, and D. A. Haber, "Increased prevalence of EGFR-mutant lung cancer in women and in east asian populations: Analysis of estrogen-related polymorphisms," *Clinical Cancer Research*, vol. 14, pp. 4079-4084, Jul 2008.

[171] S. Maheswaran, L. V. Sequist, S. Nagrath, L. Ulkus, B. Brannigan, C. V. Collura, E. Inserra, S. Diederichs, A. J. Iafrate, D. W. Bell, S. Digumarthy, A. Muzikansky, D. Irimia, J. Settleman, R. G. Tompkins, T. J. Lynch, M. Toner, and D. A. Haber, "Detection of mutations in EGFR in circulating lung-cancer cells," *New England Journal of Medicine*, vol. 359, pp. 366-377, Jul 2008.

[172] S. J. Lewis, N. M. Cherry, R. M. Niven, P. V. Barber, and A. C. Povey, "GSTM1, GSTT1 and GSTP1 polymorphisms and lung cancer risk," *Cancer Letter*, vol. 180, pp. 165-171, Jun 2002.

[173] C. J. Der and G. M. Cooper, "Altered gene products are associated with activation of cellular rasK genes in human lung and colon carcinomas," *Cell*, vol. 32, pp. 201-208, Jan 1983.

[174] L. Jia and G. A. Coetzee, "Androgen receptor-dependent PSA expression in androgen-independent prostate cancer cells does not involve androgen receptor occupancy of the PSA locus," *Cancer Research*, vol. 65, pp. 8003-8008, Sep 2005.

[175] http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE2350 (accessed on June, 20th, 2009)

VITA

VITA

Qian You received her B.E. of Computer Engineering and Technology from Chu Kezhen Honors College, Zhejiang University, China, in 2004. She entered the doctoral program of computer science in Purdue University in 2005. She was a Purdue Research Foundation Graduate Assistant in 2009-2010. Her research interests include visual analytics, machine learning and bioinformatics. She has actively engaged in interdisciplinary research and has coauthored more than 10 research articles. After graduation, she joined Amazon.com.