VISUAL ANALYSIS OF PATTERNS AND SUBSTRUCTURE INTERACTIONS IN COMPLEX CHEMICAL COMPOUNDS

A Thesis

Submitted to the Faculty

of

Purdue University

by

Ruimin Gao

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science

May 2020

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL STATEMENT OF COMMITTEE APPROVAL

Dr. Yingjie Chen, Chair

Department of Computer Graphics Technology

Dr. Tim E. McGraw

Department of Computer Graphics Technology

Dr. Yu Zhu

Department of Statistics

Approved by:

Dr. Nicoletta Adamo-Villani

Graduate Program Professor

ACKNOWLEDGMENTS

I would like to express great appreciation and gratitude to my advisor, Professor Yingjie Chen, for his support, guidance, and inspirations throughout all these years.

I would like to thank my committee members, for offering valuable suggestions and input. Also, I am in debt to all the faculties and schoolmates in the Computer Graphics Technology department for helping me achieve success.

Last but not least, I offer my sincere appreciation for my beloved parents, for their love and encouragement while I'm studying abroad. And my best friend Patrick Teall, for his kindness and support.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Gao, Ruimin M.S., Purdue University, May 2020. Visual Analysis of Patterns and Substructure Interactions in Complex Chemical Compounds . Major Professor: Yingjie Chen.

Substructure composition can determine the activity of compounds. To understand the potential factors affecting the compound activity, users need first to analyze patterns and how substructures interact with each other. In addition to the traditional statistical approach, interactive visualization can bring significant value with intuitive analysis and involvement of human judgment. This research study presents an interactive visualization system for visual pattern analysis. With various visualizations and data manipulation, the system enables users to explore patterns and their relationships, as well as discover how substructures interaction may affect the activity of compounds. Users will be able to derive insights and better understand the data effectively and potentially benefit from the analysis in fields such as drug discovery.

CHAPTER 1. INTRODUCTION

1.1 Statement of Problem

The chemical compound dataset can be vast and complex to process and extract valid information. Generally, a compound is a thing composed of two or more separate elements. In the biochemistry field, a compound is defined as a substance consisting of two or more substructures chemically bonded together. Each distinct compound could have various compositions of substructures. In this research project, the presence or absence of various substructures in each compound forms the dataset. A compound can be either active or inactive. Different combinations of substructures can determine the activity of each compound (active or inactive). The combination is a pattern. As the patterns vary in numbers of substructures and composition, the correlation between compounds and substructures might be very complex. The purpose is to explore how the presence or absence of different substructures and their interactions simultaneously affect the activity of one compound. Meanwhile, extracting sets of patterns that discriminate active or inactive compounds and exploring the relationship between patterns are required. The structure of overall patterns and the pattern-subpattern relationship is of high interest for analysts to find positive or negative factors.

When building models to solve problems, a statistical approach might not perform adequately, especially for working on large-scale and complex datasets. As the purpose of data visualization is to help understand data in a more directly perceivable way, the visual analytic approach can be applied to help understand the data and refine the statistical model. Therefore, it is crucial to find out how to apply data visualization on understanding data mining models and methods when extracting and analyzing patterns and networks of such complex

compound-substructure dataset. In this project, the system applied some data mining techniques such as frequent itemset for helping people to explore patterns and networks. The dataset is extracted from U.S. National Library of Medicine database PubChem. However, the more critical problem is how to visualize the data mining process and utilize the pattern network exploration.

1.2 Research Question

How can data visualization be used to help to understand and analyze patterns and their interactions among complex chemical compounds?

- What are patterns? How can they be extracted from datasets?
- What is the structure of patterns? What are the relationships between patterns?
- How can patterns and their networks be visually represented?
- How can different patterns and the interactions of substructures impact the activity of compounds? How can these impacts be evaluated?

1.3 Scope

The research study could provide a more perceivable method for collaborating data mining approaches with information visualization for constructing solutions for particular problems. For data analysis, it specifically emphasizes pattern extraction and network analysis in analyzing complex chemical compound datasets with this collaborative approach.

While using the usual data mining approach to solve problems, including processing data and building models to fit data, it can be difficult for analysts to understand the data and techniques involved directly. The main focus of this research study is to supply a relatively perceivable visualization approach for

analyzing complex data when typical data mining approaches cannot fit and elucidate the data in particular cases.

1.4 Significance

Researchers have done some studies in visualizing data mining processes and models, even the visualization of pattern mining and association rules. However, unlike the usual transaction-like datasets, this research project focuses on a chemical compound dataset that contains two significant properties: active and inactive for different compounds. To be more specific, the dataset in this project is a sparse one-and-zero matrix that indicates whether or not the substructure exists in the compound, and the compound can be active or inactive. Different compositions of substructures (patterns) determine the compound activity. The goal of this project is to build a visualization system to help analysts understand the compound composition and pattern relationship, creating new opportunities for visual pattern mining with other core properties (activity) involved.

1.5 Assumptions

In the research of applying visual analytics in searching for patterns of substructures that can affect the activity of compounds, some assumptions serve as preconditions for this research.

- Different impacts for substructures. The first assumption is that different substructures have different impacts on deciding the activity of compounds. In this case, if different composition substructures do not imply different impacts on compounds, then patterns will not decide the activity of compounds.
- Dependent impact for each substructure. The second assumption is that the extent of the impact on compounds from substructures is not independent in terms of each substructure, which also means that the activity of a compound

might not only be affected by a single substructure, but also a set of substructures. If the impact on compounds is independent, only a single substructure can determine the activity of compounds. Then there will not be patterns that the system might extract and explore. Then the relationship and network of patterns will not help researchers and analysts deciding what the potential factors that affect the activity of compounds are.

- Correlated patterns. The third assumption is that patterns are correlated. If the patterns are independent of each other, there will not be a subspace for patterns that have different impacts on the activity of compounds. So that is another necessary assumption.

1.6 Limitations

The limitation of this research is that it mainly focuses on the exploration of the network and patterns. It includes:

- First, this research will explore the dimensions of factors that might affect the compounds. These factors can be multi-dimensional. For example, patterns of two substructures will have an impact on the activity of compounds, patterns of three or more substructures will also have an impact on the activity of compounds, and those impacts can be different.

- Second, this research will include the hierarchies and networks of patterns. The relationship and network of patterns might be hierarchical or directional, like a graph structure. The factors that might affect the compound might not be one certain pattern.

- Third, this research will include the exploration of subspaces where the patterns are. There are set-subset or pattern-subpattern relationships between different patterns. The impact of those patterns can differ from the pattern

itself to its subspace. With the need for querying for the relationship between patterns, the subspaces with patterns will be another aspect to explore.

- The most important limitation of this research is that the insights users could generate from exploring the system are based on existing observations. They are likely to have a bias. Therefore, this research study focuses on explorative analysis. It is not built for decision making, but rather providing insights and suggestions for future experiments.

1.7 Delimitations

Even though there might still be some aspects of this research, this research still has some delimitations:

- As the focus will be on how the different components of substructures affect the activity of compounds, this research will not include other factors that might affect the activity of compounds.
- This research only takes the impact of substructures as the main factor that contributes to the activity and property of compounds.

1.8 Definitions

In the broader context of thesis writing, this research defines the following terms:

Compound : A substance consisting of two or more substructures chemically bonded together. It can be represented with a list of substructure indices present in the compound, such as {1, 6, 272 ... 900, 1001}.

Substructure : The basic element constitutes each compound. Each distinct compound could have various compositions of substructures. Each

substructure is represented with its index number from the sparse matrix column, such as 1, 272, and 581.

Pattern : In a dataset, if a set of items or substructures frequently co-exist together or are strongly correlated, it can be a pattern, such as {272, 900, 518}, represented as a set of present substructure indices.

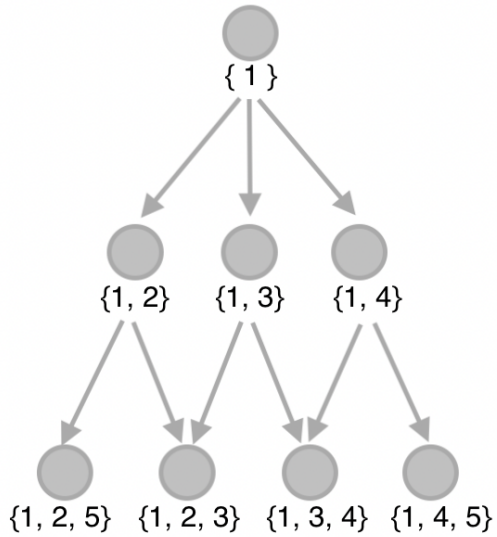


Figure 1.1. Example of Directed Acyclic Graph of Patterns

Parent-child pattern relationship : The relationships between different patterns form a DAG (Directed Acyclic Graph). A DAG is a finite directed graph with no directed cycles. They are not like tree structure as some of the child nodes could have multiple parents. The link (edge) between two pattern nodes implies a direction. The order of patterns is determined by the subset-superset relationship, like the Hasse diagram that represents a finite partially ordered set. (e.g. {x, y} $\rightarrow$ {x, y, z}, where parent node is {x, y} and child node is {x, y, z}) The direction of the link can not reverse or point back at a parent.

The parent-child pattern relationship exists between a sub-pattern and a super-pattern. The parent is a subset of the child, and the child is the

super-set of the parent. The size difference between parent and child can only be 1. (e.g. pattern {1, 2} is a parent/subset/sub-pattern of pattern {1, 2, 3})

Active ratio : For selected patterns, active ratio is defined as the proportion of active compounds the selected pattern appears in, versus the overall number of active compounds. If the overall active compound number is A, overall inactive compound number is I, then the selected pattern is represented with p. (In this case, the overall compound number is equal to $A + I$) The active ratio is defined as:

$$A_p/A$$

Inactive ratio : For selected patterns, inactive ratio is defined as the proportion of inactive compounds the selected pattern appear in, versus the overall number of inactive compounds. The inactive ratio is defined as:

$$I_p/I$$

a/i score : Activity of one specific pattern. It is defined to describe whether the pattern tend to appear more in active compounds or inactive compounds. This is a basic indicator to evaluate the pattern impact. The a/i score is defined as:

$$A_p/(A_p + I_p) \tag{1.1}$$

Pattern impact : The pattern impact can be positive or negative. Being positive means the pattern appear more in active compounds comparably (a/i score $>$ 0.5); Being negative means the pattern appear more in inactive compounds (a/i score $\leq$ 0.5)

Support : How frequently the pattern occur in the dataset. In this case, it represents the proportion of patterns that contains X with respect to all patterns. The support of X with respect to overall dataset T is:

$$supp(X) = \frac{|X \subseteq t; t \in T|}{T} \tag{1.2}$$

Confidence : How often the rule (X implies Y) that has been found true. In this case, it means the proportion of patterns that contains both X and Y with respect to patterns that contains X.

$$conf(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)} = \frac{|X \subseteq t; Y \subseteq t|}{x \subseteq t} \tag{1.3}$$

Frequent itemset : It can also refer to frequent patterns, which satisfy the condition: supp(X) $\geq$ minimum support threshold. For example, {beer}: 60%, {milk}: 70%, min_supp = 70%, therefore, frequent-1-itemsets will be {milk}.

Association rules : I is a set of n binary attributes (items), D is a set of transactions in a dataset, each t in D contains a subset of items in I,
A rule:

$$X \Rightarrow Y, \text{ where } X, Y \subseteq I$$

This means X is associated with Y, therefore formed an association rule.

Network : Similar to a computer network, a network in data visualization is defined as two or more data nodes linked together. In this project, we involved graph topology and DAG (Directed Acyclic Graph), which only provides one path between any two nodes. The graph can help us understand the hierarchical structure of all pattern nodes.

Parallel Coordinate with Scatter plots : It is a typical combination of both Parallel Coordinate and scatter-plots, usually for high-dimensional data visualization.

Each axis of the parallel coordinate represents one dimension or attribute of each data entry. Moreover, the coordinate of each scatter plot is based on every two dimensions from the parallel coordinate axis. Users can analyze data in both a multi-dimensional and a bi-dimensional way.

Compound-substructure dataset : The dataset consists of compounds data with the presence or absence of various substructures in each compound. Different combinations of substructures determine the activity of each compound (active or inactive). The combination is a pattern. As the patterns vary in numbers of substructures and composition, the correlation between compounds and substructures might be high-dimensional.

High-dimensional statistics: High-dimensional statistics studies more than multivariate data with higher dimensions.

CHAPTER 2. REVIEW OF RELEVANT LITERATURE

This chapter provides a review of the literature relevant to applying visualization in data analysis of patterns and networks.

The chemical compound dataset can be vast and complex to process and extract valid information. When it comes to data related to medicine, the scale of effectiveness and components affecting the effectiveness are challenging to generate or even analyze the given data. Statistical approaches are usually used for searching for patterns and networks among the elements. However, many models cannot describe the data and illustrate a good story in ordinary life. Especially for complex high-dimensional medical datasets with a subtle relationship between elements, it is hard to find lurking components or patterns that affect the result. As visual data exploration gets the user directly involved in the data mining process, a visual analytic approach can be applied to help understand the data, support the data mining process, and even refine the statistical model (Keim, 2002; Keim, Mansmann, Schneidewind, & Ziegler, 2006). There are many studies on applying visual analytics in data mining and the analysis process, such as visualizing clustering, classes, and feature selection (Krause, Perer, & Bertini, 2014; Strehl & Ghosh, 2003). However, there is a gap in applying visual analysis on patterns and networks with associations, especially in high-dimensional datasets.

Therefore, this research tried to build a method to involve visual analytics effectively in adjusting and evaluating the statistical model. It will eventually assist in extracting patterns and analyzing complex, large-scale medical datasets. The following literature review will focus on exploring the topic and research gap in this field.

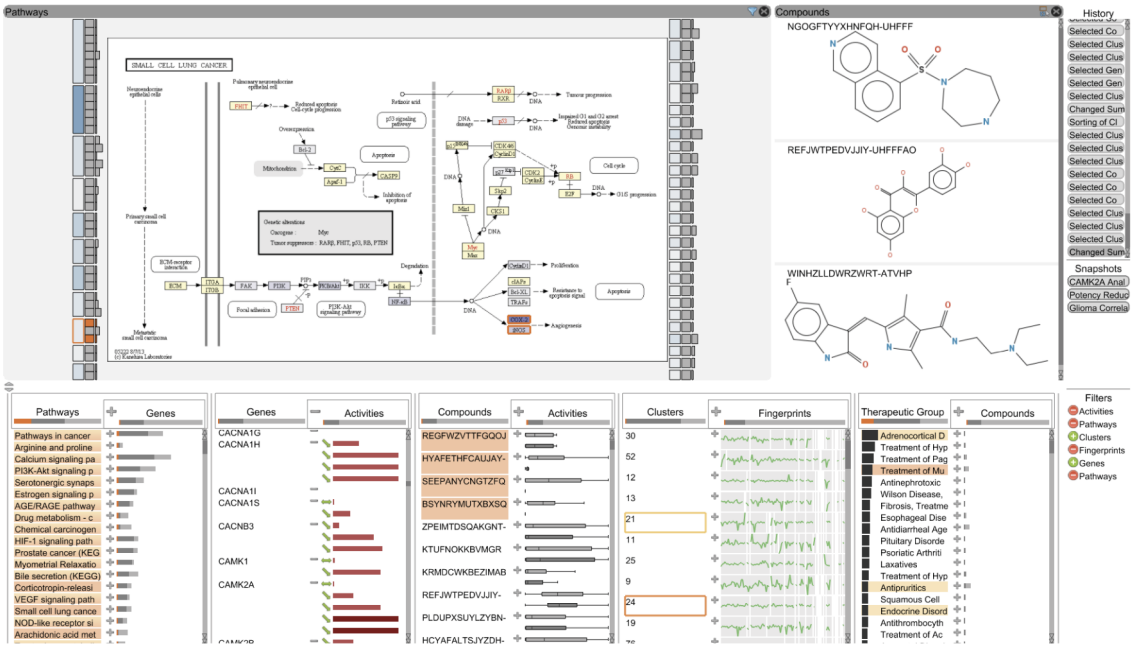


Figure 2.1. ConTour: Data elements in columns relationship view (bottom). detail views of selected pathway and chemical structures of compounds (top)

2.1 Drug Discovery with Multi-relational datasets

While this research study focuses on visual analytics of compound patterns and substructure interactions, many research studies have been conducted regarding multi-relationship data mining and visualization. In the biochemistry field, to discover drugs, large scale data analysis is fundamental. Exploration and evaluation of potentially useful compounds based on relational datasets are necessary. In the paper by Partl et al. in 2014, an interactive visual analytics tool named ConTour is created to solve this problem. The system puts all elements in a column and utilizes interactions to uncover the relationships. Users can select one or more items to highlight and re-order other elements. It also provides filters for drilling down the large scale datasets, as well as productive sorting strategies. The system enables interactive nesting of columns and display attributes and also the connection level to other datasets. It also has detailed views to present attributes and their correlated data. ConTour can be used for biochemists to explore the chemical

compounds. Other research studies for solving multi-relational or graph exploration problems include Jigsaw's list view, for visualizing relationships (Stasko, Görg, & Liu, 2008) (Görg et al., 2010); and GraphTrail for heterogeneous network analysis and exploration history (Dunne, Henry Riche, Lee, Metoyer, & Robertson, 2012). The other techniques used in the ConTour is faceted browsing, the current works include InfoZoom (Spenke & Beilken, 2000) and FOCUS (Spenke, Beilken, & Berlage, 1996).

2.2 Visual data mining

Lots of studies concerning the combination of the visualization and data mining process have been conducted in the past. In the research article by Daniel A. Keim (2002), he proposed a detailed classification of information visualization, visualization techniques, visual data mining techniques determined by data type and interaction involved. Daily life produces an enormous flood of data with lots of parameters, which leads to the high dimensionality of the data. Even though people believe recording and processing these data can help them find valid information, it is still a difficult task. As people want the data mining process to be as effective as possible, it requires humans to be involved in the exploration process and combine the flexibility and creativity of human knowledge with the tremendous computation capability computers can provide. Visual data exploration requires no deep understanding of complex statistical algorithms or parameters. It is rather intuitive and perceivable (Munzner, 2008).

2.2.1 Visual data mining techniques

Then Keim also classified visual data mining techniques. The classification has three dimensions: visualization data type, visualization techniques, and interaction techniques. Usually, a system will be a combination of multiple visualization and interaction techniques. In his further studies on challenges in

visual data analysis, he talked about the visual analytics overviews, concepts and the scope, and the technical research challenges in the area (Keim et al., 2006). There is a gap between the rapid data collection and storage process and the lack of ability to analyze these data. Visual Analytics can be described as "a science of utilizing interactive visual solutions to augment analytical process" (Cook & Thomas, 2005). Visual analytics focuses on utilizing the integration of human judgment to handle a large, multifarious and dynamic amount of information with visual representation and interaction techniques involved in the analysis, While visualization itself serves as "a graphical representation of data or concept" (Ware, 2020). This combination of related research areas of visualization, data mining, and statistics turns visual analytics into a dynamic and diverse field of research. For the critical decision-making process, human factors tend to be more flexible, creative, and have strong background knowledge. While computers have incompatible computing ability and storage capacity, the combination of both will perform very well in the data analysis process. The decision-makers can focus more on analysis with their intuitive cognition and perception. It is also the specific advantage of visual analytics.

However, technical challenges exist in this process as well. It is challenging to explore a large amount of data adequately. Some data still becomes useless. It also affects the limit of display capability for visualizations. Dimension reduction or data reduction methods are often required as only a small portion of the data can be displayed. Visual scalability becomes the critical factor that brings the rapidly growing data on the screen in an appropriate way. A dynamic process, an adaption of visual analytics on real-world application, interpretability, and more fields involved are also challenges of visual analytics.

2.2.2 Multivariate patterns

There is another research paper on the combination of visual analytics and the data mining process. It focuses on a multivariate patterns visualization system (Guo, Chen, MacEachren, & Liao, 2006). With the assistance of the methodology tools and techniques such as sorting, clustering, and visualization, pattern discovery in multivariate dimensions has become much easier for data analysts. The article has covered five parts: a map, re-organized matrices, a parallel coordinate plot. It also includes geographic small multiple displays and a two-dimensional cartographic color design method. The visualization system supports an overview of complex patterns and also allows users to focus on individual patterns and detailed views with various interactions. Similar work such as a human-centered exploration solution suite also illustrates the combination of computation and visualization (Guo, 2003). It aims to enable performing multivariate clustering and abstraction, using a parallel coordinate plot to visualize the multivariate patterns, visualizing the spatial-temporal variations of multivariate patterns, and supporting exploring patterns from different aspects and at different levels through human interaction (Marozzi, 2015).

2.2.3 Visual predictive modeling

As those studies regard visualization more likely as an interpretation of the data, a significant amount of data scientists are intrigued in understanding the data and predictive probabilities associated with them. Josua Krause, Adam Perer, and Enrico Bertini (2014) proposed a systematic way to use visual analytics for feature selection in predictive modeling. The process of predictive modeling consists of dividing datasets into groups and using feature construction techniques for feature vector definition. Then users need to define parameters for cross-validation to make sure that the result is statistically robust. After it, users can extract informative features with feature selection algorithm, and then evaluate the level of prediction of

the model with selected classifier. Feature selection usually requires that analysts try multiple types and the output of algorithms that are often hard to interpret. This research designed a system for predictive modeling, specifically for understanding how features are ranked across the whole process. The system also allows users to build models interactively. It included three main tasks: comparison of feature selection algorithms, comparison of classification algorithms, and new feature sets manual selection and testing. The system provides a visualization system of a large group of features and the modeling algorithms usage of them. It uses a visual design that features are the main components of the visual representation. Each visual object shows a feature, and the information obtained from the algorithms is reflected through its design and layout. The system also uses a case study concerning medical records of patients for clinical researchers to predict the outcomes. This research provides insights into applying visual analytics in the data mining process with visual representations involved in the analysis process rather than only in data interpretation. As the focus is on predictive modeling, specifically on the visualization of feature selection, the potential gaps still exist in this visual analytical field. People also used bibliographic coupling and cocitation for problems related to data mining (Kermarrec & Moin, 2012).

2.3 High-dimensional data visualization

Another focus of this research project is high-dimensional data visualization. For high-dimensional datasets, one of the issues is dimension reduction. High-dimensional space is far beyond human cognition and imagination. Visualizing it is another big issue. Many works are related to these aspects.

2.3.1 Interactive high-dimensional data visualization

Andreas Buja, Dianne Cook, and Deborah Swayne (1996) proposed an interactive data visualization taxonomy for high-dimensional data analytics. The

tasks include locating Gestalt, structuring queries, and compare. Various types of interaction manipulations were utilized to assist the tasks: focusing, linking, and view arranging. Then they provided a high-level introduction to a visualization system for multivariate data as the system emphasizes focusing, linking, and views arranging, corresponding to high-dimensional data focus, brushing linked scatterplot, and conditional plots matrices. A data visualization taxonomy is developed to set up guidelines of disordered techniques and provides interpretation and further explanation of the purposes behind. It can also intrigue and encourage people to discover new techniques. What we see from the data is decided by focusing. The rendering method plays a great part in illustrating what focusing means. Dynamic queries are used to search for information. They help users with overwhelming information by providing an overview of data (Shneiderman, 1994). Linked views are very effective in guiding users to visually construct it and also see the response in a visual way. Flexible view arrangements bring out values of effective comparison. For example, pairwise variable plots can be arranged with the scatterplot matrices to make more sense of the data.

2.3.2 Parallel coordinates

As Keim proposed several challenges in visual analysis, the high dimensionality of data has always been an issue for data analysis and visualization (Keim, 2006). In 1991, Alfred Inselberg and Bernard Dimsdale proposed a new visualization design that breaks the dilemma of visualizing and displaying high-dimensional data (Inselberg & Dimsdale, 1991). To visualize multi-dimensional geometry, they designed a new visualization form named parallel coordinates. The visualization maps multi-dimensional data to two-dimensional sets. Some geometrical properties similar are used to present planar images of hypersurfaces. The visualization displays the geometrical properties to represent how close the point is to the boundary. Parallel coordinates yield graphical

representations of multi-dimensional relations rather than just finite point sets. Later on, Alfred Inselberg 2002 again discussed visualization and data mining of high-dimensional data. Parallel coordinates use geometrical properties to represent multi-dimensional data relationships. The overall problem of high-dimensional data querying and presentation for multivariate relationships has been simplified into discovering patterns in a two-dimensional representation system. The example used is a country's economic dataset. The parallel coordinates are used to analyze and locate the multivariate relationships with hypersurfaces. It focuses on visualizing the high-dimensional data and analyzing multivariate relations without too much cost of losing data or hidden information. The two-dimensional patterns of parallel coordinates are straightforward to display and explore visually. The advantages of parallel coordinates are displaying data with information, visual queries, and good interaction. They all support users to search for patterns and relationships in the variables of the data. (Graham & Kennedy, 2003).

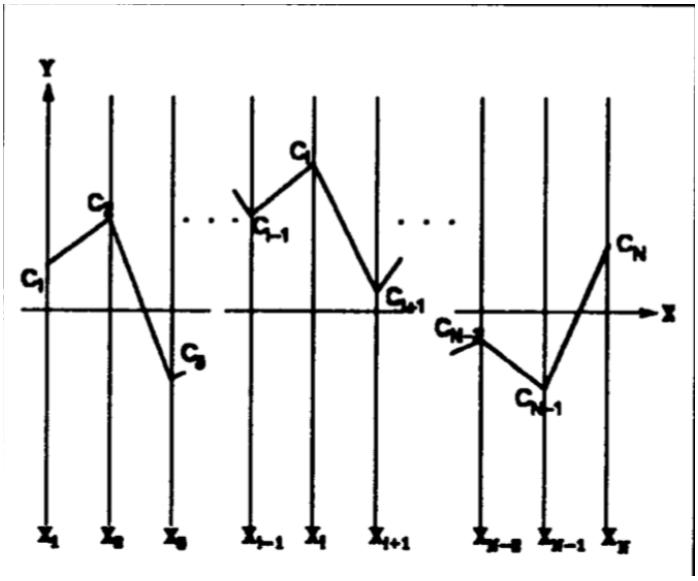


Figure 2.2. Parallel Coordinate: Parallel axes for R^N

2.4 Pattern Mining

2.4.1 Frequent Patterns and Association Rules

C. K. Leung, Kononov, Pazdor, and Jiang proposed a visual analytic system for frequent pattern visualization and analysis: PyramidViz. The system used a visual display to represent frequent patterns, showing their frequency and relationships. (C. K. Leung, Kononov, et al., 2016) Other visualization exploration tools include frequent itemset mining along with association rule. (Blanchard, Guillet, & Briand, 2007) In another article by C. K. Leung, Carmichael, et al., they combined the HCI (Human-Computer Interaction) research along with data mining research and analyzed a few existing visual analytics systems for frequent pattern mining. The systems all have different visualization forms for frequent patterns and vary in focus and benefit. FpMapViz focuses on the hierarchical structure of the patterns, and the visual representation is treemap-like (C. K.-S. Leung, Jiang, & Irani, 2011). RadialViz has an orientation-free radial layout (C. K.-S. Leung & Jiang, 2012). PyramidViz is also hierarchical but focuses more on the individual growth of a pattern (C. K. Leung, Kononov, et al., 2016). They all use color as a visual cue to represent attributes such as frequency.

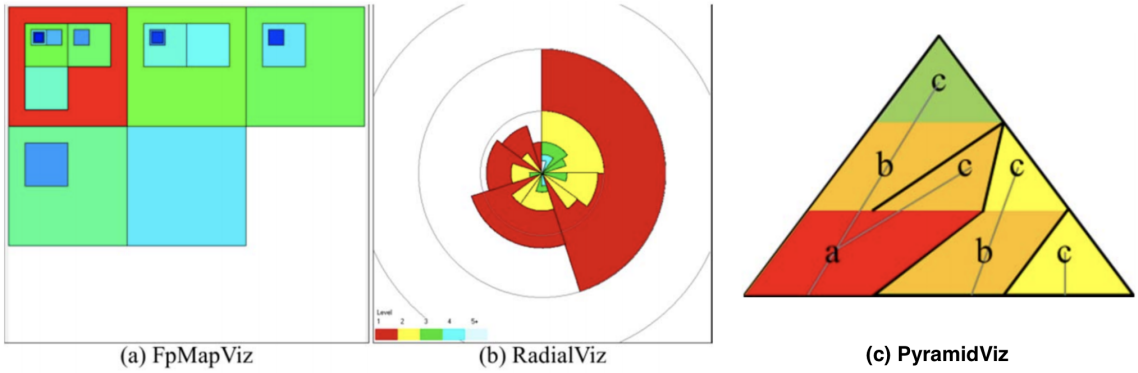


Figure 2.3. Different Frequent pattern visualizations

2.5 Summary

This chapter provided a review of the literature relevant to applying visualization in data analysis of patterns and networks. In general, there is a research gap of applying visual analytics on the high-dimensional data mining process, specifically for pattern and network exploration with associations. Many issues of visualization concerning high-dimensional data and a combination of data mining and visualization, this research paper will propose a system that applied visualization in the high-dimensional data analysis process. The system can help understand data and model, explore the patterns and relationships of associated elements, and finding lurk components or patterns that affect the result. This research aims to involve a visual analytical approach in supporting data mining and analyzing processes, eventually provide an overview of patterns and networks in the data. There will be a case study specifically on a chemical compound dataset that shows how different patterns of substructures affect the activity of chemical compounds, with the exploration of the patterns and their network.

The next chapter provides the framework and methodology to be used in the research project.

CHAPTER 3. FRAMEWORK AND METHODOLOGY

This chapter provides the framework and methodology used in the research study and the development of the pattern exploration system, including a brief introduction to the frameworks and technology used during the implementation, data mining methodology applied in the pattern exploration system. Then it will uncover the underlying design disciplines that guide the entire development process, utilizing the theory of seven stages of visualizing data (Fry, 2008) and interactive dynamics taxonomy for visualization (Heer & Shneiderman, 2012).

3.1 Frameworks

The frameworks used in this pattern analysis system consist of three parts: data manipulation, UI framework, and visualization components. For easier data processing, cleaning, and manipulation, we choose Python as the back-end language. First, Python has one of the largest communities for data cleaning, mining, and analysis, providing a significant number of packages and tools. Second, the community already has multiple well maintained and supported full-stack frameworks, such as Django and Flask. We choose Flask for the system implementation as it is very light-weight and less complex. For front-end system development, we choose React.js and its related toolkit. React.js is easily one of the most popular SPA (Single Page Application) UI framework. It provides an easy-to-manage structure and extensive features that benefit small-scale applications, especially with visualization components included. A variety of visualization tools and libraries can be integrated with the React framework. We choose D3.js to create complex customized visualization and Echarts.js for the fast development of existing common chart types.

3.1.1 Data Manipulation

The initial intention of developing a visualization pattern exploration system is to make pattern exploration and analysis more intuitive and perceivable. While trying to solve a high-dimensional data problem, the traditional statistical approach or data mining methods might be less intuitive without visual assistance. The insights generated during any stage of the analysis process should be preserved and presented whenever the data analysts want. The data handling for the system became very important.

The back-end data handling utilized Pandas. Pandas is an open-source library that provides high-performance data structures and data analysis tools. Most data parsing and calculation algorithms in our pattern exploration system use Pandas. The other data mining package used is Orange3-Associate for mining frequent itemsets and association rules. The implementation is a FP-growth frequent pattern mining algorithm (Han, Pei, & Yin, 2000) with bucketing optimization (Agarwal, Aggarwal, & Prasad, 2000) for conditional databases of few items. We use this algorithm to extract frequent itemsets and regard them as frequent patterns.

3.1.2 UI frameworks

As this research study mainly focuses on how to utilize visualization to assist data analysis, specifically pattern mining, the interface and design of the presentation of data became inevitably crucial to achieving our goal. A few of the benefits of using React.js as UI framework is that it is swift, scalable, and easy to use. It allows users to create reusable UI components, serving as the view layer in the MVC model (Model-View-Control). The React components are self-contained; they have their properties and state, which makes managing the overall structure of the system relatively simple.

However, with D3.js implementing visualization components, things can be difficult. One of the most significant challenges developers usually encounter with D3 + React application is to make these two tools work seamlessly together. The nature of both D3 and React is to manipulate DOM (Document Object Model) elements. D3 integrates data point in each DOM element, and update element properties accordingly. While React controls its own virtual DOM tree for much more efficient comparison and update, in this case, developers have to delegate the control of DOM elements to either one of them. In this pattern exploration system, the visualization component DOM alternation is owned by D3.js for flexibility purposes. While React is in charge of state management, data fetching, and parsing, and management of other parts of the UI.

3.2 Data Mining Methodology

The fundamentals for visual data mining should be how data was parsed, calculated, and presented. Moreover, for our pattern exploration system, the main focus is to help data analysts better understand the compound dataset and explore its network. The core data mining method used in this system is: **Association Rules**. Association rules are statements with probability indicating the relationship between items in a large dataset. Association rules have a wide range of applications, including exploration of correlation in chemical compound datasets. It uses data mining models to find co-occurrence of items (Rouse, 2018). The association rule is one of the most commonly used methods to explore relationships and patterns in a dataset. However, as our chemical compound data has a unique property: compound being active or inactive, co-occurrence by itself became less descriptive of the complex relationships hidden in our chemical compound dataset.

In our system, we only used a few terminologies from association rules as we want to focus on the network and the parent-child / subset-superset relationship. The pattern exploration system will have a support filter to define frequent itemset,

then generating set-subset relationships. (e.g. Pattern {1, 2} is the parent of pattern {1, 2, 3}) In this case, the pattern that has no child or super-pattern with given support will be closed-pattern or max-pattern. Its connected parent patterns and extended parent patterns will have larger support.

- **Pattern**: a pattern is a set of items, sub-sequences, or substructures that frequently occur together (strongly correlated).
- **Frequent Itemset**: for a set if items X in a dataset, if the support of X is greater than minimum support threshold user provided, we call this itemset frequent itemset.

$$supp(X) \geq supp_{min} \tag{3.1}$$

- **Closed-pattern**: a pattern X is closed if the pattern X is frequent and there is no super-pattern $Y \supset X$ have the same support.
- **Max-pattern**: X is the max-pattern if X is frequent and there is no frequent super-pattern $Y \supset X$. (In this case, we do not care about the support of parent-pattern. Their support will be larger than super-pattern.)

3.3 Design Process

Before we step into solving our problem, there have always been gaps between different fields: data mining, visual design, and data visualization. While visual design cannot process thousands or even millions of data pieces, data mining can do so, but it disconnects the interaction with data. To build something that tries to solve a data problem, we need to use a universal process to close the gaps between individual disciplines and focus on how data is understood, interpreted, and displayed. How data is represented has the same level of importance as the data itself (Fry, 2008). In the book "Visualizing Data" by Ben Fry, the seven stages of visualizing data process are defined as follows:

Acquire Get data.

Parse Transform and order data into structures feasible.

Filter Only keep data of interest.

Mine Provide a statistical or data mining context to the data.

Represent Visual encode the data. Map data into various visual cues.

Refine Improve visual representation or interaction.

Interact Allow control of data or visualization.

The process itself is a workflow, with each stage connected. It separates information visualization into multiple phases while keeping the overall workflow coherent. However, there is a problem that can be triggered by isolating the tasks and delegate them to different individuals. The single source of truth should be the original data, no matter what format it is. The information it conveys can be lost during or between the transition. The final representation and outcome will be derived from the data mining stage and might not directly reflect the initial problem. (Fry, 2008) In this case, we want to have control of all stages and follow the entire path until we finally reach the understanding of the data. This seven-stage process will guide us on how to design and implement the pattern exploration problem.

3.3.1 Variables

When it comes to statistical analysis of the problem, choosing the appropriate variables is extremely crucial for measuring different metrics and aspects of the data. In the pattern exploration system, we defined and used the variables as follows:

Support : [0, 1]. It describes how frequently the pattern appears in the dataset. In this case, it represents the proportion of patterns that contains X for all patterns. The support of X for overall dataset T is:

$$supp(X) = \frac{|X \subseteq t; t \in T|}{T} \tag{3.2}$$

A/I Score : [0, 1]. The activity of one specific pattern. It is defined to describe whether the pattern tends to appear more in active compounds or inactive compounds. A/I Score is a basic indicator to evaluate the pattern impact. The A/I Score is defined as the ratio of active ratio versus the sum of both active and inactive ratio:

$$A_p/(A_p + I_p) \tag{3.3}$$

A/I Score Difference : [-1, 1]. It describes how the pattern impact changes between two connected patterns (parent-child relationship). With the new substructure added, the child pattern might have a larger or smaller A/I Score than the parent pattern. e.g., Pattern {1, 2} has A/I Score of 0.5, its connected child pattern {1, 2, 3} has A/I Score of 0.4, the A/I Score difference between these two patterns will be - 0.1, and it can be negative based on the direction.

$$aiScore_{diff} = aiScore_{child} - aiScore_{parent} \tag{3.4}$$

Level : Level is the number of distinct substructures in a pattern. (e.g. pattern {1, 2, 3} has level of 3)

Impact factor : It is a term for substructure that significantly increases or decreases the A/I Score upon addition from parent pattern to child pattern. e.g. Pattern {1, 2} has A/I Score of 0.5, its connected child pattern {1, 2, 3} has A/I Score of 0.4. The A/I Score difference threshold was set to ± 0.1. So, substructure 3 will be a negative impact factor that converts a pattern with

A/I Score of 0.5 to a pattern with A/I Score 0.4. The significance is defined by users (A/I Score Difference).

These variables mostly serve as filter metrics to narrow down and extract points of interest, also indicating individual values of each pattern and its parent/child patterns.

3.3.2 Measure for Success

The criterion for what constitutes a "success" of this research study is that we can derive insights from the data with the assistance of the designed visualization system. Therefore, we will set up multiple use cases to showcase how the system can be utilized to answer the questions and sub-questions from the problem statement.

3.4 Summary

This chapter provided the framework and methodology to be used in the research study. More details of how the seven-stage process guides the design and implementation of the pattern exploration system will be revealed in the next chapter.

CHAPTER 4. DESIGN AND IMPLEMENTATION

To derive insights from data and answer the questions raised from the problem statement, we need to truly understand what the goal is and build a solution capable of addressing the issues and potentially driving the data analysis and finding points of interest. The goal of this research study is to enable users to explore how the data visualization system can help understanding patterns, their relationships, and networks in a complex chemical compound dataset. For intuitive representation, visualization will be an essential part of the system, guiding users to understand what patterns are and their relationships between each other. Besides the visual part, we also need to integrate proper interaction and the ability of data formatting, filtering to update the view of data.

The design and implementation of the system will follow the seven-stage rule of visualizing data (Fry, 2008). However, the rules will not be followed slavishly as the design process theory should be utilized to fit into the needs and context of our problem and system.

4.1 Design Process

One of the most crucial question in the problem statement is: what are patterns and what is their relationships? To dive deeper into the network of patterns, we need to understand the definition of a pattern.

A pattern is a set of substructures. While substructures constitute a compound, the combinations of substructure derived from these substructures can be different patterns. The more this pattern is present in different compounds, the more meaningful the pattern can be. However, a pattern does not only hold attributes such as support but also has a new attribute A/I Score created to

indicate how frequent the pattern is present in both active and inactive compounds. Moreover, this allows us to explore the correlation between patterns and how they might affect the activeness of a compound.

The other aspect of the question is the relationships between the patterns. The researchers have used association rules to find the correlation between different itemsets. Association can be one dimension of the relationship. However, this relationship addresses more on the correlation between individual itemset. We need a higher scale view of the relationships in a large number of patterns. Moreover, that leads to the definition of a network. A pattern network should show not only the direct relevance of the patterns but also the hidden attributes. Therefore, we define this direct relevance as a parent-child or subset-set relationship. e.g. Pattern {1, 2} is the parent of pattern {1, 2, 3}, they share a direct connection. The network will be an overview of many connections.

After understanding the essential problem, the following stages of design and implementation are necessary to push the process forward.

4.2 Early Experiments

Some early experiments were conducted, including a matrix view of a compound-substructure relationship. The initial dataset is a sparse matrix (U.S. National Library of Medicine, 2016), where each column represents a substructure, each row represents a compound, and the value of cells represents the presence (1) or absence (0) of the substructure in the compound. The system aims to present a matrix-based view to present different types of relationships. We also used collaborative filtering to construct a self-to-self relationship for compounds and substructures. A network can be represented by an adjacency matrix, where each cell (i, j) represents an edge from vertex i to vertex j, meaning the substructure j is present in compound i. For example, the left-top view in figure 4.1 shows the primary matrix, where each column is a substructure, and each row is a compound.

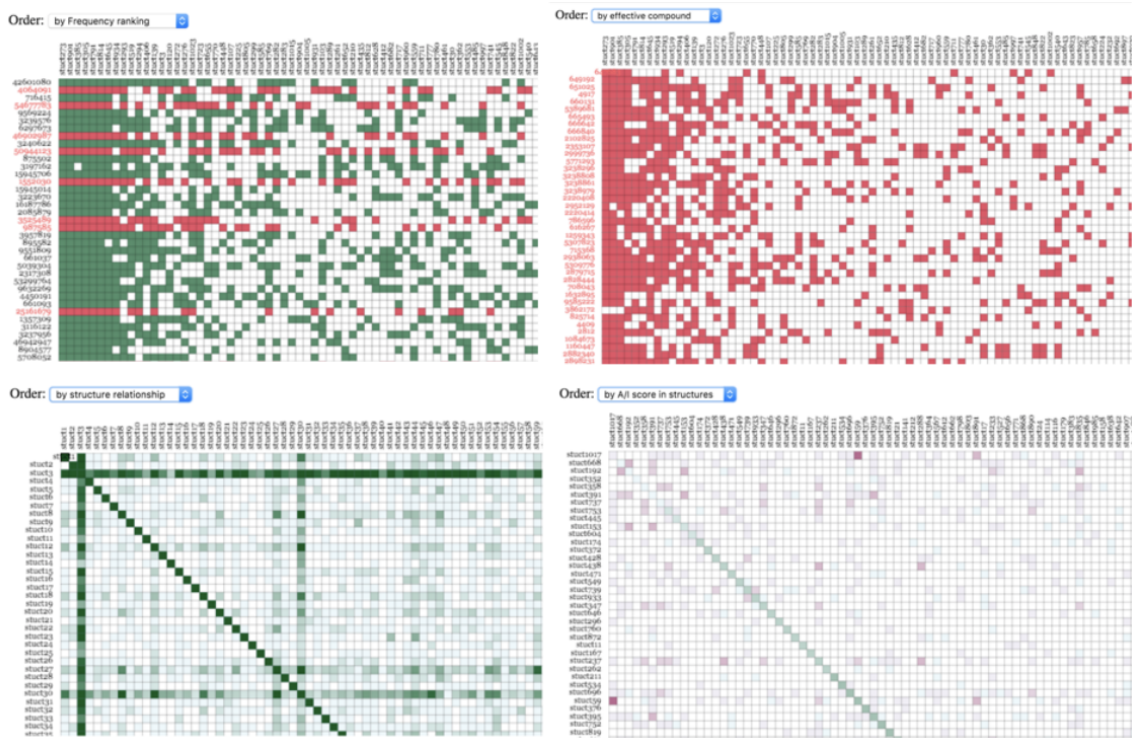


Figure 4.1. Matrix-based view for different type of relationships

A green row represents an inactive compound, and a red row represents an active compound. (This color scheme is conflicting with the final design, it will be reversed at the end)

We used the adjacency matrix to represent the compound-substructure relationship then use the collaborate filter to construct the inner relationship between compounds or substructures. Here we also created a new attribute A/I score based on the co-occurrence of elements. A/I score calculation: Use conduct of two matrices, one is the standard compound-to-substructure matrix, the other uses activity of compound to mark the sign of current value (active: + inactive: -). We use $U = A^T A$ to multiply two matrices. The value of $U(i, j)$ is the sum of all the compounds the substructure i and j have in common. The more positive, the more active compounds they have in common. (In the heat map it tends to be redder, the

value of white elements are around 0, and the value of green ones tend to be negative) an element representing the common feature of these two components.

As our goal is to see the effectiveness of composition (pattern) of multiple structures and how they affect the activeness of compounds, we can go through the two-dimensional heatmap and explore which binary combination of structures have a better and more substantial effect on compounds. For example, in the bottom-right picture in figure 4.1 substructure 1017 and substructure 59 has a deep red color of the element, which means they have a high A/I score, and a lot of active compounds have them existing together. Through exploring the heatmap, the binary relationship, how substructure affect the compounds can be revealed quickly.

However, there are many limitations to this matrix-based visualization. First, it is a two-dimensional visualization. It only enabled exploration of a two-dimensional network and the potential relationship between two substructures or two compounds. In order to see more of the whole network and the relationship between multiple substructures (patterns), we need to make some future improvements. Also, the color scheme does not make sense anymore if we adopt the concept of active/inactive or positive/negative, as green is more common to represent positive items. Therefore, we will update the color scheme in the later design and experiments and reverse the representation.

4.3 Data Pre-processing

Data pre-processing corresponds to acquire, parse, and filter stages in the design process. It includes two major parts: how the data is defined and how the data is parsed to fit into our use cases.

The original dataset was extracted from U.S. National Library of Medicine database PubChem for chemoinformatics research. It is a matrix of the presence of substructures in each compound, where each column is an individual substructure or a set of substructures, and each row represents a compound with different

activeness. The value of each cell can be 0 or 1, indicating the absence (0) or presence (1) of a substructure identifier in a compound. There are 2615 rows and 1024 columns in this dataset, and the top 663 rows are active compounds.

Table 4.1. *Example of Original Dataset*

cpdid	s1	s2	s3	...	s1023	s1024
647683	0	0	0	...	0	0
649192	0	0	0	...	1	0
651025	0	0	1	...	1	0
...						
44601947	1	0	1	...	0	0
974700	0	0	1	...	0	0

4.3.1 Data Parsing and Transformation

The original dataset has 2615 x 1024 cells, almost 2.7 million in total. However, the dataset itself is a vast sparse matrix, with way more absences (0) than presences (1). A sparse matrix is a matrix in which most elements are zero. One of the common issues of the sparse matrix is the storage of data, which also leads to unnecessary computations due to the significant number of zeros. To reduce the memory cost of this redundant storage for a sufficiently large matrix, we need to compress the sparse matrix. There have been various methods for sparse matrix compression, such as List of lists (LIL), Coordinate list (COO), and Compressed sparse row (CSR). We adopted the method of List of Lists, where each row contains a list of entries of column index and the value. The column index can be sorted for a unique reference and faster lookup. As most data values of the compound dataset are merely substructure presences (1), only column indexes will be stored in each row list to represent the data value. The sparse matrix dataset is parsed into a list of lists, where each list represents a compound containing substructure indexes to indicate the constituent of each row.

The choice of sparse matrix or compressed compounds list still depends on cases of data usage. For frequent computing itemsets with Orange3-association library, we use function *frequent_itemsets(A, support)*, where input A presents the sparse matrix, and input support represents the support threshold.

After compressing the sparse matrix, the data still need further transformation to serve as the source of truth properly. As the major focus for the pattern exploration system is pattern and network, the data will be shaped accordingly to fit into the use cases. The basic process is as in figure 4.2:

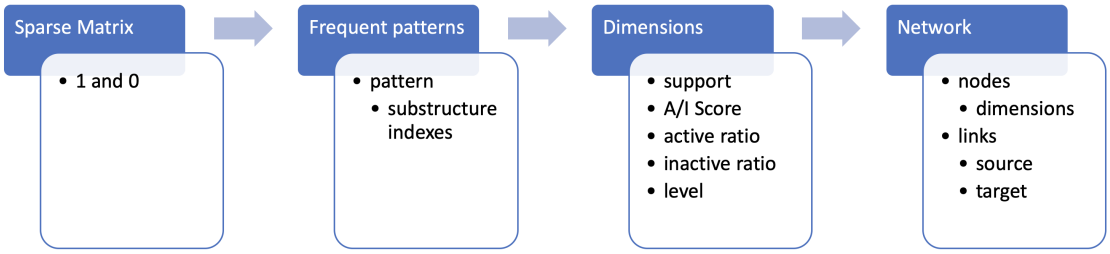


Figure 4.2. Basic process of Data Transformation

1. Convert the sparse matrix with substructure presences and absences into an array of arrays. It will serve as the input for calculating frequent patterns.

2. Generate frequent patterns with user input [support]. Each pattern will contain a combination of distinct substructures.

3. Calculate dimensions or enriched attributes for the patterns generated previously. The dimensions of each pattern are support, A/I Score, active ratio, inactive ratio, and numbers of elements in the pattern. The calculations follow the equation defined in the introduction chapter.

4. Generate network from patterns. Network data consists of nodes and links, where each node represents a pattern with dimension attributes, and each link

represents a parent-child relationship. The parent-child relationship has to be a direct ancestor and descendant. The level difference has to exactly 1. (e.g. pattern {1, 2} and pattern {1, 2, 3}). Users can provide input [A/I score] and [A/I score difference] to filter nodes and links.

4.4 Visualization Components

Visualizing components is corresponding to the Represent stage in the seven-stage design process. For a system mostly utilizing visualization to assist in the understanding of the problem, the visualization component design and implementation have a significant effect over the entire data analysis process. To design the visualization components, we need to start from understanding the problem, defining the visual language, and then putting all pieces together into the visual form. This section will cover the process of how we encode the data into visual language and the prototype and implementations of each component.

4.4.1 Visual Encoding

The definition of visual encoding is to map data values to visual representation. It is essentially a translation from numeric data, metrics into more intuitive visual languages, such as shapes, colors, and sizes. The human brain is excellent in finding connections and patterns. Once the data is wired to its visual representation, the brain can proceed with finding information or insights. However, choosing the correct form of visual cues is utterly important. The visualization should be appropriately encoded to enable users to connect and relate. (Yau, 2013)

In the book *Data Points: Visualization That Means Something* by Yau, the most common visual cues are Position, Length, Angle, Direction, Shapes, Area, Volume, Color saturation, Color Hues. While all visual cues can be used to present data in some way, their scale and accuracy of expressing quantitative information are different. (Cleveland & McGill, 1985) On the scale from accurate to generic,

the visual cues are ordered from length, slope, angle, to area, color intensity, volume, and color hue. They all vary in how precise the information is represented, which is also related to human perception.

In the pattern exploration system, the visual cues are used everywhere for the visualization component design. One of the most important visual cues used across the entire system is Color Hues representing A/I Score, as shown in figure 4.3. The reason we choose color hues to represent A/I Score is that A/I Score is one of the essential attributes of a pattern, although color hue is very generic rather than accurate. While other visual representations can represent patterns, such as different shapes, color hue can stay coherent across all other representations. The other factor is that even human eyes are not sensitive at grasping the quantitative information from color hue but are excellent distinguishing the differences. The system also reduced the A/I Score value mapping from [0, 1] to [0.3, 0.7] since A/I Score for most frequent patterns lie in this range.

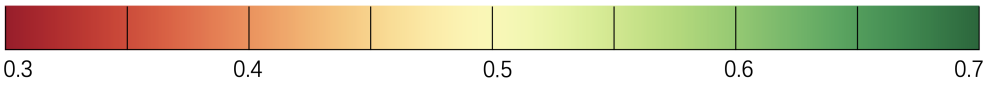


Figure 4.3. Color Hue mapping of A/I Score (Red-Yellow-Green)

Another visual encoding is applied in designing the pattern network view. As the basic unit of the visualization, each pattern is represented as a circle (shape), its radius has a linear correlation with support (area), its color is coded with A/I Score color hue mapping (color hue). See figure 4.4. The positioning of nodes in a network is related to the number of substructures inside each pattern (position). As shown in figure 4.6, each vertical cluster represents a group of patterns with the same pattern length. From left to right, the group level gradually increases one by one. In this case, patterns in the left cluster should always be the parents to patterns in the right cluster. The other factor contributing to the node positioning is the pattern correlation. Links connecting parent-child relationships will also have a force to pull

correlated parent-child nodes together. In figure 4.6, many red nodes are located at the bottom of each cluster. Users can easily spot the correlations between patterns with low A/I Score due to the links pulling these nodes closer.

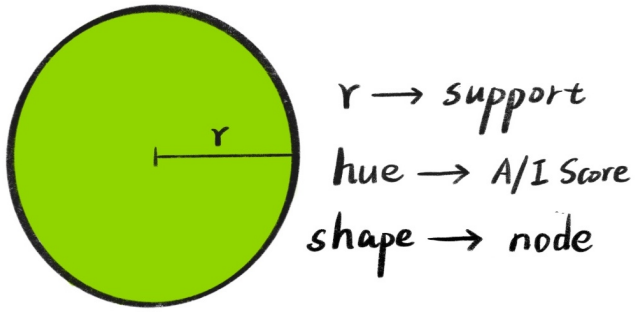


Figure 4.4. Visual Encoding for Pattern Element

The visual encoding for links is relatively simple. The link represents connection; therefore, we use color hue to encode the relationship between the parent-child nodes. Upon addition of a new substructure, if the A/I Score of child pattern significantly increases or decreases from parent pattern, the new substructure is regarded as an impact factor. Moreover, the link between them will turn green or red to indicate increasing or decreasing the A/I Score. Users can define the significance of A/I Score difference.

4.4.2 Network Visualization

One of the standard visualizations for networks is the node-link diagram. It represents the data values and their connections with different visual treatments. Nodes represent patterns, and links connecting the nodes represent their relationship. Visualizing the pattern network is essential to understand the relationship and deriving insights. With the ideas in mind, the next step is to sketch and finalize the definition of the network. There are many existing network visualization software, such as Gephi, Graphvis, and Cytoscape. However, they all

lack the capabilities of advanced customization. To have features such as unique layout and customized visual encoding, we need to build the node-link diagram with more generic tools to enable a use-case based network visualization design and implementation fully.

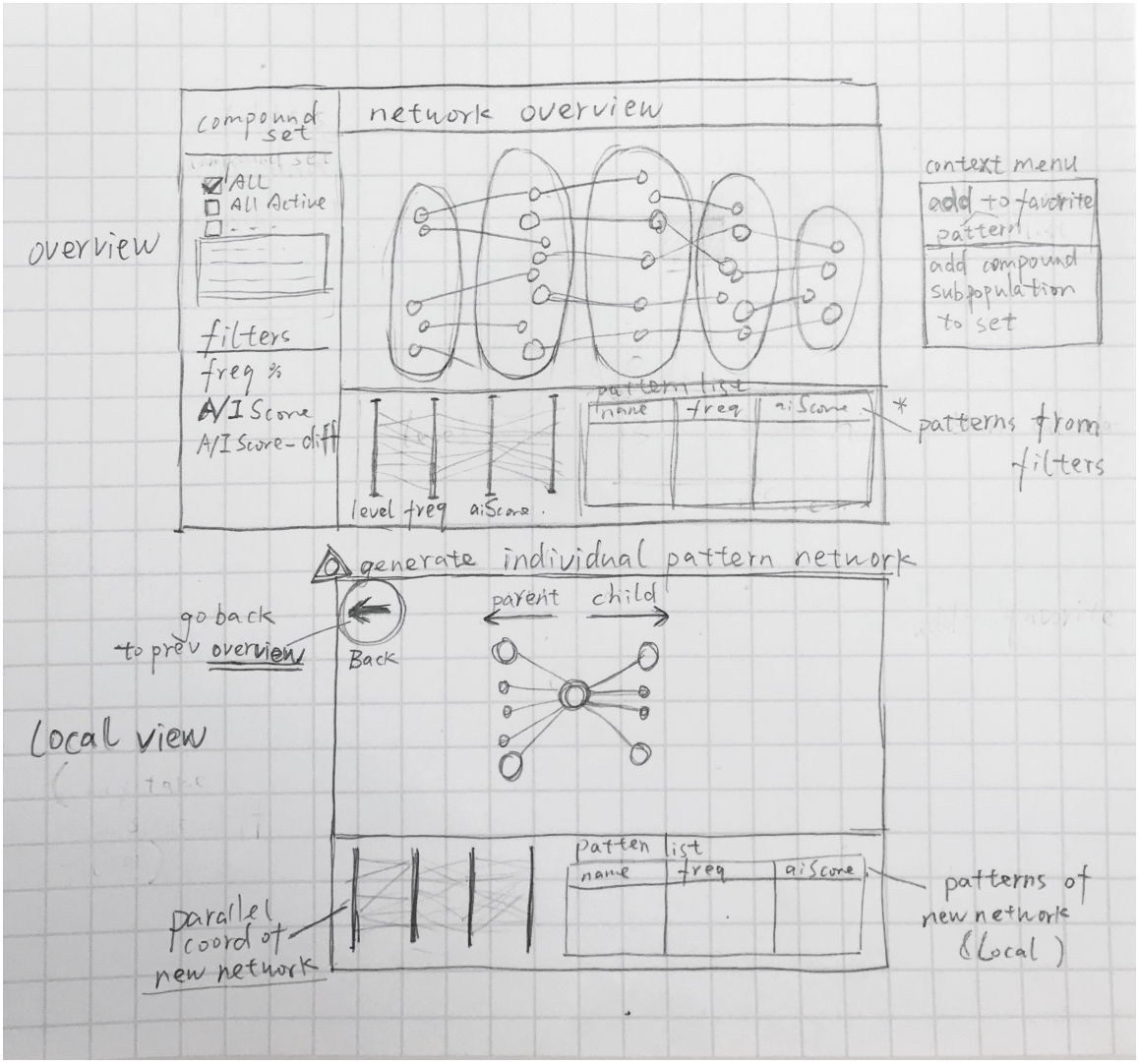


Figure 4.5. The Sketch of Network Overview and Sub-network Local view

In figure 4.5, the initial sketch of network visualization contains three parts: side panel with data source and filters, network visualization, and bottom panel with parallel coordinates and pattern list. As the filters constrain the data, the network can grow overwhelming. Therefore, the parallel coordinate is removed from

the network view and added in its tab. The pattern list is also removed from this view. The only view left is the network visualization panel.

The network visualization consists of pattern nodes and their relationship. With link connecting parent-child nodes, the network's basic layout is a sequence of vertical clusters. Each cluster represents a specific level of patterns. For example, the leftmost cluster only has patterns of 1 substructure. The cluster on its right has patterns of 2 substructures. The link does not connect grandparent/grandchild nodes for a reason to reduce the visual complexity. If a rightmost pattern is frequent, then all its parent and grandparent patterns (sub-patterns) will be frequently based on the max-pattern definition.

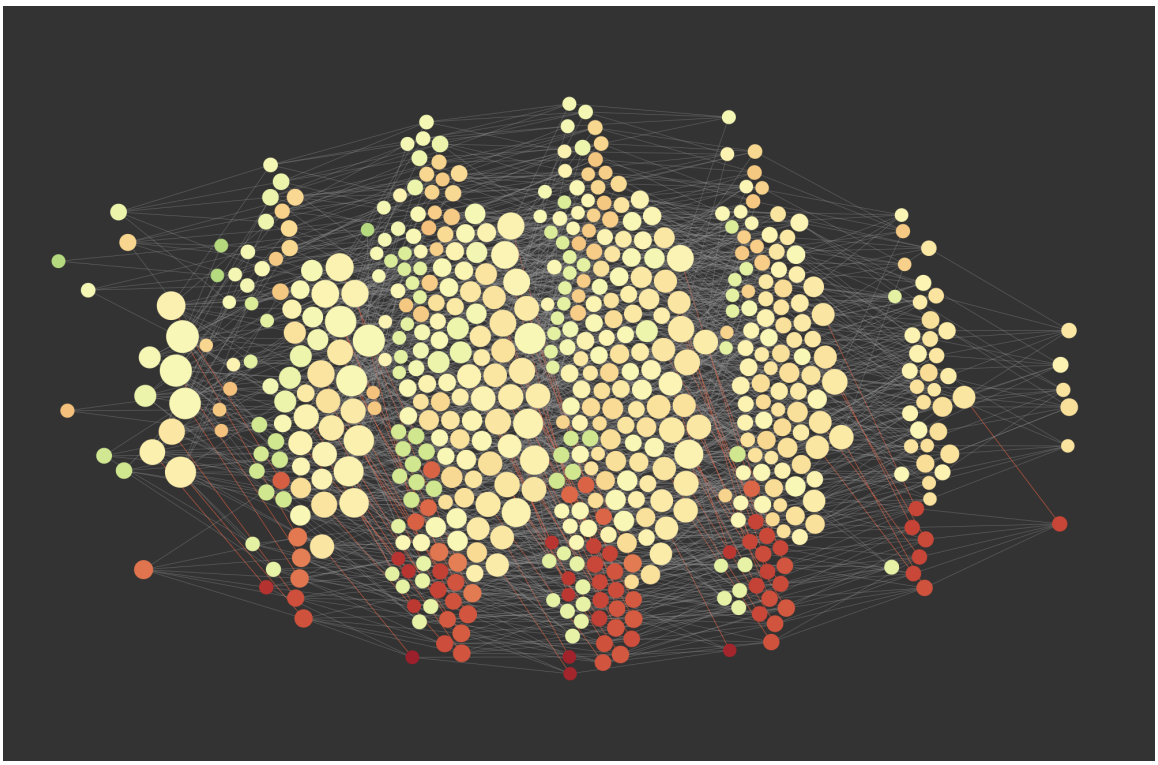


Figure 4.6. Pattern Network Overview (support ≥ 0.4)

There are two main views of pattern network visualization: network overview and sub-network sectional view. The overview provides a high-level view of the pattern network. Users can filter the data by conditions such as support and A/I

Score Difference. Then the pattern network will be displayed according to the result. The overview can be used to understand the primary network and find points of interest. However, one of the limitations is that due to the nature of graph visualization, it can be extremely complex and hard to find insights. The constrain applied on overview is less strict in general, which can lead to potential loss of details or hidden information.

To solve this issue, users can generate sub-network from a pattern of interest in the overview. Once users found a pattern and would like to explore more, they can choose to focus on the specific pattern and expand all parent and child nodes, as well as their parent and child nodes (e.g., figure 4.7). The expanding process is less constrained as the sub-network is only related to the pattern of interest and thus less complex compared with network overview. It also focused on a smaller scope rather than being universal. Users can always go back to the overview from the sub-network view.

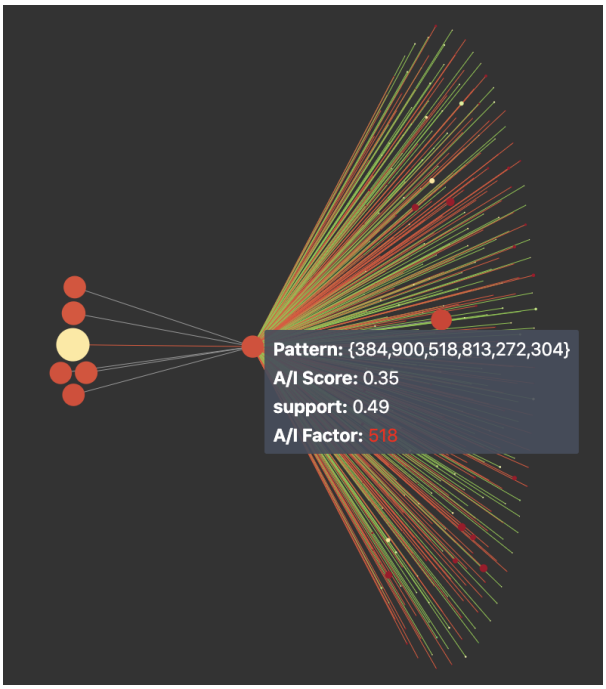


Figure 4.7. Sub-network View of pattern {384, 900, 518, 813, 272, 304}

4.4.3 Pattern Distribution

Even with the network component as the centerpiece of the system, other components are still contributing to the data analysis process. Aside from the pattern relationship, other visualization components will focus more on patterns distribution.

One of the distribution visualization components is simply a bar-chart of patterns, as shown in figure 4.8. The x-axis is level, meaning the number of substructures in a pattern. The y-axis is the number of patterns. Different color represents patterns with A/I Score < 0.5 (in red) and patterns with A/I Score ≥ 0.5. This chart quickly revealed the fact that most frequent patterns tend to present more in inactive compounds. Furthermore, mid-length frequent patterns are more dominant in amount, which also proves the combinations of different substructures can lead to more dynamic results. User input can alter the data source of the chart. For example, the figure 4.8 is based on patterns that match user input support ≥ 0.3.

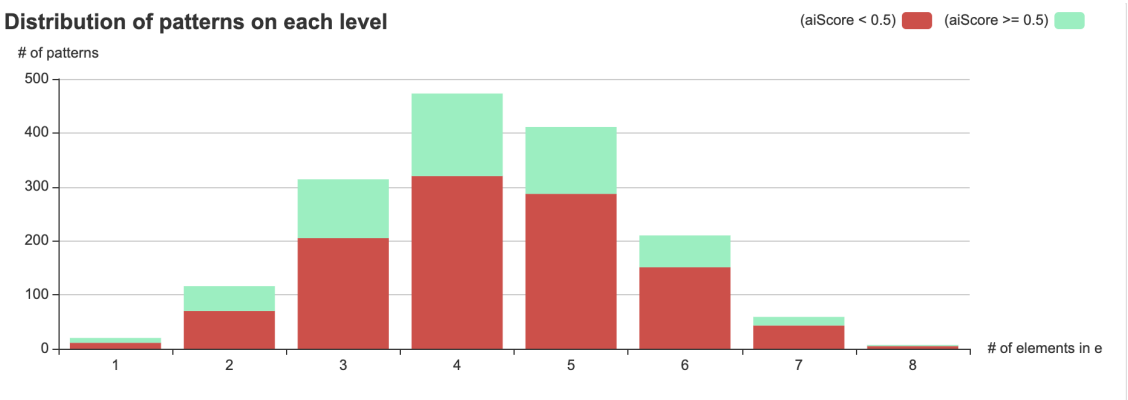


Figure 4.8. Distribution of Patterns at each level

Another component for understanding pattern distribution is a heatmap, with numbers of patterns distributed based on their support and A/I Score. Heatmap is a common visualization of representing numeric data values with colors. The most commonly used color schemes can be different color hues, saturation, or

brightness. Usually, higher data values are represented with a darker / deeper color. In figure 4.9, the distribution of patterns lies in two categories: support and A/I Score. This heatmap is generated with the entire dataset and no user input. Thus it is an overall descriptive distribution.

The x-axis is support, and the y-axis is A/I Score. The color varies among different cells. Based on support, lower support threshold contains more patterns, and while some patterns have the same support, A/I Score 0.5 seems to be a very distinct line that distributes most patterns with a wide range of support. The other fact that can be derived from the graph is that most patterns still distribute at a lower A/I Score range.

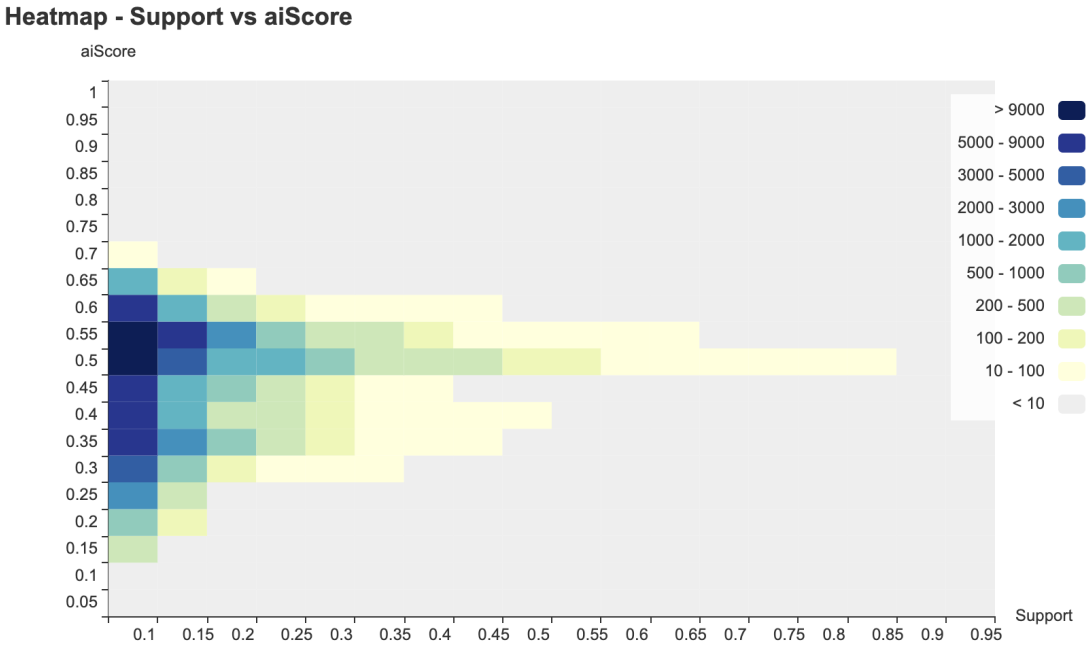


Figure 4.9. Heatmap - Support vs A/I Score

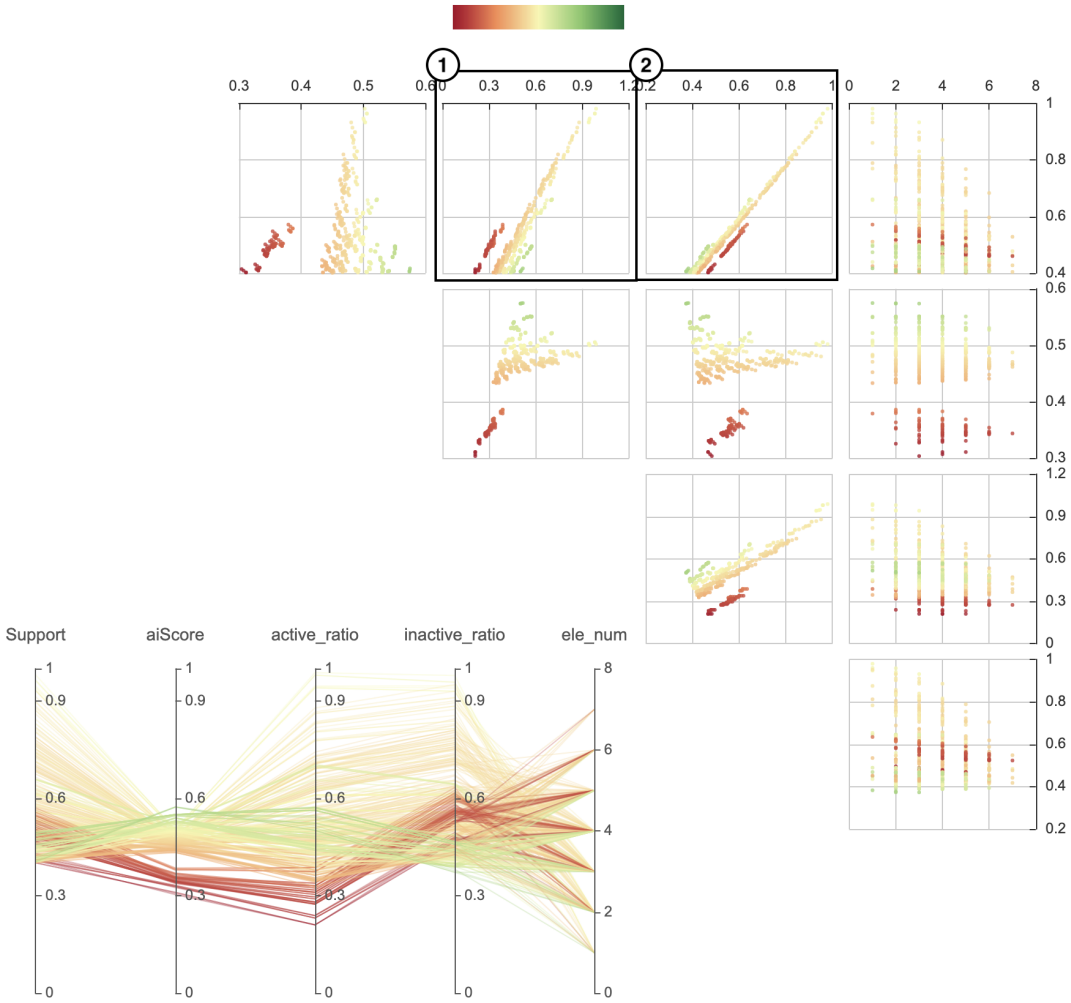


Figure 4.10. Parallel Coordinate + Scatter plot (support ≥ 0.4)

4.4.4 Parallel Coordinate + Scatter plots

Parallel Coordinate is no doubt one of the most common visualizations for multivariate data. It allows the comparison of several individual dimensions that can be numerical or categorical.(Inselberg & Dimsdale, 1991) Each line represents a data entry, and the points of a line represent the attributes. In this system, we first parse the sparse matrix into a more diverse list of patterns with multiple dimensions such as support, A/I Score, and level. Then these patterns can easily be presented as lines on the parallel coordinate.

In addition to the parallel coordinate, the scatter plot can also support the understanding of correlations between different dimensions. Each plot has a different combination of the y-axis and the x-axis. The color is encoded with A/I Score of a pattern, corresponding to the network view.

4.5 Filters

The design of filters is corresponding to the Refine stage of the process. The filter is the most crucial part of generating data for analysis. To understand our problem with a broader scope and analyze patterns in different conditions, we need to enable data filtering based on various use cases. The data obtained after the filter will update across all component views.

Support Support is one of the key user inputs to filter patterns. If the support of a pattern is larger than the input, the pattern node will show.

A/I Score Range A/I Score range is for filtering pattern node. If a node is in the user input range, the nodes will show.

A/I Score Difference A/I Score Difference is for filtering links. The filter will check the A/I Score on both parent pattern and child pattern, if the difference is higher than the user input, then show the links, as well as nodes on both sides regardless of whether they match other conditions.

A/I Score Difference threshold (color encoding) Even with the same name, this filter is very different than A/I Score Difference. It is only used to color encode network links. For a link connecting parent-child nodes, if the difference is higher than the threshold, then color the link. If the child node has a higher A/I Score than the parent node, the link will be green; Otherwise, the link will be red. The threshold is also the filter for calculating the impact factor. All links highlighted will be the data source for the amount of impact factor transforming pattern to be more active or inactive. (higher or lower A/I Score)

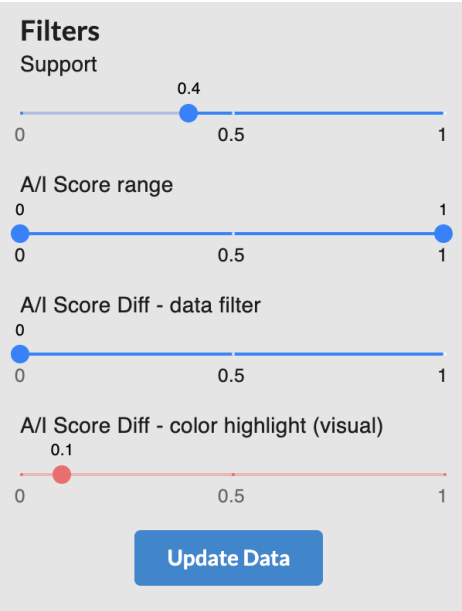


Figure 4.11. Filter example - Side panel

4.6 Interaction

Interaction for a visualization system is extremely crucial. A static image such as infographics can provide a certain amount of knowledge. However, visual analysis is more than simply reading and deriving insights from a single picture. It is an iterative and repetitive process of view creation, exploration, and refinement. The analysis should allow users to derive insights from a dynamic exploration process. (Heer & Shneiderman, 2012) The taxonomy of interactive dynamics is proposed by Heer and Shneiderman. It provides a guideline for evaluating and creating a visual analysis tool, and it will be utilized to guide the interaction design and implementation of this pattern exploration system. There are three categories in the taxonomy, each with multiple sub-categories. However, this section will only cover the interactions that fall into the taxonomy definitions. We will list all interactions in the pattern exploration system, and then fill the taxonomy category with the interactions or components for evaluation.

- Data & View Specification: visualize, filter, derive

- View Manipulation: select, navigate, coordinate, organize

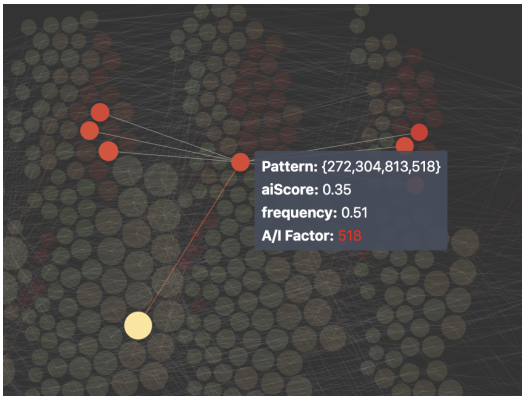
4.6.1 Network Interactions

Filter : Filter can be a part of the interaction flow. Users usually need to choose a set of filters or simply use the default filter to get network data. Then the network will update itself according to the filter. The overview filter includes support, A/I Score Range, A/I Score Difference, and A/I Score Difference color encoding. The sub-network view also includes most of the filters except for A/I Score Range.

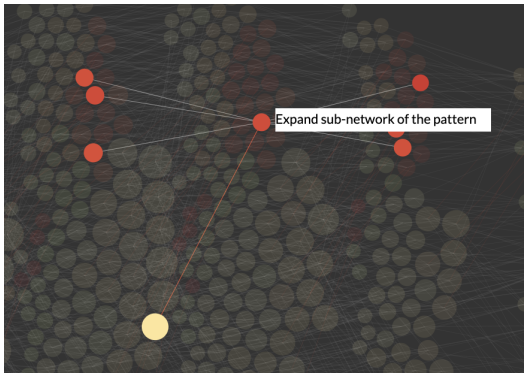
Hover : Hover is mostly used to show additional information, such as tooltip. When users hover on a pattern node, the tooltip will show. The tooltip will include the pattern's substructures, support, and A/I Score. If the hovered pattern is a child end of a colored link, the tooltip will also show impact factor (substructure) id. The impact factor will have the same color with a link to indicate whether the impact factor makes the child pattern more active. If there is more than one colored link connected, the tooltip will show a list of all impact factors. The hover also highlights the pattern's direct parent and child nodes, as well as the connecting links.

Click : Click is to highlight all nodes connecting the selected pattern. Aside from direct parent and child nodes, it will also highlight the entire corresponding sub-patterns and super-patterns. Thus, the click interaction can show the local network of a selected pattern in the overview network.

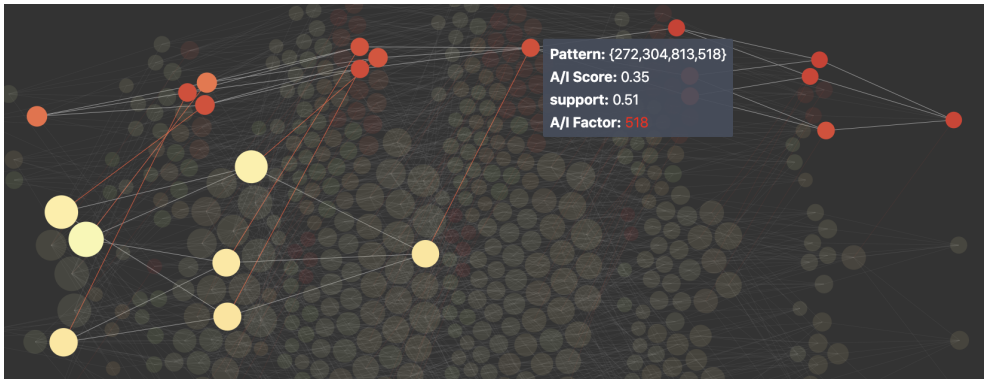
Right-click : Right-click is used to expand the sub-network of the selected pattern. It will show a drop-down with the "expand sub-network" action. Once users clicked on the expand action, a new sub-network view of the selected pattern will replace the network overview. The selected pattern will become the center of the network. Any network expansion direction will be based on whether



(a) Hover to show Tooltip and highlight



(b) Right-Click to expand



(c) Click to highlight connected pattern network

Figure 4.12. Network View Interaction examples

they are on the parent side (left) or the child side (right). Other overview interactions still apply in the sub-network view.

double-click : Double-click only applies to the sub-network view. It is to expand all possible direct parent nodes or direct child nodes of the selected pattern. If the selected pattern is the parent of the center node, it will expand the possible parents of the selected pattern. Otherwise, it will expand children of the selected pattern.

Background click : Background click is to restore the graph to the initial state. When a user clicks on background, all highlighted nodes and links will be cleared and restored to their initial state, along with all other visual

components. This interaction is necessary for a visualization component as it provides users a flexible interaction flow.

4.6.2 Parallel Coordinate and others

Brush-select : Brush-select will highlight all pattern lines that fall in the selected range. The selection also applies to scatter plots. Users can brush on any axis of parallel coordinate to highlight data of interest. To cancel the brush selection, simply click on the brush-selected axis.

Hover : Hover on other visualization components are very similar to network interaction. In scatter-plot, users can hover on pattern dot to see the pattern substructures, x-axis attribute value, and y-axis attribute value. In pattern distribution, users can hover to see the corresponding data value of selected visual elements.

4.6.3 Components and Interaction evaluation

visualize Visualizing data by choosing visual encoding. It includes the A/I Score Difference color encoding filter.

filter Filter out data to focus on relevant items. It includes the entire filter interactions on the side panel.

derive Derive values or models from source data. It includes the statistical distribution visualization components derived from filtered data.

select Select items to highlight, filter, or manipulate. It includes hover and click (select) across different visualization components to highlight and also visually filter patterns of interest.

navigate Navigate to examine high-level patterns and low-level details. It includes hover on an individual pattern to get details, also the network overview and

sub-network switching interaction. (right-click on a pattern, click on the back button)

coordinate Coordinate views for linked, multi-dimensional exploration. It includes parallel coordinate brush-select highlight links and also dots on scatter plots.

organize Organize multiple windows and workspaces. It includes different views and tabs for users to switch back and forth.

4.7 Summary

This chapter has covered the design and interaction process, guided by the seven-stage visualization design process and interactive dynamics taxonomy. Next, we will dive deep into various use cases to better understand how this system can solve our problems and answer the questions: what are the patterns, and what is their network?

CHAPTER 5. USE CASES ANALYSIS

To truly understand how this pattern exploration system can answer the questions raised initially, we need to revisit the problem statement. The main goal of this research study is to understand how data visualization can help to explore and analyze patterns and their network in a complex chemical compound dataset. This chapter will demonstrate the pattern analysis process, also address the role of data visualization, and how does it assist the pattern analysis process. Many use-cases will guide the process of pattern analysis, network generation, and system interaction.

5.1 Pattern Exploration

The flow of user interacting with the system usually start with filtering data. The system has a set of default filter:

- support ≥ 0.4
- A/I Score $\in [0, 1]$
- A/I Score Difference ≥ 0
- A/I Score Difference (color encoding) ≥ 0.1

Aside from the default input, users can update the filters at any point of analysis. The filter update will lead to the change of data that feeds into all visualization views. The filter is an essential part of the system. It serves as the generation tool for the source of truth throughout the entire pattern analysis process.

To start analyzing a set of patterns, users can update filters or use default filters to generate distinct patterns. Each pattern will include a set of substructures,

as well as attributes such as support and A/I Score. For example, in figure 5.1, with the default filters, the network is generated with the filtered data. Seven vertical clusters showed in the view; each represents a level of pattern length (from 1 to 7) and has link connections between each other. The connection will apply a force to push correlated nodes closer.

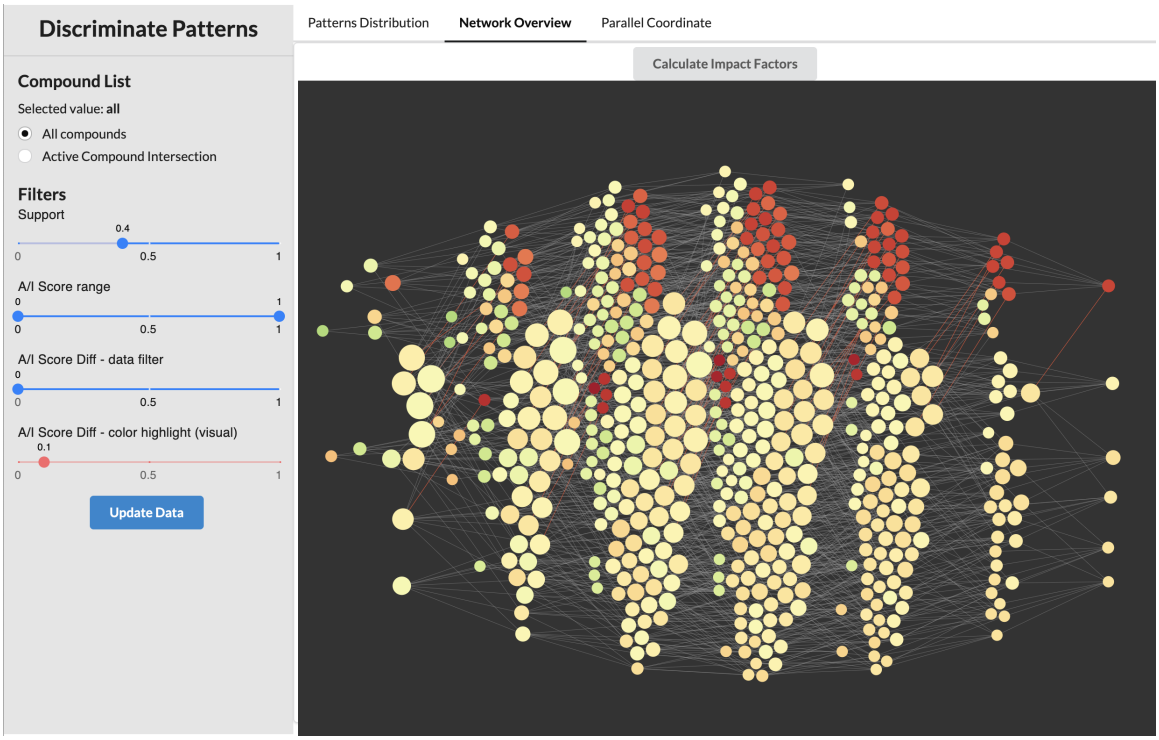
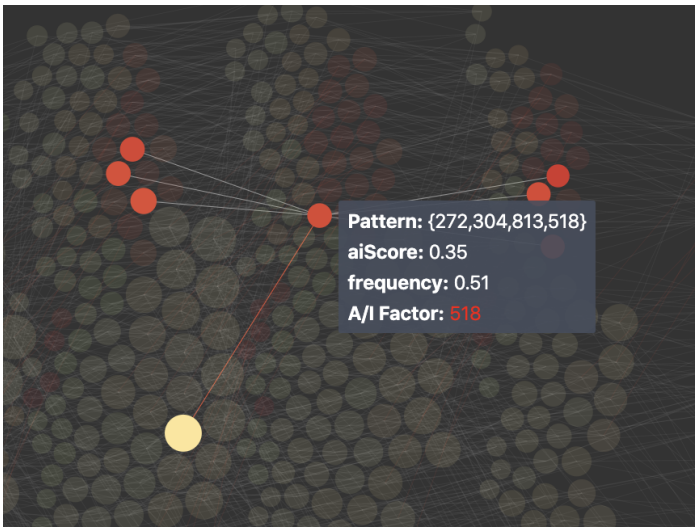


Figure 5.1. System Overview - Network tab

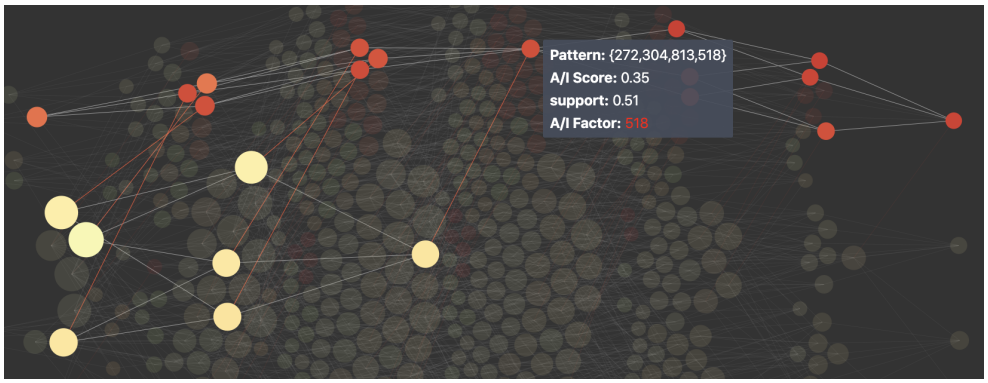
Users can observe the majority of patterns are nearly yellow-colored (A/I Score around 0.5). However, many red nodes (low A/I Score) distributed around the same height in the graph; they all fall in different vertical clusters (level) with horizontal links connecting each other. Similar to these red nodes, several light-green nodes (high A/I Score) also distributed slightly lower than or near the red nodes. The positioning implies a tendency that the correlated patterns will locate closer to each other due to the force of links.

From the figure, many links are colored red based on the A/I Score Difference (color encoding) filter input. The color highlight on the connection

implies that the A/I Score difference between parent and child nodes has surpassed the given color threshold. This color threshold is different from the A/I Score difference data filter as the data filter modifies data source. Updating the data filter will lead to a complete refresh of the network view. On the contrary, the A/I Score Difference color filter serves only as a visual treatment of links, thus merely updating link color in the graph. It does not alter the data underneath.



(a) Hover on Pattern {272, 304, 813, 518}



(b) Click on Pattern {272, 304, 813, 518}

Figure 5.2. Example of Interaction with a pattern

Users can explore and investigate patterns of interest by interacting with the network, such as hovering or clicking the pattern nodes. A tooltip will pop up when

hovering on a pattern node, with detail attributes of the pattern such as pattern substructures, A/I Score, support and impact factor (if the pattern is the child end of a colored link). It will also highlight all direct parent nodes, child nodes, and links in between. Users can keep hovering and unhovering to discover insights. For example, user is hovering on the pattern {272, 304, 813, 518} in figure 5.2(a). The A/I Score of this pattern is 0.35, and support is 0.51. The node color is red due to a low A/I Score. Most of its connected parents and children are also red, except for a yellow node on the left bottom corner in the figure. Users can notice the link between the hovered node and the yellow nodes is also red.

Users can highlight the sub-network of the pattern {272, 304, 813, 518}, and keep highlighted state of its connected nodes by clicking on the pattern, as shown in figure 5.2(b). Users can observe that most connected nodes are red in this highlighted sub-network. Many yellow nodes also connect with the nodes in this network, with red links in between. It indicates a drastic decrease of A/I Score upon the addition of some substructure. If users hover on the yellow node connected to the selected pattern {272, 304, 813, 518}, the node will be {272, 304, 813}. The tooltip in figure 5.2(a) also shows an impact factor "518" highlighted in red. If users hover on those yellow patterns and their child nodes connected with red links, most tooltips will also show "518" as an impact factor for the colored links.

Back to the overview network, if users try to hover and click on those red nodes and investigate, users can discover that most of these red nodes are connected by a common substructure: 518. As a pattern, {518} is actually in the graph on the leftmost cluster. If users click on the pattern {518}, all of its corresponding sub-network will be highlighted in the graph. Users can discover that almost every super-pattern of {518} has a low A/I Score by red color. The presence of "518" and the clustering of low A/I Score seem to have a strong correlation. In addition to the previous observation on the red links and their connected nodes, substructure "518" seems to have affected activeness of many patterns.

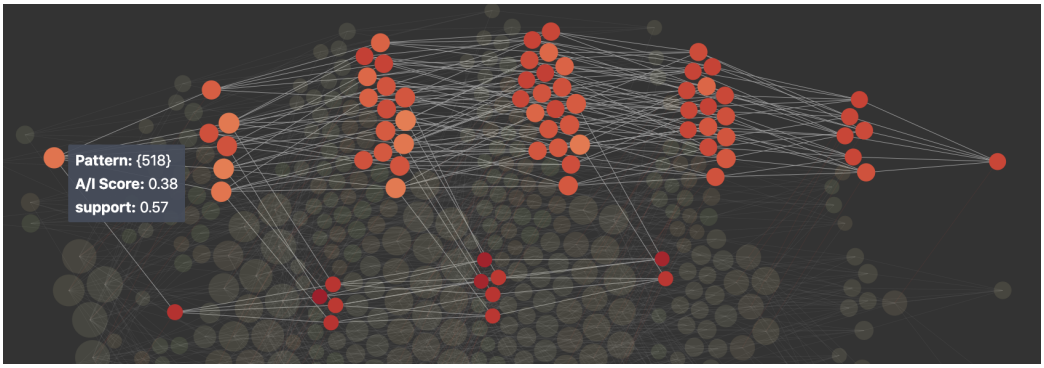


Figure 5.3. Click on pattern {518}

Similarly, users can observe the patterns with a high A/I Score. The support filter is ≥ 0.3 instead of the default 0.4 to expand the pattern population. From the network generated, users can observe many green colored links is associated with the addition of substructure 275. For example, in figure 5.4, if user click on pattern {272, 384, 275, 900, 790} and highlight the corresponding sub-network, user can observe that most connected nodes are light green. Many yellow nodes also connect to these green nodes, with green links in between. This indicates the increase of A/I Score upon addition of substructure 275, as shown in the tooltip. If users investigate other colored links in this highlighted sub-network, substructure 275 is marked as an impact factor for most of the links.

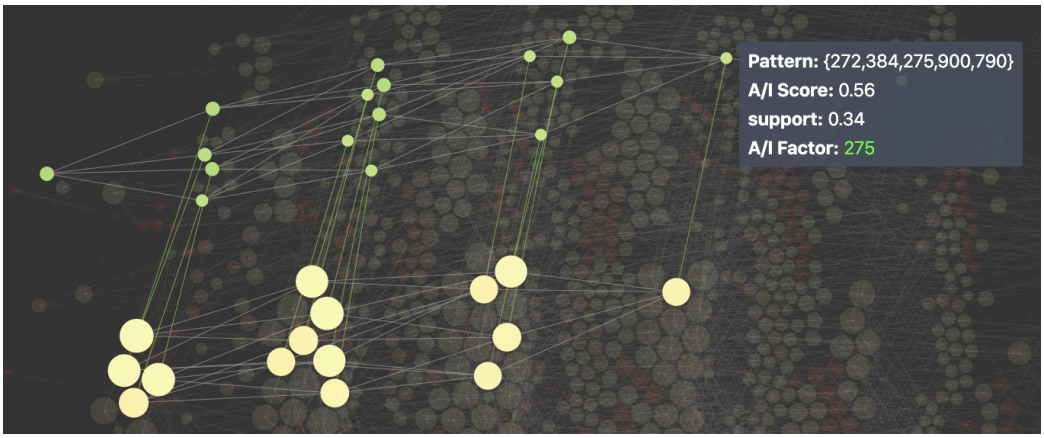


Figure 5.4. Sub-network View of {272, 384, 275, 900, 790}

The example above is not the only use case that shows the sign that substructure 275 might affect the activity of patterns. If users investigate most green nodes, 275 seems to be present in most of them. In figure 5.5, by clicking on pattern {275}, users can discover that almost every super-pattern of {275} has a high A/I Score by green color. In addition to the previous observation on the green links and connected nodes, substructure 275 seems to have affected the activeness of many patterns.

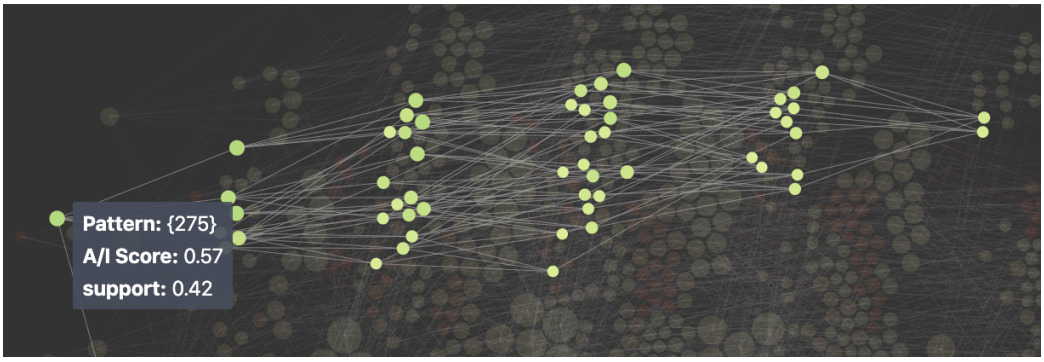


Figure 5.5. Click on pattern {275}

Therefore, users can derive insights from exploration and observation. Substructure 518 could be a significant negative impact factor among substructure interaction:

1. Most patterns with substructure 518 have low A/I Score, meaning they tend to present more in inactive compounds with a higher ratio.

2. Most patterns with substructure 518 have neutral parent patterns; their connected red links implied a drastic decrease in A/I Score. The direct cause of the A/I Scores decrease is the addition of substructure 518.

3. Most patterns affected have high support (≥ 0.4). High support also implies the significance of the patterns and substructure 518.

Similarly, substructure 275 could be a significant positive impact factor among substructure interaction:

1. Most patterns with substructure 275 have a high A/I Score, meaning they tend to present more in active compounds. It can be verified by calculating the presence ratio of substructure 275 in active/inactive compounds. The active ratio is 0.52, while the inactive ratio is 0.38.

2. The connected green links between neutral patterns and green patterns implied a drastic increase in A/I Score. The direct cause of the A/I Scores increase is the addition of substructure 275.

3. Most patterns affected have high support (≥ 0.3). High support also implies the significance of the patterns and substructure 275.

Following a similar exploration path in the system, users can discover and derive insights about other pattern relationships and substructure interaction.

5.2 Active Compound Intersection

The other dataset used in the system is from only active compounds. The algorithm calculates the intersections among active compounds; each intersection represents a pattern. Then it adds up the total presences of the patterns among all compounds, referred to as occurrence. As many long patterns present only in a few compounds, the algorithm filters out the patterns with occurrence ≤ 2 to eliminate rare cases and increase the performance. The patterns generated from active compounds intersection serve as references where users can compare the patterns and impact factors between all compounds and only active compounds. It has less constraint than the main network overview as the dataset only focused on patterns in active compounds.

In figure 5.6, users can observe that most patterns in this dataset are more active or neutral, based on the main color tone of the graph (green and yellow). Many links are colored, implying the A/I Score change is (≥ 0.1). In this active compound intersection dataset, a large number of colored links have been detected

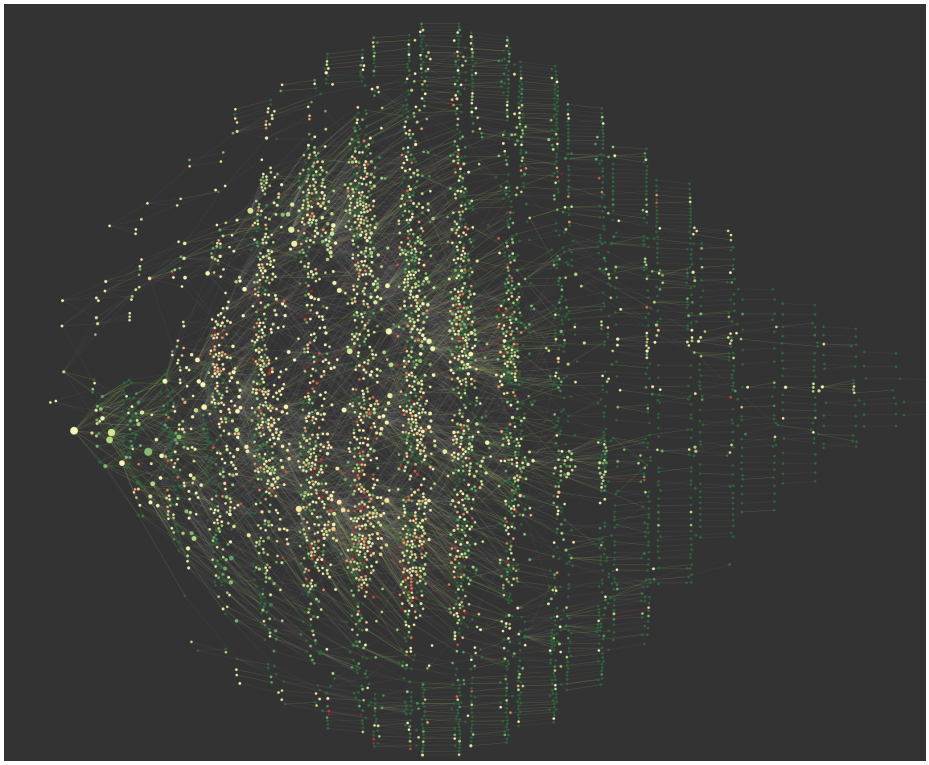


Figure 5.6. Patterns from Active compound intersection

compared with network overview, as the intersection dataset is less constrained on support. Thus, many corner cases could be discovered.

Same with network overview, users can hover and click on pattern to see details and corresponding sub-network. As the patterns are intersections of compounds, sometimes a pattern may have a grandparent/grandchild but no related parent/child. If users click on a pattern, the graph will still calculate and highlight all of its super-patterns and sub-patterns, even without links connecting them.

Users can observe the corresponding sub-network for selected patterns and derive insights. For example, in figure 5.7, the patterns related to {275} are mostly green, showing a strong tendency that these super-patterns of {275} has high A/I Score. In contrast, the corresponding sub-network in figure 5.8 shows the opposite, where most related patterns of {272, 900, 518} are red or yellow. These pattern distributions reflect the insights we generated earlier: substructure 275 mostly

Figure 5.7. Corresponding sub-network for pattern {275}

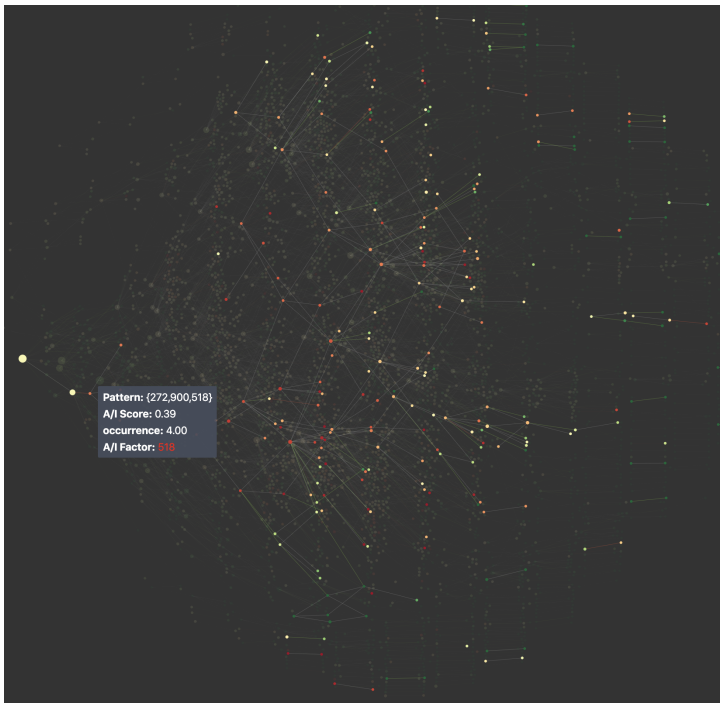


Figure 5.8. Corresponding sub-network for pattern {272, 900, 518}

interacts with high A/I Score patterns by increasing A/I Score upon addition. While substructure 518 is mostly present in low A/I Score patterns.

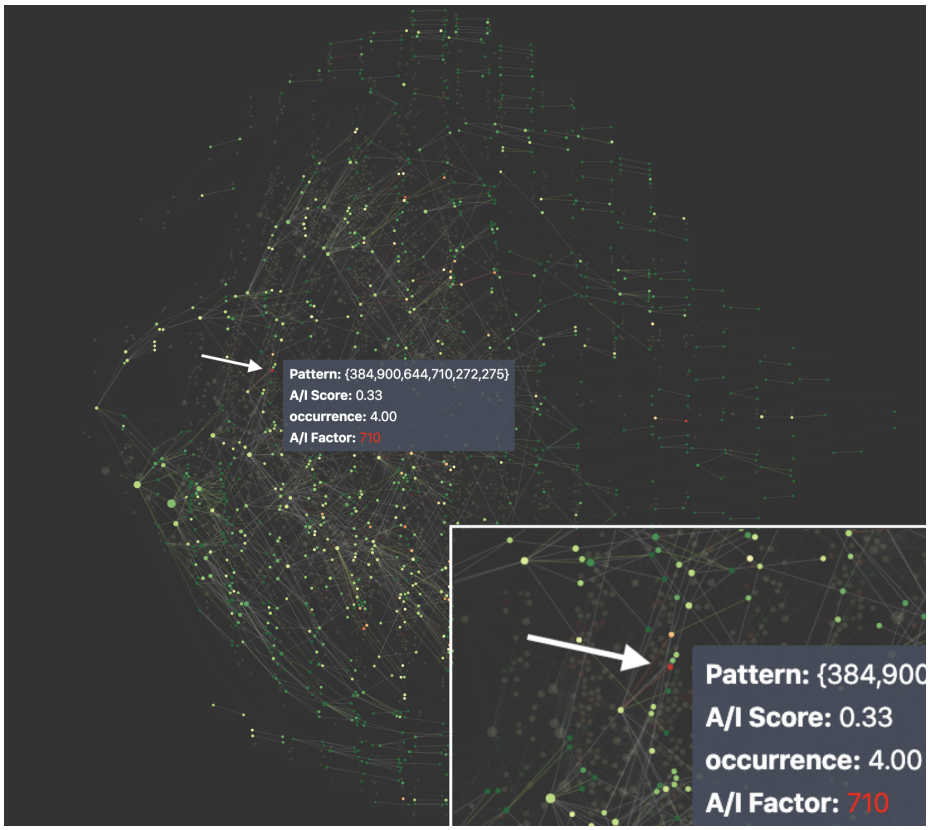


Figure 5.9. Pattern {384, 900, 644, 710, 272, 275}, 710 as negative impact factor

While the pattern network of an impact factor could be dominantly positive (green) or negative (red), users can still find many patterns and links with the opposite behavior. For example, figure 5.9 shows that even substructure 275 affects A/I Score of many patterns to be higher; some substructure can still alter its impact in the opposite direction. Users can hover on the red patterns and see what the negative impact factors are. In figure 5.9, users can hover on one of the red pattern {384, 900, 644, 710, 272, 275} in the highlighted sub-network and discover substructure 710 as an impact factor. In this graph, users can detect 710 more than once. Users can investigate on substructure 710 further to find the correlated patterns and their interactions with other substructures.

Figure 5.10. Pattern where substructure 518 and 275 co-exist

So far, two of the most frequent impact factors have been discovered: substructure 275 and 518. As they have conflicting impacts on patterns, what will happen if they are both present in the same pattern? To answer this question, users can investigate patterns where multiple conflicting impact factors coexist. For example, users can discover a pattern where both substructure 275 and 518 are present, as shown in figure 5.10. From the figure, we can see the pattern has two parent patterns with colored links in-between. The red parent has substructure 518, and the green parent has substructure 275. While both of them have relatively low or high A/I Score, their common child pattern's A/I Score is much closer to 0.5 and has a yellow color. Other patterns with the presence of both 275 and 518 also tend to have neutral A/I Score. Therefore, users can derive the possible insight that the A/I Score of a pattern could be neutralized by the presence of conflicting impact factors such as 518 and 275.

However, with the A/I Score neutralized, the pattern can still be affected by other impact factors. In figure 5.11, even the pattern {384, 900, 933 ... 272, 304, 275} has been neutralized by the presence of both 518 and 275, the A/I Score still increased/decreased from parent to child, upon addition of substructure 552 or 1002. Users can further investigate these impact factors and see how they affect other patterns and the interactions between other substructures.

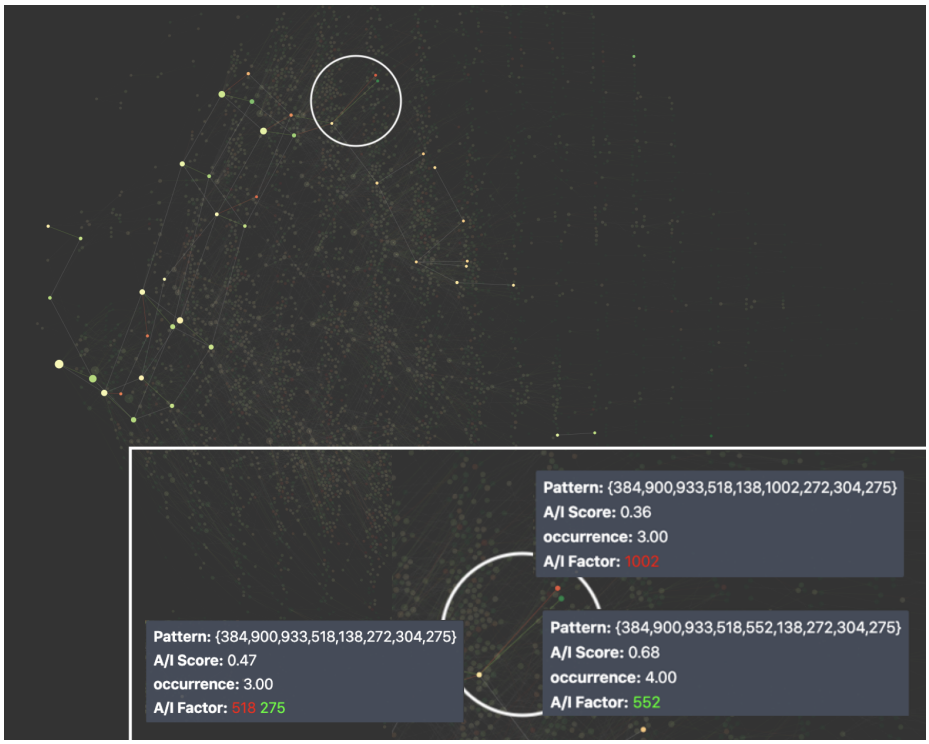


Figure 5.11. Other impact factors affecting neutralized pattern

Another discovery is that from selecting a pattern to explorer the corresponding sub-network, many patterns tend to be neutral, meaning that they have A/I Score close to 0.5 and high occurrence. It implies that the substructures associated, such as 272 and 900, are less likely to affect the activity of a pattern. If users try to calculate impact factors from a network, for example, the active compound intersection dataset, even with a very high occurrence, substructure 272 has only been detected as an impact factor once. While some significant impact factor, such as substructure 275, has been detected in 82 green links (positive

impact). Users can later categorize these substructures or patterns to narrow down the scope while searching for patterns of interest.

5.3 Calculate Impact factors

The other method to find the impact factors is to calculate them from the given network data. Compared with the usual exploration method, where users need to find colored links and summarize the impact factors manually, calculating impact factors will go through the entire network data, locate the colored links, and find the impact factors. The output of the calculation consists of two arrays: positive factors and negative factors. Each array has substructure indices ordered by the occurrence, which is the number of colored links related to the impact factor in the given network data. Each substructure will also have its general presence ratio in both active compounds and inactive compounds.

Users can calculate Impact Factors after the pattern network is generated from given filters. For example, in the network filtered by support ≥ 0.3 and A/I Score Difference color threshold ≥ 0.06, positive and negative impact factors are calculated as shown in table 5.1 and 5.2. Users can discover that substructure 275 might have a positive impact on many patterns and increased their A/I Score. It also has a very high presence ratio in active compounds (0.52) than inactive compounds (0.38). On the other hand, substructure 518 might have a negative impact and decreased A/I Score for many patterns. It has a high presence ratio in inactive compounds (0.63) compared with ratio active compounds (0.39). This observation also matches the insights derived from manual pattern exploration. Both methods can coordinate and assist each other for users to find patterns of interest and investigate the substructure interaction and impact.

As the output of calculation relies on the input data and A/I Score difference color threshold filter, users can update data sources to compare and derived insights from different results. For example, table 5.3 and 5.4 are the results generated with

Table 5.1. *Positive Impact Factor (in network overview)*

substructure	occurrence (in network)	active ratio	inactive ratio
275	54	0.52	0.38
138	8	0.57	0.47

Table 5.2. *Negative Impact Factor (in network overview)*

substructure	occurrence (in network)	active ratio	inactive ratio
518	255	0.39	0.63
271	39	0.35	0.45
654	15	0.19	0.36
722	11	0.18	0.37

Table 5.3. *Positive Impact Factor (in active compound intersection dataset)*

substructure	occurrence (in network)	active ratio	inactive ratio
275	82	0.52	0.38
544	73	0.24	0.15
740	43	0.26	0.15
...	...	...	...

Table 5.4. *Negative Impact Factor (in active compound intersection dataset)*

substructure	occurrence (in network)	active ratio	inactive ratio
518	87	0.39	0.63
271	23	0.35	0.45
651	9	0.14	0.25
654	7	0.19	0.36
...	...	...	...

active compound intersection dataset, with A/I Score Difference threshold ≥ 0.1. From the table, users can discover that substructure 275 and substructure 518 are still the most significant impact factors. Substructure 271 and 654 were also present in the negative impact factors table from the earlier use case (table 5.2).

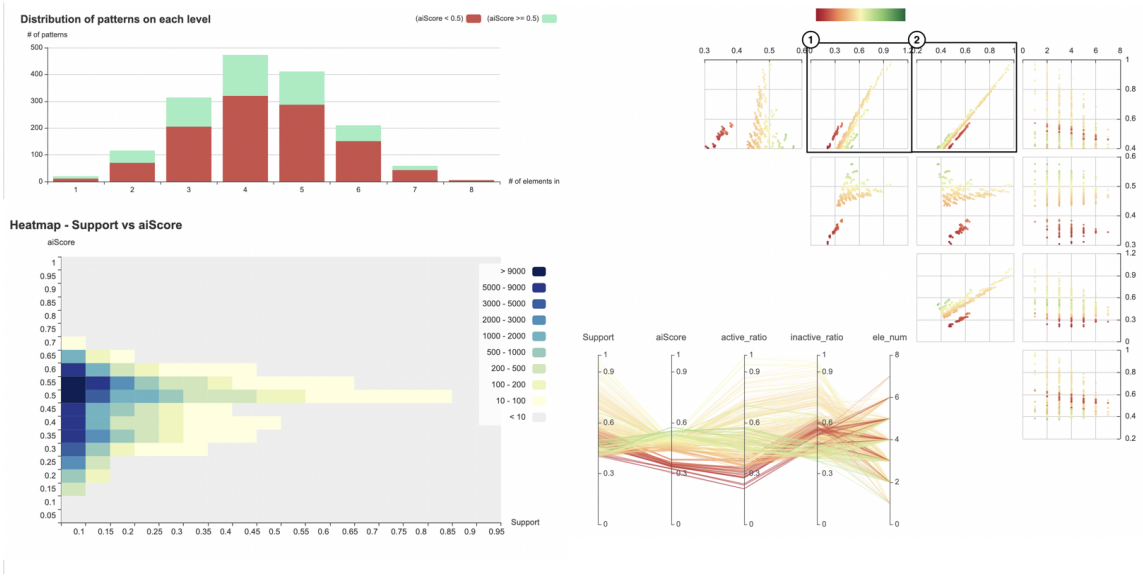


Figure 5.12. Other Visualization elements

5.4 Insights from other visualizations

Users can also generate many insights from other basic visualizations. For example, in the parallel coordinate + scatter plot view, each scatter plot has a different combination of the y-axis and the x-axis. As shown in figure 5.12, plot 1 and 2, support is linear with active ratio and inactive ratio. However, the support - inactive ratio relationship is much closer to linear (plot 2), while the support - active ratio relationship is more disperse (plot 1). There are visually significant layers separating patterns into different groups: patterns with low A/I Score (red dot line), patterns with A/I Score close to 0.5, and a wide range of support, patterns with higher A/I Score (green dot line). Users can hover on the pattern dots to further investigate.

In the graph of pattern distribution bar-chart, users can observe that most patterns have medium sizes, such as size 4 or 5. Also, from the heatmap of support-A/I Score, users can observe that most patterns distribute in the small support range based on the darker color, such as 0 to 0.1. Moreover, most patterns

tend to have neutral A/I Score (close to 0.5), but more patterns have A/I Score smaller than 0.5 from the dense colored grids below y (A/I Score) = 0.5.

CHAPTER 6. DISCUSSION AND FUTURE WORK

6.1 Advantages of Visual Analytics

Visual analytics was created to solve problems unsolvable with automatic or visual analysis alone (Keim, Mansmann, & Thomas, 2010). In the paper by Keim et al., it describes the comparison and combination of traditional automatic analysis and visualization. Automatic analysis is best at solving analytical problems that can be approached by quantitative methods. It is usually preferred than interactive visual analysis because user behavior can be costly and unpredictable. However, automatic analysis cannot solve some real-world problems that are vaguely defined. Some of the problems also require dynamic adaption of the analysis process. On the other hand, visualization methods apply human knowledge and intuition. They are best for small datasets but could fail as the scale goes too large.

The goal of visual analytics is to combine the strength of both methods: utilize computational efficiency from automatic analysis while also integrate human intuition and knowledge. Compared with the typical automated or data mining approach, the advantages of this visual analysis system for pattern exploration are listed below.

First, the initial problem we are trying to solve was vaguely defined. As the goal is to find a sub-optimal solution and derive potential insights, automated algorithms might not give precise results. To understand the patterns, networks, and substructure impacts, users have to rely more on the explorative analysis rather than purely automated analysis. The focus of the visual analytics in this system is exploration, during which users can utilize their knowledge and intuition to derive insights and refine the solution.

Second, visual analytics in this system can provide more contexts during the analysis process. For example, users can calculate impact factors and get the list of substructures, including their occurrence, ratio in active/inactive compounds. Compared with this approach, users can also derive similar insights from exploring the pattern networks, but with more abundant contexts. Users can discover that 275 could be a positive impact factor from both automated calculation and manual observation. However, the visualization also provides contexts to support further discovery, such as the patterns affected by impact factors, the relationships between the patterns, and the sub-network of the selected pattern. Thus, users can utilize their intuition to discover more insightful knowledge.

Third, visual analysis in this system allows the dynamic adaption of the analysis process. The analysis process needs to be dynamic and adjustable to accommodate users' goals. In this system, users can alter data source feeding into the visualization and dynamically interact with the system at any given time based on the use cases. Users can switch between any stages of the analysis process, including update filters to modify data, choose patterns of interest and investigate, switch between network overview and sub-network.

To summarize, Visual analytics combined both efficiency and effectiveness for solving large-scale and complex problems. In this research study, it utilized the benefits of visual analytics and better-served users to derive insights through the pattern exploration system.

6.2 Conclusion

In this research study, a pattern exploration system utilized data visualization to support analyzing and understanding patterns and how substructure interaction affects the activity of chemical compounds. The visualization views consist of network visualization, pattern distribution, and parallel coordinate. The system contributes to assisting data analysts in extracting

patterns from the data, filter patterns to locate points of interest, and understand how substructure interaction can activate or deactivate the compound. Through the system, users can discover insights through a visual exploration of patterns and their network. For example:

- Most patterns with substructure 518 have a low A/I Score (red)
- Most patterns with substructure 275 have a high A/I Score (green)
- These impact factors can drastically affect A/I Score of patterns (e.g., 518 – negative, 275 – positive)
- There are many neutral substructures that do not affect the A/I Score, but their related patterns could be affected by other impact factors (e.g., substructure 272, 900)
- Pattern A/I Score could be neutralized when multiple conflicting impact factors co-exist. (e.g. 275 + 518)
- Patterns whose A/I Score has been neutralized can still be affected by other impact factors. (e.g., 275 + 581 co-exist, A/I Score can still drastically increase or decrease upon addition of 552 or 1002)

This work enabled users to analyze patterns and substructure interactions in compounds and bridge the gap between statistical approach and visual analysis. The insights users generated by interacting with the system can also assist or suggest future exploration. For example:

- Users have discovered 518 and 275 as strong impact factors. Following the same exploration path, users could find other impact factors.
- Users can categorize the substructures based on their impact. For example, substructure 272 and 900 could be categorized as neutral factors.

- Users can highlight the correlated sub-network of a known impact factor (positive or negative) and investigate the nodes/links that have opposite behavior. For example, in the active compound intersection dataset, users can highlight all related patterns of {275}. As most of the correlated patterns are green, users can investigate those red nodes/links to discover what substructures interacted with these patterns and changed their A/I Score.

- For patterns neutralized by the coexistence of known conflicting impact factors (e.g., 275 + 518), as shown in figure 5.11, users can further investigate other impact factors affecting these patterns and the interactions between substructures.

Users can generate insights from the pattern exploration system. However, there are also limitations to this research study. The dataset used in this project is aggregated. The insights generated are mostly based on existing observations and, therefore, could be biased. This visual analytics system focuses on explorative analysis and assists users in generating possible insights. The system is built to provide suggestions for future exploration and experiments. Further validation and statistical support are required to support decision making.

This research can support users to find patterns or substructure interactions that determined the compound activity. Through the possible insights derived from the system, users can potentially discover the hidden relationships between patterns and substructures. Users can utilize this information to find the collections of substructures to be tested as compound compositions. Therefore, this is potentially beneficial for the drug discovery field.

6.3 Future Work

Although this research study provides a system for visual pattern exploration, further work and evaluation still need to be conducted to refine the data analysis process as well as the user experience while using the system.

First, the variety of filters and performance of data filtering and calculation can be improved. Filtering is the first step to generate views and start the analysis. However, the filters only include properties related to support and A/I Score. More various metrics can be brought in for further data filtering. In addition to variety, the performance of the algorithm can also be improved for a shorter wait time.

Second, users can provide a specific set of substructures for generating pattern data. Currently, users can not query for all pattern nodes that contain one or a set of substructures. With this feature, users can potentially derive more insights related to the given substructure set and expand the exploration process.

Third, further statistical analysis and visualization components can be brought in for impact factors. The impact factors indicate how substructure interact with each other to activate or deactivate compound potentially. In the current system, impact factors are only calculated during pattern network exploration. More work can be conducted around how to visualize and analyze the impact factors, such as individual statistics and visualization panels.

Last, the system can have an annotate feature to mark and save patterns of interest for further analysis. Annotate is also one of the typical interactive dynamics for the visualization system. In this system, users can expand the sub-network of a pattern of interest. However, the sub-network will not be stored, and users have to perform redundant actions such as switching between overview and sub-network back and forth. Users should be able to annotate the patterns and store any insights associated with the smoother analysis process.

REFERENCES

REFERENCES

Agarwal, R. C., Aggarwal, C. C., & Prasad, V. (2000). Depth first generation of long patterns. In *Proceedings of the sixth acm sigkdd international conference on knowledge discovery and data mining* (pp. 108–118).

Blanchard, J., Guillet, F., & Briand, H. (2007). Interactive visual exploration of association rules with rule-focusing methodology. *Knowledge and Information Systems*, *13*(1), 43–75.

Buja, D., A.and Cook, & Swayne, D. F. (1996). Interactive high-dimensional data visualization. *Journal of Computational and Graphical Statistics*, *5*(1), 78-99.

Cleveland, W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, *229*(4716), 828–833.

Cook, K. A., & Thomas, J. J. (2005). *Illuminating the path: The research and development agenda for visual analytics* (Tech. Rep.). Pacific Northwest National Lab.(PNNL), Richland, WA (United States).

Dunne, C., Henry Riche, N., Lee, B., Metoyer, R., & Robertson, G. (2012). Graphtrail: Analyzing large multivariate, heterogeneous networks while supporting exploration history. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1663–1672).

Fry, B. (2008). *Visualizing data: Exploring and explaining data with the processing environment.* " O'Reilly Media, Inc.".

Görg, C., Tipney, H., Verspoor, K., Baumgartner, W. A., Cohen, K. B., Stasko, J., & Hunter, L. E. (2010). Visualization and language processing for supporting analysis across the biomedical literature. In *International conference on knowledge-based and intelligent information and engineering systems* (pp. 420–429).

Graham, M., & Kennedy, J. (2003). Using curves to enhance parallel coordinate visualizations. In *Information visualization, 2003. iv 2003. proceedings. seventh international conference* (p. 10-16). IEEE.

Guo, D. (2003). Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization*, *2*(4), 232–246.

Guo, D., Chen, J., MacEachren, A. M., & Liao, K. (2006). A visualization system for space-time and multivariate patterns (vis-stamp). *IEEE Transactions on Visualization and Computer Graphics*, *12*(6), 1461-1474.

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *ACM sigmod record*, *29*(2), 1–12.

Heer, J., & Shneiderman, B. (2012). Interactive dynamics for visual analysis. *Queue*, *10*(2), 30–55.

Inselberg, A., & Dimsdale, B. (1991). Parallel coordinates. In *Human-machine interactive systems* (p. 199-233). Springer US.

Inselberg, A., & Lai, P. L. (2002). Visualization and data mining for high dimensional data. In *Russian summer school in information retrieval* (p. 142-184). Springer International Publishing.

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, *8*(1), 1-8.

Keim, D. A., Mansmann, F., Schneidewind, J., & Ziegler, H. (2006). Challenges in visual data analysis. In *Tenth international conference on information visualization (iv'06)* (p. 9-16). IEEE.

Keim, D. A., Mansmann, F., & Thomas, J. (2010). Visual analytics: how much visualization and how much analytics? *ACM SIGKDD Explorations Newsletter*, *11*(2), 5–8.

Kermarrec, A. M., & Moin, A. (2012). *Data visualization via collaborative filtering.* (https://hal.inria.fr/hal-00673330)

Krause, J., Perer, A., & Bertini, E. (2014). Infuse: interactive feature selection for predictive modeling of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1614-1623.

Leung, C. K., Carmichael, C. L., Hayduk, Y., Jiang, F., Kononov, V. V., & Pazdor, A. G. (2016). Data mining meets hci: data and visual analytics of frequent patterns. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 289–293).

Leung, C. K., Kononov, V. V., Pazdor, A. G., & Jiang, F. (2016). Pyramidviz: visual analytics and big data visualization for frequent patterns. In *2016 ieee 14th intl conf on dependable, autonomic and secure computing, 14th intl conf on pervasive intelligence and computing, 2nd intl conf on big data intelligence and computing and cyber science and technology congress (dasc/picom/datacom/cyberscitech)* (pp. 913–916).

Leung, C. K.-S., & Jiang, F. (2012). Radialviz: an orientation-free frequent pattern visualizer. In *Pacific-asia conference on knowledge discovery and data mining* (pp. 322–334).

Leung, C. K.-S., Jiang, F., & Irani, P. P. (2011). Fpmapviz: a space-filling visualization for frequent patterns. In *2011 ieee 11th international conference on data mining workshops* (pp. 804–811).

Marozzi, M. (2015). Multivariate multidistance tests for high-dimensional low sample size case-control studies. In *Statistics in medicine* (p. 1511-1526).

Munzner, T. (2008). Process and pitfalls in writing information visualization research papers. , 134-153.

Partl, C., Lex, A., Streit, M., Strobelt, H., Wassermann, A.-M., Pfister, H., & Schmalstieg, D. (2014). Contour: data-driven exploration of multi-relational datasets for drug discovery. *IEEE transactions on visualization and computer graphics*, *20*(12), 1883–1892.

Rouse, M. (2018). *Association rules (in data mining).* Retrieved from `https://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining`

Shneiderman, B. (1994). Dynamic queries for visual information seeking. *IEEE software*, *11*(6), 70–77.

Spenke, M., & Beilken, C. (2000). Infozoom-analysing formula one racing results with an interactive data mining and visualisation tool. *WIT Transactions on Information and Communication Technologies*, *25*.

Spenke, M., Beilken, C., & Berlage, T. (1996). Focus: the interactive table for product comparison and selection. In *Proceedings of the 9th annual acm symposium on user interface software and technology* (pp. 41–50).

Stasko, J., Görg, C., & Liu, Z. (2008). Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization*, *7*(2), 118–132.

Strehl, A., & Ghosh, J. (2003). Relationship-based clustering and visualization for high- dimensional data mining. *INFORMS Journal on Computing*, *15*(2), 208-230.

U.S. National Library of Medicine. (2016). *Chemical compounds.* (data retrieved from U.S. National Library of Medicine, `https://pubchem.ncbi.nlm.nih.gov/`)

Ware, C. (2020). *Information visualization: perception for design.* Morgan Kaufmann.

Yau, N. (2013). *Data points: Visualization that means something.* John Wiley & Sons.