# BioBridge: Bringing Data Exploration to Biologists

by

Joseph R. Boyd

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Bioinformatics and Computational Biology

by

---

May 2014

APPROVED:

---

Professor Matthew O. Ward, Major Thesis Advisor

---

Professor Samuel Politz, Thesis Reader

---

Professor Elizabeth Ryder, Head of Department

**Abstract**

Since the completion of the Human Genome Project in 2003, biologists have become exceptionally good at producing data. Indeed, biological data has experienced a sustained exponential growth rate, putting effective and thorough analysis beyond the reach of many biologists. This thesis presents BioBridge, an interactive visualization tool developed to bring intuitive data exploration to biologists. BioBridge is designed to work on omics style tabular data in general and thus has broad applicability.

This work describes the design and evaluation of BioBridge's Entity View primary visualization as well the accompanying user interface. The Entity View visualization arranges glyphs representing biological entities (*e.g.* genes, proteins, metabolites) along with related text mining results to provide biological context. Throughout development the goal has been to maximize accessibility and usability for biologists who are not computationally inclined. Evaluations were done with three informal case studies, one of a metabolome dataset and two of microarray datasets.

BioBridge is a proof of concept that there is an underexploited niche in the data analysis ecosystem for tools that prioritize accessibility and usability. The use case studies, while anecdotal, are very encouraging. These studies indicate that BioBridge is well suited for the task of data exploration. With further development, BioBridge could become more flexible and usable as additional use case datasets are explored and more feedback is gathered.

**Acknowledgments**

I would like to express my gratitude to my advisor, Matthew O. Ward, for his time, advice, and understanding. I would also like to thank the faculty of the BCB program for their support, specifically Elizabeth Ryder and Carolina Ruiz. I am grateful to my thesis reader, Samuel Politz. Additionally, I would like to thank my domain expert users: Kirsten Tracy, Leanne Ahronian, and Jeremy Shea, for volunteering their time.

# Contents

**1 Background** **1**

1.1 Introduction . . . . . 1

1.2 Related Work . . . . . 4

1.2.1 Biological Data Sources . . . . . 5

1.2.2 Towards Biological Relevance . . . . . 6

1.2.3 Networks and Pathways . . . . . 6

1.2.4 Ontologies . . . . . 10

1.2.5 Text Mining . . . . . 14

1.2.6 General Exploratory Visualization . . . . . 15

1.2.7 Specialized Exploratory Visualization . . . . . 17

1.2.8 Usability . . . . . 20

**2 Methods** **22**

2.1 Input Data Requirements . . . . . 22

2.2 Functional Requirements . . . . . 23

2.3 System Design . . . . . 24

2.4 Primary Visualization . . . . . 26

2.5 Control Window . . . . . 35

**3 Results and Discussion** **38**

3.1 User Studies . . . . . 39

3.1.1 Metabolite Dataset . . . . . 39

3.1.2 First Microarray Dataset . . . . . 40

3.1.3 Second Microarray Dataset . . . . . 41

3.2 Hypothetical Use Example . . . . . 43

3.2.1 Targeted Validation . . . . . 43

3.2.2 Open Exploration . . . . . 48

**4 Conclusions** **52**

**A Supported Interactions** **55**

A.1 Primary Visualization . . . . . 55

A.2 Control Window . . . . . 55

**B Text Mining** **58**

B.1 Document Retrieval . . . . . 58

B.2 Text Processing . . . . . 60

# List of Figures


1.1 BioBridge's primary visualization. . . . . 3

1.2 Network visualization strategies in Cytoscape as presented in [10]. . . 7

1.3 Pathway visualizations: ProMeTra and iPath as presented in [10]. . . 9

1.4 DAVID's ranked list summary of enrichment[8]. . . . . 11

1.5 DAVID is used regularly and has become increasingly popular since release (http://david.abcc.ncifcrf.gov/). . . . . 12

1.6 NetGestalt integrates GO and network data[30]. . . . . 13

1.7 PivotSlice[40] visualization . . . . . 16

1.8 INVISQUE[17] sense-making system . . . . . 18

1.9 GenAMap for gene association analysis[6]. . . . . 18

1.10 Toponome visualization[24] . . . . . 20

2.1 From input dataset to integrated visualization . . . . . 27

2.2 Glyph layout algorithm intuition. . . . . 29

2.3 Glyph layout algorithm pseudo-code. . . . . 30

2.4 Profile glyph construction. . . . . 30

2.5 Primary visualization with terms placed. . . . . 34

2.6 Control window components. . . . . 37

3.1 Gadd45-a centered view. . . . . 44

3.2 Targeted validation of terms. . . . . 45
3.3 From targeted validation of entity to open exploration of terms. . . . 46
3.4 Open exploration of terms. . . . . 49
3.5 From BioBridge into DAVID. . . . . 50

B.1 Text mining for BioBridge input. . . . . 59

# Chapter 1

# Background

## 1.1 Introduction

High-throughput techniques are at the heart of modern biology. Such techniques allow tens of thousands of biological interactions to be rapidly and economically characterized. Making sense of these interactions is critical to generating new biological discoveries. Contemporary biological problems, such as diabetes and cancer, are likely the product of many separate entities[1][32]. Tackling these problems will require biologists to discover and pursue many insights into their data. Extracting insights from ever larger data sets presents a challenge to trained data scientists, and is often daunting to biologists [21].

Data exploration is a critical step in data analysis and biological data is no exception. The human genome project ushered in the era of big data[2] for biology, putting effective exploration of biological data beyond the reach of biologists who do

[1]Throughout this thesis, the term entity will be used where no particular type of biological component is intended. Genes, RNA, proteins, and metabolites are all possible entities.

[2]Big data is a computer science term and industry buzzword for types of data that are difficult to work with as a result of some combination of size (Netflix's movie recommendation matrix), speed ("flash trading" on the stock-market), and complexity (biology!).

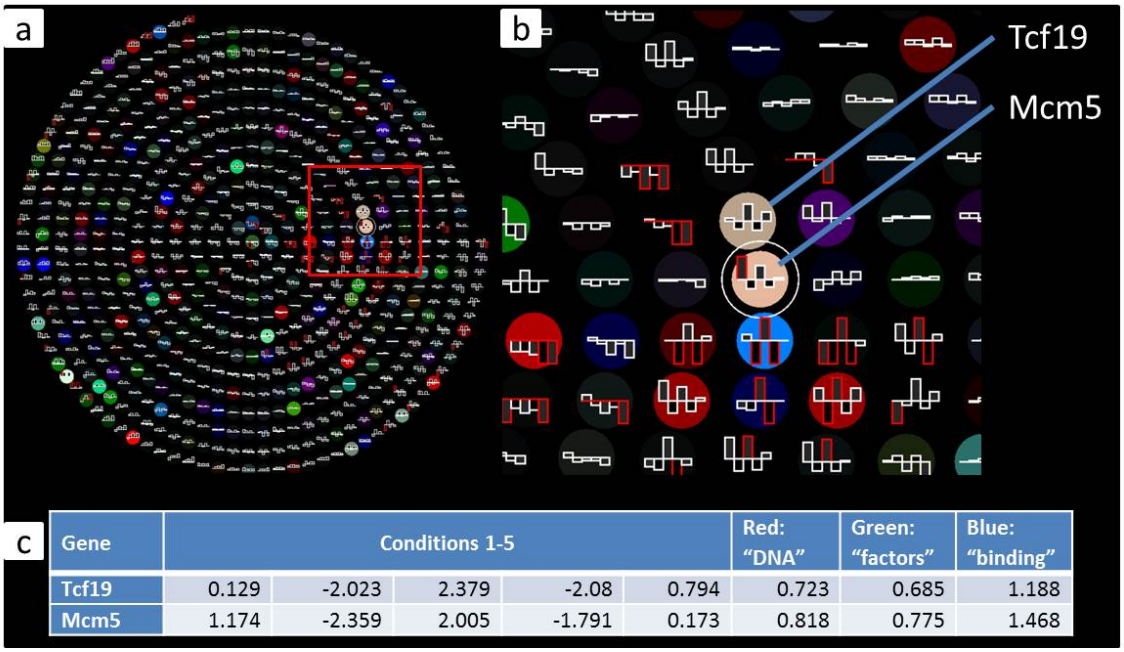
not happen to be data scientists. The tool described here, BioBridge, is an attempt to shorten the figurative distance between biologists and their data. The goal is to provide biologists direct access to their omics style datasets, minimizing the data processing that occurs outside of the biologist's control and maximizing their ability to explore and interact with their data.

To achieve these broad goals, it is informative to first consider the factors that hinder accessibility to data for biologists. The primary factor is size; large datasets necessitate databases, sophisticated statistics, and other sundries that are mundane for computer scientists but alien to many biologists. Size is the root cause, but not the first obstacle encountered by the intrepid biologist attempting to analyze their data.

Provided a biologist selects the appropriate analysis tool for their task, chances are they will encounter significant difficulty navigating the software dependencies, specialized data formats, and possibly even having to compile the tool in question. Biologists want to spend their time interpreting their data and conducting follow-up experiments, not troubleshooting.

This thesis describes the design process and features of BioBridge. The overarching goal in developing BioBridge was to maximize accessibility and usability for the non-computational biologist user group. Accessibility means the effort required to import data, run BioBridge, and interpret the results. Usability means the effort required to learn to use the tool to explore a dataset. Exploration entails investigating specific entities, verifying expected relationships, and discovering novel ones.

BioBridge is an interactive visualization tool that integrates text mining results into a 2D layout of glyphs representing biological entities. A glyph is a graphic that conveys one or more data values via visual attributes such as shape, size, color, and position[36]. BioBridge utilizes profile glyphs, which are similar to bar graphs,




| Gene | Conditions 1-5 | | | | | Red: "DNA" | Green: "factors" | Blue: "binding" |
|---|---|---|---|---|---|---|---|---|
| Tcf19 | 0.129 | -2.023 | 2.379 | -2.08 | 0.794 | 0.723 | 0.685 | 1.188 |
| Mcm5 | 1.174 | -2.359 | 2.005 | -1.791 | 0.173 | 0.818 | 0.775 | 1.468 |

Figure 1.1: BioBridge's primary visualization.
a)BioBridges Entity View integrates text mining results with user's experimental results, allowing them to rapidly locate genes (represented by glyphs) that react similarly to experimental conditions (represented by the bars of the glyphs) and have similar descriptive text (mapped to colored circles). Glyphs are arranged in concentric circles around a gene of interest at the center; glyphs that are more similar to the gene of interest are located closer to the center. b) zoomed view of the red box in (a), the; genes Tcf19 and Mcm5 are seen to have similar glyph profiles (with the exception of condition 1) and the terms binding, factors, and DNA in common. Perhaps the specifics of experimental condition 1 have altered the relationship between these two genes in a biologically interesting way. In fact, Tcf19 is a transcription factor required for later cell cycle stages and Mcm5 is part of the helicase machinery necessary to initiate DNA replication. c) provides a tabular view of experimental data (positive values are up-regulated, negative values are down-regulated) and text mining results (higher values indicate the term is more strongly associated with the entity) used in the visualization. BioBridge is designed to allow users to rapidly view different layouts of biological entities, searching for underlying biological relationships mined from the text of PubMed abstracts.

except there are many of them. The primary visualization is shown in Figure 1.1. Text mining results from PubMed abstracts are used as an approximation of known biological functions and interactions for a given entity. The terms thus derived are

presented within the data visualization to provide context, which helps to validate expected relationships and evaluate possible relationships. The design process began by considering various tasks that would be useful to biologists exploring relationships between entities within their data and attempting to interpret the underlying biology significance.

Evaluation of BioBridge was conducted via three case studies on evaluator provided datasets, one of metabolites and the other two of gene transcripts. User feedback was encouraging and generally focused on usability. Evaluations of the most recent version of BioBridge indicate that a tool such as BioBridge is most useful to biologists at the beginning of the data exploration process.

This stage of data analysis is particularly challenging and labor intensive for the non-computational biologist. Lacking accessible alternatives, they are often limited to programs such as Microsoft Excel for exploring their data. The chief contribution of this work is improving access to data exploration for non-computational biologists. User feedback has also revealed several interesting areas for future development efforts.

In summary, identification of functional requirements was aided by focusing on three use case datasets, the design process of the primary visualization and supporting control window was iterative and driven by user feedback, and the potential utility of BioBridge is demonstrated by a hypothetical work flow and evident in user feedback on the final version.

## 1.2 Related Work

The state of visualization in modern biology is well described in Gehlenborg *et al.*'s excellent 2010 review [10]. There are many visualization tools available and

they all face the challenge of presenting clear, meaningful, and integrated visualizations. In this context, relationships must be perceivable to be clear, reflect actual biological processes to be meaningful, and communicate with existing resources to be integrated. These tasks must all be addressed in the big data context of high-throughput techniques. Given the underlying size and complexity of the biological systems under studying, many visualizations rapidly become overwhelming, no matter how clear they are at a reasonable scale. The three challenges of clarity in the face of complexity, maintaining biological relevance, and effectively integrating available information, shaped the design of BioBridge and provide a useful framework for discussing related work.

However, before continuing, a slight expansion on the three concepts just introduced is warranted. Since the stated intention is to produce a tool centered on the biologist user, meaningfulness, or biological relevance and accuracy, is really the ultimate goal. Effective integration of available biology resources is necessary to discover biological meaning and a clear visualization is necessary to effectively convey biological meaning. Therefore, it is natural to first discuss the types of data sources available and how they may be used to gain biological insight.

### 1.2.1 Biological Data Sources

Biological databases are diverse. While the topic, data formats, and incorporated tools vary widely, conscious effort has been made to support integrated access. There are several generalized databases that serve as hubs to provide common terminology for more specialized databases. BioBridge draws on the contents of KEGG[3] [14, 15],

[3]The Kyoto Encyclopedia of Genes and Genomes(KEGG) is a frequently used resource for understanding high-level biological functions, chiefly through pathway oriented analysis and visualization.

and the NCBI[4] databases PubMed, Gene, and GEO[28]. KEGG links metabolites to enzymes as well as directly to article abstracts in PubMed. For microarray data, relevant experimental data are retrieved from GEO[9, 4], associated with gene entries in NCBI's Gene database, and then abstracts that PubMed directly links to genes are retrieved.

### 1.2.2 Towards Biological Relevance

There are three broad approaches for aiding biologists in understanding biological relationships. Network or pathway type approaches are already familiar to biologists and closely match the systems being modeled. Gene ontologies provide a standardized language for describing biological entities and are often used in statistical enrichment type analyses. Lastly, a lot of work has been done to apply text mining and natural language processing (NLP) techniques to sources such as biological abstracts and journal articles. These three categories are not independent of one another. For instance, text mining has been used to automatically assemble biological pathways[34, 23] and pathway membership and gene ontology information are frequently combined in the same analysis[8, 39].

Networks and pathways, ontologies, and text mining approaches will each be discussed further. Some representative systems for both general data exploration and for highly specialized biological data exploration will then be discussed.

### 1.2.3 Networks and Pathways

Network and pathway visualizations draw on established mathematical graph theory and visualization practices. These kinds of visualizations were already commonly in

[4]The National Center for Biotechnology Information(NCBI) maintains dozens of biological databases, many of which play central roles in their respective fields.

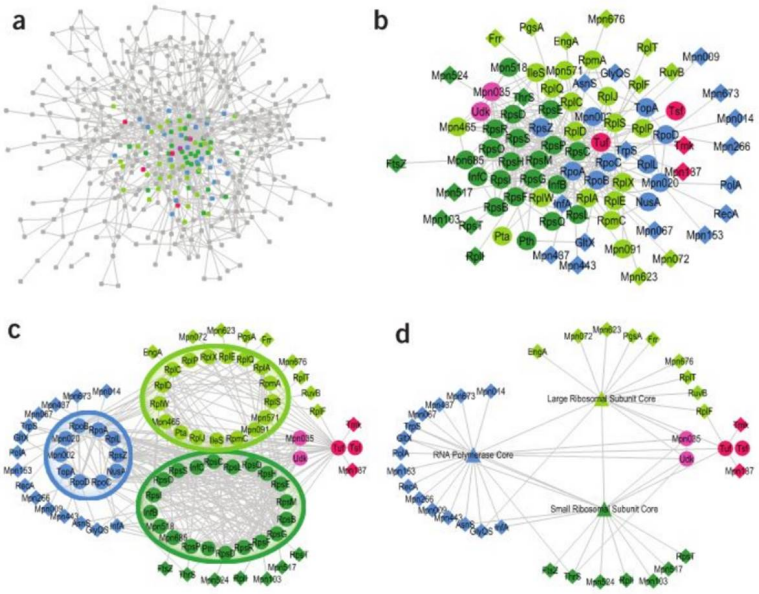



Figure 1.2: Network visualization strategies in Cytoscape as presented in [10]. Cytoscape[29] supports interactive network refinement. a)This modestly sized network of approximately 400 nodes is already cumbersome to work with. The network is limited to the colored nodes in remaining views. b)Employs the same layout strategy as a) with a limited subnetwork. Now one can see color and shape are being used to highlight significant groups of nodes. These attributes are controlled by the user. c)Employs special layout rules for nodes that are part of large complexes, polymerase and ribosomes in this case. d)Makes the visualization even clearer by collapsing complexes into meta-nodes; this is also user controlled.

use prior to the high-throughput era so they are already familiar to biologists. The Gehlenborg *et al.* 2010 review provides characteristic examples of contemporary network (Figure 1.2) and pathway (Figure 1.3) visualizations. Both visualizations have two primary limitations: they do not scale well to large and complex systems and their conventionally static nature does not convey dynamic biological processes well.

Though similar, networks and pathways have distinct meanings in the context of biological visualization. The most common type of network encountered is the protein-protein interaction network. These result from high-throughput methods

that are typically noisy, containing both false positives and false negatives. In lay terms, these networks contain information of the form "protein A sticks to protein B", where proteins are the nodes and these sticky relationships are the edges of the network. These relationships are undirected, if A sticks to B, then B sticks to A.

In contrast, relationships in pathways are directed; a path from A to B does *not* imply a valid path from B to A. Pathways also typically reflect the consensus across many experiments rather than the result of a single experiment. Comparing Figures 1.2 and 1.3, one can see that network visualizations tend to have a cluttered appearance and pathway visualizations have a clearer structure. This is a consequence of the differences just described.

Cytoscape is one of the most successful network visualization projects[29]. It is open-sourced and was released in 2003. It still receives regular core updates[31] and boasts an active plugin development community[27]. In general, Cytoscape is quite accessible for a typical biologist; this is discussed further in the Usability section. Figure 1.2 shows the same network at different stages of analysis. Cytoscape's basic functionality allows numerical data and other attributes to be mapped onto a network. Cytoscape also provides several layout methods and some analysis functions. More specialized functionality is provided by numerous plugins. Two Cytoscape plugins warrant further discussion as they are particularly popular and tie directly to gene ontology and text mining: BiNGO[20] and AgilentLiteratureSearch[34].

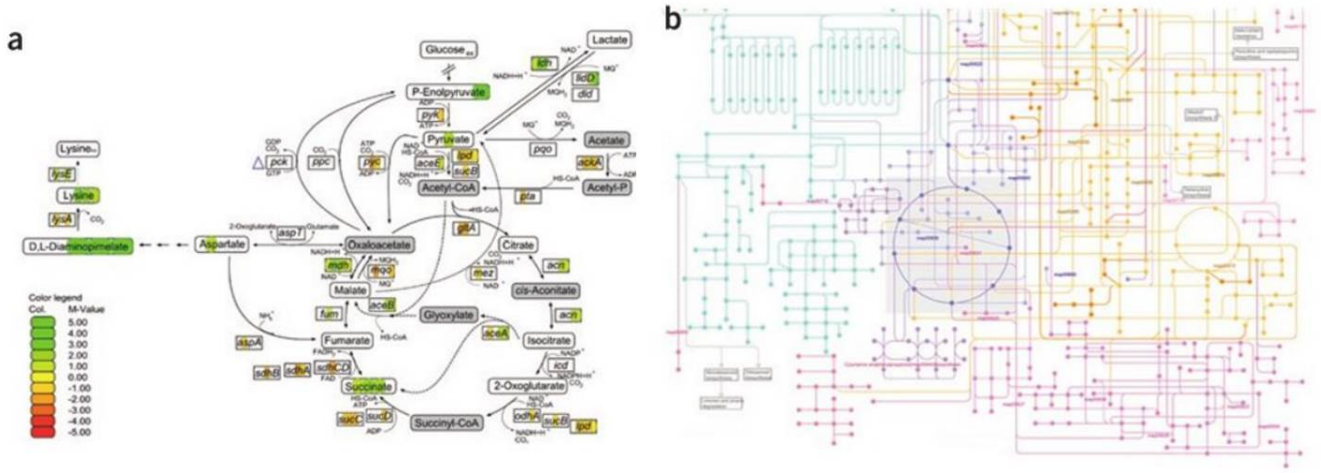



Figure 1.3: Pathway visualizations: ProMeTra and iPath as presented in [10].

a)A detailed pathway view of the citric acid cycle generated by ProMeTra[22]. Here, log fold-change values of metabolite concentration in several backgrounds are mapped to color, green indicates increase and red indicates decrease. b)A much higher-level view of a pathway generated by iPath[18, 38]. Different colors indicate relevant sub-pathways. Note the critical importance of node layout to limit unnecessary edge crossings.

BiNGO is the most popular plugin available for Cytoscape by a factor of three. BiNGO employs many of the features found in other gene ontology enrichment tools; its creators feel its main distinction is being integrated into Cytoscape[20]. This allows users to interactively construct gene list queries by selecting groups of nodes as well as viewing the results within the context of the network. The strength of this kind of direct interaction for rapidly forming and evaluating queries was one of the key features that BioBridge was designed to include.

AgilentLiteratureSearch[34] (ALS) is one of the more popular plugins available for Cytoscape. Unlike BiNGO, ALS does not help to characterize an input network but instead forms a network as the result of a text search. For example, a user may use a list of protein names as search input. ALS will then conduct a metasearch using various search engines parsed for proteins that interact with the search input and the nature of those interactions. Each node represents a protein and each edge represents an interaction in the resulting network displayed in Cytoscape. ALS then supports interactions with the network, such as adding output proteins to the input search to extend the results and viewing the source text for verification. While BioBridge does not employ this kind of sophisticated interaction mining, ALS does demonstrate that biologically relevant information is contained in publicly available documents such as abstracts.

### 1.2.4 Ontologies

Many tools that aim to integrate biological meaning utilize gene ontologies. In general, ontologies are hierarchical categorical divisions of objects and are useful for organizing information in large systems. In biology, the modern use of gene ontologies began with the GO (Gene Ontology) project in 2000[2]. Gene ontologies ensure a common vocabulary is used across all organisms, allowing knowledge

| Category | Term | RT | Genes | Count | % | P-Value | Benjamini |
|---|---|---|---|---|---|---|---|
| UP_SEQ_FEATURE | signal peptide | RT | | 50 | 32.3 | 6.5E-7 | 4.2E-4 |
| SP_PIR_KEYWORDS | signal | RT | | 50 | 32.3 | 8.6E-7 | 2.8E-4 |
| UP_SEQ_FEATURE | disulfide bond | RT | | 45 | 29.0 | 1.2E-6 | 4.0E-4 |
| SP_PIR_KEYWORDS | disulfide bond | RT | | 46 | 29.7 | 1.7E-6 | 2.7E-4 |
| GOTERM_CC_FAT | extracellular region | RT | | 40 | 25.8 | 6.9E-6 | 1.5E-3 |
| GOTERM_CC_FAT | extracellular region part | RT | | 24 | 15.5 | 3.8E-5 | 4.0E-3 |
| GOTERM_MF_FAT | oxygen binding | RT | | 6 | 3.9 | 3.8E-5 | 1.4E-2 |

Figure 1.4: DAVID's ranked list summary of enrichment[8].

Ranked lists are the primary output of gene set queries in DAVID. Results are sorted by p-value. The Annotation Cluster column indicates the source of the term being enriched. Columns from left-to-right are: 1)the source of the annotation term 2)the term, with links to annotation source if available 3) link to related terms (RT) 4) Bar chart of percent of genes in query list with term 5)the count of genes in query list with term 6) percent of genes with term 6)p-value, significance of term 7)q-value, false discovery rate.

gained from studying one organism to be transferred to other organisms. The GO project maintains 3 ontologies: biological process, molecular function, and cellular component.

DAVID (Database for Annotation, Visualization, and Integrated Discovery) takes a list of genes as input and integrates biological knowledge from multiple sources to graphically summarize enrichment of various categories within the input gene list[8]. DAVID relies on multiple data sources of categorical descriptors of genes to glean biological meaning. It integrates data from the Gene Ontology database as well as additional information regarding protein domains and biochemical pathway membership. Findings are presented to biologists primarily by ranked lists of terms. Seen in Figure 1.4, these lists include information to foster confidence in observations, such as annotation source, membership support, and p-value. DAVID

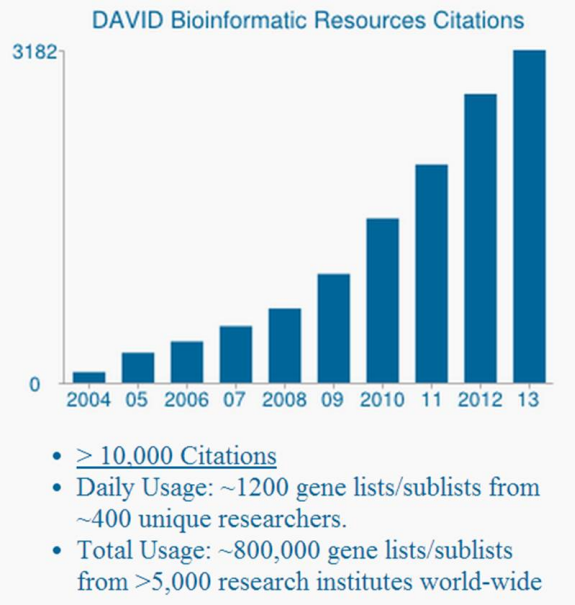



Figure 1.5: DAVID is used regularly and has become increasingly popular since release (http://david.abcc.ncifcrf.gov/).

provides a clustered heat map[5] style membership matrix and also links out to sources such as KEGG for more specialized visualizations (pathway visualization in the case of KEGG). DAVID has found wide acceptance among biologists, receiving a steadily increasing number of citations per year since its release (Figure 1.5).

Though its primary visualization(ranked lists) is quite simple, DAVID embodies the principles described in the Gehlenborg 2010 review. It is well integrated with established resources such as Gene Ontology and KEGG, allows users to explore the biological relevance of results, and its ranked list visualization is undeniably clear. The input for DAVID is a list of genes of interest. Consequently, prior to using DAVID a biologist must apply some analysis strategy to determine which genes are interesting; this may be a challenge for a non-computational biologist. The relevance of any results from DAVID are dependent on the quality of this input list. BioBridge

[5]A heat map is a common matrix visualization that uses color to represent values. After applying clustering to the underlying matrix, the heat map reveals "hot" areas where groups of rows have similar values in the same set of columns.

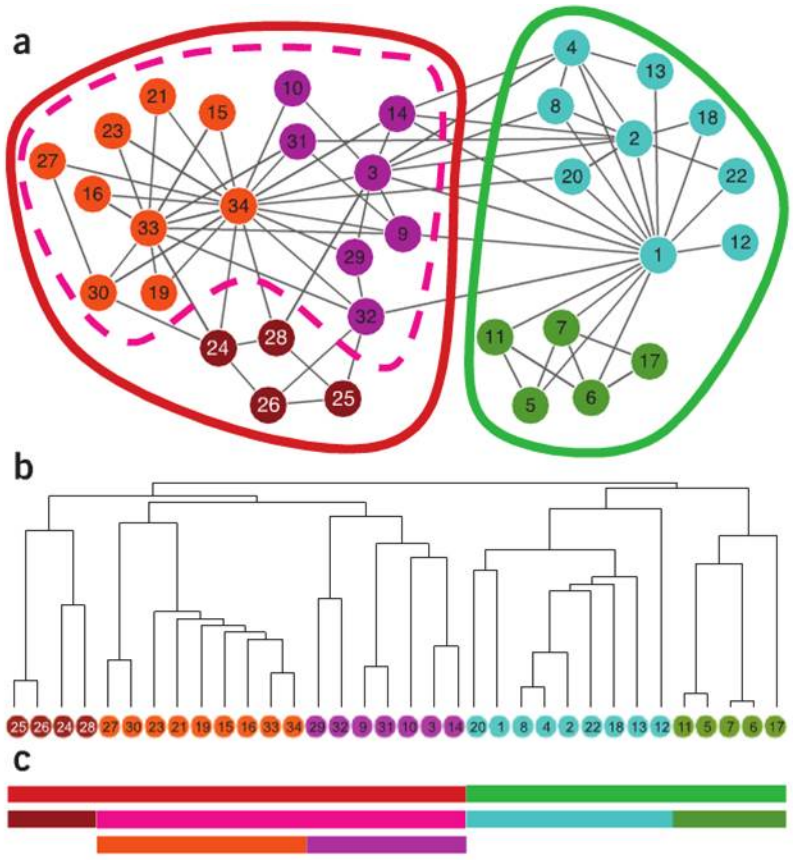



Figure 1.6: NetGestalt integrates GO and network data[30].
This relatively simple hypothetical network is converted to a linear model by hierarchical clustering. a)Highly interconnected regions are joined into different levels of clusters. These different clusters are indicated by color, the lowest level by node color and the higher levels by colored boundaries. b)The resulting hierarchical tree structure. c)The linearized tree visualization that is used by NetGestalt.

differs from DAVID since it removes this preprocessing burden from the biologist. BioBridge is also intended for exploration of entire experiments whereas DAVID is useful for validating lists of interesting genes and linking to other resources to aid interpretation.

WebGestalt[39] has many similarities to DAVID. It also takes a list of genes as input and integrates heterogeneous data sources such as GO and KEGG to find enriched biological characteristics.

NetGestalt[30], see in Figure 1.6, is an interesting recent utilization of the GO resource. To address the scalability issue associated with networks, it employs a strategy in which nodes are hierarchically clustered based on their connections. This renders the network to a linear structure divided into modular components, which are then analyzed using GO terms. NetGestalt demonstrates the power of integrating multiple types of data sources.

### 1.2.5 Text Mining

Text mining and natural language processing are areas of active development regarding biological articles and abstracts. Due to the sheer amount of literature available, computational techniques are critical in aiding human interpretation. On average, a paper is deposited in PubMed every 2 minutes[1]. Different text mining systems have different goals: some aim to extract formalized biological entity interactions[7], others attempt to validate and aid in the development of curated models[26], and some, such as PathText[16], are designed to help researchers analyze their data.

PathText is described by its creators as a "text mining integrator for biological pathway visualizations[16]." It utilizes the full text of biology papers to automatically annotate pathway models. These models are then augmented with the ability to retrieve relevant sections of papers to selected nodes. Manual annotation is also supported. PathText is particularly relevant since it integrates data from multiple databases with text mining techniques modeled to a network. However, it is limited by one of its most attractive features. Since full article text is central to its text mining strategy, PathText may only be used internally due to licensing issues[16]. For BioBridge, abstracts downloaded from PubMed are the only source of text. While full-text may be preferable from a data mining perspective, abstracts often distill the critical findings, are more structured, and through PubMed, are freely and easily

accessible. Another key difference between BioBridge and PathText is that BioBridge's visualization is derived from user data; PathText's pathway visualization is drawn from established models.

### 1.2.6 General Exploratory Visualization

PivotSlice[40] is a generally applicable visualization system designed to reveal implicit and explicit relationships between items. In a biological context, protein-protein interactions are one type of explicit relationship. Similarly, two genes correlating with one another throughout an experiment or sharing GO terms are examples of implicit relationships. PivotSlice's primary visualization technique (Figure 1.7) is a 2-dimensional table of cells, where each cell contains a subset of items, split by a different facet[6] for each axis.

Within each tabular cell, items are represented as different nodes arranged by a method the user considers meaningful. By combining attribute-derived and user-manipulated item layouts, different types of implicit relationships are revealed. Edges drawn between nodes reveal explicit relationships. By interactively controlling how each axis is subsetted, how nodes are arranged, and what explicit relationships are drawn, the user discovers various trends in the dataset.

Akin to PivotSlice, BioBridge's similarity based layout method reveals implicit relationships. BioBridge's visualization of text mining results reveals additional implicit relationships among items, although some relationships in the source text are certainly explicit. BioBridge may benefit from employing a faceted approach to text mining; a biologist may only wish to only see terms related to diseases or only

[6]The term "facet" has particular meaning in the visualization field. Each facet is an orthogonal way of viewing the data. The three primary gene ontologies are good biological examples of facets; cellular component, molecular function, and biological process each define different subsets of the same genes.

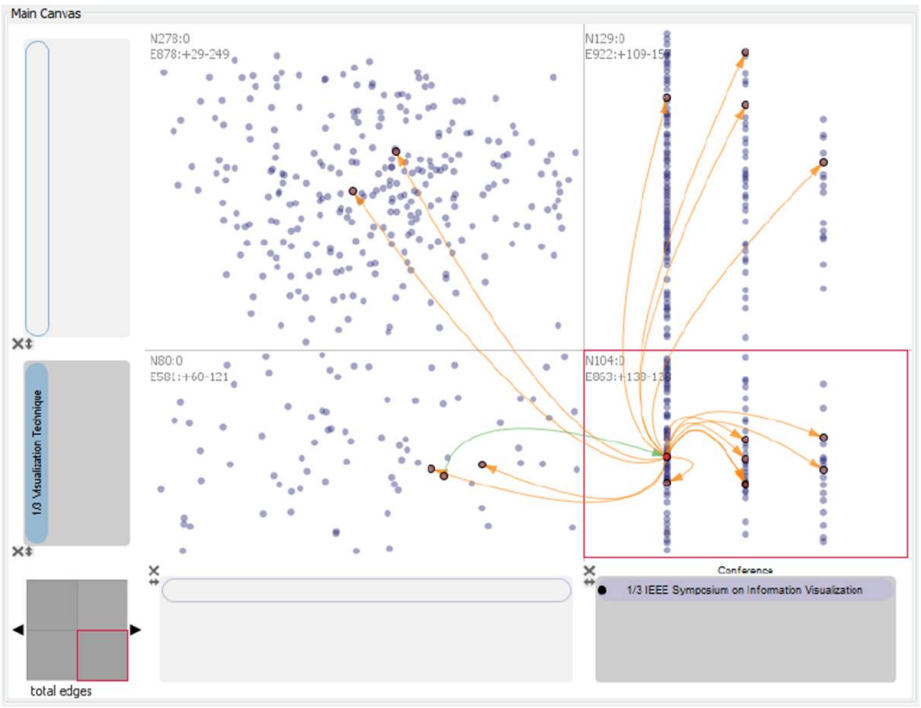



Figure 1.7: PivotSlice[40] visualization

PivotSlice utilizes facets to organize data and reveals direct relationships within data. The dataset visualized here is of journal articles. The y-axis is split by the keyword facet and the x-axis is split by the conference facet. 3 keywords of interest and 3 conferences of interest have been specified. The region outlined in red contains all articles with the keyword "Visualization Technique" that were presented at *InfoVis*, *CHI*, or *UIST*. Conference is used to align the nodes into 3 vertical groups. It is unclear what values were used for other x and y values of nodes. Orange arrows indicate the selected node cites those papers, the single green arrow indicates a citing paper.

terms describing cellular localization.

The INVISQUE[17] system is also generally applicable. The authors identify INVISQUE as a sense-making system. Sense-making is a formalized way of thinking about data exploration; people attempt to fit data into frames or mental models. When data do not fit into an individual's current frame, the frame is adjusted and the data is reconsidered. INVISQUE is designed to allow users to rapidly adjust and consider frames to better understand their data.

INVISQUE utilizes the metaphor of arranging index cards representing journal articles according to selected metadata attributes (Figure 1.8). For instance, the index cards may be sorted left to right by publication date, grouped by author, and given a vertical position according to number of citations. INVISQUE allows the user to maintain and arrange various such views simultaneously, accompanied by a description of each view and a scatter-plot summarizing a specified attribute. This process is designed to be highly intuitive, visualizations are rearranged by direct manipulation with the mouse, much the same way a user would physically rearrange a real-world workspace to help think about a problem.

The visual elements of INVISQUE are quite different from BioBridge and yet the two strive to support the types of tasks, namely data exploration and hypothesis generation. Both use position and proximity to help users to understand their data. INVISQUE allows users to control how their data are filtered and which aspects are summarized. BioBridge would benefit from the addition of filtering functions for entities and terms as well as quantitative summarization options.

### 1.2.7 Specialized Exploratory Visualization

GenAMap[6] is a visual analytics system that incorporates several different visualizations for genetic polymorphism association analysis (Figure 1.9). GenAMap is relevant because it is an exploratory visualization tool developed to aid geneticists in coping with the large amount of data association analysis produces. The basic strategy of GenAMap is to provide an overview visualization, a heat map of genes clustered by GO terms, that can be used to browse the dataset and make selections for generating more specialized visualizations. Each visualization can be used to answer different types of questions. A standard gene association network view presents the same information as the heat map in a more intuitive way, but cannot

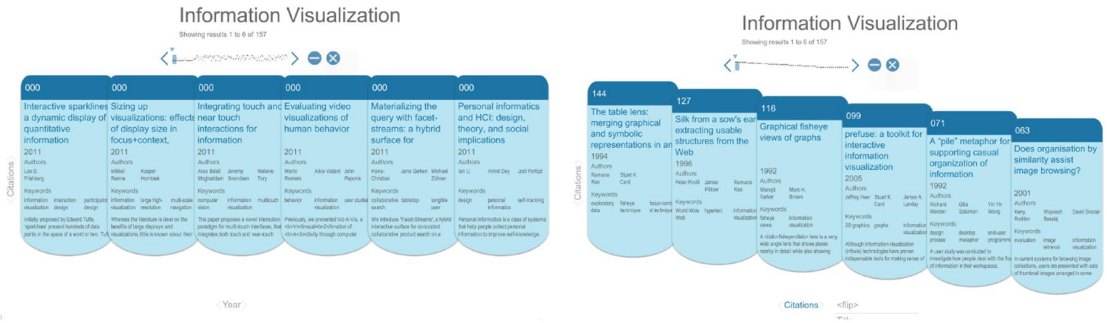



Figure 1.8: INVISQUE[17] sense-making system

Two index card arrangements of the same dataset of journal articles that a user has arranged next to one another. Both views only display articles that contain the phrase "Information Visualization". On the left,articles are sorted by publication date, the first 6 results are displayed in detail. On the right, articles are sorted by number of citations; number of citations is also mapped to relative height of the index card. Both views have a summary scatter-plot where number of citations is the y-value. As one might expect, older articles tend to have more citations.

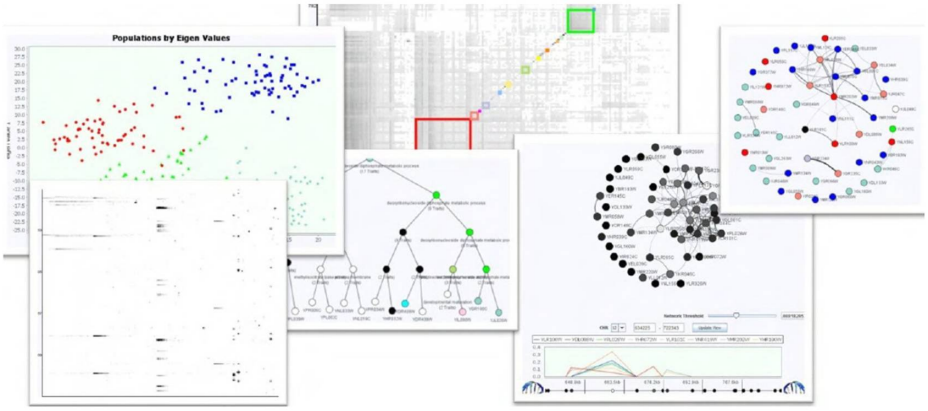



Figure 1.9: GenAMap for gene association analysis[6].

Via a scalable heat map visualization (top center), the user selects clustered sets of genes, indicated by colored boxes. These sets of genes are then linked to several domain relevant visualizations, such as gene ontology hierarchies seen bottom center and genome position association seen bottom right.

scale as well as the heat map. An association tree view relates functional information to genomic regions. Whereas GenAMap is specialized for genetic association data, BioBridge is generally applicable to any tabular dataset.

Livengood, Maciejewski, Chen, and Ebert developed a system for analyzing sets

of metabolomic data[19]. More specifically, their system aids in the comparison of four dimensional GCxGC-MS[7] datasets; the four dimensions are retention time 1, retention time 2, mass, and intensity. This is one of the first domain specific visualization system to allow multiple GCxGC-MS samples to be viewed simultaneously. Livengood *et al.*'s system supports exploratory analysis by allowing the researcher to rapidly generate different views (such as mean of sample, mean difference between samples, and standard deviation of samples).

Though this system is clearly highly domain specific compared to BioBridge, both systems support rapid view generation to facilitate exploration. Livengood *et al.* also identify better compatibility with existing tools as a primary aim of future work; BioBridge is already quite compatible since it has tabular data as input and output. That said, BioBridge would still greatly benefit from directly linking to commonly used online resources such as DAVID.

Many of the visualizations discussed so far utilize position and proximity to help convey relationships. An interesting example of a visualization where position and proximity are not variable options is presented in S. Oeltze *et al.* 2012[24]. The authors present a visualization tool for toponome data. The toponome describes the distribution of proteins throughout different cell types. Toponomic data consists of binary protein presence vectors for each voxel[8] derived from layered confocal microscopy images, where hundreds of proteins are assayed in 3-dimensions in dozens of cells simultaneously. See Figure 1.10 for an example toponome visualization.

The central task in toponomic analysis is to identify and describe the cell types present in a given microscopic view based on which protein combinations are present and how they are distributed. There are many thousands of possible protein com-

[7]GCxGC-MS separates a sample twice by gas chromatography before identifying molecular components by mass spectrometry.

[8]A voxel is the 3-dimensional equivalent of a pixel.

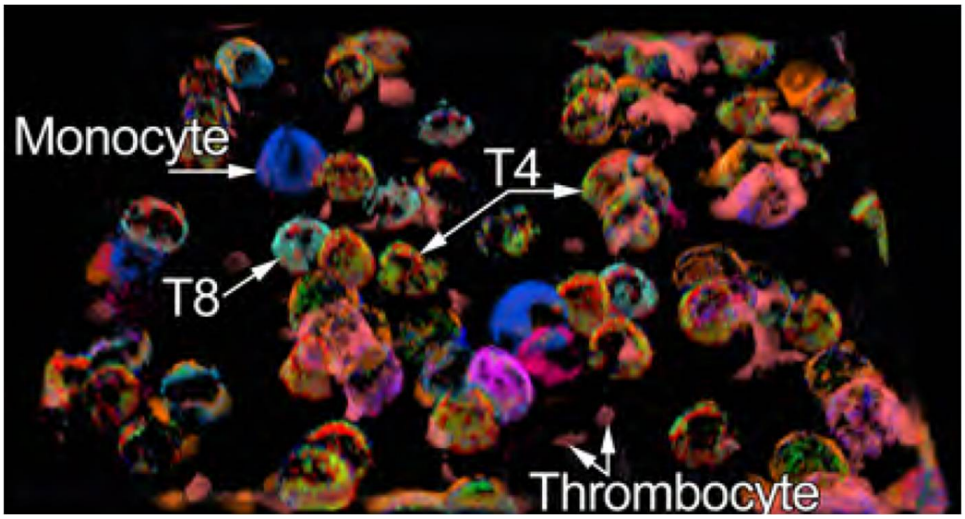



Figure 1.10: Toponome visualization[24]
This is a typical toponome view of a few dozen individual cells. Each color used in this view represents a unique combination of proteins. Several different cell types are readily visible, characterized by the type and distribution of proteins present. For instance, monocytes are nearly entirely blue whereas T4 cells seem to characterized by a mixture of red and yellow regions with some green spots.

binations; in practice, about 30 protein combinations are relevant to analysis. The toponome view presented by the authors assigns a unique color to each protein combination, then renders colored voxels accordingly. Similar visualizations have already been used for toponome data, the authors primary contributions have been in increasing usability and integration with other analysis tools.

Despite their obvious differences, BioBridge and this toponome visualization system use color in a similar way, to represent similarity when entities are not nearby one another. Both tools also support data exploration tasks by enabling users to rapidly investigate multiple visualization settings and views.

### 1.2.8 Usability

Usability for biologists has emerged as a priority for data analysis tools. This can be seen in release announcements for new tools as well as development efforts for

older tools[3]. Explicet[25], a tool released in 2013, touts its accessibility to non-bioinformaticians. The group goes so far as to state: "Practical use of Explicet in several lab settings has been found to reduce requirements for personnel with bioinformatics/computational expertise.[25]" Tools such as Cytoscape have also had significant feature expansions focused on usability[31]. BioBridge integrates the principles and concepts of the above and extends them.

# Chapter 2

# Methods

In this section, first the supported input dataset formats are described. Next, the basic task requirements of biological data exploration are outlined. Finally, the design of BioBridge and how it addresses these requirements is discussed.

## 2.1 Input Data Requirements

BioBridge is targeted at any dataset that can be expressed in tabular form. Table rows represent biological entities. Biological entities correspond to a particular flavor of omics: transcribed genes for microarray or RNAseq data, proteins for proteomics, metabolites for metabolomics, and so on. One column of entity IDs is needed as input for text mining procedures; additional columns are for optional descriptive text and required numeric data. These numeric data columns are used to construct and layout glyphs; generally the numeric data represent entity behavior and function in a series of different experimental conditions. Any supplied descriptive text is not incorporated in the primary visualization but is available in the supporting data table view. For descriptions of example of datasets, please see the case studies subsection.

Text mining is conducted outside of BioBridge. There are many parameters that go into text mining procedures. In its simplest form, each term can be considered a binary variable; documents either contain the term or they do not. More sophisticated approaches use techniques such as normalization according to various parameters and methods to ensure that terms with differing suffixes are counted together. BioBridge simply requires a table of non-negative term values for every entity; larger term values indicate the term is more significant for the given entity.

BioBridge readily accepts comma delimited (.csv) and Microsoft Excel (.xls) files. Comma delimited is a widespread format for exchanging dataset across applications. Excel is ubiquitous in biology labs and many programs are capable of producing .xls or .csv files.

## 2.2 Functional Requirements

The following tasks as essential for biologists attempting to interpret relationships between biological entities:

1. **Find related entities** - there are two ways to begin investigating possibly related entities:

   (a) **Open exploration**: navigating within the dataset, rapidly querying possible relationships, and

   (b) **Directed search**: given prior knowledge, a particular entity is thoroughly investigated, either as validation or to discover unexpected relationships.

2. **Locate highly related entities** - related entities are expected to have similar text mining results. In an undirected process the user wants to rapidly see

groups of entities that have a real biological relationship.

3. **View term trends across dataset** - rather than locating a particular entity, the user wishes to find all entities that have a given term in common.

4. **Review and combine observations** - biologists may suddenly find a previous observation more interesting given a later observation or need to make observations across multiple queries.

5. **Communicate results** - biologists need to be able to communicate observations to other biologists as well as transfer data to other applications they may routinely use.

## 2.3 System Design

This section outlines the visualization and interaction decisions designed to address the previously described task requirements.

BioBridge is implemented in Java[11], primarily due to Java's platform independence. Visualizations are realized through Processing, a free and open source graphics library for Java. Excel reading and writing is accomplished by Java Excel API. Text mining corpora are retrieved from freely available abstracts on PubMed[28] and processed in Weka [13], an open source data mining and machine learning suite with a Java API. Weka is a particularly attractive choice as it similarly emphasizes usability for domain specialists over dedicated data miners. Weka provides a comprehensive GUI that opens up the future possibility for biologists to take control of their own text mining input to BioBridge.

**Dependency versions and websites:**

1. Java 7 version 1.7.0_55 available at www.oracle.com/technetwork/java

2. Processing version 1.5.1 available at processing.org

3. Java Excel API version 2.6.12 available at jexcelapi.sourceforge.net

4. Weka version 3.6.11 available at www.cs.waikato.ac.nz/ml/weka

In keeping with an accessibility focused design philosophy, BioBridge uses a 2-dimensional visualization. While there can be perceptual benefits to having a 3rd dimension, often obfuscation issues and interaction complications counteract benefits gained[5]. Many of the remaining design decisions are informed by how biologists approach data interpretation.

When interpreting experimental data, biologists are tasked with finding related entities and hypothesizing explanations for those relationships. Relationships between entities may be direct (protein A binds to protein B), indirect (protein A and B share substrate X), more systemic (protein A and B are both in pathway Y), or even very general (protein A and B are associated with disease Z). Many such relationships have been discovered and described in the biology literature and evidence for known and novel relationships is likely present in a given dataset. Indeed, *de novo* interpretation of biological datasets is dubious at best; validation by integrating previous observations increases confidence in new findings. The primary visualization of BioBridge aims to expose novel relationships as well as use prior knowledge to provide context and confidence.

The primary visualization (Figure 1.1a) allows users to rapidly locate genes that react similarly to experimental conditions (mapped to Glyphs) and have similar descriptive text (mapped to color). Figure 1.1b shows the zoomed view of the red box in (a). The genes Tcf19 and Mcm5 are seen to have similar glyph profiles (with the exception of condition 1) and the terms binding, factors, and DNA in common. Perhaps the specifics of experimental condition 1 have altered the rela-

tionship between these two genes in a biologically interesting way. In fact, Tcf19 is a transcription factor required for later cell cycle stages and Mcm5 is part of the helicase machinery necessary to initiate DNA replication. Figure 1.1c provides a tabular view of experimental data (positive values are up-regulated, negative values are down-regulated) and text mining results (higher values indicate term is more strongly associated with entity) used in the visualization. BioBridge is designed to allow users to rapidly view different layouts of biological entities searching for underlying biological relationships mined from the text of PubMed abstracts.

A control panel GUI (Figure 2.6) contains additional functions to control and support the primary visualization. This control panel consists of four primary elements: a) a menu bar that provides access to visualization functions, the most significant of which are load, save, find entity, find term, place terms, and help; b) keys to aid interpretation of glyphs and the term colors; c) a list of shared terms based on user interaction with the primary visualization; d)and a dynamic table that tracks queried entities and terms. This table displays details that are not present in the primary visualization as well as a history of queried entities and terms. The table and the primary visualization can both be saved to record and communicate observations.

## 2.4 Primary Visualization

The primary visualization displays three types of data. It presents the input numerical data as glyphs, arranging them so that similar glyphs are near one another. The prevalence of a particular term across the entire dataset is indicated by colored circles behind the glyphs. Lastly, regions where entities strongly share terms are indicated by placing text between the relevant glyphs. All three components

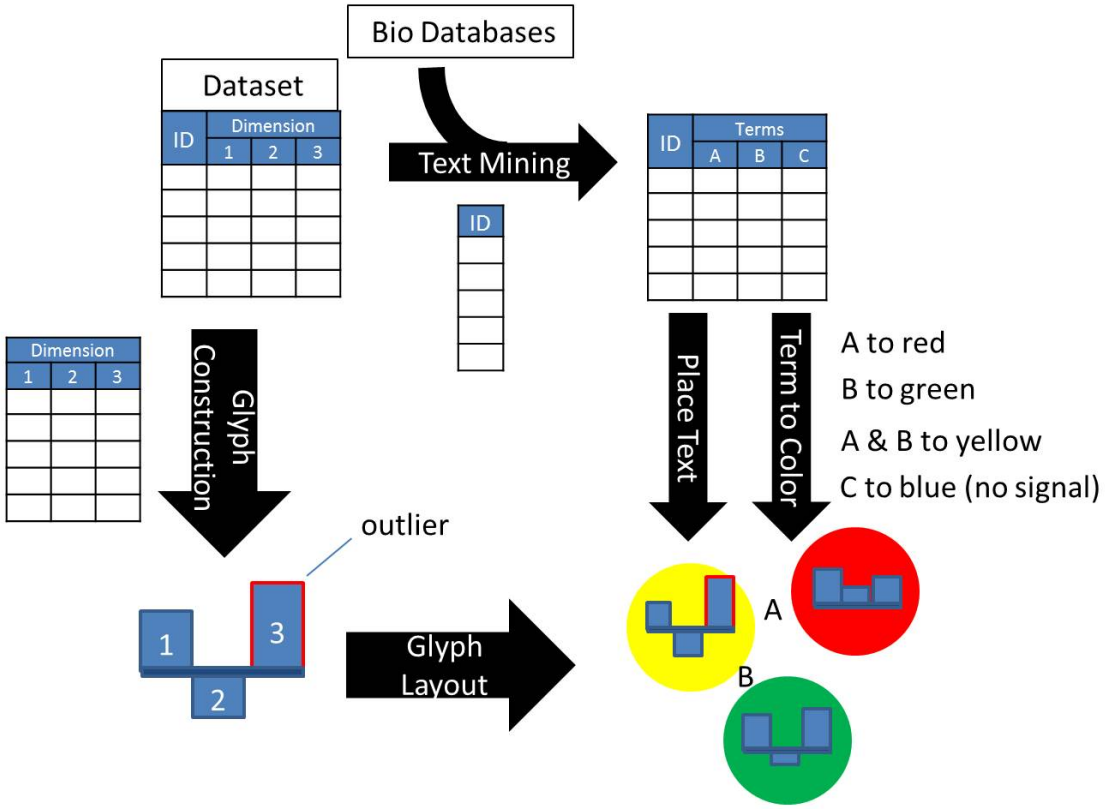



Figure 2.1: From input dataset to integrated visualization
The IDs from the input dataset are used for text mining of biological databases, notably abstracts stored in PubMed. The numeric dimensions of the dataset are used in glyph construction. Profile glyphs are used as they indicate positive and negative values well. Bars outlined in red indicate outlier values. All outliers are drawn at the same maximum bar height; this increases the ability of glyphs to effectively convey non-outlier values. Glyphs are then assigned positions by a glyph layout algorithm, detailed in Figure 2.3. Text mining results are used to place text between glyphs that share terms and also to color glyph backdrops as the user visualizes various terms across the dataset. Here, A is mapped to red and B is mapped to green; where the two co-occur, yellow results.

of the primary visualization as well as supported interactions will be discussed in the following paragraphs. A schematic overview of how the primary visualization is constructed is presented in Figure 2.1.

As Figure 2.1 shows, IDs from the input dataset are used for text mining of biological databases, notably abstracts stored in PubMed. The numeric dimensions

of the dataset are used in glyph construction. Profile glyphs are used as they indicate positive and negative values well. Bars outlined in red indicate outlier values. All outliers are drawn at the same maximum bar height, either above or below the zero line as appropriate. Treating outliers this way increases the ability of glyphs to effectively convey non-outlier values. Glyphs are then assigned positions by a glyph layout algorithm, overviewed in Figure 2.2 and detailed in Figure 2.3. Text mining results are used to place text between glyphs that share terms and also to color glyph backdrops as the user visualizes various terms across the dataset. In Figure 2.1, "A" is mapped to red and "B" is mapped to green; where the two co-occur, yellow results.

Profile glyphs were selected early on for use in the primary visualization. Profile glyphs are similar to bar graphs; they are both composed of several bars whose lengths indicate values of attributes. This allows users to make value comparisons across several dimensions simultaneously[37]. Profile glyphs are particularly suited for representing fold-change type data, since they clearly differentiate between positive and negative values.

Many biological datasets are expressed as fold-change values relative to a control; all case study datasets are composed of fold-change data. Fold-change values are calculated by dividing experimental values by a control or baseline value. Consequently, 0.5 indicates an entity's signal has decreased by half, 1.0 means it has not changed, and 2.0 indicates a doubling. Fold-change values for experimental values that fall below the control value range from 0 to 1 while for experimental values greater than control, fold-change may range from 1 to infinity. This discrepancy makes comparison between the two problematic. The common solution is to apply a log transform, rendering a half fold decrease as the negative of a two fold increase. This log transform of the dataset is what was used to construct profile glyphs;

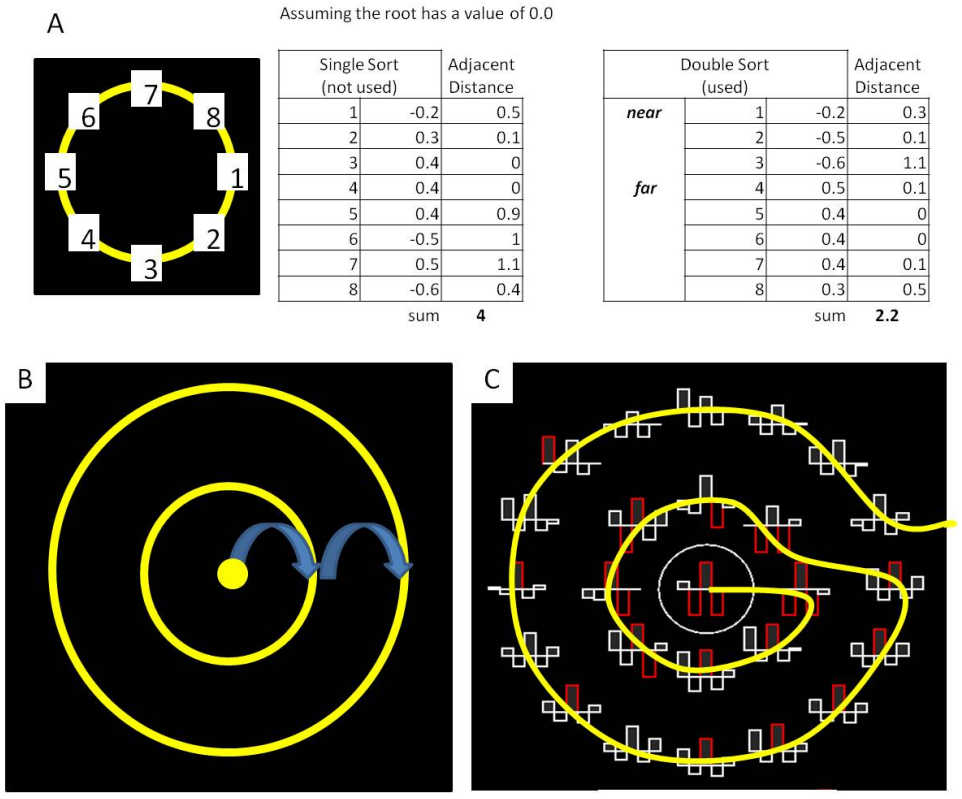



Figure 2.2: Glyph layout algorithm intuition.
The glyph layout algorithm uses a layer based strategy built around a specified root entity at the center. a) Each layer is sorted independently. BioBridge uses a double sorting method based on the most similar entity to the root ***near*** and the entity most dissimilar to ***near***, ***far***. This increases the similarity of adjacent glyphs within each layer (reduces the distance between values). b) Layers are populated starting from the center. c) The net result is that glyphs are placed following the yellow line from the root.

construction of profile glyphs from raw experimental values is shown in Figure 2.4.

The presence of outliers in the dataset was also considered. Each bar of the profile glyph represents the value of one dimension, bars above the baseline indicate positive values and bars below the baseline are negative. Originally the height of each bar was normalized to the maximum absolute value of each dimension. This created a situation where one extreme outlier could force all other values for a particular dimension into a small range of bar heights. The solution employed was

```
Given root instance r,  and unplaced instances unplaced
sort unplaced by distance to r -> list sorted
Repeat until all instances placed as glyphs
{
    first instance in sorted -> near
    glyphs that fit in first layer -> nL1
    instance within range 2:nL1 of sorted, that is most distant from near -> far
    split 1:nL1 into two groups, closer to near than far and closer to far than near.
    sort groups by distance from near  and far respectively.
    append far sorted to near sorted and overwrite 1:nL1 of sorted
    create and place glyphs for 1:nL1 of sorted in circle outside last layer
    remove 1:nL1 from sorted.
    near -> r
}
```

Figure 2.3: Glyph layout algorithm pseudo-code.

This algorithm uses one full sort and two sub-sorts per layer to arrange glyph in concentric circles around a specified root glyph (selected via direct interaction with the primary visualization or indirectly in the control window) )based on glyph similarity. Dissimilarity is measured by euclidean distance in this implementation. The number of glyphs that fit per layer is calculated based on the layer's circumference and a parameter that sets a minimum distance between glyphs. This algorithm results in high similarity between neighboring glyphs within layers, as well as good similarity between layers, particularly in the direction where members of ***near*** are placed.

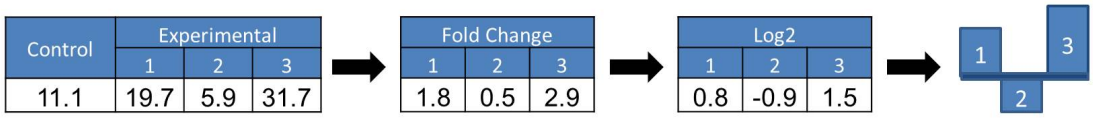



Figure 2.4: Profile glyph construction.

BioBridge's glyphs are designed to handle positive and negative values. This is particularly useful when datasets can be expressed as the fold change from control to experimental conditions. Typically, each experimental value is expressed relative to an appropriate control. A log transform is then applied. This allows equal magnitudes of under and over expression to be the negative of one another.

to calculate an outlier threshold value, 2.5 standard deviations from the mean, for each dimension. Normalization is based on this outlier threshold rather than the

maximum value. Only values within the outlier threshold are normalized; outliers are assigned the maximum bar height.

Normalizing based on outlier threshold did increase the dynamic range of most glyphs, but also could mistakenly convey high values still within the outlier threshold as equal to possibly extreme outlier values. To prevent this kind of confusion, bars representing outliers are outlined in red instead of the standard white. This improves perception of extreme values across the dataset.

The visualization layout design is based on the concept that related entities will tend to have similar numeric profiles. Many biologically related entities react similarly, or at least simultaneously, to experimental conditions. For example, pathways are broadly suppressed, or the proteins that function in a complex are transcribed by the same transcription factor. Since entity similarity is a primary concern, position is used to convey it. Position is a very effective graphical feature for encoding numerical data[35]. The glyph layout algorithm is shown in Figure 2.3.

This layout algorithm uses one full sort and two sub-sorts per layer to arrange glyphs in concentric circles around a specified root glyph (selected via direct interaction with the primary visualization or indirectly in the control window) based on glyph similarity. Dissimilarity is measured by Euclidean distance in BioBridge, but any distance measure could easily be substituted. The number of glyphs that fit per layer is calculated based on the layer's circumference and a parameter that sets a minimum distance between glyphs. This algorithm results in high similarity between neighboring glyphs within layers, as well as good similarity between layers, particularly in the direction where members of ***near*** are placed.

The combination of the similarity based glyph layout algorithm with the perception power of glyphs enriches the types of observations users can make. Since glyphs convey data values for several attributes simultaneously, entire regions of the visual-

ization can be rapidly characterized by the user. This allows for observations about the variability within and between different regions of the dataset. When combined with visualization of the text mining results, the user can begin hypothesizing which biological functions are significant in the dataset.

For visualizing terms across the dataset, a shaded circle is drawn behind each entity's glyph. Circle shading is determined by the entity's value for the terms being visualized. The current implementation only supports three terms simultaneously, each mapped to shades of red, green, or blue.

These terms are mapped directly to RGB values. For instance, if term "A" were mapped to red, entities completely unrelated to "A" would have an R value of 0. Likewise, the most associated entities would have an R value of 255. The remaining entities would fall somewhere in between. With RGB coloring, red and green make yellow, red and blue make purple, equal values of all three together make gray or white, and so forth.

The use of RGB values is strangely suited to biology. Immunofluorescence is a technique for visualizing cellular localization of proteins. In the images it produces, red and green are often used to visualize these proteins; where both entities are present, yellow is seen.

The potential for white background circles occurring when all three mapped terms have particularly high values required a tweak in the profile glyphs. Rather than simple white bars on a black background, white outlines with black fill are used. The white outlines are still readily perceivable and the glyphs are no longer lost against a white term circle.

Revealing regions of highly shared terms is accomplished by automatically querying the visualization between sets of closely placed glyphs and placing a representative term where aggregate text mining scores are high (Figure 2.5a). By default,

the query set size is 3. To be placed, terms must have a signal above a specified threshold for all 3 instances in the query set. Only the top ranking term is placed at a given position.

Term placement (Figure 2.5a) draws attention to regions where entities share biological characteristics and the terms displayed give the user some sense of what those characteristics may be. In Figure 2.5b the term "breast" appearing between glyphs characterized by extreme experimental values may serve as an entry point for further exploration. Indeed, querying this region returns cancer related terms such as "tumor", "cancer", and "therapeutic". Figure 2.5c shows there are still occasional term to term overlaps in the current implementation; "repeat" and "ubiquitin" mutually obscure one another in this case.

When prototyping term placement there was a lot of obfuscation, both due to terms overlapping with terms and terms overlapping with glyphs. To alleviate this issue, terms are no longer placed horizontally, but perpendicular to the radius at any given point. This entirely relieved the term to glyph overlaps and most of the term to term overlaps. To further reduce term to term overlaps, the radius at which each term is drawn alternates every other term. With these changes, and the ability to increase the threshold for term placement, the term obfuscation problem is reduced to tolerable levels.

Direct interaction with the primary visualization is focused on supporting exploration of terms beyond those displayed. Placing text mining terms between glyphs is effective for revealing areas where glyphs possess common terms, but there is no guarantee that the most interesting term is displayed. The introduction of the ability to specifically query within the visualization addresses this shortcoming. Results of these user queries are output to a sorted list in the control window. Since exploring these terms is a primary task of BioBridge, queries are accomplished with a

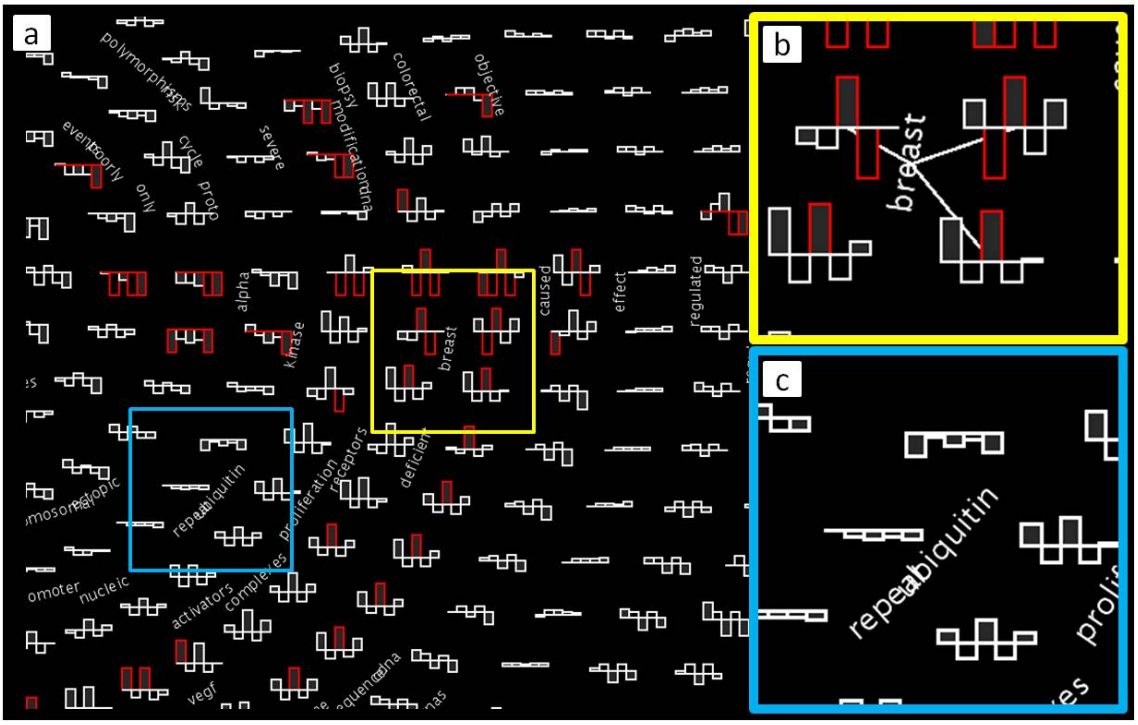



Figure 2.5: Primary visualization with terms placed.
a) Primary visualization with terms placed where their aggregate value is high for the 3 nearest glyphs. b) Detail view of yellow box in a). The term "breast" appears between glyphs characterized by some extreme values. A user query of this region (the 3 lines explicitly indicate which glyphs are queried) returns additional cancer related terms. c) Detail view of blue box in a). Attempts to eliminate obfuscation have not been completely successful; "repeat" and "ubiquitin" mutually obscure one another in this case.

single left-click of the mouse in the same area. Clicking and holding offers a preview of which entities will be queried, allowing users to be certain they are conducting the intended query. A user may also wish to further explore terms that define single entities. A double-left-click accomplishes a query of the single nearest glyph.

The other set of interactions is focused on navigating the visualization. This entails two sub-tasks: panning and zooming. A right-click and drag accomplishes panning; this style of dragging the view around the visualization is common in applications such as Google Maps. Zooming in and out is accomplished through the mouse-wheel. This is a common functionality for the mouse-wheel, particularly

when scrolling up and down is not appropriate. Thus far, no users have requested keyboard based panning and zooming controls. Should this issue arise, solutions that implement common keyboard navigation strategies based on the arrow keys or WASD would need to be explored.

## 2.5 Control Window

The primary functions of the control window are to present details not revealed in the primary visualization, record observations, and provide indirect interaction with the primary visualization. These functions are fulfilled by a results list (Figure 2.6a), a menu bar (Figure 2.6b), a term-to-color key (Figure 2.6c), and a dynamic table (Figure 2.6d).

In practice, the term results list is the most frequently used element of the control window. This is probably because the term results list is the element most tightly linked to the primary visualization. It presents the results of queries and controls the assignment of terms to colors. Query results are presented as a ranked list of terms, with aggregated values indicated. The values do not have an inherent meaning but are instead dependent on the text mining process. By selecting items on the list, users may add terms to the dynamic table and visualize selected terms across the entire dataset as colored circles.

The menu bar provides access to additional visualization functionality. Following user expectations, File is on the far-left and includes saving, loading, and exiting functions. Similarly, Help is on the far-right. The remaining menu items, Find, Visualization, and Table are listed from left to right in order of anticipated frequency of use. Find is used to locate specific entities or visualize specified terms. Visualization includes functions with broad effects, such as turning term placement

on or off and resetting the camera's position and zoom level. The Table menu bar item is primarily concerned with controlling what is added to the dynamic table and removing undesired items.

In the specific case presented in Figure 2.6, the user has queried a single entity, Crip1, and been presented with the list of terms in a). The values are listed following the terms. Browsing this list allows the user to roughly characterize Crip1. The terms "cysteine", "rich", and "intestinal" relate to the entity's full name, Cysteine-rich intestinal protein. The terms "myc" and "hybrid" derive from the use of myc-tagging, a protein purification method in which hybrid proteins are created. Basic kinds of descriptors such as these are frequently at the top of individual query results, with more interesting functional information further down the results list. The user may wish to use the find commands in b) either to locate an entity they expect to be related to the terms being viewed or to add a new term that they think may be interesting. c) tracks which terms are mapped to which color and the gradient aids in interpretation. Any terms that have been mapped are also tracked in table d). This table also displays precise values, both for the most recently queried glyphs and any other previously queried glyphs. Note the most recently queried entity is the last row and most recently mapped terms are the last three columns.

There are two types of keys included in the control window. The color key serves two functions. It indicates which terms are currently mapped to which color and provides an RGB color gradient to aid the user in interpreting mixed colors. The glyph key presents an enlarged view of whichever instance is currently selected and helps the user to remember which dimension each bar represents. Both keys only show a current snapshot of the visualization state. In contrast, the dynamic table serves as a record of all user queries.

The dynamic table is a standard tabular visualization that presents the data

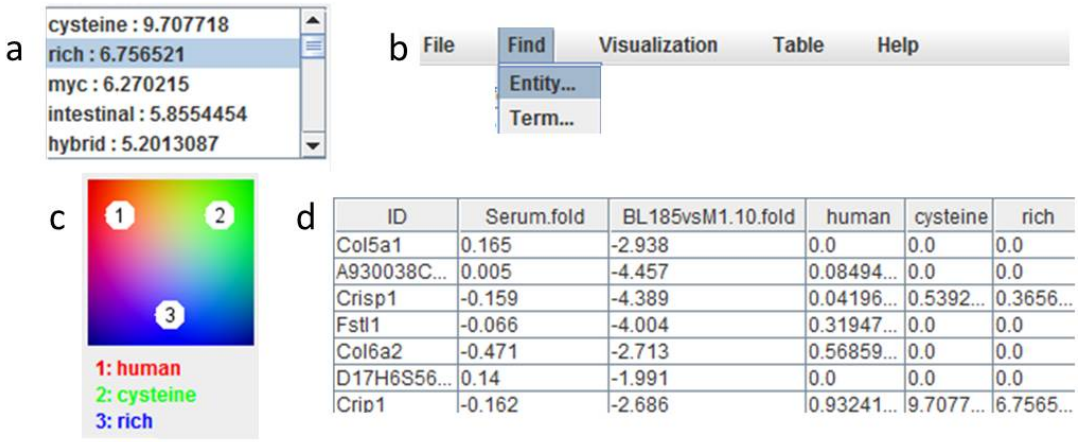



Figure 2.6: Control window components.
a) The sorted list of results from the previous user query. b) The menu bar provides access to additional visualization functionality. Here the user wants to find a specific entity in the primary visualization. Selecting Find-Entity will open a series of dialog boxes to complete the operation. c) The color key keeps track of which terms are currently mapped to which color and helps the user to interpret RGB gradients. d) The dynamic table records recently queried entities and terms and precisely represents the underlying numerical data used in glyph construction and layout.

shown in the glyphs. As the user queries entities in the primary visualization, these entities are added as rows to the dynamic table. Any terms searched for or visualized across the dataset are also added to the table as columns. Currently, there is not a cap to how many columns or rows the dynamic table can grow to hold. Consequently, the table can grow to an unmanageable size. The current solution is to provide reset type functions that restore the table to its original number of columns and rows. The dynamic table may also be saved at any time, allowing review or further analysis in another program. Methods for controlling and editing the table contents are present in the Table menu bar item.

# Chapter 3

# Results and Discussion

Evaluations were conducted via three case studies. Each case study centered on an evaluator supplied dataset of actual experiment data, one of metabolomics data and two of microarray data. Evaluators were biology graduate students who were near completion of their Ph.D.. Case studies were initiated at different stages of the development process, from initial prototype to essentially completed. Each case study entailed multiple informal observation and interview sessions at successive stages of development. During the case studies, biologists were observed finding expected relationships between entities as well as discovering potentially interesting novel relationships.

After describing the use case studies and how they have shaped the development of BioBridge, a hypothetical case study that illustrates key features of BioBridge will be presented.

## 3.1 User Studies

### 3.1.1 Metabolite Dataset

This case study was the first one initiated. Entities in this dataset are metabolites, small molecules that are energy sources and building blocks for larger processes. Entity IDs are KEGG IDs. KEGG[14] is a popular database focused on presenting a systems view of biology, connecting genes, proteins, compounds, reactions, and diseases. Only 200 entities out of 300 had attached KEGG IDs so the full dataset could not be utilized. KEGG also provides relevant PubMed IDs; these were used to generate the text mining corpus. Some additional text columns are included in this dataset to aid interpretation by providing chemical names and sub and super pathway membership. The numerical dimensions of this data set are measured concentrations of metabolites in different *Drosophila* mutant backgrounds.

Since this was the first dataset evaluated, both the prototype visualization and the text mining strategy employed were being tested for the first time. Even if the visualization had been completely functional, it was clear that decisions made during text mining dramatically impacted usefulness of the final product. My initial approach focused on assembling the set of enzymes associated with each metabolite, and then using the PubMed articles that KEGG provides as a reference to generate the text corpus. This approach weighted my text mining input towards papers focused on chemical mechanisms or purification techniques. These were not the user's areas of interest.

Eventually, it was discovered that directly searching PubMed for the metabolite's chemical name provided a more function- and interaction-rich text corpus. The number of abstracts for each metabolite varied widely, from a handful for less studied metabolites to thousands for ubiquitous metabolites such as glucose or ATP. This

imbalance was addressed by normalizing based on the number of abstracts for each metabolite, a common text mining strategy. Once I had these more suitable text mining results, the evaluator's interest increased noticeably. They could readily see the observations they had already seen using other analysis tools; this increased their confidence while using BioBridge and encouraged them to continue exploring to find new relationships.

After several internal iterations geared towards usability, the evaluator was impressed with the ease of use and speed at which they could check different metabolites and verify expectations. A lot of time had already been spent analyzing this particular dataset, laboriously going through the hard copy of an Excel worksheet with a highlighter. In comparison, this evaluator said that BioBridge would have made this initial overview sort of analysis faster and more enjoyable.

### 3.1.2 First Microarray Dataset

In both microarray case studies, entities are gene transcripts. Although gene transcripts may or may not be translated into proteins, their over or under expression is still evidence of a cellular response. This dataset had already gone through significant preprocessing; it was limited to about 700 genes that correlated with the experimental conditions. Entity IDs are gene symbols, which were directly input to abstract searches in PubMed. No additional text information was provided with this dataset. The numerical data for this data set consists of fold change values between a non-invasive hepatocarcinoma cell line and isolated cell subpopulations with higher metastatic activity.

This case study began with a much more developed version of BioBridge. The first challenge encountered was a consequence of the increased dataset size. All glyphs could no longer be expected to fit on the screen simultaneously at a useful

size. This precipitated the addition of panning and zooming capabilities. The evaluator for this dataset also said they had trouble finding entities as they thought of new relationships to check. This made the addition of a Find functionality a top priority.

As in the metabolite dataset, the first entities investigated were those the biologist thought would be interesting. Again, being able to make expected observations made the biologist much more confident and comfortable when first exploring their dataset. Fortunately for us, this dataset had not already been analyzed as exhaustively as the metabolite dataset. As the biologist verified expected relationships, they also discovered new entities to investigate.

These new entities of interest tended to fall in two groups: those that were already known to function in the pathway the biologist was interested in, but outside of their specific area of study, and those with entirely separate known functions. Both groups were interesting to the biologist. The first type often triggered a response similar to "This makes sense because..." which would reinforce the biologists understanding of entity relationships. The latter type of observation had two distinct types of outcome. If the entity appeared isolated, *i.e.* did not share any terms with surrounding entities, it would often be disregarded as coincidence. However, if there did appear to be some underlying biological significance, they made a note to further investigate the relevant entities outside of BioBridge.

### 3.1.3 Second Microarray Dataset

Preparations for this case study began after completing the previous two. As such, this presented a prime opportunity for evaluating the completed user interface from the perspective of a new user. Therefore, the first evaluation session began with only a broad explanation of BioBridge's purpose and the user was allowed to interact with

the interface without guidance.

In general, the functionality of the control window portion of BioBridge performed well. The user was able to open a dataset and begin querying it without guidance. Some specialized functions, such as placing terms within the visualization, were unknown to the user so they did not think to look for them. Many of the primary visualization interactions had a similar problem. This category of interface problems related to unknown features could probably be addressed by displaying periodic tips or hints, perhaps only until the user demonstrates awareness of the feature.

The biologist user for this study was more computationally inclined than the previous two. Consequently, they expressed interest in some of the aspects of BioBridge that do not support user control at this time.

The first such issue they raised dealt with adjusting the similarity measurement used in the glyph layout algorithm. BioBridge currently uses a Euclidean distance measure where all dimensions are weighted equally. This biologist expressed interest in different distance metrics, such as Manhattan distance, and adjusting the weighting of the dimensions.

The other issue raised was controlling the text mining procedures. After discussing the text mining I had conducted, as described in Appendix B, this user wanted to try various alternatives. This interest seemed to stem from general curiosity and desire for more control as opposed to an identified flaw or weakness. Most of the changes the user proposed dealt with the document retrieval portion of the pipeline. This would be one of the impactful portions to control, but also the most time consuming for the user to experiment with. Retrieving large sets of abstracts from PubMed can take a few hours, compared to seconds or a few minutes for the remainder of the pipeline. The user mainly wanted to specify PubMed search

parameters such as species and date range.

## 3.2 Hypothetical Use Example

This section is intended to demonstrate how the different features of BioBridge work together and to be representative of how evaluators approached data exploration using BioBridge.

The dataset for this example is the first microarray dataset described. Since observed biological relationships will be discussed in more detail here, a more detailed description of the dataset follows. The experiment underlying this dataset is attempting to identify genes related to cell migration in hepatocellular carcinomas. Control values in this dataset are transcript expression levels in a non-metastatic tumor cell line, BL185. Two subpopulations of cells were isolated from BL185 that had either significantly higher *in vitro* migration or invasion activity, BL185-M1 and BL185-I1 respectively.

Though the supplied dataset contained additional numerical dimensions, this analysis will focus on BL185 vs. BL185-M1 and BL185 vs. BL185-I1, therefore glyphs will contain only two bars. Remember, the fold change data has been log transformed so that bars above the zero line represent increased expression and bars below the zero line represent decreased expression.

### 3.2.1 Targeted Validation

As observed in use case studies, most users want to validate that particularly well known entities are behaving as expected. However, the initial visualization is arbitrarily constructed around the first entity in the dataset. Therefore, the first task is to locate an entity already thought to be interesting and revisualize the data.

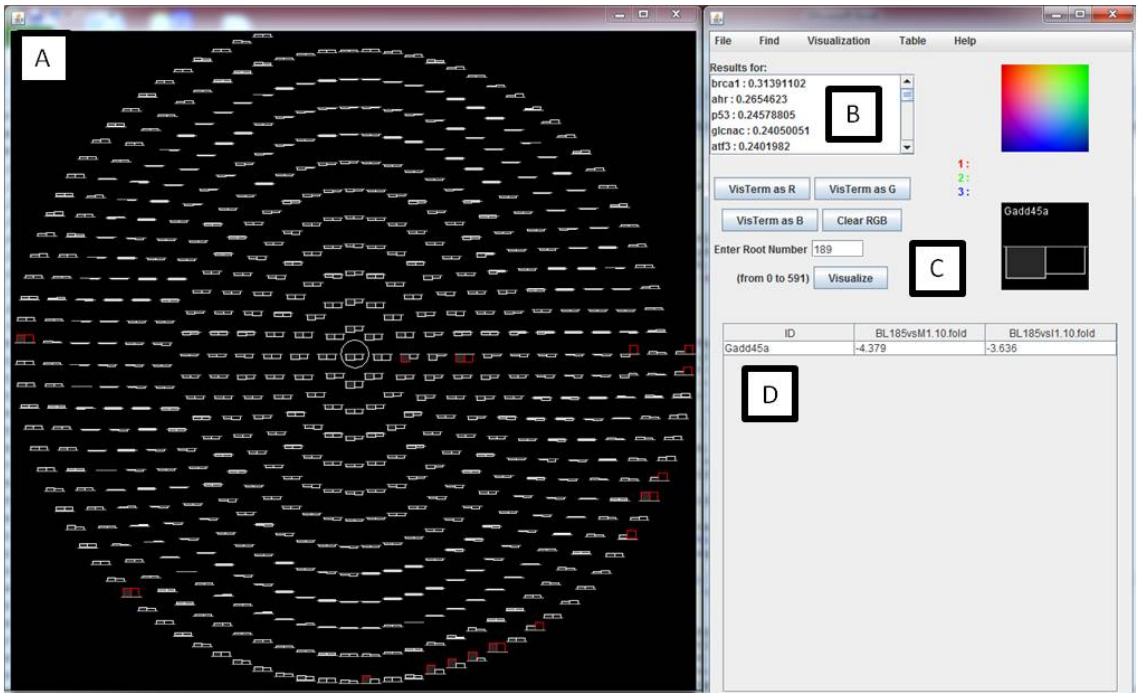



Figure 3.1: Gadd45-a centered view.
a) The visualization constructed around Gadd45-a. b) The query results for Gadd45-a by itself. c) When Gadd45-a was selected, its ID number was automatically entered here so the user could easily regenerate the visualization around it. Gadd45-a was also drawn in the detail box to the right. d) Since Gadd45-a was queried, its information was added to the dynamic table.

The protein Gadd45-a has been shown to inhibit cell migration in other cell types and to be down-regulated in the majority of hepatocellular carcinomas[12]. This observation, motivated Brian Lewis' lab at the University of Massachusetts Medical School to conduct this microarray experiment and it was coincidentally the basis for a WPI major qualifying project by Sally Trabucco in 2010[33]. Consequently, Gadd45-a is a natural first choice for validation.

To locate Gadd45-a, the user would use the Find Entity function located on the Control Window's menu bar. This is as simple as typing in the first few letters of the gene's symbol and selecting the desired gene from a drop down list of matches. This circles Gadd45-a in the visualization and sets up the Visualize button to generate a

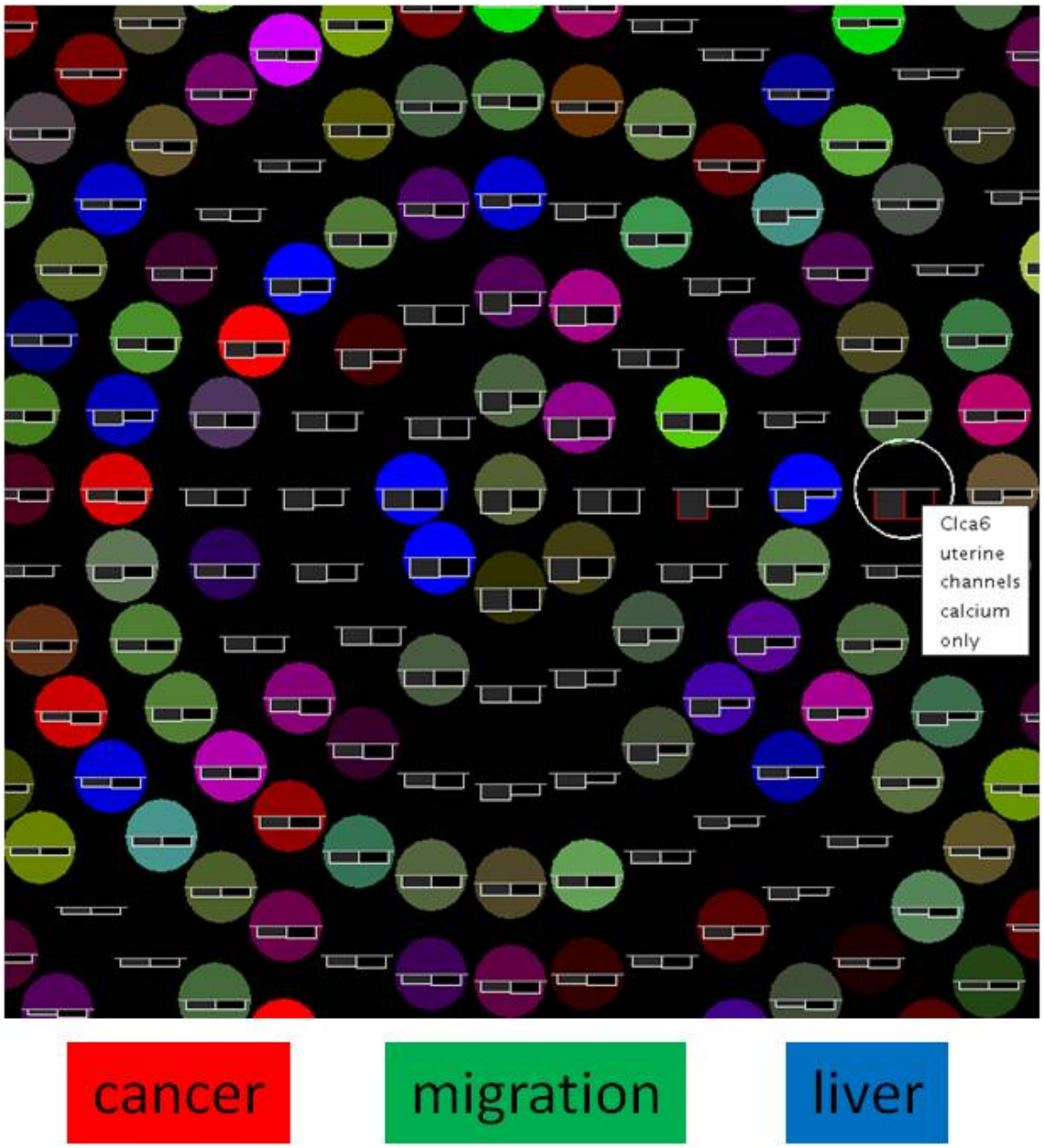



Figure 3.2: Targeted validation of terms.
A zoomed in view of the center of Figure 3.1. The three terms "cancer", "migration" and "liver" are mapped to red, green, and blue, respectively. These terms were selected by the user due to details of the biological experiment that generated the dataset. The gene Clca6 is particularly strongly down-regulated but is unrelated to these terms. This may prove interesting following additional research into Clca6.

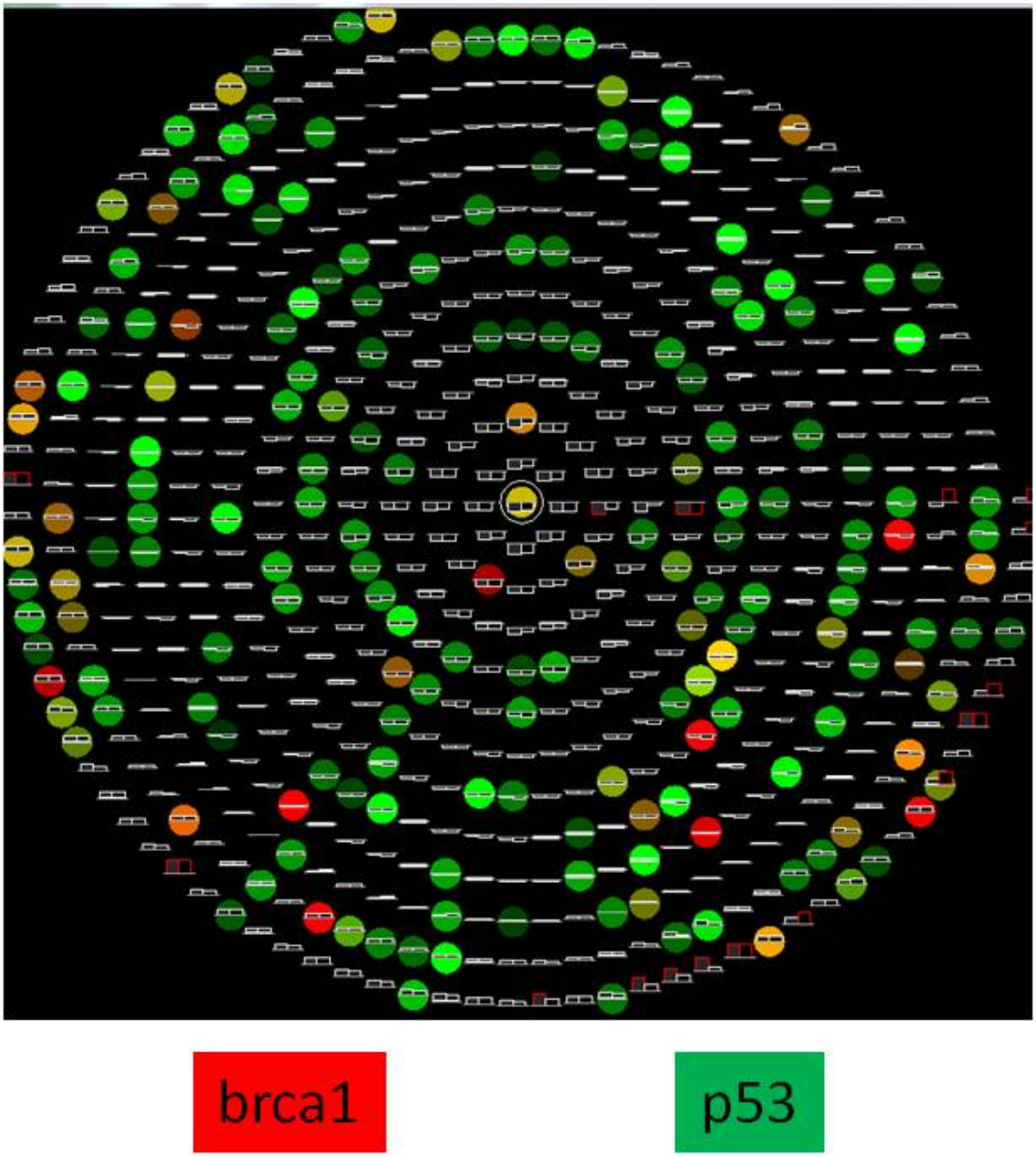



Figure 3.3: From targeted validation of entity to open exploration of terms. The terms "brca1" and "p53" are mapped to red and green. These terms happen to also be gene names that are prevalent in cancer biology. "p53" has a wider distribution than "brca1". It would be interesting to investigate the two entities with similar expression to Gadd45-a that also have overlap of "brca1" and "p53".

new layout centered on Gadd45-a. Clicking Visualize generates the desired Gadd45-a centered view, seen in Figure 3.1.

From this new view, the user can quickly see that Gadd45-a is indeed down-regulated in both BL185-M1 and BL185-I1. In fact, it is evident that Gadd45-a is one of the most down-regulated genes. From here, the user had several options for moving forward.

One option is to continue their targeted approach by characterizing specific terms across the visualization. The terms "cancer", "migration", and "liver" are reasonable choices. This is accomplished by using the Find Term function and selecting different colors for each term. The resulting visualization is in Figure 3.2.

A second, more exploratory, option is to visualize terms that are related to Gadd45-a. Two gene names, "brca1" and "p53" are both near the top of the term list for Gadd45-a and they both happen to be frequently associated with cancer. Rather than using Find Term, the user can visualize them by simply selecting them in the results list and assigning them a color. The resulting visualization is in Figure 3.3.

The third option is to switch entirely to a exploratory approach. This process is characterized by rapidly querying interesting entities and testing many different term mappings while on the lookout for interesting associations. For example, Clca6 is quite noticeable as the only entity with negative outlier values for both dimensions (Clca6 is being hovered over in Figure 3.2 and visible in the other Gadd45-a views). A biologist would definitely want to investigate this further. Clca6 happens to be a member of the calcium-activated chloride channel protein superfamily. Visualizing "calcium" and "channel" reveals two other similar genes that have some association with these terms, Mtm1 and Cth; it would be interesting to know why.

### 3.2.2 Open Exploration

The process of open exploration may be initiated on its own, but as was just shown, it often emerges naturally during targeted validation. One of BioBridge's strengths is that it creates many opportunities to make unanticipated observations that lead to open exploration type investigations. BioBridge also allows easy switching between open exploration and targeted validation.

Resetting BioBridge to its initial view (top of Figure 3.4), some general observations about the dataset can be made. This view happens to be centered on a gene that is up-regulated in both BL185-M1 and BL185-I1. One can readily observe there are handful of genes with extremely up-regulated expression, these appear near the center as a result of their similarity to the central gene. One can also see that expression level in BL185-M1 correlates strongly with expression level in BL185-I1. The few cases where only one is strongly over or under expressed, may be interesting in their own right. It is more difficult to make observations about genes that are down-regulated since they are pushed to the outer layers of the visualization by the glyph layout method.

Placing significant terms may aid interpretation of this view; this is accomplished through the Place Terms function located under Visualization on the menu bar. Without zooming in to see details, one can already say there are some regions that tend to have significant terms and some that do not (Figure 3.4 top). For instance, there are very few terms placed in the region characterized by strong up-regulation; it will probably be difficult to say much about the up-regulation of these genes using BioBridge alone. The region to the left of the center displays many terms. Zooming in to get a better view, the word "cycle" appears often. This is interesting enough to prompt the user to map "cycle" across the dataset (Figure 3.4 bottom).

During the process of exploring this region, the dynamic table assembles a list of

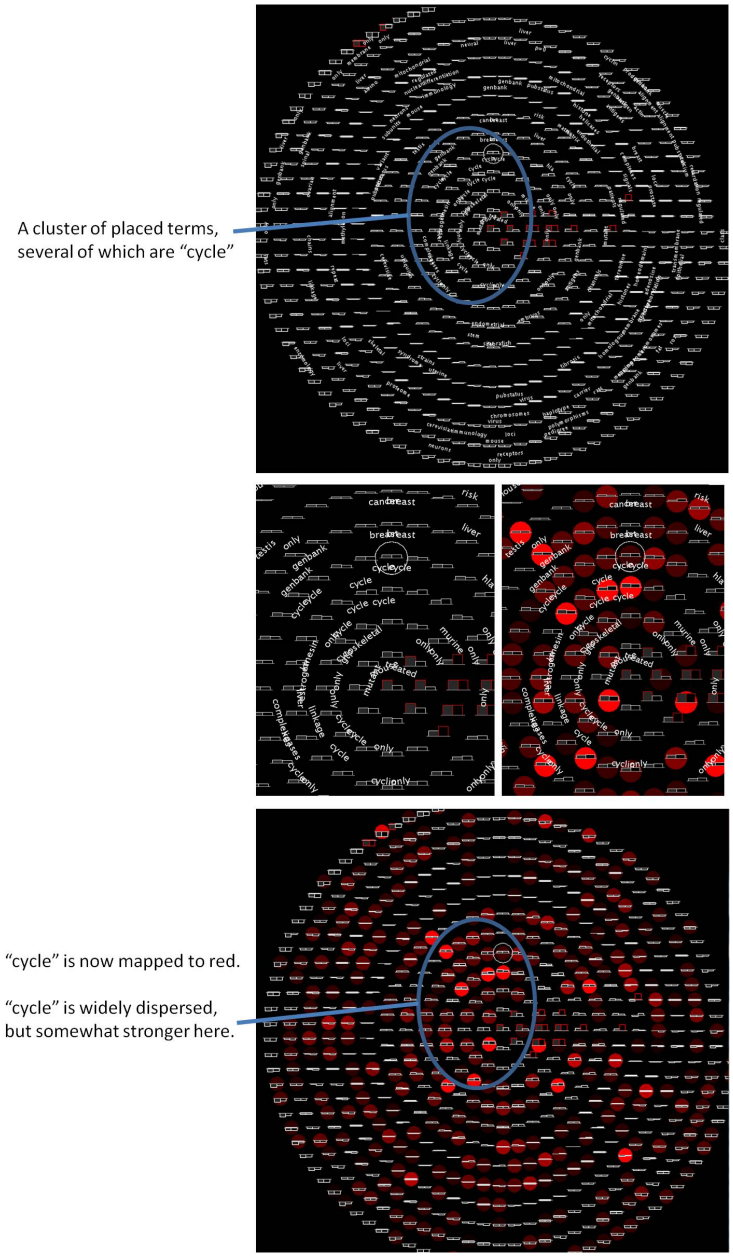
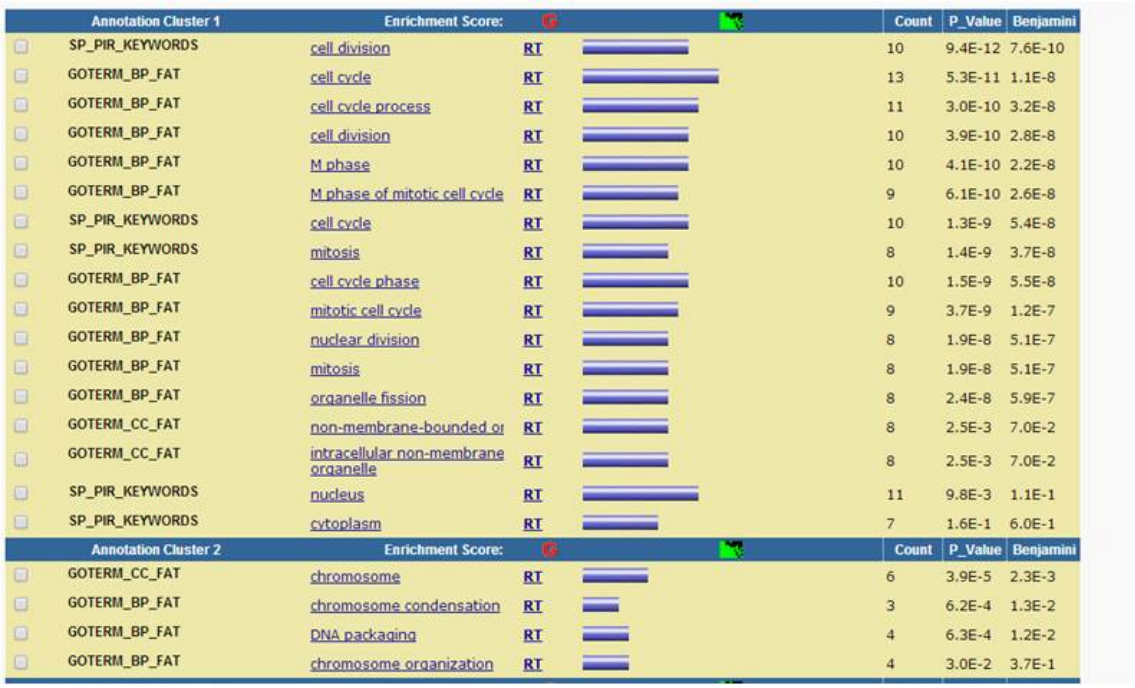



Figure 3.4: Open exploration of terms.

Using the Place Terms function, terms have been placed wherever glyphs have enough text mining results in common. Often, these terms form some regions where they are noticeably dense and others where they are noticeably sparse. Only the single strongest term is displayed so further queries are necessary. In this example the term “cycle” is conspicuously prevalent in the indicated region. This caused the user to visualize “cycle” across the dataset.



| Annotation Cluster 1 | Enrichment Score: | | | Count | P_Value | Benjamini |
|---|---|---|---|---|---|---|
| SP_PIR_KEYWORDS | cell division | RT | | 10 | 9.4E-12 | 7.6E-10 |
| GOTERM_BP_FAT | cell cycle | RT | | 13 | 5.3E-11 | 1.1E-8 |
| GOTERM_BP_FAT | cell cycle process | RT | | 11 | 3.0E-10 | 3.2E-8 |
| GOTERM_BP_FAT | cell division | RT | | 10 | 3.9E-10 | 2.8E-8 |
| GOTERM_BP_FAT | M phase | RT | | 10 | 4.1E-10 | 2.2E-8 |
| GOTERM_BP_FAT | M phase of mitotic cell cycle | RT | | 9 | 6.1E-10 | 2.6E-8 |
| SP_PIR_KEYWORDS | cell cycle | RT | | 10 | 1.3E-9 | 5.4E-8 |
| SP_PIR_KEYWORDS | mitosis | RT | | 8 | 1.4E-9 | 3.7E-8 |
| GOTERM_BP_FAT | cell cycle phase | RT | | 10 | 1.5E-9 | 5.5E-8 |
| GOTERM_BP_FAT | mitotic cell cycle | RT | | 9 | 3.7E-9 | 1.2E-7 |
| GOTERM_BP_FAT | nuclear division | RT | | 8 | 1.9E-8 | 5.1E-7 |
| GOTERM_BP_FAT | mitosis | RT | | 8 | 1.9E-8 | 5.1E-7 |
| GOTERM_BP_FAT | organelle fission | RT | | 8 | 2.4E-8 | 5.9E-7 |
| GOTERM_CC_FAT | non-membrane-bounded or | RT | | 8 | 2.5E-3 | 7.0E-2 |
| GOTERM_CC_FAT | intracellular non-membrane organelle | RT | | 8 | 2.5E-3 | 7.0E-2 |
| SP_PIR_KEYWORDS | nucleus | RT | | 11 | 9.8E-3 | 1.1E-1 |
| SP_PIR_KEYWORDS | cytoplasm | RT | | 7 | 1.6E-1 | 6.0E-1 |
| **Annotation Cluster 2** | **Enrichment Score:** | | | **Count** | **P_Value** | **Benjamini** |
| GOTERM_CC_FAT | chromosome | RT | | 6 | 3.9E-5 | 2.3E-3 |
| GOTERM_BP_FAT | chromosome condensation | RT | | 3 | 6.2E-4 | 1.3E-2 |
| GOTERM_BP_FAT | DNA packaging | RT | | 4 | 6.3E-4 | 1.2E-2 |
| GOTERM_BP_FAT | chromosome organization | RT | | 4 | 3.0E-2 | 3.7E-1 |

Figure 3.5: From BioBridge into DAVID.
These are some of the results returned by DAVID after inputting the entities indicated in Figure 3.4. Only the top two annotation clusters of six are shown. After querying entities in BioBridge, inputting them into DAVID was done by copy-pasting from the dynamic table.

all entities queried. This list may be saved for later use or copy-pasted into a tool such as DAVID. In fact, inputing the 26 genes represented in this "cycle" enriched area into DAVID's annotation clustering method results in 6 clusters, the top 2 of which are presented in Figure 3.5. The most significant annotation cluster represents about 10 of the genes present and is characterized by cell cycle and mitosis functional terms. The p-values calculated by DAVID are certainly optimistic considering the term "cycle" was the reason for this query. However, this does validate BioBridge's use of text mining since the two tools are in agreement here.

This hypothetical analysis shows that BioBridge supports both targeted validation and open exploration. The biology expert can locate entities and terms they believe to be interesting, observing expected relationships. During this process, the

biologist is presented with views that provide the opportunity to observe and begin exploring unexpected relationships. The ability to easily switch back and forth between targeted and open analysis is a key feature of BioBridge.

# Chapter 4

# Conclusions

In this thesis, the design and evaluation of the BioBridge interactive visualization tool was presented. It is intended to aid non-computational biologists by making data exploration accessible and intuitive. BioBridge integrates user supplied tabular data with text mining results and allows users to explore relationships within and between the two. BioBridge is targeted at tabular datasets; these are extremely common in biological data. Indeed, tabular omics-style data is already central to biology research and additional high-throughput methods continue to be developed.

BioBridge was designed by considering the fundamental tasks biologists perform when attempting to interpret new datasets. While the evaluated datasets fall short of big data, they are of sufficient size that they are not readily approachable without computational methods. A primary visualization window presents a similarity based arrangement of glyphs representing biological entities along with textual information in the form of locally shared terms and as well as selected terms across all data instances. This primary window supports direct manipulation mouse-driven interactions for both querying and navigation. A secondary control window provides a familiarly structured GUI to enhance data presentation and manipulation.

Biologist users were incorporated early on in the design process and initial user feedback on BioBridge was encouraging. BioBridge is capable of facilitating observations made using other methods as well as exposing possible novel insights missed by less exploratory methods. Perhaps most encouraging, evaluators used words like cool and fun to describe BioBridge. If BioBridge simply makes initial data analysis and exploration more accessible, easier, and faster for non-data science experts, it is a significant contribution.

There are several future avenues of improvement for BioBridge. In rough order of priority:

1. Conducting formal evaluations to more rigorously validate design choices.

2. Streamlining the text mining pipeline, ideally to the point where a text mining novice could make informed modifications to corpus assembly and term vector calculation parameters.

3. Increasing integration with other analysis tools. Ideally, a user could select groups of entities and directly analyze them in another tool such as DAVID. This can be accomplished by copy-pasting from the dynamic table currently, but a direct method would be ideal.

4. It should be easy to replace text mining with gene ontology or pathway membership information. Either of these could be expressed as a table of binary memberships, allowing them to simply replace the term vector table that results from text mining as input for BioBridge. This has not been attempted but would be useful.

5. Exploring strategies to address scalability issues via use cases of thousands of entities and dozens of dimensions. Entity scalability could likely be increased

with a combination of clustering and more flexible filtering. Dimensional scalability will require reconsideration of glyph design and probably some algorithm implementation improvements.

6. Developing and comparing alternative glyph layout methods for the primary visualization. The challenge here would be spacing; equally spaced glyphs make placing text much simpler, easier to interpret, and eliminates the possibility of glyphs obscuring one another. Additional glyph layout strategies based on established dimensional reduction techniques, such as principal component analysis (PCA) or multidimensional scaling (MDS), are worth considering.

7. Placing text on its own is a rudimentary solution. More sophisticated methods such as delineating regions of term similarity warrant consideration, and if effective, would reduce the previously mentioned need for regular glyph spacing.

8. The choice of direct RGB mappings limits the visualization to 3 simultaneous term visualizations across the dataset. More nuanced color mapping strategies could increase this number; as could increasing the complexity of the shaded circles to employ differently shaded pie slices or concentric circles.

Though a top priority, conducting formal evaluations is time consuming. It may be wiser to implement some of the other improvements listed prior to investing the time required to formally evaluate BioBridge.

In closing, BioBridge is a proof of concept for developing tools that biologists find accessible and intuitive. The use case studies presented, while anecdotal, are very encouraging. If maintained and updated, BioBridge could become more flexible and usable as additional use case datasets are explored and more feedback is gathered.

# Appendix A

# Supported Interactions

## A.1 Primary Visualization

1. Left mouse button
   (a) Single click – query 3 nearest glyphs
   (b) Double click - query single nearest glyph and set as focus
   (c) Click and drag - preview 3 nearest glyphs, does not query
2. Right mouse button
   (a) Click and drag - move/pan camera
3. Scroll wheel
   (a) spin - zoom camera in and out

## A.2 Control Window

1. File

(a) Open - opens a standard file explorer dialog to load a new dataset to analyze. If loading the dataset fails, file format requirements are presented

(b) Save Image - saves the current view in the primary visualization as a .tiff file

(c) Save Table - saves the current dynamic table as a .csv file

(d) Exit  closes the program

2. Find

(a) Entity - opens a series of dialog windows for the user to specify an entity name to locate

(b) Term - as above, but for terms

3. Visualization

(a) Place Terms – automatically queries between all adjacent sets of 3 entities, placing the highest ranking term provided it is above the input threshold

(b) Clear Terms – removes all placed terms

(c) Reset Camera – returns camera to original position, where root entity is at center

4. Table

(a) Rows: Remove Selected – removes any rows that are currently selected in the dynamic table

(b) Rows: Clear All – removes all rows

(c) Columns: Remove Selected - removes any columns that are currently selected in the dynamic table

(d) Columns: Clear All – removes all columns

5. Help

   (a) General – explains the general purpose of BioBridge and the basic functionality

   (b) Controls – lists supported mouse actions and keyboard shortcuts

   (c) Data Input Formats – describes formating requirements of the users data and supporting term vector file

# Appendix B

# Text Mining

While text mining is not part of BioBridge's functionality, the results of text mining are an essential input. As such, it is worth discussing relevant issues and describing the procedures used for the case studies described in the results section. The overall process is outlined in Figure B.1.

## B.1 Document Retrieval

Retrieving relevant documents is a prerequisite for effectively performing subsequent text mining operations. The goal is to retrieve a series of documents for each entity to be visualized. For all datasets discussed here, this involves downloading a set of abstracts from PubMed. Locating these abstracts can be accomplished a number of ways. The primary strategy employed relied on PubMed's search features and integration with other NCBI databases. For metabolites, alternate strategies were explored that utilized the curated list of reference articles provided by KEGG for each metabolite. These reference articles primarily deal with compound purification and identification; this was not the type of information the evaluator was interested in. This highlighted the importance of carefully considering document retrieval

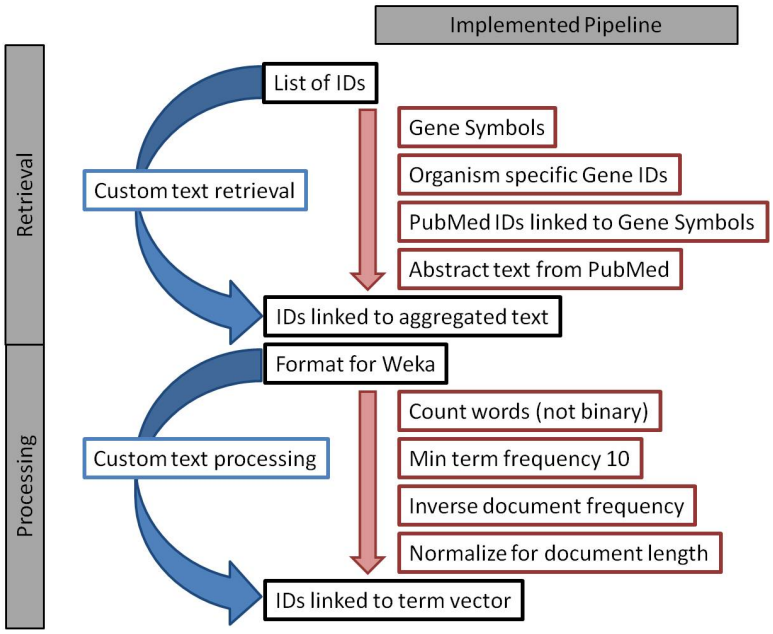



Figure B.1: Text mining for BioBridge input.
Text mining consists of two stages, an initial retrieval stage and a final processing stage. The implemented pipeline is outlined on the right in red. On the left (in blue), are alternative application specialized or possibly user controlled text mining routes. A specialized text mining process only has to rejoin the main pipeline as aggregated text or a term vector table to be compatible with BioBridge.

methods.

For the microarray datasets, searching for abstracts was straightforward since the majority of microarray probes are associated with gene symbols. Given this list of gene symbols, the next decision to make was how to formulate PubMed searches. After some experimentation, the Gene Name and Organism tags were used to specifically retrieve a gene ID that ties directly to PubMed articles. There are many other potentially relevant ways to conduct a PubMed search for text mining input. One can imagine filtering by date, sets of organisms, or simply searching for the gene's symbol, to maximize results, being of interest in certain contexts. Regardless, the result is a list of PubMed IDs that may then be downloaded from PubMed as abstracts and assembled as a document corpus for input to the processing

stage.

## B.2 Text Processing

Once abstracts have been retrieved, the task is to convert documents of text to numerical term vectors. After some initial processing that removes extraneous formatting and punctuation, the remainder of processing is done in a freely available data mining program called Weka that is offered by the University of Waikato. Weka may be downloaded from http://www.cs.waikato.ac.nz/ml/weka/. Installation and usage documentation is also provided. The decision to rely on Weka was made because its GUI makes it fairly easy to start using. This makes it possible for the text mining pipeline to be put under direct control of users at this point. While an automated option will still be provided, the third evaluator expressed a lot of interest in controlling the text mining process. The remainder of this section describes the general text mining procedures applied to the use case datasets as well as the reasoning behind them.

Weka provides a function for converting text to bag-of-words term vectors. Unlike some natural language processing methods, bag-of-words does not preserve any sentence structure or context but only tallies each occurrence of a word. There are several parameters that may drastically affect the process of converting text to bag-of-words. Firstly, users are unlikely to be interested in terms that occur in most or all of the documents. Obvious examples are "the" and "and" but terms such as "gene" are likely candidates in the biological domain. Therefore, each term is divided by the percentage of documents it appears in. This magnifies the impact of infrequent terms and diminishes the impact of uninteresting terms. It is also generally appropriate to discard exceptionally rare terms. While these terms may be

important to a specific entity, they add a lot of noise to the dataset and would not help in finding similarities among entities. Therefore, a minimum frequency threshold of 5 documents was applied for keeping each term. No maximum threshold was applied. Instead, an inverse document frequency (idf) transform was applied to ensure overly common words had very low scores.

Normalization based on the amount of text available for a given entity was also applied. This is because the number of articles related to each entity is highly variable. Using the metabolite dataset to illustrate, there are many thousands of articles that reference glucose or ATP, but most of the metabolites have fewer than 100 relevant articles. This raises two issues, there will be far more terms associated with the over represented entities and the counts of these terms will be much higher. This is addressed by normalizing text mining results for each entity relative to the size of the source text.

# Bibliography


[1] Sophia Ananiadou, Sampo Pyysalo, Junichi Tsujii, and Douglas B Kell. Event extraction for systems biology by text mining the literature. *Trends in biotechnology*, 28(7):381–390, 2010.

[2] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[3] J Christopher Bare, Tie Koide, David J Reiss, Dan Tenenbaum, and Nitin S Baliga. Integration and visualization of systems biology data in context of the genome. *BMC bioinformatics*, 11(1):382, 2010.

[4] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. Ncbi geo: archive for functional genomics data setsupdate. *Nucleic acids research*, 41(D1):D991–D995, 2013.

[5] Doug A Bowman, Ernst Kruijff, Joseph J LaViola Jr, and Ivan Poupyrev. *3D user interfaces: theory and practice*. Addison-Wesley, Boston, 2004.

[6] Ross E Curtis, Peter Kinnaird, and Eric P Xing. Genamap: visualization strategies for structured association mapping. In *Biological Data Visualization (BioVis), 2011 IEEE Symposium on*, pages 87–94. IEEE, 2011.

[7] Jan Czarnecki, Irene Nobeli, Adrian M Smith, and Adrian J Shepherd. A text-mining system for extracting metabolic reactions from full-text articles. *BMC bioinformatics*, 13(1):172, 2012.

[8] Glynn Dennis Jr, Brad T Sherman, Douglas A Hosack, Jun Yang, Wei Gao, H Clifford Lane, Richard A Lempicki, et al. David: database for annotation, visualization, and integrated discovery. *Genome biol*, 4(5):P3, 2003.

[9] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[10] Nils Gehlenborg, Seán I O'Donoghue, Nitin S Baliga, Alexander Goesmann, Matthew A Hibbs, Hiroaki Kitano, Oliver Kohlbacher, Heiko Neuweger, Reinhard Schneider, Dan Tenenbaum, et al. Visualization of omics data for systems biology. *Nature methods*, 7:S56–S68, 2010.

[11] James Gosling. *The Java language specification*. Addison-Wesley Professional, Boston, 2000.

[12] Laura Gramantieri, Pasquale Chieco, Catia Giovannini, Michela Lacchini, Davide Treré, Gian Luca Grazi, Annamaria Venturi, and Luigi Bolondi. Gadd45-a expression in cirrhosis and hepatocellular carcinoma: relationship with dna repair and proliferation. *Human pathology*, 36(11):1154–1162, 2005.

[13] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[14] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[15] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, 40(D1):D109–D114, 2012.

[16] Brian Kemper, Takuya Matsuzaki, Yukiko Matsuoka, Yoshimasa Tsuruoka, Hiroaki Kitano, Sophia Ananiadou, and Jun'ichi Tsujii. Pathtext: a text mining integrator for biological pathway visualizations. *Bioinformatics*, 26(12):i374–i381, 2010.

[17] Neesha Kodagoda, Simon Attfield, BL Wong, Chris Rooney, and Sharmin Choudhury. Using interactive visual reasoning to support sense-making: Implications for design. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2217–2226, 2013.

[18] Ivica Letunic, Takuji Yamada, Minoru Kanehisa, and Peer Bork. ipath: interactive exploration of biochemical pathways and networks. *Trends in biochemical sciences*, 33(3):101–103, 2008.

[19] Philip Livengood, Ross Maciejewski, Wei Chen, and David S Ebert. A visual analysis system for metabolomics data. In *Biological Data Visualization (BioVis), 2011 IEEE Symposium on*, pages 71–78. IEEE, 2011.

[20] Steven Maere, Karel Heymans, and Martin Kuiper. Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–3449, 2005.

[21] Vivien Marx. Biology: The big challenges of big data. *Nature*, 498(7453):255–260, 2013.

[22] Heiko Neuweger, Marcus Persicke, Stefan P Albaum, Thomas Bekel, Michael Dondrup, Andrea T Hüser, Jörn Winnebald, Jessica Schneider, Jörn Kalinowski, and Alexander Goesmann. Visualizing post genomics data-sets on customized pathway maps by prometra–aeration-dependent gene expression and metabolism of corynebacterium glutamicum as an example. *BMC systems biology*, 3(1):82, 2009.

[23] Chikashi Nobata, Paul D Dobson, Syed A Iqbal, Pedro Mendes, Junichi Tsujii, Douglas B Kell, and Sophia Ananiadou. Mining metabolites: extracting the yeast metabolome from the literature. *Metabolomics*, 7(1):94–101, 2011.

[24] Steffen Oeltze, Paul Klemm, Reyk Hillert, Bernhard Preim, and Walter Schubert. Visualization and exploration of 3d toponome data. In *VCBM*, pages 115–122, 2012.

[25] Charles E Robertson, J Kirk Harris, Brandie D Wagner, David Granger, Kathy Browne, Beth Tatem, Leah M Feazel, Kristin Park, Norman R Pace, and Daniel N Frank. Explicet: graphical user interface software for metadata-driven management, analysis and visualization of microbiome data. *Bioinformatics*, 29(23):3100–3101, 2013.

[26] Carlos Rodríguez-Penagos, Heladia Salgado, Irma Martínez-Flores, and Julio Collado-Vides. Automatic reconstruction of a bacterial regulatory network using natural language processing. *BMC bioinformatics*, 8(1):293, 2007.

[27] Rintaro Saito, Michael E Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, Samad Lotia, Alexander R Pico, Gary D Bader, and Trey Ideker. A travel guide to cytoscape plugins. *Nature methods*, 9(11):1069–1076, 2012.

[28] Eric W Sayers, Tanya Barrett, Dennis A Benson, Evan Bolton, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Scott Federhen, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 38(suppl 1):D5–D16, 2010.

[29] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[30] Zhiao Shi, Jing Wang, and Bing Zhang. Netgestalt: integrating multidimensional omics data over biological networks. *Nature methods*, 10(7):597–598, 2013.

[31] Michael E Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, and Trey Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432, 2011.

[32] Michael Snyder, Sherman Weissman, and Mark Gerstein. Personal phenotypes to go with personal genomes. *Molecular systems biology*, 5(1), 2009.

[33] Sally Trabucco. Gadd45-a and metastatic hepatocellular carcinoma cell migration, 2010. Worcester Polytechnic Institute, Major Qualifying Project.

[34] Aditya Vailaya, Peter Bluvas, Robert Kincaid, Allan Kuchinsky, Michael Creech, and Annette Adler. An architecture for biological information extraction and representation. *Bioinformatics*, 21(4):430–438, 2005.

[35] Matthew Ward, Georges Grinstein, and Daniel Keim. *Interactive data visualization: foundations, techniques, and applications*. AK Peters, Ltd., Natick, MA, 2010.

[36] Matthew O Ward. A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization*, 1(3-4):194–210, 2002.

[37] Matthew O Ward and Benjamin N Lipchak. A visualization tool for exploratory analysis of cyclic multivariate data. *Metrika*, 51(1):27–37, 2000.

[38] Takuji Yamada, Ivica Letunic, Shujiro Okuda, Minoru Kanehisa, and Peer Bork. ipath2. 0: interactive pathway explorer. *Nucleic acids research*, 39(suppl 2):W412–W415, 2011.

[39] Bing Zhang, Stefan Kirov, and Jay Snoddy. Webgestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic acids research*, 33(suppl 2):W741–W748, 2005.

[40] Jian Zhao, Christopher Collins, Fanny Chevalier, and Ravin Balakrishnan. Interactive exploration of implicit and explicit relations in faceted datasets. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2080–2089, 2013.