BioLinker: Bottom-up Exploration of Protein Interaction Networks

Tommy Dang*
Department of Computer Science
Texas Tech University

Paul Murray†
Department of Computer Science
University of Illinois at Chicago

Angus Forbes‡
Department of Computer Science
University of Illinois at Chicago

Abstract—Systems biologists and cancer researchers require interactive visualization tools that enable them to more easily navigate and discover patterns at different levels of the biological hierarchy of signaling pathways. Furthermore, biologists are often interested in understanding and exploring the causal biochemical links between processes. When exploring the literature of particular biological pathways or specific proteins within those pathways, biologists find it useful to know the contexts in which biochemical links are active and, importantly, to be aware of potential conflicts when different experiments introduce alternative interpretations of the function of a pathway or biochemical reaction. We introduce *BioLinker*, a interactive visualization system that helps users to perform bottom-up exploration of complex protein interaction networks. Five interconnected views provide the user with a range of ways to explore pathway data, including views that show potential conflicts within pathway databases and publications and that highlight contextual information about individual proteins. Additionally, we discuss system details to show how our system manages the large amount of protein interactions extracted from the literature of biological pathways.

Index Terms—Protein interaction network, Dynamic network visualization, Potential conflict matrix.

1 Introduction

Molecular and systems biologists are tasked with the comprehension and analysis of incredibly complex networks of biochemical interactions, called pathways, that occur within every cell. It quickly becomes unwieldy to represent even a subset of all interactions between proteins, complexes, and biochemical reactions within a pathway, resulting in a visually cluttered "hairball" of tangled edge crossings. It is highly desirable to have interactive navigation tools to more effectively inspect the complex biological networks. In this paper, we propose *BioLinker*, a system that facilitates a bottom-up exploration of protein interaction networks. A user can initiate an exploration from a small subnetwork that encircles a particular protein of interest and then iteratively expand the network on demand by choosing individual elements. *BioLinker* provides multiple views that each display a different aspect of the currently selected subnetwork, such as the contexts of the protein interactions and the publication data related to the discovery of these interactions. These views are all linked to provide important supplementary information through brushing and linking.

We also describe the underlying infrastructure that allows *BioLinker* to load and filter the large amount of biological data at interactive rates. Our database currently contains over 3.5 million protein interactions extracted from more than 290,000 publications. The complete API functions are available at http://ccrg-data.evl.uic.edu/index-cards/explorer.

2 Related Work

Pathways may contain dozens or even hundreds of biomolecules. Reactions between these biomolecules include state transitions, such as activation, and frequently involve multiple inputs and outputs. While some reactions can be usefully thought of as taking place one after another, in a linear series of steps, others are more accurately modeled as feedback loops [6] where the outputs from one stage inhibit or activate the inputs to another stage. Furthermore, a system-level understanding of a complex pathway involves a more holistic understanding than depicting individual interactions may provide. Visualizing complexity while also presenting a high-level overview of such protein interaction networks is a significant challenge [17].

To tackle these challenges, numerous visualization techniques have been proposed. *PathwayMatrix* [7] represents binary relations between pairs of proteins and biomolecules. Various strategies are used to reorder proteins in the matrix; this enables the identification network structures which is difficult to extract using traditional node-link diagrams with force-directed layouts, such as *PCViz* [3] or *VisANT* [11]. However, a drawback of matrix representations of networks is that it is difficult to trace a path between nodes. In contrast, *RectionFlow* [6] emphasizes both the structural and causal relationships among proteins, complexes, and biochemical reactions within a given pathway. Other hybrid forms, such as *BioFabric* [14] and *Compressed Adjacency Matrices* [9], also attempt to reduce the number of edge crossings while enabling the traceability of network flows.

Our scientific understanding of many signaling pathways is incomplete, both with respect to participating cellular components and their conditions (or *context*, e.g., a pathway being active only during "late phases of tumorigenesis"). Recent efforts in text mining focus on extracting causal mechanisms and contexts from research abstracts and scientific papers [20–22]. These efforts generate a large amount of data (millions of protein interactions together with context knowledge and expressions of uncertainty). Moreover, aligning this extracted evidence to existing databases of biochemical pathways, such as *Pathway Commons* [4] and *Wiki Pathways* [13], is a daunting task.

In this paper, we propose *BioLinker*, a system that facilitates a bottom-up exploration of protein interaction networks. A user can initiate an exploration from a small subnetwork that is connected to a particular protein of interest and then iteratively expand the network on demand by choosing individual elements. *BioLinker* provides multiple views that each displays a different aspect of the currently selected subnetwork, such as the contexts of the protein interactions and the publication data related to the discovery of these protein interactions. Each of these views— *overview*, *main network view*, *context view*, *conflict matrix view*, and *publication view*— are all linked to provide important supplementary information through brushing and linking.

3 Overview of Visualization Tasks

We worked closely over a period of twelve months with three computational and molecular biologists to identify and prioritize visualization tasks in visualizing complex protein interaction networks. Our in-depth interviews with these experts led us to identify six important tasks that were not currently well-supported in existing visualization tools. The six visualization tasks, described in more detail later in the paper, include:

- **T1** Starting with specific protein, the visualization should allow users to iteratively expand the network on demand.
- **T2** Overlaying cancer genomics data onto the network.
- **T3** Finding the paths between two proteins based on user-specified number of hops.
- **T4** Comparing protein interaction sub-networks by context.

*e-mail: tommy.dang@ttu.edu
†e-mail: pmurra5@uic.edu
‡e-mail: aforbes@uic.edu

Manuscript received 31 March 2007; accepted 1 August 2007; posted online 2 November 2007.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

T5 Displaying publication data on request.
T6 Enabling conflict detections in literature regarding protein interactions.

The process of developing *BioLinker* was an iterative process that relied on feedback from the domain experts at various stages of the design and implementation. Many of the ideas that were included were first suggested by one of the domain experts, or emerged organically through conversations with them. The contributions of this paper are centered around building the software system that can handle the increasingly large amount of generated data and allow users to perform the above visualization tasks.

4 METHODS

In this section we describe the primary components of the *BioLinker* visualization. *BioLinker* uses a client-server architecture to handle large amount of data from various databases. The server contains four components:

- **Publication database**: We currently store over 290,000 PubMed articles on our server.
- **Index card database**: An *index card* captures information, such as participants and interaction type, extracted from text in an article. Index cards are stored in structured data format to ensure that systems provide results in a single format that is computable but also human readable and editable.
- **Comparing index cards**: For every two given indexes, we identify if there are potential conflicts between them. A comparator script is deployed on our server for this purpose.
- **Genomic alteration database**: *BioLinker* accesses the cBioPortal [10] for Cancer Genomics through its web API. The portal was originally developed at Memorial Sloan-Kettering Cancer Center.

The client side contains five components (depicted in Fig. 1):

- **Overview/Protein Selector**: This panel allows users to select an initial protein.
- **Main view**: Within this view, users can iteratively expand the network by clicking on individual protein names.
- **Context view**: This view shows statistics about the contexts of proteins (including species, cell type, organism) for the selected index cards in the main view.
- **Publication view**: This view shows the discoveries related to how particular proteins interact with each other (as encapsulated by the index cards) and how our understanding of these interactions has changed over time.
- **Conflict matrix**: The index card comparisons are presented in form of an adjacency matrix.

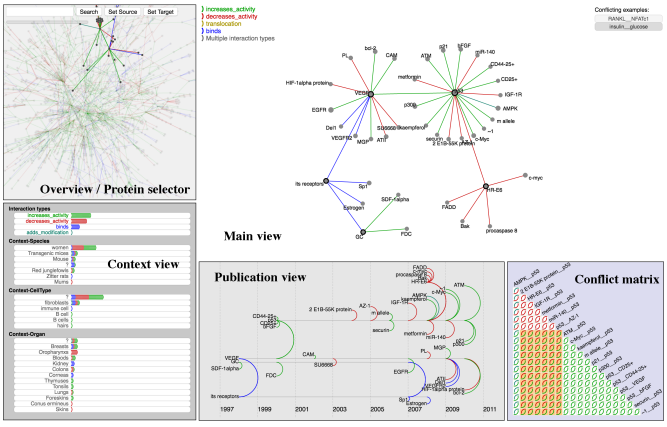

Fig. 1. Client side of *BioLinker* visualization has multiple views.

These views are all interconnected through brushing and linking. For example, when a user selects a link (index card) between two proteins in the main view, the publication view shows the associated publication information (such as paper id, paper title, authors, journal name, and publication date) while the context view shows the species name, organ, and cell type for that protein interaction.

4.1 Publication database

PubMed Central (PMC) is a free full-text archive of biomedical and life sciences journal literature. We currently store over 290,000 PMC publications on our server. PMC uses XML tags to encode information about journal article submissions. These XML elements have specific style rules associated with them. There are many elements in a PMC publication, and we show the most important subset of these XML elements (as requested by our expert users), including the paper title, authors' names, authors' contact information and affiliations, publication year, journal name, and external link to the actual paper when it is available online. An example of PMC paper information stored on our server can be found at http://ccrg-data.evl.uic.edu/index-cards/api/NXML/PMC1174968.

4.2 Index card database

The index card database is provided by the REACH Biomedical Information Extraction research group at University of Arizona. Their text extraction system automatically "reads" research papers and abstracts to construct causal models of biological pathway data and to facilitate the exploration and analysis of these models [20, 22]. For each paper presented to the REACH reading system, a set of index cards is produced that capture the interactions reported in the paper along with the experimental evidence presented in the text as well as the relationships between the extracted interactions and the provided BioPAX model.

An index card is a JSON object containing the following fields: participants, interaction type (binds, adds_modification, removes_modification, translocates), context (cell line, cell type, organism, tissue type), pmc_id (the paper the result came from), evidence_text (the text from the paper containing the information in the result.

Given that index cards are output in the JSON format, a MongoDB database is used to store index cards along with their associated publications (as PMC XML documents) and participants. The use of MongoDB also allows for a flexible document schema, which works well with the highly variable nature of index card documents. Since MongoDB does not support XML directly, PMC XML files are stored as binary objects. Index cards are stored in their raw form, and each index card is associated with the PMC XML document that it was extracted from. Each participant is also extracted from each index card and stored in a separate MongoDB collection, allowing for queries over participants.

A web API was created to serve index card, publication, and participant data. The API exposes a flexible query structure, allowing end users to query relations between index cards, participants, and PMC XML data in a variety of ways. An example of index cards on our server can be found at http://ccrg-data.evl.uic.edu/index-cards/api/IndexCards/findOne.

4.3 Index cards comparator

When there are multiple connections between two proteins/complexes, it may fall into one of the two following interesting circumstances: (1) If they have the same color (same interaction type), these are supporting evidences in different publications which confirm the interaction between two elements. (2) If they have the opposite colors (opposite interaction types, such as *increase_activity* versus *decrease_activity*). Domain experts are very interested in the latter case where there are conflicting evidences about the knowledge obtained from different publications (and usually in different years). We will show examples of the second circumstance below when we describe the conflict matrix and the publication view.

The above circumstance is just one specific example of conflicting evidences in literature about the interactions between proteins/complexes. We have deployed a more general comparator to detect potential conflicts on our server. Conflicts may happen between index cards of the same interaction type (for example, translocation but in the opposite direction) or different sets of participants (for example, one is a subset/member of another protein family participant).

4.4 Genomic alteration database

The cBioPortal [10] is hosted at and maintained by the Memorial Sloan Kettering Cancer Center. It provides access to data by the Cancer

Genome Atlas as well as many carefully curated data sets. Currently, cBioPortal contains data from 147 cancer genomics studies (as of December 2016) and more than 17,000 tumor samples [1]. *BioLinker* accesses cBioPortal for Cancer Genomics through its web API.

4.5 Overview / Protein Selector

This panel provides an overview of thousand index cards sampled from millions of index cards in the database based on specified filtering conditions, such as protein interaction within the *alphat* cell line. Users can start by entering a protein name into a search box (as depicted in the left panel of Fig. 2). This will perform a request to load the selected protein and its immediate neighbors from our index card database. As users iteratively expand the subnetwork in the main view, the overview keeps track of the expanded sub-network over the overall context (as depicted in the right panel of Fig. 2).

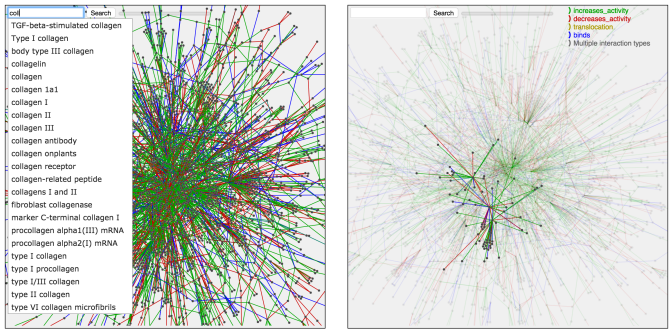

Fig. 2. Protein selector (using the search box) in *BioLinker*.

4.6 Main View

A subnetwork encircling the selected protein is initially shown in the main view. Users can iteratively expand the network by simply clicking on protein names (visualization task **T1**). A query of relevant index cards (containing neighboring proteins) is sent to the index card database per user request. Node (protein) sizes are computed based on the number of direct neighbors. Edges (index cards) are color-encoded by interaction types, such as green for *increase_activity* and red for *decrease_activity*. When users mouse over a protein name, a pop-up window displays protein information together with statistics of its immediate neighbors (if expanded) as depicted in Fig. 3.

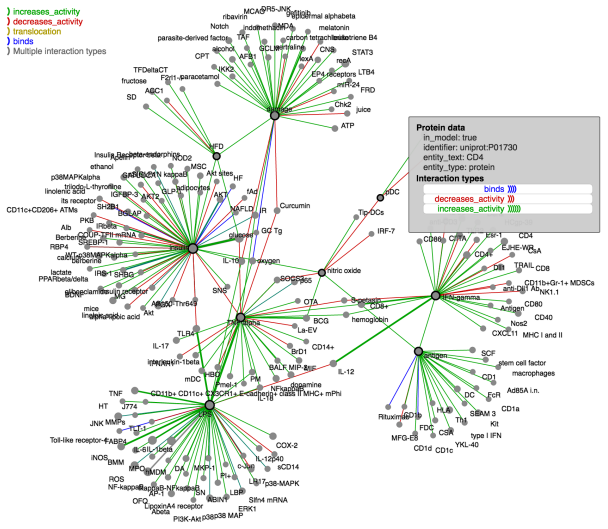

Fig. 3. Main view in *BioLinker*: displaying protein information together with statistics of its immediate neighbors on mousing over.

BioLinker supports finding paths between selected proteins (visualization task **T3**). Fig. 4 shows an example. Users specify source node (*PIK3CA* protein), target node (*TRAF6* protein), and the maximum number of hops in between source and target (5 hops). *BioLinker* displays all possible paths under that condition. The source node is pinned to the left while the target node is pinned to the right of the visualization. The shortest path from *PIK3CA* to *TRAF6* goes through two hops (*Akt* and *NF-kappaB*). By stretching source and target nodes to both sides of the force-directed layout, the shortest path is usually the horizontal path from left to right. This feature is useful in visualizing shortest path and is not available in popular visualization tools such as ChiBE 2 [1], Cytoscape [18], and PCViz [3].

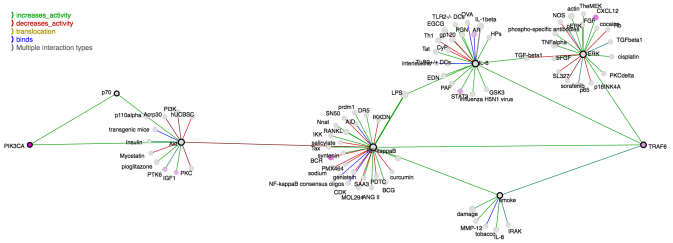

Fig. 4. Main View in *BioLinker*: Visualizing all possible paths from *PIK3CA* (left) to *TRAF6* (right) going through 5 or fewer hops.

BioLinker also supports overlaying cancer genomics data onto the network (visualization task **T2**). In Fig. 4, purple nodes are proteins with high copy number alteration in the Bladder Urothelial Carcinoma study (TCGA, Nature 2014). Notice that both source and target nodes are highly altered in this cancer study. *BioLinker* accesses this cancer study on cBioPortal through its web service interface (described in Section 4.4)

4.7 Context View

Working definition of context in text: An instance of context is a single assertion about species (such as human, mouse, and yeast), organ, tissue type, cell-line that holds across some region of text; any mechanism components mentioned in that region acquire that context assertion as a property.

Systems biologists and cancer researchers are frequently interested in understanding the contexts of biochemical reactions and comparing protein interaction sub-networks by context (visualization task **T4**). The left panel in Fig. 5(a) shows stacking plots [8] in our context view for the network in the main view (right panel). In particular, we show a 2-degree separation network of a selected protein (*antigen*) which is located in the center of the main view. Fig. 5(b) depicts brushing and linking of two views. We have selected *mouse* in the *species* category. Other context categories are updated accordingly. In the main view (on the right), we notice that all protein interactions in *mouse* are between *antigen* and its immediate neighbors, but not the second degree separated neighbors.

4.8 Publication View

We use TimeArcs visualization [5] to show the discoveries of these index cards by publication year. A request to load relevant publications for new index cards is sent to the PMC publication database on our server every time the protein network is expanded. Fig. 6(b) shows an example of publication view of the graph in Fig. 6(a). In particular, the time axis goes from left (2004) to right (2012). An arc connects two proteins/complexes at a particular time (based on when the interaction was discovered/ publication year). The link colors encode interaction types.

Mousing over a protein name displays all publications related to that protein. As depicted in Fig. 6(c), mousing over an arc displays publication data associated with an index card, such as paper title, authors' names, authors' contacts, affiliations, publication year, journal name, and external link. Users can go to the actual paper by clicking on the provided link. Notice that the evidence (the actual text in this paper where the protein interaction was extracted from) is highlighted in a different color (the color used to encode that interaction type). The user can also request to show the publication chart over time of a

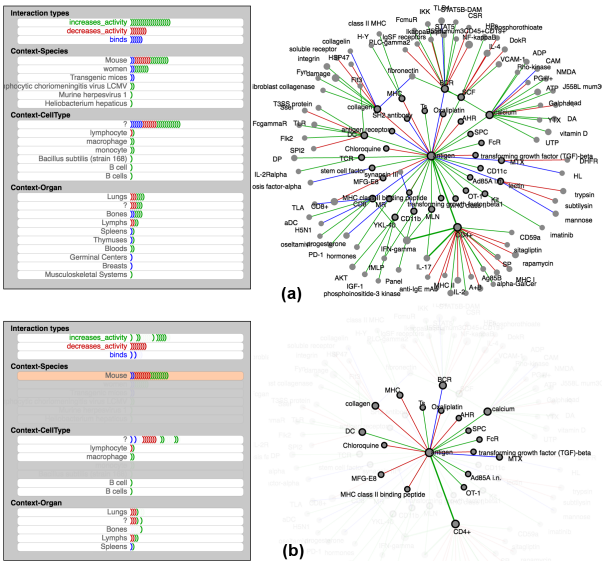

Fig. 5. Visualizing and interacting with context view in *BioLinker*: (a) Stacking plots (left) by species, cell type, and organ of the 2-degree separation network of *antigen* (right) (b) Selecting *mouse* in context view.

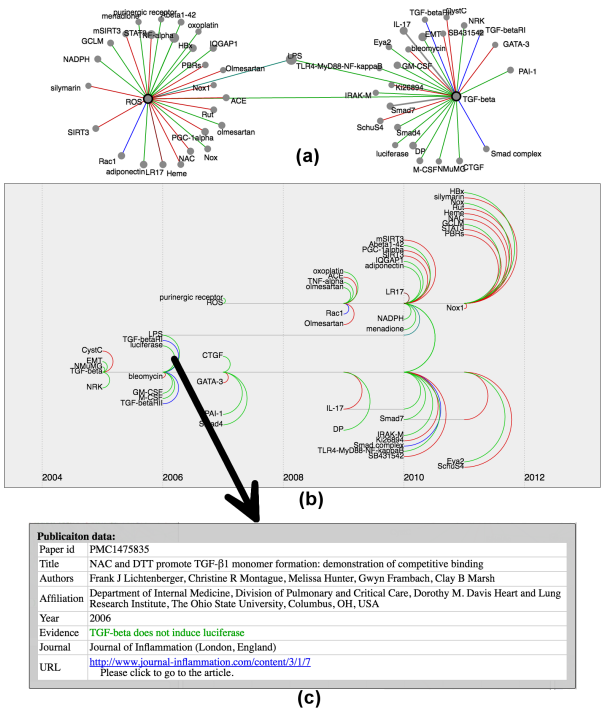

Fig. 6. Visualizing and interacting with publication view in *BioLinker*: (a) Sub-network of *TGF-beta* and *ROS* proteins in the main view (b) Publication view of the same data where time axis goes from left to right (c) Publication pop-up window when mousing over an arc in the publication view (at the origin of the arrow).

selected protein in our entire PMC publication database. (The earliest publications archived by PMC relevant to biological pathways date back to more than 60 years ago.)

4.9 Conflict Matrix

We use the index card comparator script on our server to detect potential conflicts between index cards in the main view. In Fig. 7, *TGF-beta* is selected and iteratively expanded. The index card comparison results are presented in the form of an adjacency matrix. Index cards in the conflict matrix are ordered by interaction type. In each cell, we draw an arc symbol for each interaction (colored by type) of the two participating index cards as depicted in the left panel of Fig. 7(a). The cell backgrounds are colored by the results returned from the comparator script on our server (orange are index cards with potential conflict). In Fig. 7(b), we inspect a potential conflict cell in the matrix (left). Two corresponding index cards are highlighted in the publication view (right). As depicted, the two index cards have the same participants (*TGF-beta* and *Smad7*), but opposite interaction types (*increase_activity* versus *decrease_activity*). This indicates conflicting knowledge obtained from two different publications in 2010 and 2011. We now look further into details of publication data for each index card to verify these conflicting information as shown in Fig. 7(c).

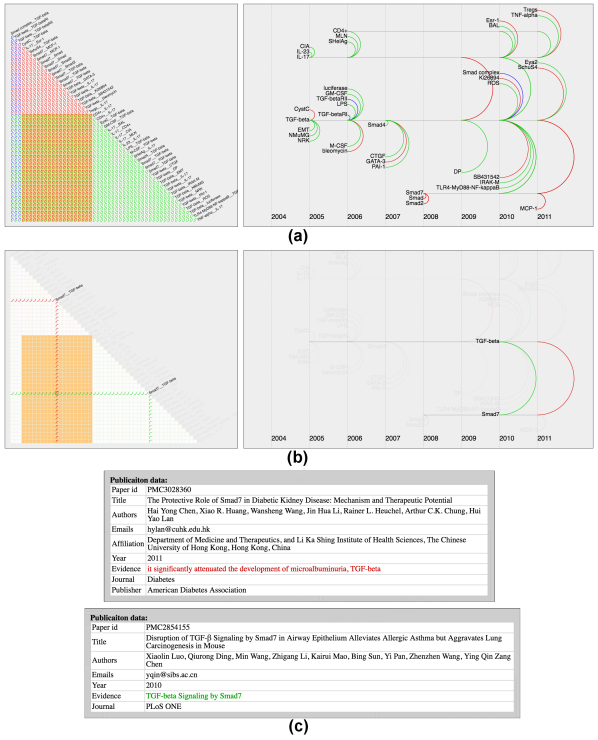

Fig. 7. Visualizing and interacting with conflict matrix in *BioLinker*: (a) Conflict matrix (left) and publication view (right) of the *TGF-beta* sub-network (b) Linking conflict matrix (left) and publication view (right) on mousing over a potential conflict cell (between two compared index cards) in the matrix (c) Publication data of the two highlighted index cards in (b).

5 Results and Discussion

5.1 Comparison of *BioLinker* to related tools

The matrix representation avoids edge crossings for dense networks. However, a drawback of all matrix representations is that paths between nodes are difficult to identify and trace. There are many visualization tasks related to tracing flows [6] through protein interaction network (see visualization tasks **T1**, **T3**, and **T4**). Therefore in this section, we compare *BioLinker* to related node-link representations (see Table 1) with respect to the visualization tasks defined previously in this paper.

	T1	T2	T3	T4	T5	T6
PCViz [3]		✓	✓			
ChiBE 2 [1]		✓	✓			
Cytoscape [18]	✓		✓			
Extended LineSets [15]	✓					
ReactionFlows [6]			✓			
BioLinker	✓	✓	✓	✓	✓	✓

Table 1. Comparisons of protein network visualization tools on six tasks.

This section does not mean to survey all visualization tools for biological network analysis [12, 16, 19]. Instead, we try to cover more recent tools on the selected visualization tasks, from the common tasks in network visualization (**T1**, **T2**, and **T3**) to more recent demands from domain experts due to the emerging availability of high volume data (**T4**, **T5**, and **T6**).

5.2 Implementation

BioLinker is implemented in D3.js [2]. The application, source code, sample data, and demo video are provided via our GitHub project repository, located at http://github.com/CreativeCodingLab/BioLinker.

REFERENCES

[1] O. Babur, U. Dogrusoz, M. Cakir, B. Aksoy, N. Schultz, C. Sander, and E. Demir. Integrating biological pathways and genomic profiles with chibe 2. *BMC Genomics*, 15(1):642, 2014.

[2] M. Bostock, V. Ogievetsky, and J. Heer. D3 data-driven documents. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2301–2309, 2011.

[3] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, Ö. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander. Pathway commons, a web resource for biological pathway data. *Nucleic acids research*, 39(suppl 1):D685–D690, 2011.

[4] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, . Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, 39(suppl 1):D685–D690, 2011.

[5] T. Dang, N. Pendar, and A. G. Forbes. Timearcs: Visualizing fluctuations in dynamic networks. *IEEE Conference on Visualization (EuroVis)*, 2016.

[6] T. N. Dang, P. Murray, J. Aurisano, and A. G. Forbes. ReactionFlow: An interactive visualization tool for causality analysis in biological pathways. *BMC Proceedings*, 9(6):S6, 2015.

[7] T. N. Dang, P. Murray, and A. G. Forbes. PathwayMatrix: Visualizing binary relationships between proteins in biological pathways. *BMC Proceedings*, 9(6):S3, 2015.

[8] T. N. Dang, L. Wilkinson, and A. Anand. Stacking graphic elements to avoid over-plotting. *IEEE Transactions On Visualization And Computer Graphics*, 16(6), 2010.

[9] K. Dinkla, M. A. Westenberg, and J. J. van Wijk. Compressed adjacency matrices: untangling gene regulatory networks. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2457–2466, 2012.

[10] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, and N. Schultz. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Science Signaling*, 6(269):pl1–pl1, 2013.

[11] Z. Hu, J.-H. Hung, Y. Wang, Y.-C. Chang, C.-L. Huang, M. Huyck, and C. DeLisi. Visant 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic acids research*, page gkp406, 2009.

[12] d. W. Huang, B. Sherman, and R. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(5):1–13, 2009-01-01 00:00:00.001.

[13] M. Kutmon, A. Riutta, N. Nunes, K. Hanspers, E. L. Willighagen, A. Bohler, J. Mlius, A. Waagmeester, S. R. Sinha, R. Miller, S. L. Coort, E. Cirillo, B. Smeets, C. T. Evelo, and A. R. Pico. Wikipathways: Capturing the full diversity of pathway knowledge. *Nucleic Acids Research*, 44(D1):D488–D494, 2016.

[14] W. J. Longabaugh. Combing the hairball with biofabric: a new approach for visualization of large networks. *BMC bioinformatics*, 13(1):275, 2012.

[15] F. Paduano and A. G. Forbes. Extended LineSets: A visualization technique for the interactive inspection of biological pathways. *BMC Proceedings*, 9(6):S4, 2015.

[16] G. A. Pavlopoulos, A.-L. Wegener, and R. Schneider. A survey of visualization tools for biological network analysis. *Biodata mining*, 1(1):1–11, 2008.

[17] P. Saraiya, C. North, and K. Duca. Visualizing biological pathways: requirements analysis, systems evaluation and research agenda. *Information Visualization*, 4(3):191–205, 2005.

[18] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[19] M. Suderman and M. Hallett. Tools for visually exploring biological networks. *Bioinformatics*, 23(20):2651–2659, 2007.

[20] M. A. Valenzuela-Escárcega, G. Hahn-Powell, T. Hicks, and M. Surdeanu. A domain-independent rule-based framework for event extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Assian Federation of Natural Language Processing: Software Demonstrations (ACL-IJCNLP)*, 2015.

[21] M. A. Valenzuela-Escarcega, G. Hahn-Powell, and M. Surdeanu. Description of the odin event extraction framework and rule language. *CoRR*, abs/1509.07513, 2015.

[22] M. A. Valenzuela-Escárcega, G. Hahn-Powell, and M. Surdeanu. Odin's runes: A rule language for information extraction. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC)*, 2016.