

Visualizing Aggregated Biological Pathway Relations

Byron Marshall, Karin Quiñones, Hua Su, Shauna Eggers, and Hsinchun Chen
University of Arizona, Department of Management Information Systems, Artificial Intelligence Lab
McClelland Hall, The University of Arizona, Tucson, Arizona 85721-0108

{byronm*,kdq,hsu,seggers,hchen}@eller.arizona.edu

ABSTRACT

The Genescene development team has constructed an aggregation interface for automatically-extracted biomedical pathway relations that is intended to help researchers identify and process relevant information from the vast digital library of abstracts found in the National Library of Medicine's PubMed collection. Users view extracted relations at various levels of relational granularity in an interactive and visual node-link interface. Anecdotal feedback reported here suggests that this multi-granular visual paradigm aligns well with various research tasks, helping users find relevant articles and discover new information.

Categories and Subject Descriptors

H.5.2 [User Interfaces]: Graphical user interfaces

General Terms

Algorithms, Human Factors

Keywords

Knowledge Aggregation, Visualization, Biological Pathway Relations

1. INTRODUCTION

Every day, more than 1,600 life science journal articles are added to the National Library of Medicine's PubMed service. As in other digital library applications, carefully crafted indexing and retrieval methodologies are needed to help users effectively harness the information in this vast and growing collection. A number of information extraction systems have been developed to more effectively utilize collections of biomedical text. These systems often use Natural Language Processing (NLP) techniques to extract biological pathway information from free text [1]. Biological pathways document how various components in a cell contribute to biochemical or cellular processes.

The Genescene system lets users view pathway relations extracted by the Arizona Relation Parser (ARP) [2] in a visual, node-link representation. Each ARP relation consists of two labeled entities and a labeled connector. Labels are name strings found in the text. Early users wanted better organized relations to reduce visual clutter. In previous work we outlined a framework for organizing

or "aggregating" extracted relations [3]. We aggregate by identifying features in the relations and using those features to combine and connect extracted information. An offline process identifies and stores features such as Substance (e.g. TP53), Function (e.g. apoptosis), Substance Type (e.g. gene vs. protein), Mutation (Yes/No/indeterminate), and Species for the entities and connectors in a set of ARP relations. Our query interface displays relevant relations from the stored set in response to user-supplied query terms.

To organize returned relations into a network, our aggregation approach establishes the equivalence of entities and connectors at different levels of granularity. Depending on the task, a researcher might want to consolidate items differently. For example, consider the two relations: "mutant p53 suppresses apoptosis" and "mutant p53 blocked E1A-induced apoptosis." At a very fine level of analysis, a researcher might want to differentiate "suppresses" from "blocked" or "apoptosis" from "E1A-induced apoptosis." These details may be key in understanding how p53 functions. At a more abstract level the single relation "mutant p53 inhibits apoptosis" might be more appropriate. Please note that effective aggregation must combine varying representations of both connectors and entities.

Our interface lets a user query for relevant relations, control aggregation of those relations, and link to source texts. Substance synonyms are used to expand query terms entered by the user to rank the relevance of stored relations. Journal, date, and other data are stored to help with relation ranking. The relation selection algorithm favors relations with identified substances or functions and is tuned to try to find connections between all returned entities rather than simply returning a "star" pattern where all links are focused on a single node. User-selected aggregation parameters control how relations are combined and connected. The aggregated relations are presented in a table view that allows the user to select or de-select individual relations. Selected relations are sent to the network view (Figure 1) to support visual exploration. A customized algorithm arranges the nodes to minimize overlap. Users can move items around, explore the details of the relations, and click to see the underlying articles. "Sessions" can be saved to store a user's current work. All settings, selections, relations, and positions are stored in a file which can be reopened later.

2. USER FEEDBACK

A pilot group of three biologists gave positive feedback: "In my head I've been trying to do what this is doing for you," "I am seeing things that I didn't know before," "I think a lot of people would get a lot of use out of this, as long as it doesn't scare them off in the beginning," and "It took me a few weeks just to find that Sin3 interacts with p53, where when you type this in [to Genescene] it's right there." In particular, we focused in on the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

JCDL '05, June 7–11, 2005, Denver, Colorado, USA
Copyright 2005 ACM 1-58113-876-8/05/0006...\$5.00.

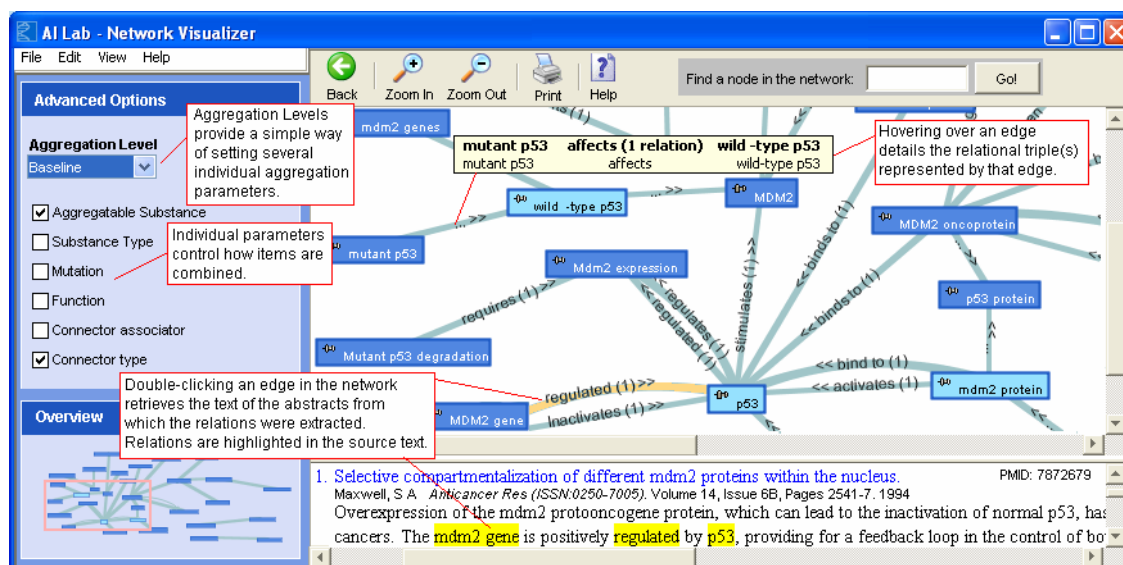


Figure 1. The Genescene Aggregation Interface

utility of relation aggregation. Comparing un-aggregated relations (“baseline”) with the highest level of aggregation (“simple pathway”) we were told: “If an undergrad were looking for some very basic information for a project then it [differences shown in baseline vs. simple pathway] doesn't matter.” Another said about a baseline network: “I like having a mess because then I can see everything -- I don't feel like I've missed something.” They went on to say “If I had a specific question then I would start there [simple pathway aggregation] -- especially if I had some existing knowledge about what's the same and what's not.”

The researchers also saw value in controlling the aggregation parameters. One user came across an example like that shown in Figure 2 and explored details of the extracted relations using the aggregation parameters. In the Figure's first network where different values for the mutation feature have been combined, an apparent conflict arises: TP53 is both activating and inhibiting MDM2. When the mutation control is checked, non-mutant TP53 is shown to activate and mutant TP53 to inhibit MDM2. The researcher put it this way. “You have to have a difference, to actually figure out what's going on. A lot of times that's where the discovery comes from, that it's doing exactly the opposite of what you expected, cause then you delve deeper into it and realize, ‘ok, here's the difference’.”

3. DISCUSSION AND FUTURE WORK

This feedback hints that allowing researchers to explore relations at different levels of granularity increases visualization utility. This paradigm is particularly important for relations extracted from free text as compared to manually curated relations because free text frequently includes rich, multi-faceted information that does not fit into a neat organizational structure. This kind of semantic extraction and display methodology may also be useful in other digital library domains. To supplement the anecdotal evidence presented here, we are already engaged in work that systematically studies the effectiveness of our aggregation methodology and the utility of the aggregation interface.

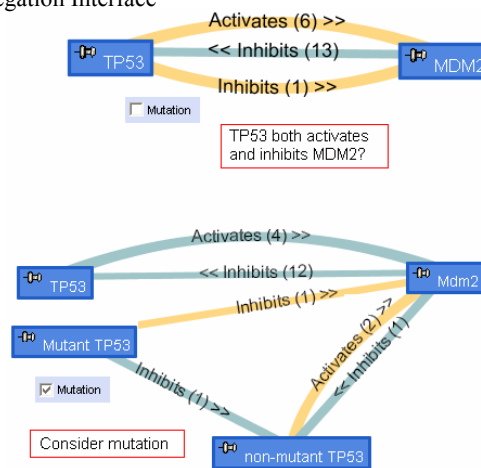


Figure 2. Differentiating Features for Analysis

4. ACKNOWLEDGMENTS

We thank our pilot users and the Genescene development team. This work was supported in part by: NIH/NLM, 1 R33 LM07299-01, 2002-2005, “Genescene: a Toolkit for Gene Pathway Analysis.”

5. REFERENCES

- [1] A. Rzhetsky et al., "GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data," J. Biomed. Inform., vol. 37, pp. 43-53, 2004.
- [2] D. McDonald, H. Chen, H. Su, and B. Marshall, "Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser," Bioinformatics, vol. 20, pp. 3370-8, 2004.
- [3] B. Marshall, H. Su, D. McDonald, and H. Chen, "Linking ontological resources using aggregatable substance identifiers to organize extracted relations," presented at Pac. Symp. Biocomput., Big Island, Hawaii, 2005.