BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 24 2010, pages 3138–3139
doi:10.1093/bioinformatics/btq597

Systems biology

GSearcher: Agile Attribute Querying for Biological Networks

Gang Su[1,2,*], Brian D. Athey[2,3] and Fan Meng[2,3,4]

[1]Bioinformatics Program, [2]National Center for Integrative Biomedical Informatics, [3]Center for Computational Medicine and Bioinformatics and [4]Psychiatry Department and Molecular & Behavioral Neuroscience Institute, University of Michigan, Ann Arbor, MI, USA

Associate Editor: Trey Ideker

ABSTRACT

Summary: GSearcher provides a highly interactive user experience in navigating attribute data associated with large and complex biological networks. The user may either perform a quick search using keywords, phrases or regular expressions, or build a complex query with a group of filters for efficient and flexible exploration of large datasets.

Availability: http://brainarray.mbni.med.umich.edu/gsearcher/

Contact: sugang@umich.edu

Received on March 31, 2010; revised on August 24, 2010; accepted on October 15, 2010

1 INTRODUCTION

Cytoscape is one of the most popular open source software for the analysis and visualization of biological networks (Shannon *et al*., 2003). In a biological network, genes and proteins are modeled as nodes and interactions are modeled as edges, which are associated with various attribute data such as gene annotation or expression level. As the size of networks and amount of attribute data increase, highly flexible and scalable search solutions become a necessity.

Cytoscape is bundled with Quick Find, Filters and Enhanced Search Plugin (ESP) (Ashkenazi *et al*., 2008). These tools all use a submit-wait workflow: the user types the query, hits the 'Enter' key to begin the search and then waits for the results to be shown in the default attribute browser. The submit-wait process must be repeated to compare different queries, to correct errors and to progressively improve a query. This process not only creates the perception of a slow search by forcing the user to wait for complete results from unsatisfactory preliminary searches, but also interrupts any coherent thought process.

A flexible and scalable toolkit for rapidly navigating biological networks is vital to speeding the search workflow, aiding researchers' thought processes and creating a more appealing experience. In many modern search engines, such as searching a song in iTunes (http://www.apple.com/iTunes/), the search result is updated instantly from the user's input without waiting for the user to hit the 'Enter' key. This interactive model enables the user to complete a query from live feedback, dramatically improving the efficiency of searches and the aesthetic appeal of interaction with the software.

Current Cytoscape search tools also have some issues that undermine the efficiency and accuracy of searching. First of all, all these tools use the default attribute browser to display results. The potential for interference made it difficult to use these tools along with other plugins that also utilize the default attribute browser, such as MCODE. Secondly, these tools only select nodes or edges matching the search criteria, without indicating where in the attribute table matches occur. The user can only locate the position of matches by manual scanning, which can be very difficult for browsing the result from a fuzzy search in a very large attribute table. Thirdly, it is difficult for the user to compare matching and non-matching attributes as non-matching attributes are always hidden by current tools. Finally, all these tools only support a subset of fuzzy matching rules. For example, wildcards currently are not allowed to be placed at the beginning of a query, which forbids the user to perform suffix-based searching.

We hope to address these issues and supplement current search functions with the GSearcher plugin, a fully interactive, highly flexible search interface that supports full JAVA regular expression (regex).

2 DESIGN AND IMPLEMNTATION

2.1 Search engine design

GSearcher is built on JDK 6. We used GlazedList (http://publicobject.com/glazedlists/) library as the underlying data model. Upon initialization, the current network's attribute data are transformed into a specific GlazedList table model for high performance searching, sorting and updating. This transformation is very efficient; GSearcher only takes 792 ms to transform a Michigan Molecular Interaction (MiMI; Tarcea *et al*., 2009) human interaction network of 11 884 nodes and 88 134 edges with 21 attribute fields on a 2.67 GHz Intel Core i7 920 PC. In comparison, ESP takes about 4 s for indexing on the same computer. The numerical primitive data types (Double, Float, Integer) are preserved; Hash/Array attributes are flattened into Strings. Subsequent searching on the same network does not require table model reloading (re-indexing).

2.2 Dynamic linking of GSearcher to network view

In order to provide interactive feedback and result sorting, browsing and highlighting, we built GSearcher's independent data browser using JXTable (http://swinglabs.org/). This browser listens to user input and updates search results interactively independent from the default browser. Attribute text which matches the query are highlighted in the table. The user may either remove non-matching attribute rows from the browser, or keep them in the browser to compare with the matching rows. Similar to the default browser, the selected rows in the result table are dynamically linked to the network view, but now the user may either 'select' or 'highlight'

*To whom correspondence should be addressed.

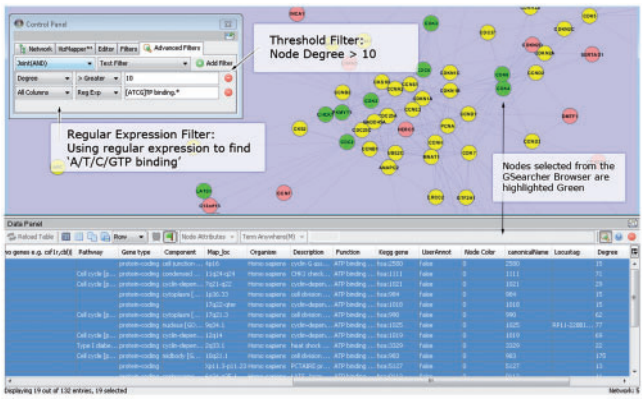

Fig. 1. A screenshot of GSearcher on MiMI human Interactome. A subnetwork is created from nodes with their attributes containing the keyword 'cyclin'. Nodes can be dynamically selected from the GSearcher browser using conditions in the two illustrated filters. These nodes are highlighted in green. In contrast, previous selections by the user or other plugins are highlighted in yellow, demonstrating how GSearcher interact with other plugins with minimal interference.

nodes or edges. When the nodes/edges are highlighted, the selection state is preserved so that search can be performed independently with minimal interference to other Cytoscape functions (Fig. 1). Rows in the browser can be sorted by a single click on the column header, either numerically or alphabetically depending on the primitive data types. Undesired attributes can be removed from the view and the search pool by hiding the corresponding columns in the browser table.

2.3 Quick search

Quick search mode enables the user to apply a single query on all or selected attribute fields. There are currently six different matching modes:

- Terms Anywhere (M): allow a match to occur anywhere in the attribute table. Multiple keywords can be submitted separated by spaces. Unlike ESP, the default operator for joining multiple terms is AND. This is more similar to typical online searching.
- Begins with Phrase: only matches a phrase at the beginning of attributes.
- Reg Exp: the query term is treated as JAVA regular expression.
- Exact: the query term must match an attribute perfectly.
- Phrase: the query is treated as a phrase with spaces preserved.
- Exclude Phrase: the query is treated as a phrase, and attributes which do not contain the phrase are highlighted.

GSearcher provides some search capability currently unavailable in Cytoscape. For example, querying 'nuclease' on the MiMI network using 'Terms Anywhere (M)' returns 116 matches, while ESP only returns 13 matches. 'endonuclease' and 'ribonuclease' were left out by ESP because suffix matching is not allowed. Using regular expression, the user can build even more flexible rules. Using 'CDC\d+' as a regular expression query will only match attributes beginning with 'CDC', followed by a number, such as 'CDC16'. The ESP syntax only allows 'CDC*' in which the wildcard cannot be refined to represent a set of characters. 'biological\−_\s+function' will match not only 'biological_function' and 'biological-function', but also 'biological function'—which allows fuzzy matching to span over spaces. '(?!ATP)binding.*' will match a binding term NOT following ATP, such as 'RNA binding'. Therefore, by incorporating regular expression, GSearcher substantially supplements current Cytoscape search functions.

2.4 Advanced search

While Quick search applies the same search criteria to one or multiple attributes, the Advanced search combines an arbitrary number of Quick Search filters which can be applied to different attributes. Filters can be joined either with AND, which indicates all filters must be satisfied, or OR, where at least one filter is satisfied. There are currently three types of filters:

- Text filter: each text filter is one implementation of Quick Search.
- Threshold filter: compare a numerical attribute with a certain threshold value, using numerical comparison operators (such as >).
- Range filter: test whether a numerical attribute is within the specified range.

The combination of filters offers users great flexibility when querying the network. Figure 1 shows an example of incorporating a threshold filter and a regular expression filter.

3 CONCLUSIONS

GSearcher provides Cytoscape users with an interactive interface and full regular expression support for building complex queries. Its added flexibility and interactivity supplement current Cytoscape search functions, help researchers navigate large attribute datasets and facilitate exploratory analysis of biological networks.

ACKNOWLEDGEMENTS

We thank the Cytoscape community for help during the plugin development, Jing Gao and Allan Kuchinsky for tests and Josh Bucker for manuscript proofreading.

Funding: National Center for Integrated Biomedical Informatics through National Institutes of Health (grant 1U54DA021519-01A1 to the University of Michigan).

Conflict of Interest: none declared.

REFERENCES

Ashkenazi,M. *et al.* (2008) Cytoscape ESP: simple search of complex biological networks. *Bioinformatics*, **24**, 1465–1466.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Tarcea,V.G. *et al.* (2009) Michigan molecular interactions r2: from interacting proteins to pathways. *Nucleic Acids Res.*, **37**, D642–D646.