BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 7 2007, pages 910–912
doi:10.1093/bioinformatics/btm032

Systems biology

NetMatch: a Cytoscape plugin for searching biological networks

A. Ferro[1,*], R. Giugno[1], G. Pigola[1], A. Pulvirenti[1], D. Skripin[1], G. D. Bader[2] and D. Shasha[3]

[1]Dipartimento di Matematica e Informatica, Università di Catania, Viale A. Doria 6, I-95125 Catania, Italy, [2]Banting and Best Department of Medical Research and Department of Medical Genetics and Microbiology, University of Toronto, 160 College St, Toronto, Ontario, Canada M5S 3E1 and [3]Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street New York, NY 10012, USA

Received on October 31, 2006; revised and accepted on January 24, 2007

Advance Access publication February 3, 2007

Associate Editor: Chris Stoeckert

ABSTRACT

Summary: NetMatch is a Cytoscape plugin which allows searching biological networks for subcomponents matching a given query. Queries may be approximate in the sense that certain parts of the subgraph-query may be left unspecified. To make the query creation process easy, a drawing tool is provided. Cytoscape is a bioinformatics software platform for the visualization and analysis of biological networks.

Availability: The full package, a tutorial and associated examples are available at the following web sites: http://alpha.dmi.unict.it/~ctnyu/netmatch.html, http://baderlab.org/Software/NetMatch

Contact: ferro@dmi.unict.it

1 INTRODUCTION

Many biological systems arise from complex interactions between components (people, organisms, cells, proteins, DNA, RNA and small molecules) (Barabási and Oltvai, 2004). Such networks are naturally modeled as large graphs, which can be analyzed using graph theoretical techniques. Locating subgraphs matching a specific topology is useful to find higher-order connectivity motifs of networks that may have functional relevance in the modeled biological system. In cell biology, it may be of interest to see whether the connectivity of genes of one functional type is similar to some characteristic shape, like a feed-forward loop. In epidemiology, acquaintance graphs between people may characterize specific patterns of disease outbreaks and can be used to optimize vaccine delivery.[1]

A small number of tools currently exist for querying networks for motifs, but none of these allow user-defined queries implemented as easy to use software. For instance, the tYNA (Yip *et al.*, 2006) Cytoscape plugin can submit a network to a remote server for detection of four common motifs and the MAVisto (Schreiber and Schwobber-meyer, 2005) and FANMOD (Wernicke and Rasche, 2006) software find over-represented network motifs of a user defined size.

A query to NetMatch is a graph, some of whose elements are constants and some are wildcards (which can match an unspecified number of elements). The query results are subgraphs of the original graph connected in the same way as the query graph. NetMatch provides an efficient graph matching algorithm (Cordella *et al.*, 2004) with extensions to handle multiple labels per node, multiple edges between pairs of nodes, and approximate queries. NetMatch has been implemented as a plugin for Cytoscape (Shannon *et al.*, 2003), an open source software platform for network visualization and analysis that is extensible through a straightforward plug-in architecture, allowing rapid development of additional features.

2 METHODS AND IMPLEMENTATION

NetMatch supports subgraph matching queries against a target network, previously loaded into the Cytoscape workspace. Approximate queries are special subgraphs that may contain: (i) nodes and edges labeled with a special wildcard symbol '?', which can match any single value of a user specified node or edge attribute; (ii) approximate paths, which are paths of length $\leq n$ or $\geq n$, where n is a positive integer, that can connect two nodes. NetMatch handles target and query graphs with multi-edges (more than one edge between two nodes), loops (edges starting and ending at the same node) and a list of attributes for each edge and node.

The searching (matching) process is carried out by using the state space representation (Cordella *et al.*, 2004) where a state is a *partial mapping* and a transition between two states corresponds to the addition of a new pair of matched graph nodes. The aim of the matching process is the determination of a mapping, which is a bijection, and consists of a set of node pairs covering all the query graph nodes. When a pair of nodes is added to the partial mapping, consistency conditions are checked. Such consistency rules allow the pruning of the search space, reducing significantly the computational cost of the matching process.

Approximate query graphs are handled by first independently processing all maximal specified subparts and then 'joining' in all possible ways the results of the subqueries. The joining process connects the subparts by all paths satisfying the approximate paths present in the query.

NetMatch can be set to interpret labeled/unlabeled, directed/undirected networks (Fig. 1). Users can express queries in NetMatch by (i) loading from an existing file, (ii) importing from the Cytoscape workspace, (iii) drawing using the NetMatch query drawing tool.

The query drawing tool (Fig. 2) allows multiple edges, zooming, moving and resizing operations, and exports drawing results directly to NetMatch. It also has a predefined set of frequently used network motifs (Milo *et al.*, 2002) for convenience. The matching

*To whom correspondence should be addressed.

[1]In the standalone version of NetMatch, user applications have ranged from cancer research to remote sensing to network security.

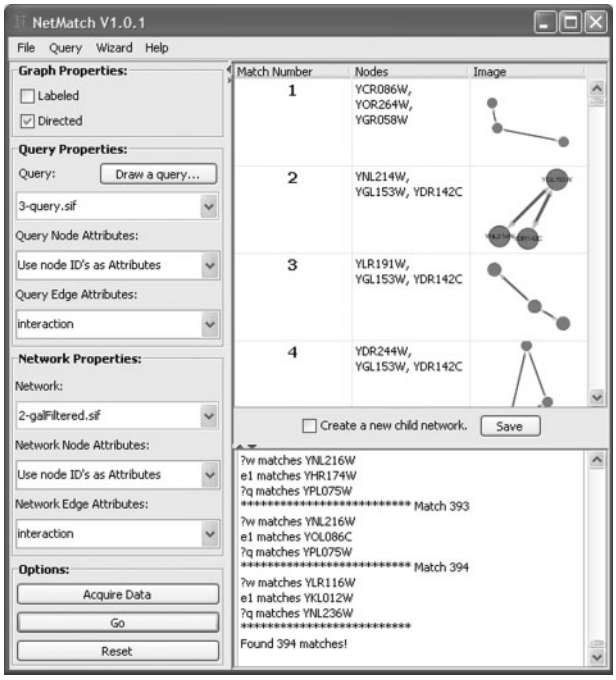


Fig. 1. The NetMatch main window showing the results of a small 3-node query. Users set NetMatch parameters on the left side of the window, then press 'Go'. NetMatch computes and displays results graphically on the right side of the window. If there are too many results, the user is warned and can choose to display all results (which may be time and memory intensive). Users can click on the results to select them for further analysis in Cytoscape or save all results to a file.

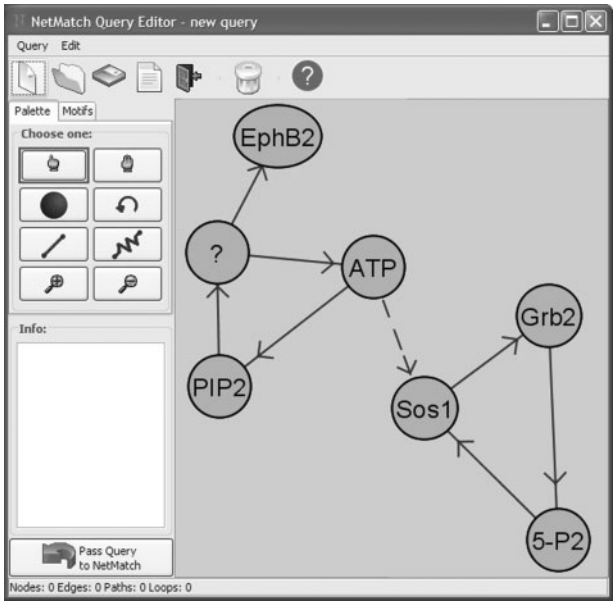
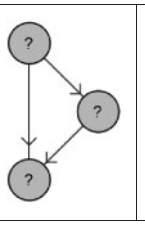
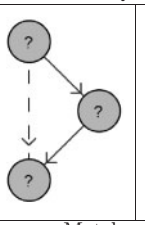
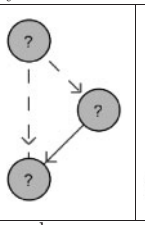
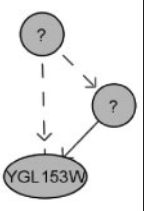


Fig. 2. The NetMatch query editor allows selection, node, edge, loop and approximate path creation and zooming. It has a number of preset motifs that can be placed on the drawing canvas in a single operation for convenient editing.

Table 1. Query examples

Query	User events
Find all feed-forward loops in a network modeling activation and inhibition control processes in the cell	Define feed-forward loop motif using NetMatch editor (using predefined motif)
Find all short paths from proteins located in the plasma membrane to proteins in the nucleus. These may represent signaling pathways	Load GO terms into Cytoscape, then define a NetMatch query: plasma membrane node, approximate path, nucleus node. Set the approximate path to ≤ 2
Find all three-chain motifs in a gene network where the genes are all significantly differentially expressed (P-value < 0.01)	Gene expression data with associated P-values must be loaded into Cytoscape. Define a three-chain motif and set each node attribute to <0.01. Search P-value node attributes on target network

Table 2. Query running time. For each query (column) the number of matches and the running time in milli seconds are reported. Queries were run against a graph of 331 nodes and 362 edges. The first query (first column) is the first example reported in Table 2. Generally, running time is related to the number of matches and increases by adding approximate paths in the query definition (the second query has an approximate path with label >0, the third query has two paths labeled >0). However, a query with two approximate paths but a specified node label has less matches (the fourth column)

Query			
			
Matches number			
22	421	1766	6
Time			
1	15	312	16

results are shown in NetMatch along with images of matched subnetworks and match information (Fig. 1). Clicking on one particular match will highlight its position in the target network in the Cytoscape main view. Any matched subnetwork can be saved and further analyzed and manipulated as a separate network in the Cytoscape workspace, using standard Cytoscape features.

As an example of a NetMatch query, kinase cascades in yeast were searched. A set of 7000 high quality protein interactions among 3000 yeast proteins and associated Gene Ontology (GO) annotation, (packaged with Cytoscape) were imported. A NetMatch query of three nodes connected in series, each node's attribute set to 'kinase activity', corresponding to the GO Molecular Function term for kinase was created. To ensure that all paths were found, NetMatch was set

to undirected mode. The query resulted in 12 matches (five unique), including a known kinase cascade in Ras signaling and complexes involved in metabolism. Other query examples together with NetMatch user events are reported in Table 1.

Concerning complexity, NetMatch deals with an NP-complete problem (graph matching). As reported in (Cordella *et al*., 2004), given an exact query with N nodes, best case analysis is $\Theta(N^2)$, whereas worst case is $\Theta(N!N)$. The average case analysis is very difficult to perform since it depends on a variety of parameters (number of attributes, nodes, edges, query frequency). However, experiments show that the algorithm used in NetMatch (Cordella *et al*., 2004) is efficient and it outperforms common alternatives. NetMatch allows any query topology including approximate queries. These queries require target graph traversal. This feature does not affect the earlier theoretical and experimental complexity. Table 2 shows practical query running time of several examples drawn from the first example reported in this table. Generally, one can say that running time is related to the number of matches. Such value increases by adding approximate paths in the query definition.

NetMatch will be extended in the future to have more powerful querying facilities, including Boolean combinations of attributes on nodes and edges, nodes and edges that do not occur and attributes for all nodes or edges along an approximate path. Suggestions for new query ability are welcomed.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges was provided by MIUR ex60% 2005 n. 2104010.

REFERENCES

Barabási,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cells functional organization. *Nature, Genetics Volume*, 101–113.

Cordella,L. *et al*. (2004) A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Trans. on PAMI*, **26**, 1367–1372.

Milo,R. *et al*. (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.

Schreiber,F. and Schwobbermeyer,H. (2003) Mavisto: a tool for the exploration of network motifs. *Bioinformatics*, **21**, 3572–3574.

Shannon,P. *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*. *http://www.cytoscape.org/*, **13**, 2498–2504.

Wernicke,S. and Rasche,F. (2006) Fanmod: a tool for fast network motif detection. *Bioinformatics*, **22**, 1152–1153.

Yip,K.Y. *et al*. (2006) The tyna platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. *Bioinformatics*, **22**, 2968–2970.