When the Web Meets the Cell: Using Personalized PageRank for Analyzing Protein Interaction Networks

Gábor Iván,1,2 Vince Grolmusz1,2*

1Protein Information Technology Group, Eötvös University, Pázmány Péter stny. 1/C, H-1117 Budapest, Hungary
2Uratim Ltd., InfoPark D, H-1117 Budapest, Hungary

ABSTRACT
Motivation: Enormous, and constantly increasing quantity of biological information is represented in protein interaction network databases. Most of these data are freely accessible through large public depositories. The robust analysis of these resources needs novel technologies, being developed today.

Results: Here we demonstrate a technique, originating from the PageRank computation for the World Wide Web, for analyzing large interaction networks. The method is fast, scalable and robust, and its capabilities are demonstrated on metabolic network data of the tuberculosis bacterium and the proteomics analysis of the blood of melanoma patients.

Availability: The Perl script for computing the personalized PageRank in protein networks is available for non-profit research applications (together with sample input files) at the address: http://uratim.com/pp.zip.

1 INTRODUCTION

The problem of finding important nodes in a large network emerged in several fields, but the best solutions to date were appeared in conjunction of the World Wide Web graph. Here the nodes are the web pages, and directed edges are the hyperlinks between the web pages. The web search engine techniques gave motivations to this question, since the important web pages, related to a web search, need to be returned first to the users of the web search service.

The most natural measure of importance of a vertex, the degree (i.e., the number of connected edges, in the case of an undirected graph) or the in-degree (i.e., the number of incoming edges, in the case of a directed graphs) is historically well established, and corresponds, e.g., in scientometry, to the number of citations to a published article. However, in the case of the web graph, the degree proved to be easy to manipulate, by simply inserting artificially a large number of referring edges into the graph.

Kleinberg’s HITS algorithm assigns quality scores to the nodes, and the quality of the referring nodes is inherited by the referred nodes, so low-quality manipulations can be filtered out. It turned out, however, that the HITS algorithm is also prone to more sophisticated manipulations, and it is not robust enough Lee and Borodin (2003).

The most successful web-page ranking algorithm, the PageRank algorithm, was developed by Page and Brin (1998), and used in the search engine of Google. The algorithm can be described as the following random walk on the graph: the walker starts at a uniformly chosen random vertex of the graph, then with probability $1 - c$ it follows a uniformly selected, random out-leading edge from the vertex, and with probability $c$ it teleports to a uniformly selected, random vertex of the graph, where $0 < c < 1$. The PageRank of a node $v$, corresponding to a certain sense to its importance, is the stationary limit probability distribution, that the walker is at the node $v$.

In applications for biological networks the stability of the PageRank is the most attractive property, since the published protein interaction networks contain numerous false positive and false negative interaction edges, even for the highest quality of data gathered for one of the most researched subjects, the yeast interactome Krogan et al. (2006), Gavin et al. (2006), Goll and Uetz (2006). Therefore network-ranking algorithms need to be stable in the case of a moderate number of false positives and false negatives.

The best stability estimation for the PageRank Lee and Borodin (2003) is given by the following inequality:

$$||\mathbf{p} - \tilde{\mathbf{p}}||_1 \leq \frac{2(1 - c)}{c} \sum_{j \in U} p_j,$$

where $i^{th}$ coordinate of vector $\mathbf{p}$ gives the PageRank of vertex $i$, and vector $\tilde{\mathbf{p}}$ gives the PageRank of the vertices after edges with endpoints in set $U$ are deleted or added. In other words, if $c$ is not too close to 0, and only the edges between less important nodes are perturbed, then the impact of this perturbation remains low to the PageRank. It is a remarkable property, since less important protein interactions are seldom mapped reliably, and the inequality shows that these errors will not accumulate to influence much the overall PageRank vector.

2 RESULTS AND DISCUSSION:

PageRank for the Analysis of Metabolic Networks: Protein-protein interaction networks are usually represented by undirected graphs. For undirected graphs PageRank is proportional to the degree of the nodes, so it does not help in choosing more important or less important nodes in the network, relative to simple degree-counting. However, metabolic graphs are directed graphs, with nodes representing biochemical reactions and a directed edge

---

*to whom correspondence should be addressed

1 The revised and improved version of this preprint is accepted for publication in Bioinformatics ©2010, Published by Oxford University Press. All rights reserved.
connects nodes \( u \) and \( v \) if reaction \( u \) has a product that is used by reaction \( v \). Therefore, the PageRank calculations may enlighten deep and robust network properties of the graph. We computed PageRank for the metabolic network of the \textit{Mycobacterium tuberculosis} (Fig S1). In that figure, the warmer colors show higher PageRanks, and the size of the nodes are proportional to their degree.

Consequently, those vertices that are warmer in color than were proportional to their degree are of special interest: they are more "important", more frequently hit by the random walker than the others with the same local network property: the vertex degree. It is a remarkable finding in the metabolic network of the tuberculosis bacterium, that a recently found important protein, the FAD-dependent thymidylate synthase (ThyX) Myllykallio et al. (2002) has the sixth largest PageRank in the network, much larger than other nodes with higher degree (Figure 1 and Table S2 in the on-line supporting material). The high PageRank may be due to the particularities of the thymidylate biosynthesis pathway in \textit{Mycobacteria Vertessy and Toth} (2009).

**Personalized PageRank for PPI networks:**

The personalized PageRank was developed for the prediction of the personal preferences in the valuation of the content on the World Wide Web Page et al. (1999). In computing the personalized PageRank, the randomized walker teleports with the probability of \( c + c' \) where \( 0 < c + c' < 1 \); with probability \( c' \) to some vertices, corresponding to the personal interest of the WWW surfer, and with probability \( c \) to the remaining vertices of "no-personal-interest".

Personalized PageRank seems to be capable to robustly evaluate the importance of the vertices of a network, relatively to some already known relevant nodes: if the random walker teleports to the important nodes with much higher probability than to any other vertices, then the resulting limit distribution will mark the nodes in the neighborhood of the relevant nodes with higher personalized PageRank. Additionally, personalized PageRank computation is scalable: it can be well approximated even for the largest networks encountered Fogaras et al. (2005).

We demonstrate here the applicability of the personalized PageRank in the evaluation of proteomics data. In proteomical PageRank, the randomized walker teleports with the probability of \( c + c' \) where \( 0 < c + c' < 1 \); with probability \( c' \) to some vertices, corresponding to the personal interest of the proteomics data. We computed the personalized PageRank for the metabolic network of the tuberculosis bacterium, that a recently found important protein, the FAD-dependent thymidylate synthase (ThyX) Myllykallio et al. (2002) has the sixth largest PageRank in the network, much larger than other nodes with higher degree (Figure 1 and Table S2 in the on-line supporting material). The high PageRank may be due to the particularities of the thymidylate biosynthesis pathway in \textit{Mycobacteria Vertessy and Toth} (2009).

**REFERENCES**


Fig. 1. Two dense subgraphs from the metabolic graph of the \textit{Mycobacterium tuberculosis}. On the left panel, large nodes correspond to large degree, but yellowish colors correspond to low PageRank. On the right panel, the small but orange-colored R06613 correspond to the KEGG reaction ID, catalyzed by the ThyX enzyme. The full figure is available as Figure S1 in the on-line supporting material.


**Acknowledgments.** The authors acknowledge the partial support of the OTKA grant CNK 77780 and NKTH project TB-INTERACTOME.