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

Gábor Iván,[1,2] Vince Grolmusz[1,2]**

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

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 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 Fogaras *et al.* (2005),Brin and Page (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:

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

where i^{th} coordinate of vector $\mathbf{p}$ gives the PageRank of vertex i, and vector $\hat{\mathbf{p}}$ 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 *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 thymidilate biosynthesis pathway in 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 analysis low concentration proteins seldom appear reliably in the results, therefore the robustness property of the PageRank computation is more than useful.

We considered the proteomics data of melanoma patients published in Forgber *et al.* (2009): 13 proteins were detected with higher levels in the plasma. We personalized the PageRank to these nodes in the human protein-protein interaction graph HPRD (c.f., Materials and Methods in the on-line supporting material for details). Table S3 contains the list of the largest rank nodes, and Figure S4 more than 8700 nodes of the human protein network, situated closer to at least one of the chosen proteins than 3 edges.

It is a remarkable result that many proteins of the largest PageRank vertices are clearly related to melanoma (Table S5 in the on-line supporting material).

Conclusions: We strongly believe that the synthesis of biology and computer science (Brent and Bruck (2006)) will open up great possibilities in the exploitation of the enormous amount of biological data. In particular, ordinary PageRank can help to evaluate important nodes and pathways in directed networks, such that metabolic networks, and the personalized PageRank may facilitate the robust analysis of large proteomics studies.

REFERENCES

Brent, R. and Bruck, J. (2006). 2020 computing: can computers help to explain biology? *Nature*, **440**(7083), 416–417.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. pages 107–117.

Fogaras, D., Rácz, B., Csalogány, K., and Sarlós, T. (2005). Towards scaling fully personalized PageRank: algorithms, lower bounds, and experiments. *Internet Math.*, **2**(3), 333–358.

Forgber, M., Trefzer, U., Sterry, W., and Walden, P. (2009). Proteome serological determination of tumor-associated antigens in melanoma. *PLoS ONE*, **4**(4), e5199.

Gavin, A.-C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**(7084), 631–636.

Goll, J. and Uetz, P. (2006). The elusive yeast interactome. *Genome Biol*, **7**(6), 223.

Krogan, N. J. et al. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, **440**(7084), 637–643.

Lee, H. C. and Borodin, A. (2003). Perturbation of the hyperlinked environment. *Computing and Combinatorics: COCOON 2003, Big Sky, MT, USA, July 25-28, 2003*, LNCS 2697, pages 272–283.

Myllykallio, H. *et al.* (2002). An alternative flavin-dependent mechanism for thymidylate synthesis. *Science*, **297**(5578), 105–107.

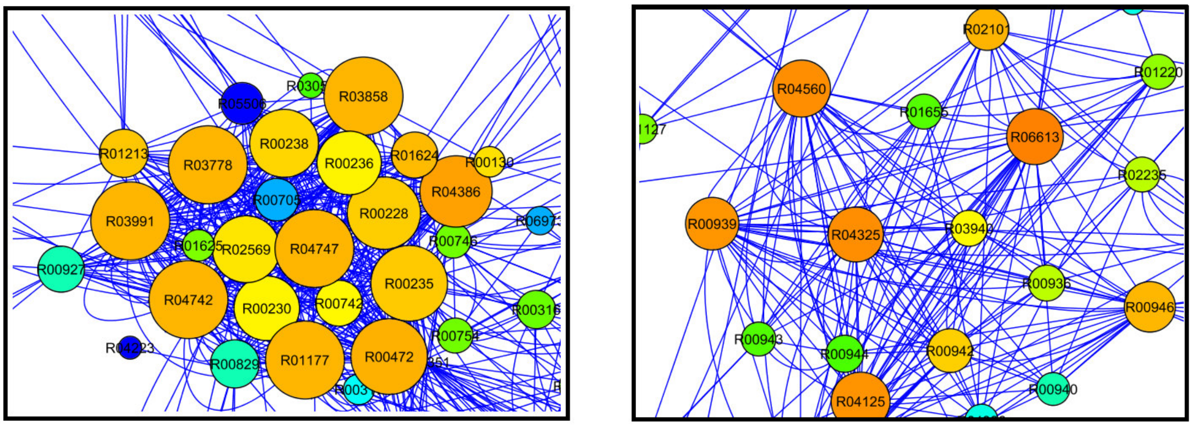

Fig. 1. *Two dense subgraphs from the metabolic graph of the Mycobacterium tuberculosis. On the left panel, large nodes correspond to large degree, but yellowish colors correspond to low PageRank. On the right panel, the small but orange-colored R06613 correspond to the KEGG reaction ID, catalyzed by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

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. *Stanford InfoLab TR 1999-66* (1999).

Vertessy, B. G. and Toth, J. (2009). Keeping uracil out of DNA: physiological role, structure and catalytic mechanism of dutpases. *Acc Chem Res*, **42**(1), 97–106.

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