BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 3 2011, pages 405–407
doi:10.1093/bioinformatics/btq680

Systems biology

Advance Access publication December 12, 2010

When the Web meets the cell: using personalized PageRank for analyzing protein interaction networks

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

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

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Enormous and constantly increasing quantity of biological information is represented in metabolic and 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.

Contact: grolmusz@cs.elte.hu.

Supplementary information : Supplementary data are available at *Bioinformatics* online.

Received on April 22, 2010; revised on October 25, 2010; accepted on December 7, 2010

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).

*To whom correspondence should be addressed.

The most successful web page ranking algorithm, the PageRank algorithm, was developed by 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 outleading 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 (Gavin *et al.*, 2006; Goll and Uetz, 2006; Krogan *et al.*, 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 of the overall PageRank vector.

2 RESULTS AND DISCUSSION

2.1 PageRank for the analysis of metabolic networks

Protein–protein interaction (PPI) 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 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

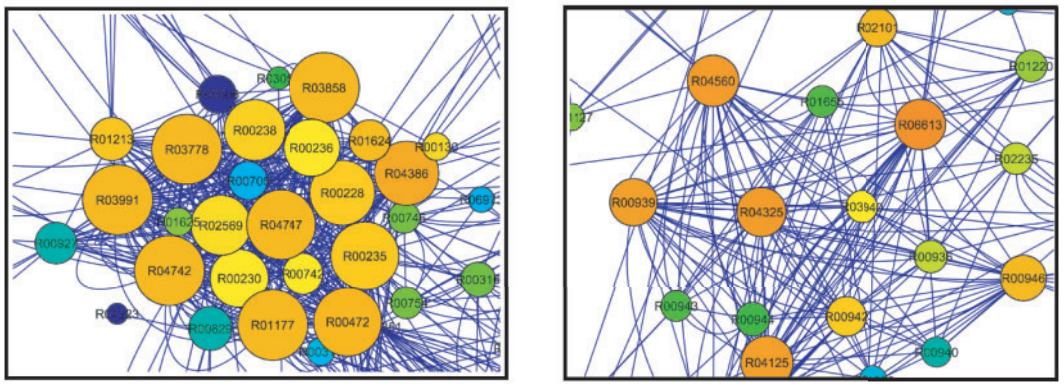

Fig. 1. Two dense subgraphs from the metabolic graph of the *M.tuberculosis*. On the left panel, large nodes correspond to large degree, but yellowish colors correspond to low PageRank. On the right panel, the small but orange-colored R06613 correspond to the KEGG reaction ID, catalyzed by the ThyX enzyme. The full figure is available as Supplementary Figure S1, where we note that R00945, functionally related to R06613, has the seventh largest PageRank. (Myllykallio *et al.*, 2002).

Mycobacterium tuberculosis (Supplementary 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 that 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 (Fig. 1 and Supplementary Table S2). The high PageRank may be due to the particularities of the thymidilate biosynthesis pathway in Mycobacteria (Vértessy and Tóth, 2009).

2.2 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, and 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 PPI graph HPRD (Prasad *et al.*, 2009) (cf., Materials and Methods in the online Supplementary Material for details). The HPRD human interactome contains 27 801 nodes, corresponding to human proteins, and 38 806 edges between these proteins, corresponding to interactions. Supplementary Table S3 contains the list of the largest rank nodes, and Supplementary Figure S4 contains more than 8700 nodes of the human protein interactome, situated closer to at least one of the chosen proteins than three edges.

It is a remarkable result that many proteins of the largest PageRank vertices are clearly related to melanoma (Table 1 and in a more complete form provided in Supplementary Table S5). More exactly, the 22 topmost ranged nodes, 10 are the nodes we personalized to (colored by light yellow on Table 1, therefore their high rank is natural), two have no apparent relation to the melanoma (colored by light gray on Table 1) and the green rows have clear relation to cancer (literary references are given in Supplementary Table S5).

One should remark that by the UniProt database, 160 human proteins are related to melanoma (Consortium, 2010). That is, 0.57% of the proteins in the analyzed 27 800 nodes in HPRD (Prasad *et al.*, 2009). This fact represents the selectivity and power of personalized PageRank computation, together with Table 1.

3 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. We should also note that the application of PageRank-like techniques are not enterily novel in biologic context: the IsoRank algorithm of Singh *et al.* (2008), for computing global network alignment between distinct protein interaction networks, has definite similarities to the PageRank algorithm.

Table 1. The proteins of the largest PageRank in the Personalized PageRank computations

PageRank	Accession number	Protein/gene name
858.89	P08107	HSP70 protein B
821.84	Q6EEV6	SUMO44
808.67	P55072	VCP
805.55	P26599	hnRNP1
801.60	P07954	Fumarate hydratese
790.41	P04075	Aldolase A
787.43	Q96EY1	HSP70 protein 9B
765.91	P06733	Enolase 1
754.35	O43852	Calumenin
729.05	P07195	LDH H
725.25	P15121	Aldose reductase
691.07	P40926	Malate dehydrogenase
592.09	Q15797	SMAD1
565.39	P02743	Serum amyloid P-component (SAP)
192.30	Q99972	Myocilin
141.15	P63104	YWHAZ
132.61	P00747	Plasminogen
130.78	P00505	Aspartate aminotransferase
125.47	P54253	Ataxin-1
116.36	P63167	Dynein light chain 1 (DLC1)
112.03	P61981	14-3-3 protein gamma
100.25	P04637	Cellular tumor antigen p53

The PageRank was personalized to the nodes of the yellow rows, coming from the melanoma data of Forgber *et al.* (2009). Green rows denote the newly found proteins of high PageRank, related to cancer, and proteins of uncolored rows have no clear correspondence to cancer. The full table with references is available in Supplementary Table S5.

Funding: The authors acknowledge the partial support of the OTKA grant (CNK 77780); NKTH project TB-INTERACTOME and EU grant no. TAMOP 4.2.1./B-09/1/KMR-2010-0003.

Conflict of Interest: none declared.

REFERENCES

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

Brin,S. and Page,L. (1998) The anatomy of a large-scale hypertextual web search engine. *Comput. Net. Isdn Syst.*, **30**, 107–117.

Consortium, U. (2010). The universal protein resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.

Fogaras,D. *et al*. (2005) Towards scaling fully personalized PageRank: algorithms, lower bounds, and experiments. *Internet Math.*, **2**, 333–358.

Forgber,M. *et al*. (2009) Proteome serological determination of tumor-associated antigens in melanoma. *PLoS ONE*, **4**, e5199.

Gavin,A.-C. *et al*. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.

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

Krogan,N.J. *et al*. (2006) Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, **440**, 637–643.

Lee,H.C. and Borodin,A. (2003) Perturbation of the hyperlinked environment. In Warnow,T. and Zhu,B. (eds) *Computing and Combinatorics: 9th Annual International Conference, COCOON 2003, Big Sky, MT, USA, July 25–28, 2003*, Vol. 2697 of *Lecture Notes of Computer Science*, pp. 272–283.

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

Page,L. *et al*. (1999) The pagerank citation ranking: bringing order to the web. *Tech. Report.*, Stanford Infolab, No. 1999–66. Stanford University.

Prasad,T.S.K. *et al*. (2009) Human protein reference database and human proteinpedia as discovery tools for systems biology. *Methods Mol. Biol.*, **577**, 67–79.

Singh,R. *et al*. (2008) Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc. Natl Acad. Sci. USA*, **105**, 12763–12768.

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