High Co-Citation of Enzymes in Interaction Networks Reliably Characterizes Their Functional Similarity

Lilla Tóthmérész[1] and Vince Grolmusz[1,2]

[1]Protein Information Technology Group, Department of Computer Science,

Eötvös University Pázmány Péter stny. 1/C, H-1117 Budapest, Hungary

E-mail: tothmereszlilla@gmail.com, grolmusz@cs.elte.hu

[2]Uratim Ltd. InfoPark D, H-1117 Budapest, Hungary.

Abstract

New methods for reliable quantitative analysis of biological network data are in high demand in today's bioinformatics and proteomics. Here we demonstrate the applicability of the co-citation, developed earlier for the analysis of scientific literature and the web graph for finding functionally similar nodes in protein-protein interaction networks in several model organisms. We have found clear correspondence between related enzymatic functions and high co-citation of proteins in interaction networks.

1 Introduction

The need for understanding the functional and structural interconnection patterns of proteins lead the scientific community to the construction of protein-protein interaction (PPI) networks from tens or even hundreds of thousands measurement data [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. These networks are widely available publicly, and facilitate an unprecedented view to the proteome.

Reliable tools for handling and analyzing large PPI networks are being developed today [12, 13, 14, 15]. One relevant question is to find *function-*

ally similar proteins in these networks, using exclusively the graph-theoretic properties of the networks.

In the present work we demonstrate that co-citation is a reliable tool for finding functionally similar proteins in PPI networks.

Co-citation was introduced by Henry Small in 1973 in the context of scientific literature analysis [16]. The co-citation of two documents was defined as the number they cited *together* in other publications.

Co-citation can be applied for identifying similar nodes in a graph or network, exclusively from the graph structure: if two nodes have high co-citation, then they can be considered similar (clearly, a node is most similar to itself, so high number of common neighbors of two different nodes may imply similarity in some sense). In the case of the web graph, this method is well known and used widely [17, 18].

One can easily define co-citation for un-directed graphs as well: The co-citation of two nodes is the number of their common neighbors in the graph.

A co-citation-like approach was applied by in biological context for verifying and repairing protein-protein interaction networks and for the prediction of complexes by [19]: there triangles were used in the analysis, with two edges from the protein-protein interaction network, and one edge derived from structural domain-domain similarity.

In our approach we use solely the network topology, and nothing else for the analysis. For the evaluation of our method, we need an independently verified functional similarity measure for the proteins involved: we apply the

Properties of the networks	number of nodes	number of edges	average degree	highest degree	number of components	number of nodes in the largest component
Homo sapiens	8138	36682	6.62	479	106	7916
Saccharomyces cerevisiae	5687	65819	16.87	968	9	5677
Homo sapiens max 90	7211	26993	5.31	87	139	6906
Homo sapiens max 120	7566	7566	5.85	119	124	7289
Saccharomyces cerevisiae, max 90	5157	39110	10.39	84	44	5099
Saccharomyces cerevisiae, max 120	5309	45842	11.97	115	36	5259

Table 1: *The summary of the networks analyzed. The data were downloaded from the IntAct [11] database. The "max 90" and "max 120" versions of the networks were gained by deleting all nodes of degree, larger than or equal to 90 and 120, respectively.*

enzyme commission number (EC number) [20] classification of the enzymes for the evaluation of similar biochemical function.

2 Materials and Methods

The PPI networks were downloaded from the IntAct database [11]: (http://www.ebi.ac.uk/intact/main.xhtml): the *Saccharomyces cerevisiae* data the on March 12, 2010, and the *Homo sapiens* data on April 6, 2010.

For the networks downloaded we computed the co-citation [16] of the vertices.

Definition 1 *For a graph vertex x let $\Gamma(x)$ denote the set of neighbors of x. The co-citation of nodes a and b is defined*

$$c(a,b) = |\Gamma(a) \cap \Gamma(b)|$$

Let A denote the adjacency matrix of the graph of the PPI network: the nodes $v_1, v_2, \ldots, v_n$ of the graph represent proteins, the edges the interactions between proteins; then the adjacency matrix of an n-vertex graph is an n x n 0-1 matrix, where in row i and column j there is a 1, if and only if node v_i is

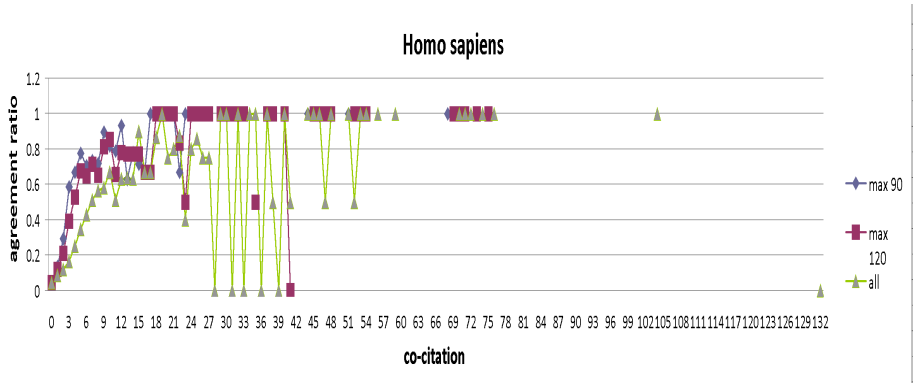

Figure 1: *Functional similarity in the function of co-citation in* Homo sapiens. *The x-axis shows the co-citation values, the y-axis shows the ratio of similar pairs/all pairs in all three versions of the graph: the original one, and in the "max 120" and "max 90" versions of it. Obviously, filtering out the large degree nodes improves considerably the functional correspondence of the high co-citation nodes in the graph.*

connected to node v_j. Clearly, the co-citation of the graph can be computed easily by the matrix product AA^T: the intersection of row i and column j of AA^T contains $c(v_i, v_j)$ if $i \neq j$.

For the evaluation of node similarity in PPI networks we applied enzyme commission numbers (EC numbers) [20]. EC numbers classify enzymes according to the reactions they catalyze. An enzyme can have more than one EC number (if it catalyzes more than one reactions), and from proteins only enzymes have EC numbers. The EC number consists of four blocks: The first block describes one of the six functional groups (oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases). The two blocks following refer to further subclasses within the main class. The last block is a consecutive number of each of the enzymes in the particular subclass [20]. Consequently, if the first three blocks of the EC numbers coincide for two proteins, it means they catalyze similar reactions. In the evaluation we took those proteins similar, that coincided in the first three blocks of one of their respective EC numbers. Since not all proteins have EC numbers, this way of evaluation does not capture the whole graph.

The SwissProt [21] and the KEGG [22] databases were used for annotating the protein accession numbers of the nodes with EC numbers [20].

3 Results and Discussion

We show in the below that co-citation is a stable measure of node similarity. Note, that stability is a crucial property if we are dealing with error-prone large networks like PPI networks [23] or the World Wide Web. Most probably the stability of the PageRank [24, 25] made it the most successful tool for finding important nodes in large networks, while other, not so stable node valuations (like the HITS algorithm [26]) are much less used today.

Theorem 1 *Let $G = (V, E)$ be an undirected graph of n vertices, and let $\tilde{G} = (\tilde{V}, \tilde{E})$ be a graph that we get after adding or deleting at most k edges. Let us denote $c(a, b)$ the co-citation of nodes a and b in graph G, and $\tilde{c}(a, b)$ the co-citation of nodes a and b in graph $\tilde{G}$. Let us denote the vector of co-citations of all possible node-pairs of length $\binom{n}{2}$ in graph G as $c = (\ldots, c(a, b) \ldots)$ and in graph $\tilde{G}$ as $\tilde{c} = (\ldots, \tilde{c}(a; b), \ldots)$. Then*

$$||c - \tilde{c}||_1 \leq 2kd_m$$

where d_m is the maximum degree of the nodes, connected by the added or deleted edges in graph G.

Proof: If we add or delete an edge between u and v, only those co-citation values can change, where one of the partners is either u, and the other is a neighbor of v or v and the other is a neighbor of u. Therefore, the sum of the absolute values of the co-citations may change by at most

$$\sum_{\{u,v\}} d(u) + d(v)$$

where the summation is taken for the u, v endpoints of the added/deleted edges. Since for all u: $d(u) \leq d_m$, the statement follows. □

Theorem 1 shows, that if the perturbed edges connect to low-degree nodes, then the total effect of the perturbation to the co-citation vector is small. This observation shows the robustness of the co-citation, since the less important parts of the interaction graphs (that contain low-degree nodes only) are frequently mapped less reliably.

3.1 Relativized version of co-citation

Clearly, co-citation need also be examined relative to the degrees of the vertices in the graph: if most of the neighbors of two nodes are common that may imply the stronger similarity than in the case when only a small fraction of them is common.

The following definition gives the definition of the relativized co-citation measure.

Definition 2 *The Jaccard coefficient of the co-citation is defined as*

$$c_J(a, b,) = \frac{|\Gamma(a) \cap \Gamma(b)|}{|\Gamma(a) \cup \Gamma(b)|}.$$

3.2 Biological evaluation of the node similarity

In the experiments, we computed similarity scores on the whole graph, but when evaluating the results, we only considered those node pairs where both nodes had EC numbers.

For test graphs, we used the human and the yeast protein interaction networks, downloaded from IntAct [11]. These organisms were chosen because

of the size and the quality of their interaction networks. We refer to Table 1 for the quantitative characterizations of the test networks.

Proteins increase by one the co-citation-values of each pairs, formed from their own interaction partners: i.e., if node a is connected to b and c, then a increases the co-citation of the b, c pair by 1. Therefore those vertices that have several hundred neighbors contributes to the co-citation of a very high number of vertex pairs. The highest degree node of the yeast network is 968 (c.f., Table 1), that means that this protein contributes 1 co-citation to each of $\binom{968}{2} = 468,028$ pairs, formed from its neighbors. We believe that these high-degree, "sticky" proteins with a high number of interacting partners are not relevant in estimating the co-citation–based functional similarity of the pairs of their interacting neighbors, therefore we also derived the"max 90" and "max 120" versions of the networks, by deleting all nodes of degree, larger than or equal to 90 and 120, respectively.

The x-axis of Figure 1 shows the co-citation values for *Homo sapiens*, the y-axis shows the ratio of similar pairs/all pairs in all three versions of the graph: the original one, and in the "max 120" and "max 90" versions of it. Obviously, filtering out the large degree nodes improves considerably the functional correspondence of the high co-citation nodes in the graph.

For the numerical analysis of Figure 1, first we filtered out the green triangle extremities of x coordinate of 104 and 132. After that, the largest data point corresponds to co-citation $n_1 = 76$ on the original green graph, to $n_2 = 75$ on the red "max 120" graph, and to co-citation $n_3 = 70$ on the blue "max 90" graph.

Amongst those nodes, that have at least half the maximum co-citation (i.e., $n_1/2$, $n_2/2$ or $n_3/2$ in the three versions of the graph, respectively), the ratio of similar pairs were:

- 0.9375 when we consider the "max 90" graph;
- 0.9545 when we consider the "max 120" graph;
- 0.8 for the original graph.

Figure S2 in the on-line material shows the co-citation values for *Saccharomyces cerevisiae*.

Figure S3 in the on-line material shows the dependence of the agreement ratio of the first three blocks of the EC numbers of the "max 90" graph of the human data set, in the function of the Jaccard coefficients c_J of the

co-citations (see Definition 2). Figure S4 in the on-line material gives the analogous graph for the yeast data.

4 Conclusions

Co-citation, as a functional similarity measure in protein-protein interaction networks was considered, and it was proven, that high co-citation of two proteins with enzymatic functions implies high functional similarity, measured by EC number coincidences. As an important example we proved, that in the human interactome, if we filter out the nodes of degree larger than 120, more than 38 co-citations imply more than 95% of functional coincidence between the nodes.

5 Acknowledgements

The authors acknowledge the partial support of the NKTH project TB-INTER and the OTKA project CNK 77780 and the EU funded project TÁMOP-4.2.1/B-09/1/KMR The authors are grateful for Gábor Iván and Dániel Bánky for help in EC number annotation and PPI network databases.

References

[1] Andrew Chatr-aryamontri, Arnaud Ceol, Luisa Montecchi Palazzi, Giuliano Nardelli, Maria Victoria Schneider, Luisa Castagnoli, and Gianni Cesareni. Mint: the molecular interaction database. *Nucleic Acids Res*, 35(Database issue):D572–D574, Jan 2007.

[2] B. Aranda, P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge, C. Derow, M. Feuermann, A. T. Ghanbarian, S. Kerrien, J. Khadake, J. Kerssemakers, C. Leroy, M. Menden, M. Michaut, L. Montecchi-Palazzi, S. N. Neuhauser, S. Orchard, V. Perreau, B. Roechert, K. van Eijk, and H. Hermjakob. The intact molecular interaction database in 2010. *Nucleic Acids Res*, 38(Database issue):D525–D531, Jan 2010.

[3] T. S Keshava Prasad, Kumaran Kandasamy, and Akhilesh Pandey. Human protein reference database and human proteinpedia as discovery tools for systems biology. *Methods Mol Biol*, 577:67–79, 2009.

[4] Andreas Zanzoni, Luisa Montecchi-Palazzi, Michele Quondam, Gabriele Ausiello, Manuela Helmer-Citterich, and Gianni Cesareni. Mint: a molecular interaction database. *FEBS Lett*, 513(1):135–140, Feb 2002.

[5] Ioannis Xenarios, Lukasz Salwinski, Xiaoqun Joyce Duan, Patrick Higney, Sul-Min Kim, and David Eisenberg. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, 30(1):303–305, Jan 2002.

[6] Gary D Bader, Doron Betel, and Christopher W V Hogue. Bind: the biomolecular interaction network database. *Nucleic Acids Res*, 31(1):248–250, Jan 2003.

[7] Samuel Bader, Sebastian Kuhner, and Anne-Claude Gavin. Interaction networks for systems biology. *FEBS Lett*, 582(8):1220–1224, Apr 2008.

[8] Javier De Las Rivas and Alberto de Luis. Interactome data and databases: different types of protein interaction. *Comp Funct Genomics*, 5(2):173–178, 2004.

[9] Michael E Cusick, Niels Klitgord, Marc Vidal, and David E Hill. Interactome: gateway into systems biology. *Hum Mol Genet*, 14 Spec No. 2:R171–R181, Oct 2005.

[10] Lukasz Salwinski, Christopher S Miller, Adam J Smith, Frank K Pettit, James U Bowie, and David Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Res*, 32(Database issue):D449–D451, Jan 2004.

[11] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler, and H. Hermjakob. Intact–open source resource for molecular interaction data. *Nucleic Acids Res*, 35(Database issue):D561–D565, Jan 2007.

[12] Igor Ulitsky, Nevan J Krogan, and Ron Shamir. Towards accurate imputation of quantitative genetic interactions. *Genome Biol*, 10(12):R140, 2009.

[13] Guanming Wu, Xin Feng, and Lincoln Stein. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol*, 11(5):R53, 2010.

[14] James C Costello, Mehmet M Dalkilic, Scott M Beason, Jeff R Gehlhausen, Rupali Patwardhan, Sumit Middha, Brian D Eads, and Justen R Andrews. Gene networks in drosophila melanogaster: integrating experimental data to predict gene function. *Genome Biol*, 10(9):R97, 2009.

[15] Michael A Fischbach and Nevan J Krogan. The next frontier of systems biology: higher-order and interspecies interactions. *Genome Biol*, 11(5):208, 2010.

[16] Henry Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24:265–269, 1973.

[17] Jeannette Janssen, Pawel Pralat, and Rory Wilson. Estimating node similarity from co-citation in a spatial graph model. In *SAC '10: Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 1329–1333, New York, NY, USA, 2010. ACM.

[18] R. Lempel and S. Moran. Rank-stability and rank-similarity of link-based web ranking algorithms in authority-connected graphs. *Information Retrieval*, 8:245–264, 2005.

[19] Bill Andreopoulos, Christof Winter, Dirk Labudde, and Michael Schroeder. Triangle network motifs predict complexes by complementing high-error interactomes with structural information. *BMC Bioinformatics*, 10:196, 2009.

[20] E. C. Webb. Enzyme nomenclature. recommendations 1984. supplement 2: corrections and additions. *Eur J Biochem*, 179(3):489–533, Feb 1989.

[21] UniProt Consortium. The universal protein resource (uniprot) in 2010. *Nucleic Acids Res*, 38(Database issue):D142–D148, Jan 2010.

[22] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 27(1):29–34, Jan 1999.

[23] Johannes Goll and Peter Uetz. The elusive yeast interactome. *Genome Biol*, 7(6):223, 2006.

[24] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *COMPUTER NETWORKS AND ISDN SYSTEMS*, 30:107–117, 1998.

[25] Hyun Chul Lee and Allan Borodin. Perturbation of the hyperlinked environment. In T. Warnow and B. Zhu, editors, *Computing and Combinatorics: 9th Annual International Conference, COCOON 2003, Big Sky, MT, USA, July 25-28, 2003*, number 2697 in Notes in Computer Science, pages 272–283. Springer Verlag, 2003.

[26] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677, San Francisco, California, 25–27 January 1998. ACM Press.