International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:10, 2009

Community Detection-based Analysis of the Human Interactome Network

Razvan Bocu, *University College Cork,* Dr Sabin Tabirca, *University College Cork*

Abstract—The study of proteomics reached unexpected levels of interest, as a direct consequence of its discovered influence over some complex biological phenomena, such as problematic diseases like cancer. This paper presents a new technique that allows for an accurate analysis of the human interactome network. It is basically a two-step analysis process that involves, at first, the detection of each protein's absolute importance through the betweenness centrality computation. Then, the second step determines the functionally-related communities of proteins. For this purpose, we use a community detection technique that is based on the edge betweenness calculation. The new technique was thoroughly tested on real biological data and the results prove some interesting properties of those proteins that are involved in the carcinogenesis process. Apart from its experimental usefulness, the novel technique is also computationally effective in terms of execution times. Based on the analysis' results, some topological features of cancer mutated proteins are presented and a possible optimization solution for cancer drugs design is suggested.

Keywords—Betweenness centrality, interactome networks, protein-protein interactions, protein communities, cancer.

I. Introduction

A. Interactome networks and their importance

THE concept of interactome networks represents a very important biological construct. It is widely used to describe the protein interactions that determine the organization and function of a biological organism. These networks feature a complex structure that makes any research endeavour in the field to be carried on with inherent difficulties. Nevertheless, a proper understanding of the general structure of the protein interactions is necessary, as they consistently influence the function of a biological organism as a whole, from the simplest to the most complex ones. Therefore, it is mandatory to discover more efficient techniques that can be applied to the study of the structure and properties of the interactome networks.

Betweenness centrality is one of the centrality measures that allows for the interactome networks to be properly analyzed, because it essentially allows for various functional protein clusters to be determined with a high degree of accuracy. A classical betweenness computation construct is the Brandes algorithm [13]. It proves to be efficient enough in practice, featuring a complexity of O(nm+n2logn), where n is the number of vertices and m is the number of edges. It normally processes a network with thousands of nodes and tens of thousands of edges in a few hours. The Newman-Girvan algorithm [1, 6] is another classical construct dedicated to computing the betweenness centrality for edges. Although it is an important algorithmic construct, it still does not perform as expected in terms of execution times, as it processes a network with thousands of nodes and tens of thousands of edges in more than ten hours on a standard Intel Pentium Dual Core machine. One of the authors' previous papers [19] proposes a new algorithm that is able to optimize the betweenness computation for a network featured by thousands of nodes and tens of thousands of edges. This approach is based on the Dijkstra's algorithm and reduces the computation time for a network with thousands of nodes by up to 90% in the worst case. The additional gain in performance is significant and the underlying mechanism was explained thoroughly in one of the authors' previous paper.

Recent contributions showed the extraordinary influence that proteins exercise on fundamental physiological processes. In this respect, this paper demonstrates that cancer affects the most important proteins in the interactome network and, as a consequence, the normal function of the organism is greatly affected. An accurate understanding of the structure and importance of proteins requires the usage of efficient analysis techniques. This paper is aimed to describe a novel interactome analysis technique that employs an efficient community detection algorithm.

The paper will briefly describe the most relevant existing works regarding the betweenness computation and community detection. Furthermore, the novel interactome network analysis technique is introduced and thoroughly described and analyzed. Also, its practical usability is assessed on real proteomic data.

B. Relevant existing works

The research workflow that produced the results that are presented in this paper is based on relevant achievements that are the consequences of a thorough and extensive research activity. Therefore, this subsection will enumerate and succinctly describe the main existing research works, which contributed to the advances proposed in this paper.

Although the scientific literature related to betweenness is not excessively extensive, there are enough papers and research projects that are worth to be mentioned. Among these, we shall select the ones that had a decisive influence on our research pathway.

One of the first extensive works on betweenness belongs to Newman and Girvan. The Newman-Girvan algorithm is one of the methods used to detect communities in complex systems.

Razvan Bocu is a PhD Researcher and demonstrator in the Department of Computer Science, University College Cork, email: razvan.bocu@cs.ucc.ie
Dr Sabin Tabirca is a researcher and Senior Lecturer in the Department of Computer Science, University College Cork

The concept of "community structure" is related to the one of clustering, though it isn't quite the same. A community consists of a subset of nodes within which the node-node connections are dense, and the edges to nodes in other communities are less dense. There are a number of alternative methods for detecting communities in networks. These include hierarchical clustering, partitioning graphs to maximize quality functions such as network modularity, k-clique percolation, and some other interesting algorithmic methods [2]. Nevertheless, we preferred to make use of the Newman and Girvan conceptual system due to its structural articulation and practical usage in many situations. The Newman-Girvan algorithm is particularly used to compute betweenness for edges (links) that connects the nodes (proteins) in a network.

The Brandes algorithm is able to compute the betweenness exactly even for fairly large networks. It proposes a more efficient algorithm based on a new accumulation technique that integrates well with traversal algorithms solving the single-source shortest-paths problem, and thus exploiting the sparsity of typical instances and, as a consequence optimizing the overall computation efficiency. The range of networks for which betweenness centrality can be computed is thereby extended significantly [13]. Moreover, it turns out that all standard centrality indices based on shortest paths can thus be evaluated simultaneously, further reducing both the time and space requirements of comparative analyses.

The Brandes algorithm has a significantly improved structure, which makes it run faster and improves its general readability and usability. As a natural consequence, the algorithm is able to enlarge the $O(n^3)$ bottleneck and requires $O(nm + n^2 \cdot \log n)$ to execute. This is a major improvement, which can prove very important for a high-scale network, featured by thousands of nodes. Before the Brandes algorithm was presented, the analysis of a large network featured by thousands of nodes and tens of thousands of edges was an almost prohibitive endeavour using sequential algorithms run on normal single-core machines. The Brandes algorithm scales sensibly better than any previous implementation of an algorithm that computes the betweenness centrality measure. As an example, processing networks with thousands of nodes was previously a challenging task, which is made an accessible one using the new Brandes algorithm.

The idea of betweenness is tightly related to the idea of shortest path computation, as it can be seen in the next section. Therefore, it is very important to compute the shortest paths in the analyzed network as efficiently as possible. Following a series of theoretical and experimental activities carried on interactome networks, it was concluded that interactome networks feature a sparse nature. Therefore, it is essential for an efficient sequential betweenness algorithm applied on interactome networks, to use a shortest path algorithm that is designed to optimize computations on sparse networks. Based on the advances accumulated during the previous stages of our current research activity, an optimized Dijkstra-based pattern was used in order to develop the novel protein network analysis technique. The following sections will describe the improved computation scheme in more detail.

II. Theoretical background

A. Basic theoretical concepts

In the most common sense of the term, a graph is an ordered pair G=(V,E), comprising a set V of vertices or nodes together with a set E of edges, which are two-element subsets of V. To avoid ambiguity, this type of graph may be described precisely as undirected and simple. Using the terminology peculiar to interactome networks, proteins are modeled as vertices and the biological links as edges.

Within graph theory and network analysis, there are various measures of the centrality of a vertex within a graph that determine the relative importance of the vertex within the graph. For example, applied to the social networks study, centrality may offer an accurate measure of how important is a person in a certain network. Moreover, centrality is an essential concept for other types of networks, such as biological networks or interactome networks. In this particular case, the centrality may measure the importance of a certain protein in the network, or the relative importance of a certain sub-community (group of proteins) in the network. In the theory of space syntax, centrality specifies how important a room is within a building or how well-used a road is within an urban network. Basically, there are four measures of centrality that are widely used in the network analysis: degree centrality, betweenness, closeness, and eigenvector centrality [1].

Betweenness is a centrality measure based on shortest paths, widely used in complex network analysis. One of the fundamental problems in network analysis is to determine the importance (or the centrality) of a particular vertex (or an edge) in a network. Some of the well-known metrics for computing centrality are closeness, stress and betweenness. Of these indices, betweenness has been extensively used in recent years for the analysis of social interaction networks, as well as other large-scale complex networks. Some applications include lethality in biological networks, study of sexual networks and AIDS, identifying key actors in terrorist networks, organizational behavior, and supply chain management processes. Betweenness is also used as the primary routine in popular algorithms for clustering and community identification in real-world networks. For instance, the Girvan-Newman algorithm iteratively partitions a network by identifying edges with high betweenness scores, removing them and re-computing centrality scores.

Betweenness centrality can be computed both for nodes and for edges. The computation technique is exactly the same both for nodes and for edges, as it involves the computation of the distance matrix for a certain node or edge. Therefore, we shall briefly describe the betweenness centrality for nodes (vertices), which is a centrality measure of a vertex within a graph. Vertices that occur on many shortest paths between other vertices have a higher betweenness than those that do not. For a graph G=(V,E) with n vertices, the betweenness $C_B(v)$ of the vertex v is given by the following formula:

$$C_B(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the number of geodesic shortest paths from

vertex s to vertex t, and $\sigma_{st}(v)$ is the number of shortest geodesic paths from vertex s to vertex t that pass through a vertex v. This may be normalized by diving through the number of pairs of vertices excluding v, which is $(n-1)(n-2)$.

B. Interactome networks and cancer study

To put in context a passage that belongs to the 17^{th} century poetry, we can state that no gene is an island. Systems biology, or more specifically network biology, is aimed by the progressive discovery that a single gene is rarely responsible by itself for the fulfillment of a discrete biological function. As an answer, contemporary biology has designed a battery of methods that allow for the surveillance of the global features of the cells, ranging from DNA, RNA and proteins to small molecules, to be conducted properly. Interactome analysis generally studies the interactions that are established between the biological molecules (especially proteins) on a global scale. It is important to note that high throughput mapping of protein interactions allows for the global screening of each organism's interactome network to be accomplished. The resulting maps that contain protein interactions are called protein networks [21].

Topological features of the protein networks have been demonstrated to reflect the functionality of the interacting genes. For example, essential genes in yeast tend to be well connected and globally centered in the protein network. Furthermore, globally centered interactions are pre-disposed to be well conserved and serve as an evolutionary backbone for the network.

The availability of high-throughput experimental data has helped the construction of increasingly comprehensive and accurate protein-protein interaction networks. Initial network studies were performed on yeast, but more complex organisms are gradually being surveyed [21]. The topology of these networks not only throws a ray of light on the complex cellular mechanisms and processes, but also offers a good insight into the evolutionary aspects that are related to the proteins involved. Studies that focus on the human interactome have thus far been technically limited due to the lack of reliable and comprehensive experimental data. To compensate for this, several computational methods have been developed with the aim of predicting protein-protein interactions.

Based on data accumulated during our current research, we report an extensive study of cancer and non-cancer proteins that is based on existing carefully validated human protein-protein interaction data. Wachi et al. (2005) have reported increased interaction connectivity in differentially expressed proteins in lung cancer tissues. However, a comprehensive study of the interaction attributes of all already discovered mutated human cancer genes has not previously been attempted. In our studies to date, we examined the connectivity of proteins known to be susceptible to mutations leading to cancer [21]. We used a clustering method aimed at highlighting proteins in centrally connected hubs that form the backbone of the overall interactome network. We show that cancer proteins display a global topology significantly different to non-cancer proteins, indicating an increased central role of cancer proteins within the interactome. Networks of interacting proteins have been already constructed for the entire human genome using an orthology-based method described by Jonsson et al. in 2006. One of the benefits of the orthologous (interspecies) approach is the resulting reduced noise in protein interaction data, which allows for certain problematic interactions to be properly detected. In short, the method identifies supposedly existing interactions based on homology to experimentally determined interactions in a range of different species. The supposedly existing interactions receive confidence scores based on two factors: the level of homology related to proteins found experimentally to interact, and the amount of experimental data available concerning the proteins under scrutiny.

C. Community detection in interactome networks

The scientific activity during the current phase of our research was based on the works that were briefly described in the previous sections. Our goal was to configure a clear and straightforward interactome analysis technique that, based on the existent biological data, is able to detect the role of proteins in a carcinogenic process.

We developed a new clustering and community detection algorithm that is based on the edge betweenness computation. After a certain number of iterations, the algorithm marks and extracts the biological links with the highest betweenness and thus, accurately determines the protein communities that are determined by the co-operation in order to accomplish a certain function or provoke a disorder, such as cancer.

Information on cancer proteins was obtained from a comprehensive census of human cancer genes that made use of various protein interaction databases, IntAct [15] being one of the most important ones. The construction of a validated human protein-protein interaction network allows an in-depth analysis of individual proteins in the context of their surroundings. Here, the network topographies of human cancer proteins were examined with the aim of uncovering intrinsic properties that distinguish proteins prone to cancerous mutations from those that are not.

We used biological data made available by the IntAct dataset and other smaller biological datasets that deal with proteomic data. We wanted to isolate the communities that exist in the already known interactome network. In order to accomplish this, we made use of a detection technique that involves two parts. First, the betweenness and the degree of each protein in the network are computed. Second, the whole interactome network is partitioned into functionally determined communities using a community detection algorithm that is based on the biological links betweenness computation. The procedure concludes with isolating the functionally-related protein communities and by exactly computing each protein's importance through the betweenness centrality measure computation.

Clustering methods have previously been shown to be useful in identifying protein interactions that take place within the same cellular process (Palla et al., 2005; Jonsson et al., 2006). This can be attributed to the fact that sub-networks of proteins involved in a defined cellular process are more heavily interconnected by direct protein interactions than would be

expected by chance (Jeong et al., 2001; Gunsalus et al., 2005). In other words, the carcinogenic process is generated by clusters of proteins that feature a central position in the protein network. As a consequence, the high adverse impact of any cancer form is, in our opinion, determined by the way the disease affects the fundamental proteins that coordinate the most essential processes in the metabolic and physiological chains. This conclusion was reached following the usage of the new analysis scheme, which computes the functionally-related clusters of proteins with a greater degree of accuracy.

III. The new interactome network analysis scheme

A. General presentation and remarks

The analysis process starts with the calculation of the absolute importance of each protein at the scale of the whole network by computing the betweenness centrality measure. This is accomplished by making use of a personal algorithm for betweenness centrality computation. It is based on a Dijkstra-like approach for the shortest path part of the computation. The algorithm has a complexity of $O(|V||E|)$, where V is the set of proteins in the network and E is the set of biological links established between them. Therefore, it runs faster than most of the existing algorithms and it is able to process all the proteomic data in the IntAct dataset in less than an hour.

As soon as the first phase is completed, the testing procedure calls the community detection module, which accurately determines the functionally-related communities of proteins. We use a personal improved version of the Newman community detection algorithm. The underlying idea is that the betweenness of the edges connecting two communities is typically high, as many of the shortest paths between nodes in separate communities go through them. As a consequence, the algorithm gradually removes the edge featuring the highest betweenness from the network, and recalculates the edge betweenness after every removal. This way, after a certain number of iterations of the edge betweenness algorithm, the network is reduced to two components, then after a while one of these components is reduced again to two smaller components, and so on, until all edges are removed. This is, basically, a divisive hierarchical approach, and the result is a dendogram. The community detection algorithm uses the same computation scheme as the one that is used in the first phase of the analysis technique and, as a consequence, it is optimized for the efficient analysis of interactome networks. Compared to the original Newman's approach, the usage of this optimized version of the edge betweenness computation generates the overall speedup.

B. Remarks on the testing procedure and analysis

The testing procedure takes into account real biological data that is part of the IntAct dataset. In order to extract the data that is relevant to cancer, we used the valuable data on protein families that is made available in the Pfam database [22]. The following pseudo code summarizes the sequence of steps that the analysis script triggers. As a consequence, it is a brief and formal description of the analysis scheme itself.

1. **Input:** A protein interaction dataset and a supplementary dataset consisting of protein families data.
2. **Output:** The absolute importance of each protein in the network measured with the betweenness centrality, together with a dendogram that accurately determines the functionally-related protein communities. Also, the relevant mappings regarding the cancer mutated proteins are established.
3. ParseInputDatasets(IntDataSet[],Pfam[])
4. TriggerAnalysisModule::ProteinAbsoluteImportance()
5. TriggerAnalysisModule::ProteinCommunityDetect()
6. MapCommunityData(PfamData[])
7. DisplayCancerProteinsInformation()
8. **end_script**

Fig. 1. Pseudo code of the novel analysis scheme

TABLE I
Execution times of the novel analysis technique

Test no.	Execution times
1	23
2	51
3	135
4	352
5	603
6	847
7	1358

Essentially, we generated seven test datasets that feature the following number of proteins and biological links: 1000 (1207), 2000 (4823), 2500 (5692), 3000 (6209), 7000 (14175), 8000 (21304), 9000 (35892). The execution times are shown in the table above and they are expressed in seconds. It is important to note that these execution times take into account the entire analysis procedure execution, including both modules: the determination of each protein's absolute importance and the isolation of functionally-related proteins.

We examined the protein communities our method determined and some interesting differences in the community size were noticed. Cancer proteins belong to more highly populated communities compared to non-cancer proteins. The explanation may reside in the fact that cancer proteins take part in more complex cellular (carcinogenic) processes than those proteins that are of lower importance in the interactome network and, consequently, have less influence on the carcinogenesis. It can also be asserted that larger protein communities feature a larger or more complicated cellular mechanism, in which cancer proteins play an important role.

Proteins identified as members of more than one protein community are of particular interest. In general, each protein community represents a distinct cellular process. Therefore, proteins that have multiple community membership may be participating in multiple processes, and can be considered to be at the intersection of distinct but adjacent cellular processes that are determined by particular protein communities, which are determined by our community detection technique. The comparison between the cancer protein populations and the non-cancer population reveals that cancer proteins reside

at community junctions at a sensibly greater extent than their non-carcinogenic counterparts. This particular feature of cancer proteins enforces their particular importance in the interactome network seen as a whole and, as a consequence, their influence on all the physiological processes and related disorders.

Previous researches are mainly based on protein degree computation, which offers a reasonable indication on each protein's importance. Nevertheless, this is only a relative importance at the scale of the whole interactome network under scrutiny. The new computation technique precisely determines the absolute importance of each protein in the network, based on the betweenness centrality computation. Additionally, the functionally-related protein communities are determined.

Existing contributions distinguish between highly connected domains in peripheral cores (locally central) and highly connected domains in central cores (globally central). We noticed that globally central proteins represent an essential backbone of the proteome, exhibit at a high degree evolutionary conservation, and are essential for the organism. It is important to note that cancerous disease provokes mutations exactly to these globally central proteins. This observation supports and extends the findings of Wachi et al. (2005), who showed that differentially expressed proteins in squamous cell carcinoma of the lung tend to be global hubs. Overall, the above findings reveal topological distinction of cancer proteins that is primarily displayed for cancer mutated proteins in exhibiting the highest betweenness centrality compared to the proteins that didn't loose their normal function. In other words, the carcinogenic process is generated by clusters of proteins that feature a central position in the protein network. As a consequence, the high adverse impact of any cancer form is, in our opinion, determined by the way the disease affects the fundamental proteins that coordinate the most essential processes in the metabolic and physiological chains.

The already gathered experimental information can be summed up into the following conclusions:

- The new proteomic data analysis technique accurately determines the functionally-related communities of proteins.
- We practically assessed the suitability and performance of the new technique on real proteomic data related to cancer and the interesting properties of the determined protein communities allowed us to infer an explanation regarding cancer evolution.
- Although the original Newman's community detection algorithm remains a milestone for every researcher interested in community detection, we sensibly optimized it. Therefore, we were able to design a faster community detection module for our proteomic data analysis technique.

C. Conclusions and future developments

The most important property of cancer proteins is their importance at the scale of the whole interactome. We used our two-step analysis technique to show that the globally central proteins are the ones that are the most affected in a carcinogenic process and are also located at the junction of the most important protein communities.

Our clustering algorithm allows us to explore protein-protein connectivity in a more informative way than is possible by just counting the interaction partners for each protein. It allows us to distinguish between central and peripheral hubs of highly connecting proteins, revealing proteins that form the backbone of the proteome. The fact that we observe an enrichment of cancer proteins in this group and also their highest betweenness centrality values indicates the central role of these proteins. The domain composition of cancer proteins may indicate why this is the case: we have shown, based on our experiments' results, that cancer proteins contain a high ratio of highly malign domains. Therefore, all cancer drugs should be designed in such a way to prevent possible mutations to these highly-important proteins or, if the disease is already on the way, to contribute to reverting back to the original proteomic structure.

The next stages of our research will involve further optimizations of the algorithms that power up the new analysis technique. Also, we intend to analyze even more biological datasets related to cancer and, possibly, other high-impact contemporary diseases.

Acknowledgment

This work is supported by the Irish Research Council for Science, Engineering and Technology, under the Embark Initiative program.

References

[1] R. Dunn et al., *The use of node-clustering to investigate biological function in protein interaction networks*. BMC Bioinformatics, 2004.
[2] D. Bader et al., *Approximating betweenness centrality*. Georgia Institute of Technology, 2007.
[3] D. Meunier and H. Paugam-Moisy, *Cluster detection algorithm in neural networks*. Institute for cognitive science, BRON, France, 2006.
[4] J. Yoon, A. Blumer and K. Lee, *An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality*. Bioinformatics, 2006.
[5] M.E.J. Newman, *Shortest paths, weighted networks, and centrality*. Physical review, volume 64, 2001.
[6] M. Girvan and M.E.J. Newman, *Community structure in social and biological networks*. State University of New Jersey, 2002.
[7] P. Holme et al., *Subnetwork hierarchies of biochemical pathways*. Bioinformatics, 2003.
[8] D. Ucar et al., *Improving functional fodularity in protein-protein interactions graphs using hub-induced subgraphs*. Ohio State University, 2007.
[9] K. Lehmann and M. Kaufmann, *Decentralized algorithms for evaluating centrality in complex networks*. IEEE, 2002.
[10] J. Griebsch et al., *A fast algorithm for the iterative calculation of betweenness centrality*. Technical University of Munchen, 2004.
[11] G.H. Traver et al., *How complete are current yeast and human protein-interaction networks?*. Genome biology, 2006.
[12] R. Bunescu et al., *Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome*. Genome biology, 2005.
[13] U. Brandes, *A faster algorithm for betweenness centrality*. University of Konstanz, 2001.
[14] B. Preiss, *Data structures and algorithms with object-oriented design patterns in C++*. John Wiley and sons, 1998.
[15] EMBL-EBI, *The IntAct protein interactions database*. URL: http://www.ebi.ac.uk/intact/site/index.jsf, 2009.
[16] C. Demetrescu et al., *The Leonardo Library*. URL: http://www.leonardo-vm.org/, 2003.
[17] University of California, *The DIP protein interactions database*. URL: http://dip.doe-mbi.ucla.edu/, 2009.
[18] Johns Hopkins University, *The HPRD protein interactions database*. URL: http://www.hprd.org/, 2009.

[19] R. Bocu and S. Tabirca, *Betweenness Centrality Computation - A New Way for Analyzing the Biological Systems*. Proceedings of the BSB 2009 conference, Leipzig, Germany, 2009.
[20] L.C. Freeman, *A set of measures of centrality based on betweenness*. Sociometry, Vol. 40, 35-41, 1977.
[21] P.F. Jonsson and P.A. Bates, *Global topological features of cancer proteins in the human interactome*. Bioinformatics Advance Access, 2006.
[22] Wellcome Trust Sanger Institute, *The Pfam protein families database*. URL: http://pfam.sanger.ac.uk/, 2009.