Protein Communities Detection Optimization Through an Improved Parallel Newman-Girvan Algorithm

Razvan Bocu
and Sabin Tabirca
Department of Computer Science
University College Cork, Ireland
Email: {razvan.bocu, tabirca}@cs.ucc.ie

***Abstract*—Proteins and the networks they determine, called interactome networks, have received attention at an important degree during the last years, because they have been discovered to have an influence on some complex biological phenomena, such as problematic disorders like cancer. This paper presents a contribution that aims to optimize the Newman-Girvan community detection algorithm through a parallel implementation that is based on the MPI C programming environment. The optimization involves a double improvement of the original Newman-Girvan algorithm, which is accomplished both at the algorithmic and programming level. The resulting parallel implementation's performance was carefully tested on real biological data and the results acknowledge the relevant speedup that the optimization determines. Moreover, the results are in line with the previous findings that our current research produced, as it reveals and confirms the existence of some important properties of those proteins that participate in the carcinogenesis process. Apart from being particularly useful for research purposes, the novel technique also speeds up the proteomic databases analysis process, as compared to some of the sequential community detection approaches, and also to the original sequential Newman-Girvan algorithm.**

***Index Terms*—Betweenness centrality, interactome networks, protein-protein interactions, Newman-Girvan algorithm, protein communities, cancer, parallel computation.**

I. Introduction

A. Basic Considerations on Protein Networks and Their Importance

Interactome networks, or, more specifically, networks of proteins, determine a fundamental biological theoretical entity. Theoretical and practical endeavours often use interactome networks-related formalisms in order to analyze the protein interactions that determine a biological network, which is essential for the proper organization and function of a biological organism. These networks exhibit a complex structure, which implies that any research activity in the field is handled with inherent theoretical and technical difficulties. Nevertheless, the dynamics and the structure of these biological networks have to be accurately understood, as they play an important role on the function of a biological organism seen as a whole, regardless their degree of structural complexity. As a consequence, it is highly required to design and implement efficient proteomic data analysis techniques that can be integrated in any research framework that study the structure and properties of the interactome networks.

The centrality measure called betweenness can be efficiently used for a proper analysis of the networks of proteins, because it basically allows for functional protein clusters to be determined with a high degree of accuracy. The aim of this paper is to present a faster implementation that performs the detection of communities in the interactome networks, based on a computationally-effective optimized Newman-Girvan algorithm.

The significant influence that proteins exercise on fundamental physiological processes has been demonstrated in a series of recent contributions. In this respect, this paper demonstrates, apart from the algorithmic optimization itself that cancer affects the most important proteins in the interactome network and, as a consequence, the normal function of the organism is greatly disturbed. An accurate understanding of the structure and importance of proteins requires the usage of efficient analysis techniques.

The paper will briefly enumerate the most relevant existing works regarding the betweenness computation and community detection. Furthermore, the optimized version of the Newman-Girvan algorithm will be described and analyzed. Also, its practical usability is assessed on real proteomic data.

B. Essential Previous Work

The Newman-Girvan algorithm, the theoretical construct that is the basis of this paper's contribution, is one of the methods used to detect communities in complex systems. A community is built up by a subset of nodes within which the node-node connections are dense, and the edges to nodes in other communities are less dense. It is important to note that there are a number of alternative algorithmic techniques for the detection of communities in networks. They include hierarchical clustering, partitioning graphs to maximize quality functions such as network modularity, k-clique percolation, and some other interesting algorithmic methods [1,2]. Nevertheless, we chose the Newman and Girvan conceptual system

as a pretext for this contribution, due to its structural articulation and its ability to be used in a wide range of practical situations [3]. The Newman-Girvan algorithm is particularly used to compute betweenness for edges (biological links) that connect the nodes (proteins) in a network.

The following sections will describe the improved algorithm in more detail, along with some theoretical considerations that are mandatory for a good understanding of the new implementation's structure and strengths.

II. Theoretical Considerations

A. Basic Theoretical Constructs

Betweenness is a centrality measure that is based on the shortest path computation, and is widely used in the complex networks analysis [4]. It deals with one of the main problems in network analysis that supposes the precise assessment of the importance (or the centrality) of a particular vertex (or an edge) in a network, at the scale of the whole network. Betweenness has been extensively used in recent years for the analysis of social interaction networks, as well as other large complex networks. They can also be used to deal with problems from other research fields, such as lethality in biological networks, the study of sexual networks and AIDS, the identification of key actors in terrorist networks, and so on. Betweenness computations naturally constitute the primary routine in popular clustering and community identification algorithms that are applied on real-world network data.

Betweenness centrality can be computed both for nodes and for edges. The computation technique is, in principle, the same both for nodes and for edges, as it involves the computation of the distance matrix for a certain node or edge. In other words, regardless of the graph component on which the algorithm is applied, it has to compute the number of shortest paths that go through a certain vertex or edge, taking into consideration all the possible pairs of vertices. Therefore, we shall briefly describe the betweenness centrality for nodes (vertices), which is a centrality measure that computes the importance of a vertex within a graph. Vertices that occur on many shortest paths between other vertices have a higher betweenness, and therefore importance, than those that do not. For a graph *G=(V,E)* with n vertices, the betweenness $C_B(v)$ of the vertex *v* is calculated through the following formula [8]:

$$C_B(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (1)$$

where σ_{st} is the number of geodesics from vertex s to vertex t, and $\sigma_{st}(v)$ is the number of geodesics from vertex s to vertex t that go through vertex v. If necessary, this measure may pass through a normalization process, by dividing it through the number of pairs of vertices excluding v, which is $(n-1) \cdot (n-2)$.

The work reported in this paper is mainly concerned with the protein communities detection, which is based on the edge betweenness computation. In line with the assertions made in a previous paragraph, the remarks on nodes (proteins) betweenness computation can be entirely translated for the edges (biological links) betweenness computation, by simply replacing the concept of node with that of edge. All the other considerations, including the betweenness calculation formula, remain the same.

B. Remarks on the Newman-Girvan Algorithm

The algorithm [2] is historically important, because it marked the beginning of a new perspective in the field of community detection and opened this topic to various other specialists, including biologists and physicists. In principle, edges that are removed and added to the dendogram are selected according to the values of measures of edge centrality, thus estimating the importance of edges according to some property or process running on the graph. The steps of the algorithm are:

1) Compute the betweenness value for every edge.
2) The edge featuring the highest betweenness value is removed.
3) The resulting graph is re-processed through the re-computation of all the edge betweenness values.
4) The iteration moves on to the next step [2)], as long as there are still un-removed edges.

The edge betweenness measure quantifies the number of shortest paths between all vertex pairs that run along the edge. It is an extension to edges of the popular concept of site (node) betweenness, introduced by Freeman in 1977 [14] and expresses the importance of edges in processes like information spreading, where information usually flows through informational links that are components of shortest paths. From a historical point of view, edge betweenness was introduced before site betweenness by Anthonisse (Anthonisse, 1971) in a technical report, which was not actually published. It is intuitive that edges, which link communities exhibit a large value of the edge betweenness, because many shortest paths connecting vertices of different communities will pass through them. As in the calculation of node betweenness, if there are two or more geodesic paths with the same extremities, which run through an edge, the contribution of each of them to the betweenness of the edge must be divided by the multiplicity of the paths, as one assumes that the information/biological signal propagates equally along each geodesic path. The betweenness of all edges of the graph can be calculated in a time that scales as $O(m \cdot n)$, or $O(n^2)$ on a sparse graph, with techniques based on breadth-first-search (Brandes, 2001; Newman and Girvan, 2004; Zhou et al., 2006).

C. The Parallel Version of the Newman-Girvan Algorithm

First, we observe that by finding all-pairs shortest paths using Breadth-First Search (BFS) starting from each vertex in the graph, the edge betweenness value can be obtained by summing pair-dependencies [13] over all the traversals. The pair-dependency is defined as

$$\delta_{st}(v) = \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (2)$$

where σ_{st} denotes the number of shortest paths from $s \in V$ to $t \in V$, and $\sigma_{st}(v)$ is the number of shortest paths from s to t which go through v. Pair-dependencies calculated from each BFS for every vertex in the graph are additive. Summations from all traversals will give us the overall vertex betweenness, from which edge betweenness can be obtained by a trivial generalization. Since BFS can be performed independently and simultaneously from each vertex in the graph, the calculation required at each iteration of finding the edge with the highest betweenness value can be done by parallelizing all-pairs shortest paths.

The vertices of the graph are evenly assigned to all the processors, but each processor has its own copy of the graph, as each slave needs to have the exact image of the network in order to perform its own piece of the job. The procedure is initiated by a host processor, and then each processor performs BFS from all the vertices assigned to it and sums up partial pair-dependencies obtained from each BFS. The partial pair dependencies are then sent to the host processor. The host processor is responsible for summing up all the partial pair-dependencies from each processor, obtaining the global pair-dependencies, and finding the edge with the highest betweenness value. The edge with the highest betweenness value is then broadcast by the host processor to all the processors in the communication world. All the processors delete the edge received in their own graph copy and start the next iteration until the termination condition is satisfied.

1) Modularity: Community detection algorithms determine the community structure with various degrees of accuracy. In this respect, the main problem that may impede the output of such an algorithm is represented by the possibly inaccurate allocation of nodes (proteins) to their respective communities. We have consistently faced this problem during the current phase of our research. When speaking about protein data sets that require a proper community structure detection, the accuracy of the algorithm's output is mandatory, as even the slightest community structure misconfiguration may lead to incorrect deductions and conclusions. We made use of a measure called modularity, which assesses the quality of the community structure determined by the algorithm. The output of Girvan and Newmans betweenness algorithm is the order of removal of the edges, which implicitly defines a hierarchical tree on the nodes of the graph. In order to determine where to cut the tree to create the clusters, the notion of modularity is used.

Suppose there are k clusters in the current iteration of the algorithm. A symmetric matrix E of size $k \times k$ is constructed according to the following procedure. An element e_{ij} in E represents the fraction of all edges that link the vertices in cluster i to the vertices in cluster j and e_{ii} represents the fraction of edges that connect vertices within cluster i. Thus, summation of row (or column) elements $c_i = \sum_{i=1}^{k} e_{ij}$ represents the fraction of all edges that connect vertices to and within cluster i. In these conditions, modularity is defined as $Q = \sigma_{i=1}^{k}(e_{ii} - c_i^2)$, which measures the fraction of the edges that connect vertices within the same cluster minus the expected value of the same quantity in the network (Newman and Girvan, 2004). For a random network with random decomposition, Q approaches 0. Values approaching Q = 1, which is the maximum, indicate strong clustering structure. The higher is the value, the stronger is the clustering structure in the network. The inclusion of modularity as a community assessment measure significantly improved the accuracy of the community detection process output. The general structure of the algorithm is summarized in *Algorithm 1*.

Algorithm 1 Algorithmic structure of the parallel algorithm

Input: a protein network structure
Output: a dendogram representing the community structure of proteins
$modularity\_is\_optimal \leftarrow FALSE$
$modularity\_is\_optimal\_change \leftarrow FALSE$
Host machine distributes input to slaves in an even manner
while all slaves have at least one edge to process and $modularity\_is\_optimal = FALSE$ **do**
  if $modularity\_is\_optimal\_change = TRUE$ **then**
    $modularity\_is\_optimal \leftarrow TRUE$
  end if
  for all proteins p in the network **do**
    Perform a breadth-first search on p
    Compute relevant pair dependencies
    Send pair dependencies values to the host machine
  end for
  Host machine receives pair dependencies and calculates the edge betweenness values
  Determines the edge(biological link) with the highest betweenness value
  Remove the edge with the highest betweenness value and add it to the dendogram
  Compute $modularity\_value$
  if $modularity\_value > 0.8$ AND $modularity\_is\_optimal\_change = FALSE$ **then**
    //Perform only one more iteration to ensure optimality
    $modularity\_is\_optimal\_change \leftarrow TRUE$
  end if
end while

Biological networks, and protein networks are no exception, exhibit community structure at an important extent. As a consequence, studies performed on many biological data sets confirmed that a protein network whose community structure is correctly determined features a modularity value a little bit greater than 0.8, with a maximum that is usually reached around 0.83. Nevertheless, once the modularity goes over the 0.8 threshold, the network is practically partitioned into meaningful communities, and any other iteration apart from

the additional one, would require computation time that is not meant to polish the already determined community structure.

Information on cancer proteins was obtained from a comprehensive census of human cancer genes [6] that made use of various protein interaction databases, IntAct [10] 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 [7].

We wanted to isolate the communities that exist in the already known interactome network. In order to accomplish this, we made use of the new parallel algorithm, which properly extracted the community structure from the analyzed protein data sets. The procedure also involves the assessment of each protein's absolute importance, through the betweenness centrality measure computation. Furthermore, the last and the most important operation involves the comparative analysis of the detected functionally-related protein communities against the biological data regarding protein communities that is contained in the Pfam database [16]. This strategy allowed us to exactly determine interesting properties of those proteins that produce and sustain the carcinogenesis. Further details will be provided in the following paragraphs and sections.

Clustering methods have previously been shown to be useful in identifying protein interactions that take place within the same cellular process [15, 18]. 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 [19, 20]. 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 inferred by the previous phase of our research [21] and was confirmed through the usage of the optimized parallel Newman-Girvan-based scheme, which computes the functionally-related clusters of proteins in a computationally effective manner and with an excellent degree of accuracy. Let us recall that the community structure's detection degree of accuracy is assessed by continuously computing the modularity measure, at every iteration.

III. Performance Assessment

A. Programming Platform

The parallel version of the community detection algorithm is implemented in MPI C [11] and run on a Beowulf class cluster of computers. Each node in the cluster provides 12 MB of cache memory for the parallel processes that run on it. Additionally, each node in the parallel world is powered up by an Intel Xeon 5420 processor, clocked at 2.5 GHz. The MPI C solution was chosen as a consequence of the robustness, flexibility and wide acceptance in the academia and industry of the MPI library. Also, the C language was selected as the backbone of the parallel implementation, because it is the closest-to-machine medium level programming language and, as a consequence, the gain in performance is noticeable.

TABLE I
Execution times of the optimized parallel algorithm

Processors	Execution time	Efficiency	Speedup
2	3647	1.54	3.09
4	2037	1.48	5.78
8	1486	0.97	7.35
10	1104	0.83	8.31
16	759	0.61	9.82
32	448	0.44	13.04

The testing procedure calls the optimized community detection module, which accurately determines the functionally-related communities of proteins. We extend the computational efficiency 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. Compared to the original Newman's approach, the usage of this optimized parallel version of the edge betweenness computation contributes to the generation of the overall speedup.

B. Remarks on the Testing Procedure and Analysis

The efficiency of the algorithm is assessed by running a battery of tests composed of six instances. The original sequential algorithm of Newman and Girvan is run as a sequential process on the same parallel cluster, in order to ensure perfect accuracy of the comparative analysis. The results are summarized in *Table I*.

In parallel computing, *speedup* S_p is a measure of how much a parallel algorithm is faster than a corresponding sequential algorithm, and is computed by the formula:

$$S_p = \frac{T_1}{T_p}$$

where T_1 is the execution time of the sequential algorithm, while T_p is the execution time of the parallel algorithm with p processors.

Efficiency is a performance metric that is defined by the following formula:

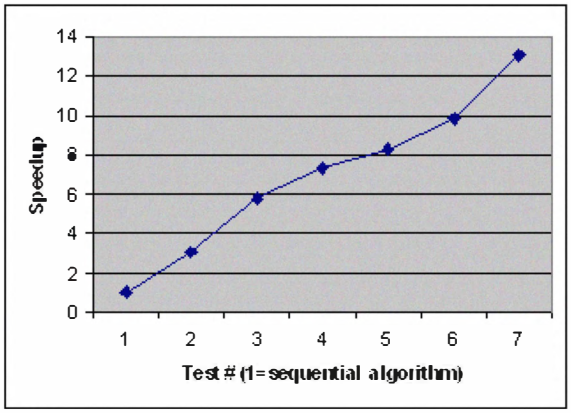

Fig. 1. Speedup induced by the optimized parallel algorithm

$$E_p = \frac{T_1}{p \cdot T_p}$$

where p is the number of processors.

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 [16].

The testing procedure made use of a compiled biological data set that features $22,573$ proteins and $1,886,753$ biological links. This size of the input data set determines a problem space that can be hardly handled properly in terms of execution times through a sequential approach.

We examined the protein communities our method determined and some interesting differences in the community sizes 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 and determines a distinct cellular process. Therefore, proteins that are part of multiple communities may generate multiple cellular 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 isolated by our community detection technique. The comparison between the cancer proteins population and the non-cancer proteins population reveals that cancer proteins reside at community junctions at a sensibly greater extent than their non-carcinogenic siblings. This particular feature of cancer proteins enforces their special importance in the interactome network seen as a whole and, as a consequence, their influence on all the physiological processes and related disorders.

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 [18]. Moreover, the findings reported in this paper support and extend the results generated by our research's previous stage. Practically, the above findings reveal the topological features of cancer proteins that are 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 parallel proteomic data analysis algorithm was designed and implemented and was found to accurately determine the functionally-related communities of proteins, together with each protein's absolute importance.
- 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 is not computationally effective for large proteomic data sets, it nevertheless remains a milestone for every researcher interested in community detection. Therefore, we sensibly optimized it, and we were able to design a faster parallel community detection module for our proteomic data analysis needs.

C. Conclusions and Future Developments

The most important property of cancer proteins is their importance at the scale of the whole interactome. The new parallel algorithm was used 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.

The resulting 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 indicates the explanation for this topological feature: 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 are used for an efficient community structure detection in protein networks. Also, we intend to analyze even more biological data sets 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] 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.

[2] M. Girvan and M.E.J. Newman, *Community structure in social and biological networks*: State University of New Jersey, 2002.

[3] D. Ucar et al., *Improving functional modularity in protein-protein interactions graphs using hub-induced subgraphs*: Ohio State University, 2007.

[4] K. Lehmann and M. Kaufmann, *Decentralized algorithms for evaluating centrality in complex networks*: IEEE, 2002.

[5] J. Griebsch et al., *A fast algorithm for the iterative calculation of betweenness centrality*: Technical University of Munchen, 2004.

[6] G.H. Traver et al., *How complete are current yeast and human protein-interaction networks?*: Genome biology, 2006.

[7] 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.

[8] U. Brandes, *A faster algorithm for betweenness centrality*: University of Konstanz, 2001.

[9] B. Preiss, *Data structures and algorithms with object-oriented design patterns in C++*: John Wiley and sons, 1998.

[10] EMBL-EBI, *The IntAct protein interactions database*. URL: http://www.ebi.ac.uk/intact/site/index.jsf, 2009.

[11] A. Grama et al., *Introduction to parallel computing, second edition*: Addison-Wesley, 2003.

[12] University of California, *The DIP protein interactions database*. URL: http://dip.doe-mbi.ucla.edu/, 2009.

[13] 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.

[14] L.C. Freeman, *A set of measures of centrality based on betweenness*: Sociometry, Vol. 40, 35-41, 1977.

[15] P.F. Jonsson and P.A. Bates, *Global topological features of cancer proteins in the human interactome*: Bioinformatics Advance Access, 2006.

[16] Wellcome Trust Sanger Institute, *The Pfam protein families database*. URL: http://pfam.sanger.ac.uk/, 2009.

[17] R. Bocu and S. Tabirca, *Sparse networks-based speedup technique for proteins betweenness centrality computation*: International Journal of Biological and Life Sciences, 2009.

[18] S. Wachi et al., *Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues*: Bioinformatics, 21, 4205-4208, 2005.

[19] G. Palla et al., *Uncovering the overlapping community structure of complex networks in nature and society*: Nature, 435, 814-818, 2005.

[20] P.F. Jonsson et al., *Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis*: BMC Bioinformatics, 7, 2, 2006.

[21] R. Bocu and S. Tabirca, *Proteomic Data Analysis Optimization Using a Parallel MPI C Approach*: IEEE Computer Society, The First International Conference on Advances in Bioinformatics and Applications, 2010.