A Bridging Centrality plugin for GEPHI and a case study for *Mycobacterium tuberculosis* H37Rv

Getulio Pereira, Preetam Ghosh, and Anderson Santos,E-mail: santosardr@ufu.br

Abstract—Bridging Centrality (BriCe) is a popular measure that combines the Betweenness centrality and Bridging coefficient metrics to characterize nodes acting as a bridge among clusters. However, there were no implementations of the BriCe plugin that can be readily used in the GEPHI software or any other software dedicated to graph-based studies. In this paper, we present the BriCe plugin for GEPHI. It is available as a third-party functionality from the native GEPHI interface as a handy plugin to add; hence, no additional download and installation process is necessary. The BriCe plugin for GEPHI is open-source, and one can access the code through the GEPHI GitHub repository. As a use case of the BriCe plugin, we analyzed the genome of Mycobacterium tuberculosis H37Rv to identify biological explanations on *why some proteins were ranked with top BriCe values?* For instance, we were able to formulate a new hypothesis combining the predicted sub cellular localization and high BriCe values concerning lipopolysaccharides (LPS) exportation. Our hypothesis provides a possible link among proteins of a glycosyltransferase group and the type VII Secretion System. The Bridging Centrality plugin for GEPHI is an easy to use tool for analyzing complex graphs and draw novel insights from graphical data.

Index Terms—Graph, Centrality, Betweenness, Bridging, GEPHI, Mycobacterium, type VII secretion system.

1 Introduction

Despite the existence of a considerable amount of complete genomes and a vast catalogue of bioinformatic tools to analyze such genomes, we are still failing to prevent the spread or even consolidate bacterial diseases worldwide [1]. Unfortunately, such an argument also applies to most illness caused by bacterial infections. We figured out in the last decade that only knowing about possible products of a genome does not guarantee a phenotypic explanation [2]. The vast extent and dimensionality of data incorporated by genome studies (e.g., transcriptome, microarray, and others), actually complicates computational analyses through machine learning algorithms which, in turn, must rely on simplification techniques for better accuracy [3]. Our limited computational efficiency in dealing with enormous amounts, variety, velocities, variability, and complexity of data [4] contribute to keeping us vulnerable to bacterial infections that prevail and provoke millions of yearly casualties. As an alternative to dealing with big biological data directly, we propose heuristic methods based on previously established knowledge about pathogens.

We can cite Betweenness Centrality (BetCe) and Bridging Centrality (BriCe) [5] as an example of heuristics applied to interaction networks. BetCe focuses on most traversed nodes no matter the number of connections. Despite BriCe also focusing on most traversed nodes, it imposes a condition on the number of connections; nodes should not be the most connected ones. On demanding less connected nodes, these become bridges in an interaction network. BriCe embeds the concept of Bottleneck nodes or Bottleneckness (Botness) [6]. Botness demands a low degree and high Betweenness Centrality for higher values, but BriCe can get higher values with degrees considered too big for Botness. The reason is that BriCe uses the Bridging Coefficient (BriCo) instead of solely the degree. The consequence is that we can note even nodes with modest Betweenness Centrality but significant Bridging Coefficient as meaningful ones for BriCe. Hence, we implemented the heuristic BriCe for complex networks in Java as a plugin for the The Open Graph Viz Platform (GEPHI) software. We used the BriCe plugin to test the hypothesis of the essentiality of the higher-ranked proteins by BriCe. The *in silico* analyses in this study involve the BriCe top-ranked proteins combined with the predicted subcellular localization [7]. The genome and the interaction network used as a case study was *Mycobacterium tuberculosis* H37Rv lineage, a reference genome for almost seven thousand *M. tuberculosis* genomes deposited at the NCBI, search tool for recurring instances of neighbouring genes (STRING) database and broadly published. We validate our hypothesis by depicting cases of proteins found in the genome of *M. tuberculosis* that act as a bridge between, at first glance, unrelated biological processes. We believe the availability of the BriCe as a plugin for a software broadly used as GEPHI and our results from the case study acts as a motivation of how BriCe, conjugated with additional data, could help us create relevant hypotheses for genomic studies.

- *P. Gosh is with the Department of Computer Science, Virginia Commonwealth University.*
- *G. M. Pereira and A. Santos are with Federal University of Uberlandia.*

Manuscript received 5 Nov. 2020; revised 9 Sep. 2021; accepted 13 Oct. 2021.
Date of publication 19 Oct. 2021.
(Corresponding author: Anderson Santos)
Digital Object Identifier no. 10.1109/TCBB.2020.3013837

2 Implementation and performance of the GEPHI plugin

We developed a BriCe plugin for the software GEPHI [8] and published it to the set of ready to install add-ons for GEPHI since Jan 2019. BriCe is a freeware software specified in Table 1. The BriCe program implements the Betweenness Centrality algorithm [9]. The main difference in our implementation with that of the native BetCe (that computes betweenness centrality of nodes) in the GEPHI software is in the number of decimal places. The BetCe values, calculated by the GEPHI software, store only six decimal places. To discriminate BriCe values of thousands of nodes in a graph, we require the java class BigDecimal. As BriCe calculated values are dependent on topological features, we cannot precisely capture the necessary scale of a network-based solely on the number of nodes/edges. To give an example of a precision scale according to topological conditions, we are going to cite an experiment on a particular network. Using several genomes with about 2500 proteins, we created two webs employing different parameters to decide if a neighborhood and phylogenetic profile are conserved. For a set of parameters, one genome achieved 120 thousand edges, and we could differentiate the top 50 BriCe ranks using a scale of 5 decimal places. For another set of parameters, the same genome achieved a web containing 920 thousand edges, and we only can discern the top BriCe ranks with a scale of 8 decimal places. For both experiments, the BriCe has a scale of 71, more than necessary for a medium-size genome. The BriCe plugin also computes and shows the centrality metrics BetCe and BriCo in a tabular form for all vertices within a graph. We calculate and include these three additional metrics to the original data laboratory of the software GEPHI. Since GEPHI itself already offers the possibility to calculate BetCe, for disambiguation, we called our BetCe of number two (Betweenness Centrality 2).

Project name	Bridging Centrality plugin for GEPHI
Project home page	github.com/santosardr/gephi-plugins
Operating system(s)	Platform independent
Programming language	Java
Other requirements	Java 1.8 or higher, GEPHI 0.9.2 or higher
License	GNU GPL
Restrictions for non-academics	licence needed

TABLE 1
Availability of the BriCe plugin for GEPHI

3 Brief description of the statistical methods

BetCe is a valuable centrality measure for complex networks. No matter what type of data is represented in an undirected graph, BetCe has the potential to unveil hidden relationships among the designated entities within a graph. It thus enables numerous types of data analyses. However, the users of this metric should understand its purpose to take proper advantage of our implementation. Our study has a biological context. However, we know that we can use BriCe in other circumstances. To demonstrate the use of BriCe in another background, we are going to explain its value over a network created by roads connecting cities, a net understood by a broader audience. If someone wants to search a network of cities to estimate the most popular traversed ones in a map, BetCe can handle such a task. First of all, we collect the set of all paths connecting a city s to another city t. This task could be monumental and hard to execute in a brute force algorithm. However, Dijkstra's algorithm can be used in this regard. Similarly, we need to gather the ways to connect cities s and t but now necessarily traversing some intermediate city v. Once we divide the number of ways of passing by v by all possible paths to reach t from s we can calculate the BetCe for the city v. Next, we need to rank cities according to BetCe to conclude the top cities on the map based on the importance of vehicle transit. Such statistics are vital, for example, to plan gas stations in a broader context. Another planning issue with gas stations could be to prioritize the main cities in this hypothetical map to be the first ones in having gas stations built. It obviously makes better sense to build the first gas stations in cities connecting large inhabited regions. The planning problem now consists of finding out cities connecting such populated areas. Such a problem motivates the use of BetCe, where higher ranked cities based on BetCe scores are related to the densely occupied regions but are not necessarily a part of these regions. The number of roads connecting cities within the dwelled areas are higher than the ones outside the cluster of towns. Our metric thus concentrates on the notion of the inverse of the Degree; this is called the Bridging Coefficient or BriCo. Finally, to create our priority places to construct the former gas stations in the map, one should multiply BetCe by BriCo generating a list of hot points vital to connect clusters of high-density populated areas; this metric is referred to as BriCe.

In bioinformatics, we can interpret these hot-points identified by BriCe as, for instance, proteins acting with a bridge role between groups of other proteins. These groups of proteins could represent clusters of players for related molecular process, and the relation dictated by top-ranked BriCe proteins. This motivates our implementation of the BriCe plugin for GEPHI. It provides the means to enumerate essential proteins capable of connecting biological processes.

3.1 Bridging Centrality calculation

$$BriCe\,(v) = BetCe\,(v)\,BriCo\,(v) \tag{1}$$

Bridging Centrality (BriCe) is a combined centrality measure for a vertex v applied to undirected graphs. It consists of the product of the Betweenness Centrality [10] and Bridging Coefficient [5]. BriCe is directly proportional to both metrics (eqn 1).

3.2 Betweenness Centrality calculation

$$BetCe\,(v) = \sum_{\substack{s,v,t \in V \\ s \neq v \neq t}} \frac{\rho_{st}(v)}{\rho_{st}} \tag{2}$$

BetCe (eqn 2) ranks vertices in a network according to ρ_{st}, the number of shortest paths between all possible pairs of vertices *s and t* passing through a vertex v, divided by ρ_{st}, the sum of all shortest paths among these pairs of vertices, even those not crossing the vertex v.

3.3 Bridging Coefficient calculation

$$BriCo\,(v) = \frac{\sum_{i \in \mathcal{N}(v)} k_i}{k_v} \tag{3}$$

Bridging Coefficient (BriCo) (eqn 3), is the other metric included in BriCe and is almost an inverse of the Degree Centrality (DegCe). BriCo considers ${k_v}^{-1}$, the inverted degree of a vertex v, divided by the inverted sum of all vertices s connected to the vertex v or $\mathcal{N}(v)$, the set of neighbors of v. Here, we simplified the formula moving the denominator from an inverted sum to a numerator.

3.4 Selection and characterization of studied proteins

We obtained the *M. tuberculosis H37Rv* Protein-Protein Interaction (PPI) network from the STRING database via direct download [11]. We next empirically decided to adopt a trustability cut-off of at least 75% over the STRING combined interaction evidence to prune interactions before further analyses. We made this choice to reduce the number of vertices and edges and also, keeping only the most reliable interaction evidence, thereby reducing false positives, which in turn could lead to fake relationships among biological processes. We converted the remaining STRING interactions into a DOT file for analyses by GEPHI producing the graph plotted in Figure 1. Figure 1 illustrates the complexity relative to what we considered a moderate interaction network, with about sixteen thousand edges and three thousand nodes. The larger circles are proteins listed in Table 2. The BriCe plugin allows us to uncover tricky relationships with significant biological meaning. Using the GEPHI visualization facilities, we can highlight the immediate connections for a node.

Figure 2 presents a graph of columns for each protein where the BriCe value determines the importance. Just a small amount of proteins resides at the very right side of the chart, the proteins we listed in Table 2. There is a significant difference between metrics of the top-ranked proteins according to BriCe and the whole set. While the first one hundred nodes/proteins account for an average degree of five edges, the entire set reaches ten edges per node, on average. Ordering all the nodes descending by column Degree gives ten proteins greater than one hundred, mean 167, and maximum degree of 281. These top-ranked proteins, according to the column Degree, are also referred to as hub nodes. Top-ranked BriCe and hub nodes are important topological features affecting the entire topology in case of removal [16]. Due to the topological importance of nodes with high BriCe/Botness and hub metrics, both features are used as starting points to align interaction networks, which are state-of-the-art for interaction network alignment algorithms [16].

We observed a unique fact when we sorted the data in descending order according to the column Degree. The top-ranked nodes, according to BriCe, got distributed along the x-axis of a figure similar to Figure 2 with no apparent standard (data not shown). After we sort the protein list according to the Degree column, the equal distribution of top-ranked nodes by BriCe along the x-axis explains the difference between Botness and BriCe metrics. One should

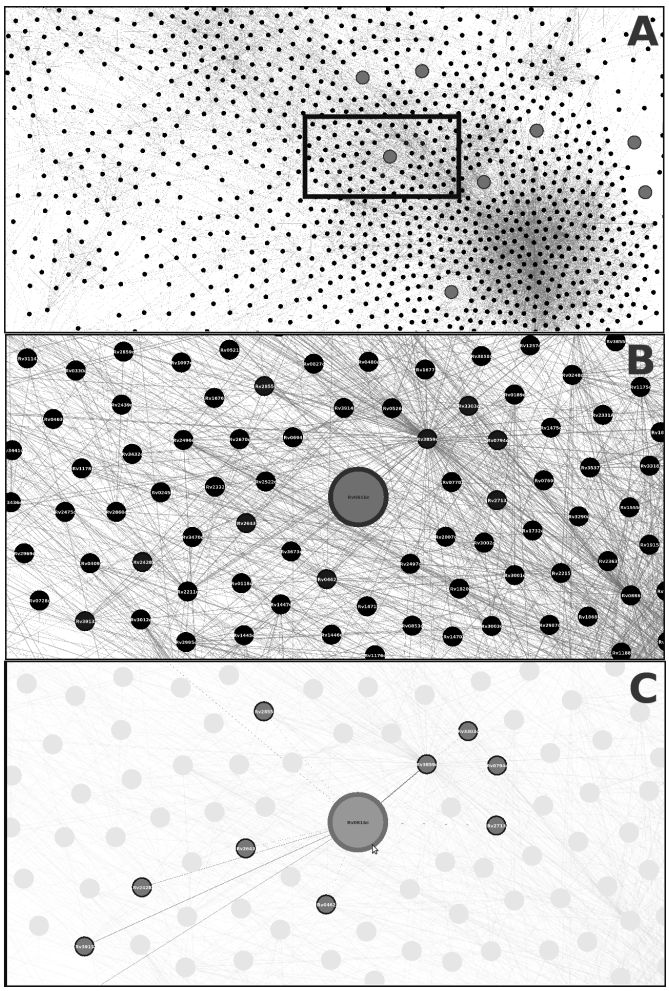

Fig. 1. Snapshots of the *M. tuberculosis* H37Rv protein-protein interaction (PPI) network. The full graph contains 3055 vertices and 15610 edges distributed in 297 communities with a modularity of 0,714 [12]. Panel A shows a partial zoom-out of the graph containing eight out of 11 top-ranked classified proteins according to Bridging Centrality (Table 2) plotted as larger circles. The vertices were distributed using the Yifan Hu algorithm [13] available in the GEPHI software. The rectangle inside the Panel A was zoomed into to create the Panels B and C showing details about the connectors to a larger vertex. We can observe high scoring vertices according to BriCe (Rv0816c) having no significant degree. However, the bigger vertices connect with other vertices having a substantial amount of connections, for instance, in Panel B, the vertex labelled Rv3859c located between 13:00 and 14:00 hours from the central vertex. Our BriCe GEPHI plugin ranked all the proteins, and we chose the top ones in decreasing order of BriCe value for literature research and queries based on the Gene Ontology (GO) [14], [15].

remember that BriCe is a more general concept than Botness. BriCe embeds the Botness concept when top-ranked nodes by Botness necessarily need a low Degree and high BetCe combination. At the same time, BriCe allows for other quantitative adjectives for Degree and BetCe. Botness accounts for just one out of four possibilities for columns Degree and BetCe values conjugations. For instance, the protein Rv0816c listed as the tenth-ranked in Table 2 has a modest BetCe value. Considering the maximum BetCe of this dataset as 0.08805, the Rv0816c BetCe corresponds to 7% of the maximum value. Compared to other proteins in Table 2, Rv0816c has a vast number of edges (eleven) that cannot be considered ”low” for Botness. Even so, its Bridging

Coefficient is proper to qualify it as one of the topmost BriCe nodes. The Rv0816c example fits an opposite example of Botness, but BriCe still embraced it. In table 2, we also have examples of low Degree and low BetCe, for instance, for Rv2942 plus high Degree and high BetCe represented by Rv0176, respectively being the first and second top-ranked proteins according to BriCe.

In table 2 there is not an instance fitting the Botness concept. We find out a protein fitting the Botness concept in the fortieth top-ranked proteins according to BriCe, out of 3055 proteins. This protein is Rv2531c, with Degree and BetCe values of 4 and 0.003601, respectively. At first glance, one can think of a 3.6 raised to 10^{-3} power a small value for BetCe, since Botness demands high BetCe values. However, the Rv2531c BetCe values are among the highest BetCe values in our *M. tuberculosis* dataset. To show that Rv2531c fits the Botness concept with low Degree (smaller than five) and high BetCe, we plotted a histogram of all BetCe values normalized by the maximum value in this dataset. To facilitate our demonstration of a high BetCe value for the protein Rv2531c, we needed to plot the logarithmic values of the normalized BetCe as depicted in Figure 3. In Figure 3, one can perceive that the normalized logarithmic value for the Rv2531c BetCe value (-3.20) stands at the right of the histogram, in the black-colored bin, more than one standard deviation (1.95) distant from the average value (-5.63).

Fig. 2. One hundred out of 3055 *M. tuberculosis H37Rv* proteins ranked according to the BriCe values, in descending order. The first one hundred ranked proteins have an average Degree of 5.37 against 10.22 of the whole genome. The maximum Degree for all the proteins is 281 against 19 for the first one hundred. Such differences in maximum and average Degree helps to explain the bridging role for the top-scored proteins partially.

We listed the top-eleven proteins according to BriCe in Table 2. We grounded our decision of which proteins to analyze further based on: (rule 1) the richness of the GO annotations and (rule 2) the number of proteins enrolled. The first candidate, Rv2942, does not pass our first rule, since there is just another one protein with some data coming from GO. The second candidate, Rv0176, did not have a single annotation in GO, even for itself. The third one is an exciting candidate, since it is related to six proteins, and there are 44 GO annotations for this set of proteins. However, we decided to focus on the fourth candidate (Rv2942) for our use case of BriCe in this work. We based our decision on the fact that there is another candidate (Rv3781) raised from BriCe score, at the eleventh position, connected to Rv2942; we chose to emphasize a pair of proteins that are both highly ranked by BriCe than a single one. Moreover, the proteins Rv2942 and Rv3781 are closely related; they are glycosyltransferases (GTs) for *M. tuberculosis H37Rv*.

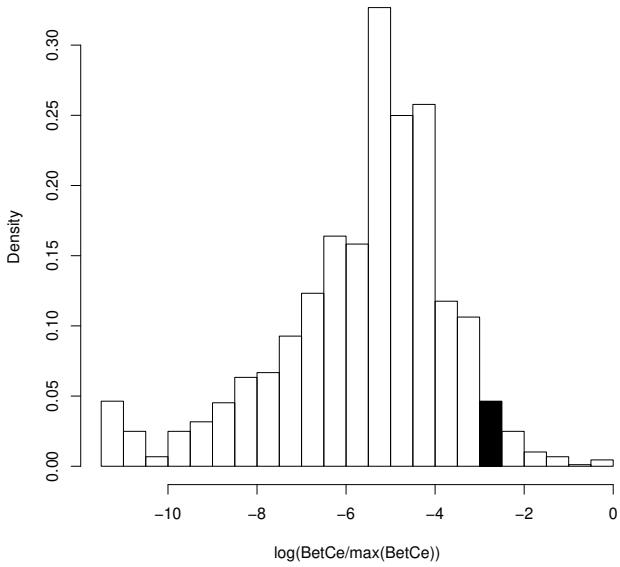

Fig. 3. We plotted this histogram normalizing the BetCe values for all *M. tuberculosis* proteins (3055) in our dataset. The normalizing factor was the maximum BetCe value from the dataset. BriCe embeds Botness because top-ranked proteins, according to BriCe, also can have low Degree and high BetCe values, as proposed by the Botness concept. The protein Rv2531c, with a Degree of four (low), standing at position forty of the top-ranked BriCe values, occupies the black-colored bin of this histogram. Despite a small apparent value for the BetCe metric for the protein Rv2531c (3.6 raised to 10^{-3}), its BetCe is located more than one standard deviation from the mean BetCe value, comprising a BetCe value that we can consider as a high one.

There are also other three proteins in this set with enough annotation in the GO database to construct a hypothesis.

Rank Number	Protein name	Degree	Betweenness Centrality	Bridging Coefficient	Bridging Centrality
1	Rv2942	2	0.00197	2.22222	0.00437
2	Rv0176	12	0.02574	0.13242	0.00341
3	Rv2641	7	0.00211	1.13880	0.00241
4	Rv3779	10	0.03047	0.07038	0.00214
5	Rv2852c	3	0.00045	3.83618	0.00173
6	Rv2037c	5	0.00331	0.51722	0.00171
7	Rv1644	2	0.00007	20.02759	0.00141
8	Rv3300c	3	0.00167	0.77064	0.00128
9	Rv1702c	5	0.00417	0.28867	0.00120
10	Rv0816c	11	0.00607	0.19040	0.00116
11	Rv3781	5	0.00406	0.28410	0.00115

TABLE 2
We depicted the top proteins ranked by BriCe in descending order. The multiplication of the value from columns Betweenness Centrality and Bridging Centrality produces the column Bridging Centrality values. Despite the proteins Rv3739 and Rv3781 not having the highest BriCe values, we opt to use these in our *M. tuberculosis* case study after we figured out both as glycosyltransferase related. In general, the top-ranked nodes have a low Degree and high Betweenness Centrality, although the small apparent values. One should remember that we can only distinguish Betweenness Centrality starting, for instance, in the fourth decimal digit or more depending on the number of nodes and edges in a network.

Although we consulted Gene Ontology (GO) in this work for annotated molecular function, biological process and sub cellular location, we also leveraged sub cellular

location predictions provided by the software SurfG plus [7]. We based our adoption of SurfG plus on our prior experience in using this software for the pan secretome of *Corynebacterium pseudotuberculosis*, strains 1002 and C231 [17]. In the Pacheco et al. work, *in vitro* experiments confirmed exportation for the majority of the full set (93 out of 137) of proteins that were predicted *in silico* as exported by SurfG plus. Other software, considered the gold standard at the time, predicted less than a quarter of the experimentally confirmed proteins. Based on this experience, we have data in achieving higher accuracy of prediction using SurfG plus for bacterial genomes. Even so, in this work, all forecasts made by SurfG plus had a match with GO annotations of sub cellular location. Moreover, since GO does not have annotations for all predicted proteins of any genome, in this work, the sub cellular localization predicted by SurfG plus serves as a fundamental data source for our rationale.

4 Application to *M. tuberculosis* H37Rv

Figure 4 is a schema that presents the main players of our hypothesis that there is a possible association among glycosyltransferases, strictly known as exporting to the inner membrane, with Type VII Secretion System (T7SS) recognised for exportation to the outer layer for the interaction. The proteins Rv3779 and Rv3781 (glycosyltransferases) are directly interacting, according to the STRING database [11]. In Figure 4 we decided to represent four out of nine proteins interacting with Rv3779 and Rv3781, plus Rv3884c. We aim to reduce the number of involved proteins and to focus on our hypothesis. The following subsections present our hypothesis detailing each player depicted in Figure 4. In the end, we intersect both players in a standard molecular process, the exportation of lipopolysaccharides (LPS) to the outer membrane.

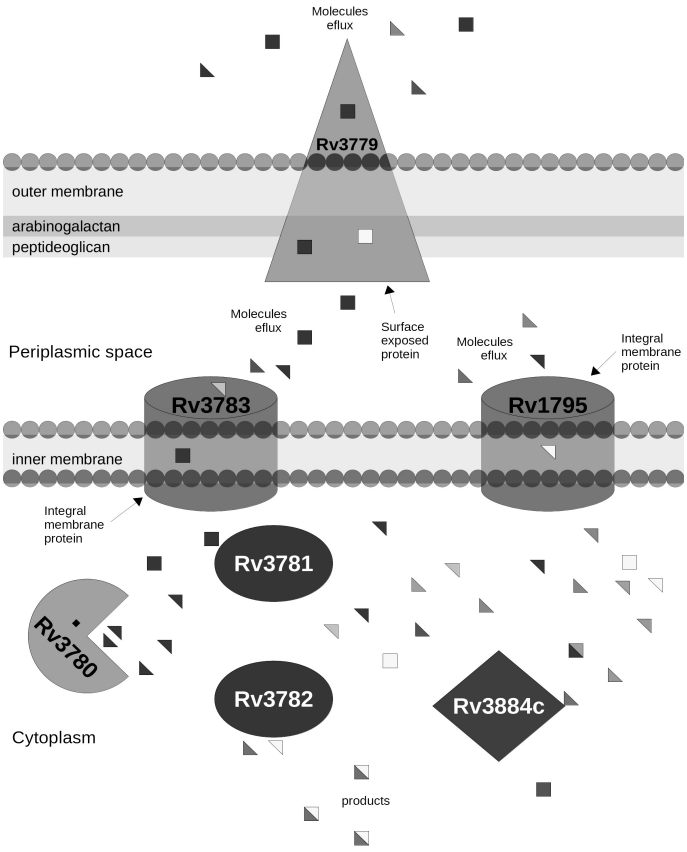

Fig. 4. Predicted sub cellular localization for proteins related to Rv3779 and Rv3781. Cylinders represent integral inner membrane, and the triangle represents a potentially surface exposed protein. We chose the other symbols for the cytoplasmatic compartment without any special meaning. This figure schematically presents the idea of a possible relationship among proteins of the Type VII Secretion System (left) and the GTs cluster (right and top). We raised predictions and annotations sustaining odds for Rv3779, belonging to the glycosyltransferase cluster, responsible for carrying out LPS for the outer membrane by interacting with the T7SS proteins. We corroborate this scenario by local sub cellular predictions, STRING interactions based predictions, annotated biological processes for all proteins involved, amino acid sequence similarity for orthologues (in *E. coli*), Gene Ontology annotations, and annotated products in the genome of *M. tuberculosis* H37Rv.

4.1 Exportation to the inner layer

The prediction of sub cellular localization for the protein Rv3779 was Potentially Surface Exposed (PSE) and Rv3781 was a Cytoplasmatic (CYT) protein. Both proteins are within the top-scored ones according to the BriCe GEPHI plugin (Table 2). One can easily perceive that the majority of proteins in Figure 4 are prone to participate in the same genome loci due to their almost consecutive locus tag numerations (Rv3779, Rv3780, Rv3781, Rv3782, and Rv3783). The proteins in Figure 4 are called glycosyltransferases, a cluster of biosynthetic cell wall genes. Glycosyltransferases are involved in glycosylation events like the synthesis of arabinogalactan, mycolic acid, and lipoarabinomannan [18]. The glycosyltransferases cluster is associated with the synthesis and transport of the polysaccharide arabinogalactan (AG) to the periplasmic space. For instance, the product UDP-galactofuranosyl transferase (GlfT1) was linked to the protein Rv3782, which is a crucial step for AG synthesis [19]; and the protein Rv3783 depicted as an integral membrane in Figure 4 is known as Wzm working with Wzt (Rv3781) for AG secretion to the periplasmic space [18].

4.2 Exportation to the outer layer

There are proteins in Figure 4 that are absent from the glycosyltransferases cluster. For instance, the Rv1795 (EccD5) is an integral membrane protein participating in an ABC transport system known as the esx-5 locus, part of a T7SS, but present only in slow-growing mycobacteria including most pathogenic species [20]. Likewise, the Rv3884c protein belongs to the esx-2 locus in *M. tuberculosis*, which is also a part of T7SS [21]. Our local sub cellular prediction categorised the protein Rv3884c as CYT. The previously mentioned proteins share an essential commonality, as they are related to the cell wall biosynthesis forging and transporting polysaccharides. The proteins Rv3884c and Rv1795 are related to composite and transportation of polysaccharides associated with lipids (lipopolysaccharides or LPS) to the outer membrane of *M. tuberculosis*. EccA2 and EccD5 are coupled with transportation of LPS to the outer membrane of the cell wall envelope through the T7SS [22].

4.3 Intersection

For Escherichia coli, a study showed nonviable mutants carrying the modified genes for the Wzt and Wzm proteins [23]. The authors demonstrated defective polymannan O-antigenic polysaccharides for mutants carrying the mutated wzt and wzm genes. The Wzt and Wzm proteins are part of the glycosyltransferases cluster, and included in *M. tuberculosis H37Rv*. The presence of proteins similar to Wzt in *M. tuberculosis H37Rv* dictated the product annotation of Rv3781 as "*Probable O-antigen/lipopolysaccharide transport*" or, putting in other words, responsible for exportation of LPS for the outer membrane. At this point, one should recall that glycosyltransferases are related to exportation only to the inner membrane. Rv3781 has 67 and 45%, respectively, of similarity and identity with Wzt protein family of E. coli. This Rv3781 feature can raise a question about a possible association among glycosyltransferases cluster and the T7SS, which sustain our hypothesis in this use case. It looks reasonable since glycosyltransferases proteins are related only to the inner membrane exportation of polysaccharides, and not exportation to the outer layer. Finally, we predicted Rv3779 (glycosyltransferases cluster) as a potentially surface-exposed protein by the SurfG plus software [7]. It can also allow one to query: *Why a group dedicated to exportation to the inner membrane should need a protein predicted to cross the outer layer, exposing part of the structure to the periplasmic space and the extracellular milieu?* We support the hypothesis of the relationships among the glycosyltransferases and the T7SS proteins due to:

(i) Their predicted sub cellular localization;
(ii) The similarity of their biological processes;
(iii) The differential BriCe value of Rv3779 and Rv3781;
(iv) Similar functional annotations among proteins at both exportation systems like, for instance, Rv3781;
(v) Also, Rv3779 and Rv3781 are possibly acting as a bridging node of *M. tuberculosis* H37Rv considering the highly probable interactions (>=75%).

Such a hypothesis could fit, for instance, as a possible explanation to the need of extra proteins necessary to complete the T7SS, an assumption well accepted according to recent works [24], [25]. In our hypothesis, the protein Rv3779 known as part of the glycosyltransferase cluster could be acting as a shipping channel for the T7SS exportation system to communicate to the outer layer of *M. tuberculosis* H37Rv.

5 Discussion

Our BriCe plugin allows us to explore new frontiers about interaction networks. In general, centrality measures applied to interaction networks can open up such new horizons. However, a metric should be meaningful according to the studied context. BriCe provides us noteworthy relationships uncovering complex and distant interactions between hundreds of thousands of interactions. Using the BriCe plugin, we extrapolate the fragile limit of the immediate neighborhood interactions commonly inspected on interaction networks by conducting a formal analysis.

A thorough analysis of an interaction network can be expensive in terms of time and memory for processing. There are several possible features to classify an interaction network. Let us assume as primary features, for instance, the number of nodes, vertices, and average degree. Table 3 lists these features for several genomes with their respective execution times for BriCe computation. In Table 3, we filtered the *M. tuberculosis* genome to keep edges with probability prediction greater than 75, 50, and 0% (all edges). Besides the three versions of *M. tuberculosis*, we also included a *Staphylococcus warneri* and a *Corynebacterium tuberculosis* genome in our analysis, two genomes we have worked with earlier. The interaction network we analyzed in this work can be considered tiny, with less than sixteen thousand edges in about three thousand nodes. It took just twelve seconds to complete the BriCe analysis. According to the average proportions of the STRING database interaction networks, and concerning the three features in Table 3, one can expect between five and ten minutes to complete the BriCe analysis on average. We performed the *S. warneri* and *C. tuberculosis* experiments in Table 3, with interaction networks created by in-house software (genppi.facom.ufu.br) which allows tweaking parameters to create interconnections. Besides, we conceived the last experiment producing a massive web that is not too beneficial for centrality measures due to few nodes being differentiated considering the number of edges incident on each node. We believe that the last experiment in Table 3 is an extreme example which is challenging to be found naturally. Even so, this extreme case did not take more than 14 minutes using a 1.3 extra GB RAM. Nevertheless, an ordinary user of our software can have BriCe analysis done, using an 8 GB RAM machine, for a few thousand nodes and hundreds of thousands of edges in a dozen minutes. For instance, the whole *M. tuberculosis* interaction network considering all edges, with probability larger than zero, took just seven minutes to compute the BetCe, Brico, and finally, the BetCe calculations.

Organism and Experiment	Nodes	Edges	Average Degree	Time in seconds
M. tuberculosis P(edge)>0.75	3055	15610	10	12
M. tuberculosis P(edge)>0.50	3877	48135	25	55
C. pseudotuberculosis full	2058	280102	272	160
S. warneri full	2468	277795	225	176
M. tuberculosis full	3967	333655	168	432
C. pseudotuberculosis extreme	2150	1132381	1053	819

TABLE 3
We analyzed three genomes using the BriCe plugin for GEPHI. The interaction network named *M. tuberculosis* full is the raw interaction network as downloaded from the STRING database. We filtered the probability of each edge with thresholds P(edge)>0.75 and P(edge)>0.50 creating the other two *M. tuberculosis* webs. We created the *Staphylococcus warneri* and *Corynebacterium tuberculosis* interaction networks using in-house software that customizes the number of edges and average degrees of an interaction network.

One can currently use our Brice plugin only in the GEPHI software, which counts as a disadvantage. However, as possible future work, we intended to also make it available as Cytoscape or R library. The fact that it is open source also allows everyone to contribute to our project. Another possible branch for our work could be researching other metrics conjugation and their biological meaning. For example, the conjugation of BetCe and BriCo metrics gives us BriCe. Perhaps, conjugating metrics like Closeness, Eigenvector, Degree, Harmonic, and Katz centrality with

BriCo or others could create a third metric possessing some biological meaning.

The innovativeness of BriCe was revealed to the scientific community in 2006 by Hwang and collaborators [5]. We interpret it as a successful case transcending the biological meaning and being rebaptized one year later in a more decisive publication venue for a broader audience [6]. We interpret our manuscript as cutting edge by spreading the news of providing easy BriCe or Botness calculations to the entire scientific and non-scientific community at a distance of a two-mouse click (one to install the additional GEPHI library and a second to run it). Our software tool makes BriCe calculations readily available for any interaction network, benefiting scientific and nonscientific communities. Until the BriCe plugin, there was no way, for instance, for a biologist using GEPHI to analyze the interaction network of an organism of interest, looking for clues about crucial nodes. Although the algorithm is not novel, supporting any undirected graph and ease of accessing BriCe calculations assure novelty.

6 Conclusion

BriCe is a mathematical relationship created by the multiplication of two other known ones, BetCe and BriCo. The BriCe plugin for GEPHI is a handy tool for graph analyses and is not restricted solely to bioinformatics research but can be useful to depict hidden relations between groups of entities in any undirected graph. In the case study presented here, BriCe helps us to develop a hypothesis on the possible relationship among two groups of proteins that were previously unrelated. Despite the lack of connection among these groups of proteins, their annotations intersect suggesting such a relation. The use of BriCe plugin for GEPHI could also help other researchers to gain novel insights about their data modelled as undirected graphs; our implementation is generic and can be readily used for centrality calculations for several other network related problems related to communication efficiency [26], [27], [28].

Acknowledgments

This study was financed in part by the Coordenao de Aperfeioamento de Pessoal de Nvel Superior - Brasil (CAPES) - Finance Code 001

References

[1] M. Pai, M. A. Behr, D. Dowdy, K. Dheda, M. Divangahi, C. C. Boehme, A. Ginsberg, S. Swaminathan, M. Spigelman, H. Getahun, D. Menzies, and M. Raviglione, "Tuberculosis," *Nature Reviews Disease Primers*, vol. 2, pp. 1607–1613, Oct. 2016. [Online]. Available: https://doi.org/10.1038/nrdp.2016.76

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular biology of the cell*. Garland Science, 2002, vol. 4.

[3] A. Drouin, G. Letarte, F. Raymond, M. Marchand, J. Corbeil, and F. Laviolette, "Interpretable genotype-to-phenotype classifiers with performance guarantees," *Scientific Reports*, vol. 9, no. 1, pp. 1–13, Mar. 2019. [Online]. Available: https://doi.org/10.1038%2Fs41598-019-40561-2

[4] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts methods, and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137–144, Apr. 2015. [Online]. Available: https://doi.org/10.1016%2Fj.ijinfomgt.2014.10.007

[5] W. Hwang, Y. Cho, A. Zhang, and M. Ramanathan, "Bridging Centrality: Identifying Bridging Nodes in Scale-free Networks," University at Buffalo, Tech. Rep., 2006. [Online]. Available: https://www.cse.buffalo.edu/tech-reports/2006-05.pdf

[6] H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein, "The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics," *PLoS Computational Biology*, vol. 3, no. 4, pp. 713–720, Apr. 2007. [Online]. Available: https://doi.org/10.1371/journal.pcbi.0030059

[7] A. Barinov, V. Loux, A. Hammani, P. Nicolas, P. Langella, D. Ehrlich, E. Maguin, and M. van de Guchte, "Prediction of surface exposed proteins in Streptococcus pyogenes with a potential application to other Gram-positive bacteria," *PROTEOMICS*, vol. 9, no. 1, pp. 61–73, Jan. 2009. [Online]. Available: https://doi.org/10.1002%2Fpmic.200800195

[8] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An Open Source Software for Exploring and Manipulating Networks," vol. 2009, Jul. 2009, pp. 361–362. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154

[9] U. Brandes, "A faster algorithm for betweenness centrality," *The Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163–177, Jun. 2001. [Online]. Available: https://doi.org/10.1080%2F0022250x.2001.9990249

[10] Y. Zhao, *Accelerating betweenness centrality computation on heterogeneous processors*, 1st ed., H. K. University, Ed. The Hong Kong University of Science and Technology Library, Jun. 2014. [Online]. Available: https://doi.org/10.14711%2Fthesis-b1301512

[11] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, J. Wyder, S .and Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, and C. V. Mering, "STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets." *Nucleic Acids Research*, vol. 47, pp. 607–613, Jan. 2019.

[12] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, pp. 100–108, Oct. 2008. [Online]. Available: https://doi.org/10.1088%2F1742-5468%2F2008%2F10%2Fp10008

[13] Y. Hu, "Efficient, High-Quality Force-Directed Graph Drawing," *The Mathematica Journal*, vol. 10, pp. 37–71, 2006. [Online]. Available: https://www.mathematica-journal.com/issue/v10i1/contents/graph_draw/graph_draw.pdf

[14] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene Ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, May 2000. [Online]. Available: https://doi.org/10.1038%2F75556

[15] T. G. O. Consortium, "The Gene Ontology Resource: 20 years and still GOing strong," *Nucleic Acids Research*, vol. 47, no. D1, pp. 330–338, Nov. 2018. [Online]. Available: https://doi.org/10.1093%2Fnar%2Fgky1055

[16] U. Ayub, I. Haider, and H. Naveed, "SAlign-a structure aware method for global PPI network alignment," *BMC Bioinformatics*, vol. 21, no. 1, p. 500, 2020. [Online]. Available: https://doi.org/10.1186/s12859-020-03827-5

[17] L. G. C. Pacheco, S. E. Slade, N. Seyffert, A. R. Santos, T. L. P. Castro, W. M. Silva, A. V. Santos, S. G. Santos, L. M. Farias, M. A. R. Carvalho, A. M. C. Pimenta, R. Meyer, A. Silva, J. H. Scrivens, S. C. Oliveira, A. Miyoshi, C. G. Dowson, and V. Azevedo, "A combined approach for comparative exoproteome analysis of Corynebacterium pseudotuberculosis," *BMC Microbiology*, vol. 11, no. 1, p. 12, 2011. [Online]. Available: https://doi.org/10.1186%2F1471-2180-11-12

[18] S. Berg, D. Kaur, M. Jackson, and P. J. Brennan, "The glycosyltransferases of Mycobacterium tuberculosis roles in the synthesis of arabinogalactan lipoarabinomannan, and other glycoconjugates," *Glycobiology*, vol. 17, no. 6, pp. 35–56, Jan. 2007. [Online]. Available: https://doi.org/10.1093%2Fglycob%2Fcwm010

[19] K. Mikusova, M. Belanova, J. Kordulakova, K. Honda, M. R. McNeil, S. Mahapatra, D. C. Crick, and P. J. Brennan, "Identification of a novel galactosyl transferase involved in biosynthesis of the mycobacterial cell wall." *Journal of Bacteriology*, vol. 188, pp. 6592–6600, Sep. 2006.

[20] L. S. Ates, R. Ummels, S. Commandeur, R. van der Weerd, M. Sparrius, E. Weerdenburg, M. Alber, R. Kalscheuer, S. R. Piersma, A. M. Abdallah, M. A. E. Ghany, A. M. Abdel-Haleem, A. Pain, C. R. Jiménez, W. Bitter, and E. N. G. Houben, "Essential Role of the ESX-5 Secretion System in Outer Membrane Permeability of Pathogenic Mycobacteria," *PLOS Genetics*, vol. 11, no. 5, pp. 1–30, May 2015. [Online]. Available: https://doi.org/10.1371%2Fjournal.pgen.1005190

[21] B. Callahan, K. Nguyen, A. Collins, K. Valdes, M. Caplow, D. K. Crossman, A. J. C. Steyn, L. Eisele, and K. M. Derbyshire, "Conservation of Structure and Protein-Protein Interactions Mediated by the Secreted Mycobacterial Proteins EsxA, EsxB and EspA," *Journal of Bacteriology*, vol. 192, no. 1, pp. 326–335, Oct. 2009. [Online]. Available: https://doi.org/10.1128%2Fjb.01032-09

[22] M. H. Touchette and J. C. Seeliger, "Transport of outer membrane lipids in mycobacteria," *Biochimica et Biophysica Acta*, vol. 1862, no. 11, pp. 1340–1354, Nov. 2017. [Online]. Available: https://doi.org/10.1016%2Fj.bbalip.2017.01.005

[23] L. Cuthbertson, J. Powers, and C. Whitfield, "The C-terminal domain of the nucleotide-binding domain protein Wzt determines substrate specificity in the ATP-binding cassette transporter for the lipopolysaccharide O-antigens in Escherichia coli serotypes O8 and O9a." *Journal of Biological Chemistry*, vol. 280, pp. 30 310–9, Aug. 2005.

[24] K. S. H. Beckham, L. Ciccarelli, C. M. Bunduc, H. D. T. Mertens, R. Ummels, W. Lugmayr, J. Mayr, M. Rettel, M. M. Savitski, D. I. Svergun, W. Bitter, M. Wilmanns, T. C. Marlovits, A. H. A. Parret, and E. N. G. Houben, "Structure of the mycobacterial ESX-5 type VII secretion system membrane complex by single-particle analysis," *Nature Microbiology*, vol. 2, no. 6, pp. 1–12, Apr. 2017. [Online]. Available: https://doi.org/10.1038%2Fnmicrobiol.2017.47

[25] E. N. G. Houben, K. V. Korotkov, and W. Bitter, "Take five Type VII secretion systems of Mycobacteria," *Biochimica et Biophysica Acta*, vol. 1843, no. 8, pp. 1707–1716, Aug. 2014. [Online]. Available: https://doi.org/10.1016%2Fj.bbamcr.2013.11.003

[26] B. K. Kamapantula, A. Abdelzaher, P. Ghosh, M. Mayo, E. J. Perkins, and S. K. Das, "Leveraging the robustness of genetic networks: a case study on bio-inspired wireless sensor network topologies," *Journal of Ambient Intelligence and Humanized Computing*, vol. 5, no. 3, pp. 323–339, Jun. 2014. [Online]. Available: http://dx.doi.org/10.1007/s12652-013-0180-0

[27] B. K. Kamapantula, A. Abdelzaher, P. Ghosh, M. Mayo, E. Perkins, and S. K. Das, "Performance of wireless sensor topologies inspired by E. coli genetic networks," *Proc. IEEE Int. Conf. Pervasive Computing and Communications Workshops*, vol. 1, pp. 302–307, Mar. 2012.

[28] A. Nazi, M. Raj, M. Di Francesco, P. Ghosh, and S. K. Das, "Deployment of robust wireless sensor networks using gene regulatory networks: An isomorphism-based approach," *Pervasive and Mobile Computing*, vol. 13, pp. 246–257, 2014.

Getulio Pereira Master's degree in Artificial Intelligence at the Faculty of Computer Science/UFU (2018). Graduated in Computer Science at the Federal University of UberlIndia (2003) and Post-Graduated Lato Sensu in Systems Development for Web and Mobile Devices at IFTM, Campus Ituiutaba (2012). Professional experience in analysis and development of large systems for the internet on the J2EE platform and consultor for relational databases. Interest in research and development in artificial intelligence, bioinformatics, computer vision, and embedded computing. http://lattes.cnpq.br/4601861527107785

Preetam Ghosh Preetam Ghosh is a Professor in the Department of Computer Science and directs the Biological Networks Lab at Virginia Commonwealth University. He obtained his MS and Ph.D. degrees in Computer Science and Engineering from the University of Texas at Arlington and a BS in Computer Science from Jadavpur University, Kolkata, India. His research interests include algorithms, stochastic modeling and simulation, network science, machine learning-related systems biology and computational epidemiology approaches, and mobile computing-related issues in pervasive grids. He published more than 170 conference and journal articles and federally funded research projects from NSF, NIH, DoD, and US-VHA. He currently serves as the Secretary/Treasurer of ACM SIGBio. https://egr.vcu.edu/directory/preetamghosh/

Anderson Santos Anderson Santos is Graduated in Computer Science (Catholic University of Minas Gerais, Brazil, 1995), Masters in Computer Science with the emphasis in Artificial Intelligence (Federal University of Minas Gerais, Brazil, 1999) and Ph. D. in Bioinformatics also at the Federal University of Minas Gerais (2012). Significant participation in the assembly and annotation of the first genome project conducted entirely in the state of Minas Gerais (Brazil) about the bacterium Corynebacterium pseudotuberculosis. Experienced in the use of Computer Science for assembly and annotation of genomes. In the year of 2012, was a post-doctoral fellow at the UFMG's Laboratory of Cellular and Molecular Genetics coordinating the work of masters and doctoral students involved in the assembly, annotation, and analysis of several bacterial genomes. I am a professor at the Federal University of Uberlandia, Faculty of Computing since the year of 2013 ministering disciplines like modeling and simulation, databases, programming, optimization and artificial intelligence. https://www.researchgate.net/profile/Anderson-Santos-7