A critical summary of 'Centrality analysis methods for biological networks and their application to gene regulatory networks'

Vedang Joshi

Candidate number: 1849432
qk18050@bristol.ac.uk

October 23, 2019

Abstract

The complexity of systems in bio-network studies requires a method of study to simplify these networks into meaningful subdivisions. Their importance can then be quantified according to predefined centrality measures, to better understand the functions of these systems. We critically examine Koschützki and Schreiber's approach to ranking genes in gene regulatory networks and note that there are other ways to rank genes apart from using correlation coefficients for different centrality measures. Through an extensive literature review, we find that Koschützki and Schreiber's paper still remains one of the few definitive studies of ranking hierarchically significant genes after 11 years from publication.

1 Introduction

Complex systems such as cellular and protein connections in biological systems can be modelled mathematically using networks. It is easier to study these networks by studying the vertex connections between points rather than analysing the elements of the network separately [1, 29]. To understand how these networks behave, a comparative analysis using different centrality measures, of structurally similar modules across different species may lead to the understanding of the evolution of different network structures [26] and contribute to the body of knowledge. This paper focuses on summarising and critiquing the different centrality measures explained in Koschützki and Schreiber's paper [15] and tries to relate, with simplifications, these centrality measures specifically the motif based centralities to gene regulatory networks (GRNs). The authors assume prior knowledge of GRNs, which is understandable considering the targeted audience for the paper. Yet for the purpose of completeness, GRNs are systems consisting of many thousands of pieces of DNA sequences wherein each piece receives and processes multiple inputs (a surjective function), in the form of regulatory proteins that recognise specific sequences within them. The end result is the increase or decrease of production of RNA and other proteins [23, 5]. This understanding is motivated by the importance of such networks in drug development and studying normal and diseased (cancerous) cell behaviours. Other evidence also suggests that this inter-connectivity between vertices in protein interaction networks is important to understanding lethality (a type of genetic interaction where the co-occurrence of two genetic events results in organismal or cellular death [9, 2]). To aid bio-mathematical research, Koschützki and Schreiber introduce

the definitions of directed graphs, motifs and other notations in graph theory. They go on to introduce the centralities degree, closeness, shortest path betweenness and motifs and how it relates to ranking the important global regulators i.e. within the top 25 (top 2% of all genes) genes. The authors have thus clearly outlined their proposals for the paper and have backed it with the relevant literature. They also split the paper into two parts, the first one, to outline the theory behind why certain centrality measures are preferred over others. The second part then uses data gathered by Salgado [25] to compare centrality measures to give the best measure(s) to analyse the data.

2 Graph theory

Koschützki and Schreiber start by describing networks being represented as a mathematical model called graphs. A directed graph G = (V, E) consists of a finite set vertices (V) and edges (E) with certain directions and for any vertex $x \in G(V)$, the set of vertices adjacent to x is called the neighbourhood N(x) of x. A *walk* is defined as a sequence of edges such that the end vertex of an edge e_i is the start vertex of an edge e_{i+1}. The walk is called a *path* if all edges and vertices are not repeated more than once. A *shortest path* is the path between two vertices with minimal length, also called a *geodesic*. The authors then define *isomorphism* as a one-to-one correspondence between the vertices of two graphs and if the direction of the edges also directionally correspond to the edges in the other graph. *Motifs* are defined as recurrent and statistically significant sub-graphs of a given graph [21, 14]. Here throughout the first section, the authors assume the readers to have no background in graph theory, which also helps to increase the clarity and structure the flow of the mathematical argument throughout the paper. All the notations used are pertinent to the comprehension of the paper.

2.1 Centrality measures

The authors then formally define centrality to be a function C(x) such that a numerical value can be assigned to any vertex x. For any two vertices x and y, x is more important than y if and only if C(x) > C(y). The different centralities are thus defined as follows according to Koschützki and Schreiber.

2.1.1 Degree centrality

Degree centrality can be defined as the number of edges connected to the vertex. As the graph is directed there exist two sub-centrality measures called indegree and outdegree. The authors make references to indegree and outdegree centrality measures later in the paper, without clearly defining them. This incoherence in the argument generates gaps in the readers flow of reading and should be added to a revised edition of the paper. For general reference, *indegree* is the number of edges directed to a vertex and *outdegree* is the number of edges that the vertex directs to other vertices [27]. The authors add that this is a local centrality measure and only the neighbourhood of the vertex of interest is taken into account. The authors then refer to the work of Freeman [7] in their paper, but as Freeman used degree centrality in social networks and not biological systems, which is the focus of this paper, it is unclear why his work has been referenced. The paper would benefit by deleting this particular reference, as it does not contribute to the flow of the argument.
Referring to the literature, the authors note that Jeong et al. [12] showed that the degree of a protein in a network correlates to its importance in the life of an organism. A much clearer connection between centrality values through degree centrality is shown by Hahn and Kern [8] as the mean centrality value for essential proteins is significantly higher than for non-essential proteins. However it is unclear why the authors have added these reviews in both the introduction to graph theory part of the paper and the main introduction, resulting in unnecessary repetition of facts.

2.1.2 Closeness centrality

Closeness centrality relates to the sums of the distances of the shortest paths in the network. The authors formally define the centrality as the reciprocal of the sum $\Sigma_{v \in V} dist(x, y)$ for x, y

$\in V$ for a graph G = (V, E). The authors then refer to the literature: Ma and Zeng [18] showed that 8 out of 10 predicted metabolites (organic and inorganic chemicals which are the reactants or products of biochemical reactions [6]) by the closeness centrality measure in *E.coli* metabolic networks, were part of an important network system in the organism. The authors thus give a succinct and brief definition of the definition, outlining their assumptions clearly, and their views are supported by the appropriate literature.

2.1.3 Shortest path betweenness centrality

By convention let $\sigma(x, y)$ be the shortest paths between vertices x and y. Let $\sigma_{\mathrm{x,y}}(v)$ be the number of shortest paths through vertex v other than x, y. The authors fail to outline their assumptions for the definition of this centrality measure. This leads to a lack of rigour in the mathematical argument for the definition. For completeness, we outline the assumptions: if $x = y$, $\sigma(x, y) = 1$; if $v \in x$, y then $\sigma_{\mathrm{x,y}}(v) = 0$. Another assumption is that the quality of connections is divided among all geodesics for each pair of vertices. By convention $0/0 = 0$ for this definition [3]. Thus, the centrality is the sum as follows:

$$\sum_{x,y \in V} \frac{\sigma_{\mathrm{x,y}}(v)}{\sigma(x, y)}$$

The measure identifies the ability of the vertex v to interpret communication between other vertices x, y.

2.1.4 PageRank

The authors deviate from the normal trend of defining and explaining the centrality measures; they instead refer the reader to "the literature for details". It shows an unsystematic approach to laying out the problem, and the definition being too "lengthy" is given as the reason for this omission. A later edition should include the following concise definition of PageRank™[4, 22]. Assuming the web to be a directed graph G = (V, E) where the V ($v_{\mathrm{i,j}} \in V$) is the pages on the web and E is the directed links between the pages. Let the rank of any page v_{i} be $R(v_{\mathrm{i}})$. Let N be the total number of pages, $M(v_{\mathrm{i}})$ be the set of pages linking to v_{i}, $L(v_{\mathrm{j}})$ be the number of links from v_{j} and the damping factor (the probability that a user keeps selecting a random page) is set to 0.31 after an analysis of biological data [10]. So the rank of v_{i} is

$$R(v_{\mathrm{i}}) = \frac{0.69}{N} + 0.31 \sum_{v_{\mathrm{j}} \in M(v_{\mathrm{i}})} \frac{R(v_{\mathrm{j}})}{L(v_{\mathrm{j}})}$$

2.1.5 Motif-based centrality

To understand motif-based centralities we simplify the definition for motifs, for general comprehension. The definition given in the text though mathematically correct, is terse and incomprehensible for the target audience of this paper.

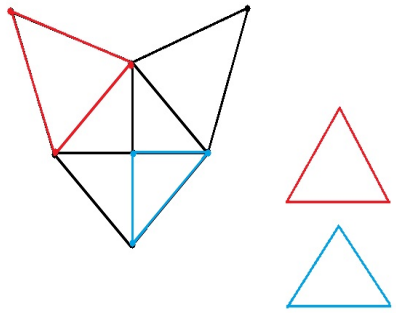

Figure 1: Motifs: Sub-graphs (here, with three vertices) with multiple occurrences in a larger network.

In a connected graph like Figure 1, the sub-graphs shown in red and blue are two of the many instances they can be observed in the larger network. The triangular sub-graphs with multiple occurrences in the network are *motifs* of the network. The authors then mention a *feed-forward loop* and use the acronym FFL without having defined it before the first use; they seem to have mentioned it in the next paragraph as an afterthought, though it is not well-defined.

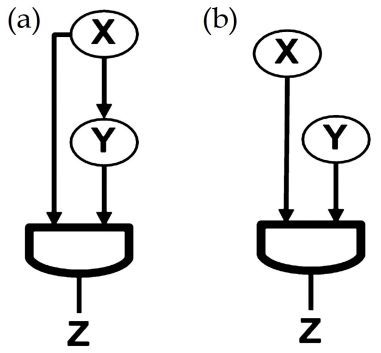

Figure 2: (a) FFL loop to synthesise Z compared to (b) a simple synthesis of Z [19].

This is indicative of a discontinuous structure of argument and is difficult to follow as the reader then has to search for where the acronym has been defined in the text. For completeness, we briefly introduce the term. A large gene regulation network can be modelled as a directed graph. That means for the vertices corresponding to the genes (g_1, g_2....g_n), the edge directed from g_i to g_j signifies a change in rate of conversion of g_j into a protein or other molecules [13]. Amongst all the sub-graphs with 3 vertices, the most frequently occurring graph is the FFL.
For the diagram above (Figure 2) in (a), gene X and gene Y both determine the production of molecule Z; this is an FFL. In (b), gene X and gene Y determine Z separately. This is a simple regulation of Z. The authors go on to describe other motif-chains such as a single input motif (SIM) which is a set of vertices being exclusively controlled by a single vertex [28].

The authors refer to literature saying that when the paper was written (in 2008), FFL motifs had been studied functionally but had not been used to rank genes. This is still valid today as the completion of motif catalogues still remains a priority for bridging the gap between the 'vertices' in GRNs and their 'edges' in 2018 [11]. Another functionality study for motifs in 2017 included relating DNA mutations to specific inherited diseases where only 1 in 5 of the genes under consideration showed certain rare motif combinations [24]. The development of DeepFold™in 2018, a neural network database to obtain motif features of proteins [17], while a comparative study, still compares the functions of gene network motifs in proteins. The authors themselves back this up saying that Wang and Purisima [30] discovered that the regulators with short half lives are part of the motifs in GRNs especially in SIMs and FFLs. This absence of literature for ranking the importance of genes highlights a need to introduce definitive ranking systems in studies in systems biology.

3 Analysis of GRNs with centralities

The authors analyse the centralities within the GRN of *E.coli*. The authors use the data by Salgado [25] and claim to use version 5.5 but from review of the source, there does not seem to be any version 5.5 as cited in the paper. This may be a typographical error, but it is still unclear which version of the data i.e. 5.0 or 6.0 is used in the paper, without further analysis. This is a serious misinterpretation of the results and prevents any repetition of results. The resulting GRN consists of 1250 vertices and 2515 edges. The authors define global regulators as genes at a high hierarchical network within the GRN, and suggest they may be important as they may be able to influence genes over a larger range. The authors propose a characterisation of 18 such global regulators amongst the top 25 (2% of all genes) genes according to the centrality measures defined in the first part of the paper (See Appendix A: Figure 3).
The authors make general observations from the table; they note that the centrality measures in the table can identify nearly 50% of the global regulators within the top 2% of genes. They add that for most of the centrality measures, the top 5 positions are occupied by global regulators, but that specific global regulators may occupy different positions on the list, for different centrality measures. We note that the gene *arcA* while occupying 4^{th} position for PageRank, occupies 18^{th} place for shortest path betweenness centrality. To make sure that the centralities do not coincide, the authors use correlation coefficients. This quantifies the degree to which a variable can predict the change of another variable, in this case, different central-

ity measures. The authors note that for the centralities with correlation coefficients above 0.9, close to 88% of the vertices in the original data have an out-degree of zero i.e. no edges directed out of the vertices. The authors also observe that the correlation coefficients for these vertices are assigned to zero in the PageRank and motif based centralities.
The authors conclude by saying that the centralities applied to the genes in the data rank each of them differently but the motif-based centrality rank most of the global regulators (15 out of 18 i.e. 83%) within the top 2% of the genes, and outperforms all other centrality measures.

4 Analysis of the discussion

The authors consolidate their position by stating that there have been different methods to analyse large sets of data, but that using centrality measures, as they have been in other fields, could be a step forward. Indeed, they show a clear connection between using motif-based centralities to identify 15 out of 18 'interesting' genes (global regulators) and they state that a biological understanding is required to analyse the results of such experiments effectively. To this end, for the audience of this paper, we aim to clearly define and explain biological concepts so as not to impede the flow of argument in the critiqued paper, but explaining the names and functions of the genes given are perceived to be beyond the scope of this article. The authors finish by mentioning that the rankings of the genes are different when different measures are taken into account. They however do not include any potential improvements to their study nor do they suggest any new methods of generalising the ranking systems. We propose that different centrality measures should be combined, and their correlation coefficients should be plotted against existing data to see if there is a difference in observations. This has already been done in energy reduction studies [16] and node importance in complex networks [31] but is yet to be seen in systems biology.

References

[1] Albrecht, M., Huthmacher, C., Tosatto, S. C. E., and Lengauer, T. (2005). Decomposing protein networks into domain-domain interactions. *Bioinformatics*, 21:220−221.

[2] Boone, C., Bussey, H., and Andrews, B. J. (2007). Exploring genetic interactions and networks with yeast. *Nature Reviews Genetics*, 8(6):437−449.

[3] Brandes, U. (2008). On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136−145.

[4] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107−117.

[5] Davidson, E. and Levin, M. (2005). Gene regulatory networks. *Proceedings of the National Academy of Sciences*, 102(14):4935−4935.

[6] Fanos, V., Antonucci, R., Barberini, L., and Atzori, L. (2012). Urinary metabolomics in newborns and infants. *Advances in clinical chemistry*, 58:194.

[7] Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, 1(3):215−239.

[8] Hahn, M. W. and Kern, A. D. (2004). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular biology and evolution*, 22(4):803−806.

[9] Hartman, J. L. (2001). Principles for the buffering of genetic variation. *Science*, 291(5506):1001−1004.

[10] Hobson, E. A., Mønster, D., and DeDeo, S. (2018). Strategic heuristics underlie animal dominance hierarchies and provide evidence of group-level social knowledge. *arXiv preprint arXiv:1810.07215*.

[11] Inukai, S., Kock, K. H., and Bulyk, M. L. (2017). Transcription factor–DNA binding: beyond binding site motifs. *Current Opinion in Genetics & Development*, 43:110–119.

[12] Jeong, H., Mason, S. P., Barabási, A.-L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41.

[13] Jin, G. (2013). Feed forward loop. In *Encyclopedia of Systems Biology*, pages 737–738. Springer New York.

[14] Juszczyszyn, K. (2014). *Motif Analysis*, pages 983−989. Springer New York, New York, NY.

[15] Koschützki, D. and Schreiber, F. (2008). Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regulation and Systems Biology*, 2.

[16] Lindmark, G. and Altafini, C. (2019). Combining centrality measures for control energy reduction in network controllability problems. In *2019 18th European Control Conference (ECC)*, pages 1518–1523. IEEE.

[17] Liu, Y., Ye, Q., Wang, L., and Peng, J. (2018). Learning structural motif representations for efficient protein structure search. *Bioinformatics*, 34(17):i773–i780.

[18] Ma, H.-W. and Zeng, A.-P. (2003). The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, 19(11):1423−1430.

[19] Mangan, S. and Alon, U. (2003). Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences*, 100(21):11980–11985.

[20] Martınez-Antonio, A. and Collado-Vides, J. (2003). Identifying global regulators in transcriptional regulatory networks in bacteria. *Current opinion in microbiology*, 6(5):482–489.

[21] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824−827.

[22] Phuoc, N. Q., Kim, S.-R., Lee, H.-K., and Kim, H. (2009). Pagerank vs. katz status index, a theoretical approach. In *2009 Fourth International Conference on Computer Sciences and Convergence Information Technology*, pages 1276−1279. IEEE.

[23] Prehoda, K. E. and Lim, W. A. (2002). How signaling proteins integrate multiple inputs: a comparison of n-WASP and cdk2. *Current Opinion in Cell Biology*, 14(2):149−154.

[24] Růžička, M., Kulhánek, P., Radová, L., Čechová, A., Špačková, N., Fajkusová, L., and Réblová, K. (2017). DNA mutation motifs in the genes associated with inherited diseases. *PLOS ONE*, 12(8):e0182377.

[25] Salgado, H., Gama-Castro, S., Peralta-Gil, M., Diaz-Peredo, E., Sánchez-Solano, F., Santos-Zavaleta, A., Martinez-Flores, I., Jiménez-Jacinto, V., Bonavides-Martinez, C., Segura-Salazar, J., et al. (2006). Regulondb (version 5.0): Escherichia coli k-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic acids research*, 34(suppl_1):D394−D397.

[26] Sharan, R. and Ideker, T. (2006). Modeling cellular machinery through biological network comparison. *Nature Biotechnology*, 24(4):427−433.

[27] Sharma, D. and Surolia, A. (2013). Degree centrality. *Encyclopedia of Systems Biology*, pages 558−558.

[28] Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics*, 31(1):64.

[29] Wang, E. (2010). *Cancer Systems Biology*. CRC Press.

[30] Wang, E. and Purisima, E. (2005). Network motifs are enriched with transcription factors whose transcripts have short half-lives. *Trends in Genetics*, 21(9):492–495.

[31] Zhang, Y., Bao, Y., Zhao, S., Chen, J., and Tang, J. (2015). Identifying node importance by combining betweenness centrality and katz centrality. In *2015 International Conference on Cloud Computing and Big Data (CCBD)*, pages 354–357. IEEE.

5 Appendix A

position	odeg	parR	katR	spb	chains	fflA	fflSum
1	**crp**	**crp**	**crp**	**hns**	**crp**	**crp**	**crp**
2	**fnr**	**ihfAB**	**fnr**	*gadX*	**ihfAB**	**fnr**	**fnr**
3	**ihfAB**	**fnr**	**arcA**	*flhD*	**arcA**	**ihfAB**	**arcA**
4	**fis**	**arcA**	**ihfAB**	**fur**	**fnr**	**arcA**	**fis**
5	**arcA**	**phoB**	**fis**	*gadE*	**fis**	**fis**	**narL**
6	**narL**	*lexA*	**hns**	**fis**	*evgA*	*modE*	**ihfAB**
7	**hns**	**cpxR**	*gadE*	**lrp**	*ydeO*	**soxS**	**hns**
8	**fur**	**soxR**	*gadX*	*rcsAB*	*gadE*	**hns**	**fur**
9	**lrp**	**fis**	**cspA**	**soxS**	**soxR**	**cpxR**	*gadX*
10	*glnG*	*evgA*	*evgA*	**fnr**	**soxS**	*fhlA*	*hyfR*
11	*narP*	*cysB*	*ydeO*	**cspA**	*torR*	*gadE*	*marA*
12	**cpxR**	*argR*	*torR*	*caiF*	*gadW*	**rob**	*flhD*
13	**phoB**	*phoP*	*gadW*	**purR**	*cspE*	*gadX*	*nagC*
14	*fruR*	**fur**	*cspE*	**narL**	**cspA**	*galR*	**soxS**
15	*modE*	*allR*	**soxS**	*marA*	*gadX*	**fur**	*modE*
16	*fhlA*	*glnG*	**soxR**	*metJ*	**hns**	*gntR*	*tdcA*
17	*lexA*	*sdaR*	**rob**	*malT*	*oxyR*	*oxyR*	*yiaJ*
18	*flhD*	*trpR*	*marA*	**arcA**	**fur**	*tdcR*	*gutM*
19	*gadE*	*agaR*	*marR*	*glnG*	*modE*	*gutM*	**ompR**
20	**purR**	*gadE*	*oxyR*	**ompR**	**narL**	*nagC*	*srlR*
21	**soxS**	**soxS**	**fur**	*Nac*	**lrp**	**narL**	*galS*
22	*argR*	**hns**	*modE*	*oxyR*	*glnG*	**ompR**	*idnR*
23	*cysB*	**lrp**	*gutM*	*hupAB*	**ompR**	*srlR*	*caiF*
24	*marA*	*tyrR*	*srlR*	*argP*	**phoB**	*argP*	*chbR*
25	*nagC*	*torR*	**narL**	*dnaA*	**cpxR**	*cysB*	**cpxR**
#global regs.	13	12	12	11	15	12	11

Figure 3: Names of the top 25 genes according to some centrality measures to find the highest number of global regulators in the top 2% of genes. Global regulators according to Martinez-Antonio et al. [20] are highlighted in bold. The last row gives the total global regulators for the particular column. **Abbreviations** odeg: outdegree, parR: PageRank, spb: shortest path betweenness, chains: motif-based centrality, fflA: FFL motif for certain important vertex A, fflSum: FFL motifs in general