 doi:10.20944/preprints202201.0170.v1

Meaningful Infrastructures in Biological Networks: Related Studies, Algorithms and Evaluation Parameters

Lionel Ngobesing Alangeh[1] and Yılmaz Atay[1,*]

[1] Gazi University, Faculty of Engineering, Computer Engineering Dept., 06570, Maltepe/Ankara, Turkey
[*] Correspondence: author: *yilmazatay@gazi.edu.tr*

Abstract: In network science and big data, the concept of finding meaningful infrastructures in networks has emerged as a method of finding groups of entities with similar properties within very complex systems. The whole concept is generally based on finding subnetworks which have more properties (links) amongst nodes belonging to the same cluster than nodes in other groups (A concept presented by Girvan and Newman, 2002). Today meaningful infrastructure identification is applied in all types of networks from computer networks, to social networks to biological networks. In this article we will look at how meaningful infrastructure identification is applied in biological networks. This concept is important in biological networks as it helps scientist discover patterns in proteins or drugs which helps in solving many medical mysteries. This article will encompass the different algorithms that are used for meaningful infrastructure identification in biological networks. These include Genetic Algorithm, Differential Evolution, Water Cycle Algorithm (WCA), Walktrap Algorithm, Connect Intensity Iteration Algorithm (CIIA), Firefly algorithms and Overlapping Multiple Label Propagation Algorithm. These algorithms are compared with using performance measurement parameters such as the Modularity, Normalized Mutual Information, Functional Enrichment, Recall and Precision, Redundancy, Purity and Surprise, which we will also discuss here.

Keywords: significative infrastructures, biological networks, normalized mutual information, recall, precision, modularity, gene ontology

1. Introduction

Today, the study of complex networks is applied in almost every area, from social to mathematical to medical fields. As the concept of complex networks grows, the idea of detecting communities within these networks also grows with it. This is because both concepts work side-by-side with each other. These complex network systems are often modelled using graph structures whereby vertices represent main nodes or points in the network and the edges represent the interaction or relationship that exists between these points. For example, in a social network such as a book club, the vertices on the graph will represent the people who are enrolled into the book club and the edges on the graph represent the relationship between these people. Let's take another example. Considering a biological network - for example a Protein-Protein Interaction Network (PPI) - the vertices are represented by proteins while the interactions between these proteins are shown on the edges.

Mathematical methods, algorithms and models have found great importance in the medical field, in biomedical research due to the increase in number of omics projects as well as the focus in collecting, analyzing and managing large volumes of biological data as well as medical data. Not to mention, network science models and algorithms have proven to be of great influence in the investigation of modularity which is considered as one of the key organizational aspects in biological systems.

Previously, biological networks were often considered with a single interaction type. That is, the network could be either protein-protein or drug-drug or DNA-DNA. But recently, these biological networks are becoming more complicated and involve interaction at multiple layers (for example, protein-drug or drug-DNA). For example, Laura Bennett et al. (2015) [5] clustered multiplex biological networks using a mathematical programming model. Also, Calderer G et al. (2021) [4] carried out their research to detect communities in large biological networks of bipartite nature. This is because each type of interaction signifies a different issue in cellular activity, so modules which correspond to cellular function can be better represented by many sources [1]. Due to this, detecting communities in these complex networks has become a center of interest for most researchers. Meaningful Infrastructure Identification is specifically helpful in improving the understanding of biological processes which consist of understudied regulatory molecules whereby the specific functions are often unknown. This concept of Meaningful Infrastructure Identification was initiated by Girvan and Newman in 2002 [2] where they defined Meaningful Infrastructure Identification as a method to identify communities, network, modules or pathways. These are internally dense connected sets of nodes.

Directed networks are networks with directed interactions. In directed biological networks, there exist different complexity measures widely used in network theory such as: Connectedness (this is a very simple complexity measure defined as the ration of edge count in the biological network to the edge count in the complete graph that consist of exactly the same count for internal vertices), the total subnetwork count, overall connectivity (represented by the overall topological Index, OI, and can be defined as the sum of all subnetworks of a given network-invariant, I), total walk count (it is directly proportional to the complexity of the network structure and improves with the degree of connectiveness of the bio-logical network) and vertex accessibility. These are some of more important measures considered when discussing directed biological networks. The most used and easy method to identify meaningful infrastructures in directed networks is to not give any importance to edge direction and apply to the directed network algorithms which are designed for undirected networks. This is an approach which is not always successful though, and also in the situation where it is successful it is worthy to note that ignoring the directions of the edges causes loss of a great deal of information which is necessary for understanding the network better. E. A. Leicht et al [39] in their research present an approach for identifying meaningful infrastructures in directed networks, by explicitly making use of the data found in the directed network. The proposed method is an improvement of the modularity maximization method which is commonly used for detecting communities in undirected networks, and they obtained efficient results similar to that of the related algorithm applied for undirected networks

Biological networks, like the signaling networks, PPI networks, co-expression networks, Gene-gene networks, signaling networks as well as metabolic biological networks, represent biological systems mathematically. PPI networks, such as that gotten from InWeb [35] where the aggregation of interactions are done from primary databases, are biological networks whereby the nodes on the network are represented by proteins, their interactions or connections are the edges and the weights of the edges represent the interaction confidence of each connection. For a signaling network such as that presented in [36], nodes represent the genes as well as the edges between them which are directed, represent the interaction between genes, that is accountable for a particular cell function. Weights of the edges here also represent the confidence value of the interaction. It is worth noting that the genes from these biological networks can also map to proteins from other networks to form a multi-layered biological network. Homology biological networks such as that presented by [37] is developed by connecting those genes which are in evolutionary relation to each other.

Neural Networks prove to be very efficient amongst other machine learning techniques. When meta-heuristic algorithms are used in combination with Artificial Neural Networks the efficiency and accuracy of prediction and classification is significantly improved. D. Devikanniga et al. (2019) [30] in their research presented some of the meta-heuristic algorithms which are improved using neural networks such as Bacterial foraging optimization algorithm-based ANN, Ant colony optimization-based ANN, Artificial bee colony optimization-based ANN, Bat algorithm-based ANN, Genetic algorithm-based ANN, Particle swarm optimization-based ANN, Social spider optimization-based ANN, the Levy flight-based ANN using Cuckoo search and many more. In these applications, the different strategies are modelled to improve the meta-heuristic algorithms [29]. As an example, the co-evolving Master-Slave and cooperative Ring Master-Slave methods improving balance between making maximum use of the Bat Algorithm

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202201.0170.v1

which is then implemented for optimizing the Artificial Neural Network structure with respect to number of nodes in layers that are not visible or hidden, number of non-visible layers, selection and weights of non-visible layers. As another example, the Genetic Algorithm is used to train an ANN with the best fitness value, and best chromosomes as well as good efficiency of the algorithm is proven by putting its result in comparison with those trained by MLP using the BPN algorithm.

As shown in figure 1, there are different kinds of biological networks from the simple empirical dolphin network, to the complex food web networks. As shown on the figure, nodes on the network with same color share similar properties with each other and can be considered to belong to the same community or as in the case of biological networks, considered to belong to the same module.

2. Literature Review

In this section, a literature review of the different works that have been done related to Meaningful Infrastructure Identification or module detection in biological networks will be presented. These works include recent works done in detecting communities in complex networks, bipartite networks, as well as the works involving the use of different types of meta-heuristic algorithms for this purpose. This literature review also includes works involving metrics based on modularity and connectivity. The related works are as follows.

Marwa Ben M'Barek et al. in their research [3], Genetic Algorithm for identifying community structures in Biological Networks, focused precisely on gene interaction networks. Their aim was to identify communities (corresponding to a set of genes or proteins) from sources of gene annotation - Gene Ontology (GO). They introduce a spectral solution crossover and mutation operator and propose a fitness score statistic dependent on a calculation of similarity and interaction value between the genes. The method used for the similarity was called the GS2 method. The results from their experiments on real data were compared with those extracted from the KEGG database corresponded to some sections for real networks that are found in other biological pathway databases. The obtained results were considered satisfactory by biological experts and prove the ability of the Genetic Algorithm to correctly identify meaningful infrastructures in biological networks. Their future works aimed at implementing multi-objective optimization to obtain better results.

Calderer G et al. (2021) [4] in their research Community Detection in Large-scale bipartite biological network structures, present a review in theory of the different approaches that are useful in identify community structures in biological networks that are bipartite. This research was based on bipartite networks due to the fact that many biological networks are bipartite in nature (such as drug gene, gene-disease, protein-gene and so on). In this work they discuss the different scores and evaluation criteria which are used for verifying these identified community structures quality. Further, they implement a couple of approaches to a drug-gene interaction (DGI) network from the Drug Gene Interaction Database (DGIdb). They then developed an bipartite network that is unweighted from the provided data to represent the gene-drug interactions. Due to the fact that every approach needs the subnetwork to be connected, they maintained the component with the larges connection (represented by up to 99% of the network with respect to nodes). As a result they discovered a subnetwork of up to 22,693 interactions. These interactions were gotten from 2,336 genes and 6,049 drugs. This was done to to show the importance in terms of strengths as well as mention the limitations of the given approaches in their implementation on much larger, bipartite biological networks. They calculated the pairwise Normalized Mutual Information (NMI) score of the community structures and they found out that these scores were similar within the range 0.6077 to 0.7746 showing that the communities share quite a large amount of similar information.

Laura Bennett et al. (2015) [5] in their work proposed a mathematical programming model for creating clusters in multiplex biological networks. The proposal had to achieve this by clustering multiple subnetworks each having an interaction type that is different and used it to discover a single representative partition of the composite community. The proposed approach, called SimMod is applied to yeast networks of genetic, co-expression and interactions for evaluation. The results obtained by SimMod are compared with two other methods: PanGIA and GenLouvain. SimMod is first evaluated and compared with the PanGIA and genLouvain, then classifying the 'combined'

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202201.0170.v1

networks and finally by clustering separate network portions. SimMod method, discovers subnetworks that averagely have a greater functional enrichment as compared to the other two-network methods mentioned. With the PanGIA finding high confidence subnetworks as a result of learning from known protein complexes, the SimMod as well as the genLouvain identify meaninful infrastructures that are more functionally cohesive when taking into consideration only the interaction density and network structure. In terms of the modelling approach used, SimMod and genLouvain are more alike, since both methods optimize modularity metric variations. SimMod does this by getting the average of the modularity across the different network pieces, while genLouvain uses a variant of modularity whereby the null model uses internetwork connections. They further illustrate that while in some situations, putting interactions of similar types together can improve the community structures functional contents. In other situations, noise can be used to blend in functional information. Hence, a proper rationale is required to employ the original datasets in a meaningful way, with respect to the problem to be solved at hand. Notwithstanding, when data specific to the given experiment is used with the proper rationale, SimMod shows great ability in identifying a good number of composite subnetworks that are funtionally enriched and which can result in the development of biological hypotheses that relates to specific experiments of clustering.

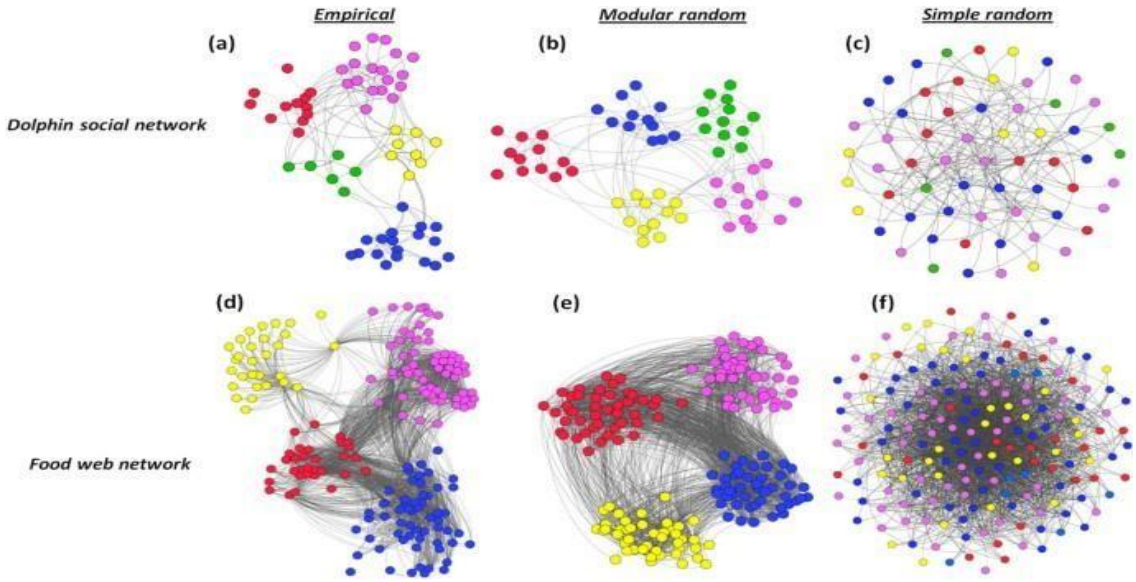

Figure 1. Simple view of complex biological networks [41]. (a) empirical network, (b) modular random network, (c) simple random network. (a), (b) and (c) are all of the Dolphin Social Network, while (d), (e) and (f) are of the Food Web Network.

Guanbo Jia et al. in 2012 [6] in their research "Community Detection in Social and Biological Networks using Differential Evolution", proposed a new algorithm called Differential Evolution based Community Detection (DECD) which makes use of the popular Differential Evolution algorithm for identifying meaningful infrastructures or communities in complex networks. Based on the standard Differential Evolution operator used for crossover, a modified binomial crossover is designed by the authors to effectively transmit necessary information concerning the community structure involved in evolution. Their proposed algorithm, unlike many other Meaningful Infrastructure Identification algorithms, doesn't need knowledge of any prior information concerning the meaningful biological community structure, and this is important for its implementation to complex real-world networks in which prior knowledge is not often available. The proposed DE based Community Detection algorithm was evaluated both with several artificial and real-world biological networks such as the American college football network, the Zachary's karate club network and the Yeast Protein-protein interaction networks. The experimental results showed that their DE based Community Detection algorithm was effective for Meaningful Infrastructure Identification in multiplex biological networks having networks with very vague community structures.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202201.0170.v1

In trying to understand complex biological systems, there is a great focus on subnetworks in which the biological entities are strongly connected. Rui Henriques and Sara C. Madeira in their research (2016) [7] proposed Bi-clustering Network (BicNET), which is a bi-clustering algorithm which could be used to identify non-trivial modules which are coherent in weighted biological networks and have good efficiency. In their work, they apply three main concepts. Firstly, they mention the importance of identifying network structures represented by symmetric, plaid, constant and order preserving bi-clustering models. Furthermore, they bring forward a method to identify these structures and to robustly handle missing as well as noisy interactions. Furthermore, they propose a new search criterion to solve bottlenecks of time and memory by performing effective exploration on the inherent sparsity in the structure of the available biological network data. The BicNet was implemented on both PPI networks and Gene-Gene interaction networks from E.coli, yeast and humans showed the efficiency of the algorithm. This method proved to be an efficient method allowing for accurately analyzing large-scale network data using an unsupervised approach for the identification of coherent network structures with homogeneity in parameters.

Maria Serban in her research [8], Exploring modularity in biological networks (2020), made the case that it is very important to mention the role that network models play in research studies, since it is also possible to systematically analyze exploratory functions of these models in bio-scientific research. Through the use of molecular as well as developmental biological examples, her research argues the fact that modeling with the same method can lead to exploratory functions including bringing in new perspectives for research, providing an alternative set of algorithms, approaches and concept for individualizing important properties of natural events, developing proof of principle explanations and demonstrations.

Kahn Rhrissorrakrai and Kristin C Gunsalus in their research, MINE: Module Identification in Networks, (2011) [9], address the problem of identifying modules in biological networks and implemented their approach on dense networks of interaction data obtained experimentally by developing a clustering algorithm which is agglomerative and has the ability to detect sets having high modularity of gene products found in molecular interaction networks that are highly interconnected. Their algorithms permit a high flexibility degree as well as user customization of results having less parameters that are adjustable. The MINE algorithm was compared with algorithms such as MCODE, CFinder, NEMO, SPICi, and MCL. It showed to outperform these algorithms in detecting clusters of higher modularity and non-exclusivity when they are applied to the C. elegans PPI networks. In general, the algorithm achieves modularity and high-level geometric accuracy for annotated functional categories.

Weiqi Chen et al. [10] in their work proposed a multi-task module identification for biological networks called MUMI. Their work was based on the fact that it is necessary to pick out overlapping found in network structures and active modules in biological networks with high significance of the topological units that have difference in dynamics that are quite significant. Experiments showed that the MUMI algorithm brings forth new discoveries and explanations to biological methods by using a combination of information gotten from active community structures simultaneously. By developing the problem as a problem of multitask learning that looks to find these two types of community structures concurrently, the model can make use of their latent complementary to achieve good search performance with respect to convergence and accuracy. Their experiments were done on several benchmark social networks and also 2 real-world biological networks such as Yeast galactose utilization pathway (YGUP).

A framework used for evaluation of clustering methods in the context of biomedicals was established by the Disease Module Identification DREAM challenge, by performing test on the association of community structures with GWAS-derived popular disease genes and trait. Gilles Didier et al. [11] in their research (2018) implemented different extensions of the MoITi software which is used to identify meaningful infrastructures in biological networks by optimizing complex network modularity. They implemented the Louvain algorithm by use of a randomized process which can take into consideration layer weights and edges, and can perform recursive clustering. These weights are arbitrarily chosen by use of the relative performance of the separate main network structures on predicting the disease subnetwork structures. This randomized procedure improves the identification of meaningful infrastructures and the results are completely dependent on the dataset of GWAS, that is chosen plus an enrichment p-value threshold.

Biological subnetworks or community structures that are isolated chemically or spatially and perform discrete functions are important as well as fundamental cellular organizational building blocks. That notwithstanding, the detection of these building blocks in highly integrated biological networks requires a good understanding and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202201.0170.v1

interpretation of the organization of these networks. In a study by Erzsebet Ravasz (2009) [12], she presents the argument that a good number biological network infrastructures are organized into many small and highly connected community structures that through a hierarchical method, combine into larger less cohesive units. She supports this argument by a visual inspection of the obtained hierarchical tree and this leads to a breakdown of metabolism (that is natural) into large community structures, that are then partitioned into more integrated and smaller substructures. As a case study, she referred to the Escherichia coli metabolic network, identifying the hierarchical modules within it and showing that the hierarchical modularity which was discovered, closely overlaps with known metabolic functions. This method is expected to be of great use in automatically discovering functionally relevant and meaningful infrastructures in many types of biological networks.

Koyel Mitra et al. [13] in their paper (2013) classify integrative approaches used for generating complex and large networks for molecular interactions for human and model organisms into four wide categories, describing their principles and reviewing their applications. These four categories are: (i) 'Active Modules' identification, which has been a very successful integrative approach. This approach consists of significant-area-search methods, network-propagation and diffusion-flow methods as well as clustering-based methods. (ii) Identification of 'Conserved modules' which help to address important questions about biological regulation and at the same time performing a prediction attempt to evolutionary principles shaping network architectures. It involves conserved interactions, pairwise network alignments, parallel alignment of multiple networks and network alignment incorporating evolutionary dynamics. (iii) Discovery of 'Composite functional' modules. This is concerned with the rationale for composite modules. (iv) Differential Network Modules, where a good number of 'differential' network analyses models have made use an experimental method in which biological infrastructures and networks substructures are compared and measured across different conditions to discover the interactions and meaningful infrastructures or subnetworks that are differentially present, absent or modified. With improvement and development in integrative pipelines for bioinformatics and expansion in capabilities for data handling and processing, a numerous combination of potential data types, species, conditions, cell states and time points should be adaptable to joint in-depth network analysis.

Modularity optimization is an important approach used for detecting community structures in complex networks. Atay et al. [14] in their research proposed some metaheuristic optimization algorithms that are based on optimizing modularity. The discussed algorithms used in this research include: Gravitational Search Algorithm (GSA), a modified Big Bang–Big Crunch algorithm (BB-BC), improved Bat Algorithm based on the Differential Evolutionary algorithm (BADE), effective Hyper-heuristic Differential Search Algorithm (HDSA), Bat Algorithm (BA) and Scatter Search algorithm based on the Genetic Algorithm (SSGA). Experiments were carried out using these algorithms on 9 different real-world complex network infrastructures and compared according to their statistical significance. The biological networks used here were the: E-coli TRN, C-Elegance MRN, Cattle PPI, Helico PPI networks. The algorithms were then compared by taking into consideration the processing time. Results showed that, BADE, SSGA, and BBBC algorithms had process times which are longer than the HDSA. Furthermore, the processing time of the HDSA was more than those of the GSA and BA and also these algorithms achieved lower success in results as compared to the HDSA.

The Disease Module Identification (DMI) DREAM challenge was an initiative created in attempt to systematically assess methods used for identification of modules on a panel of six diverse genomic networks. Raghvendra Mall et al. [15] in their research propose a method for genetic refinement that is based on concepts of combining and dividing the hierarchical tree from any Meaningful Infrastructure Identification approach for constrained Disease Module Identification in biological networks. They use the F-score evaluation metric which is calculated from several metrics or unsupervised quality such as connectivity, modularity and conductance to establish the graph partitioning quality at a given level of hierarchy. In this research they make use of a quality measure called the Inverse Confidence. This quality measure ranks and prune insignificant network structures, retrieving a candidate listing of the disease modules (DM) for complex networks. Here, the network structures which were predicted were assessed based on the total number of candidate modules that are unique and are found in correlation with complex diseases as well as traits derived from more than 200 genome-wide association study (GWAS) datasets. From experimental analysis, their proposed method detected a total of 31 Rank 9 and 16 Rank 10 DMI's and also 44 Disease Modules in the networks as compared to 60 Disease Modules detected by the DREAM challenge winner.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202201.0170.v1

Many algorithms involved in module identification use modularity as the objective function to detect communities in complex networks. Ronghua Shang et al. [16] in their research propose a Meaningful Infrastructure Identification approach that uses modularity and an improved genetic algorithm called MIGA. This algorithm uses the modularity Q as the objective function that makes the algorithm easier, making use of previously obtained information that improves the accuracy and stability of the algorithm and makes it more targeted for Meaningful Infrastructure Identification. The prior information refers to the number of discovered subnetworks. The experiments in this research were done on the following networks: Computer-generated Network, Dolphin social network, American college football, Zachary's karate club network and Books about US politics network. From the results it can be seen that the more complex the network the lesser the accuracy of MIGA. But the value of the Normalized Mutual Information obtained by the Genetic Algorithm, were not as good as those obtained by the MIGA at the end of the generations. This indicates that MIGA can identify more than half of the true sections using the Simulated Annealing Approach as well as prior information; however, without these, there is a great decrease in the Genetic Algorithm performance. MIGA uses simulated annealing as the method for performing local search that can significantly make the ability of local search better through parameter adjustment.

In a research by Sara Rahiminejad et al [17], six different Meaningful Infrastructure Identification algorithms which includes Conclude, Combo, Leading Eigen, Fast Greedy, Louvain and Spinglass used with two biological networks, where the first network is a PPI network in Saccharomyces cerevisiae (Yeast) having up to 6532 nodes and 229,696 edges while the second is a PPI network in Humans with 20,644 nodes and 241,008 edges, in order to find community structures and further analyse the obtained results with respect to functional and topological properties via the Kyoto Encyclopedia of Genes and Genomes pathway (KEGG) and Gene Ontology (GO) term enrichment analysis.

In a recent study by Mathew Matlock, Arghya Datta et al. (2018) [25], they examine a recently proposed architecture to known as wave used for forwarding information through an undirected graph in a wave of computation that is nonlinear to solve the problem of non-efficiency in information propagation over long ranges. They compare wave to graph convolution and discover that the wave learns 3 different task that are graph-based having much better efficiency. They are labelling paths connecting far-apart vertices in undirected graphs, labelling maze paths represented with images and computing circuit voltage potentials.

In a study by Xia W, Yu Q et al. (2019) [31] they aimed to identify hub genes associated with adrenocortical carcinoma progression through the use of WGCNA [38]. They used weighted gene co-expression networks as well as gene transcripts for Gene-Gene Interaction (GGI) networks downloaded from the UCSC Xena database which included the Adrenocortical Carcinoma (ACC) and other normal samples. In order to detect essential genes involved in ACC progression, they used over 2953 significant differentially expressed genes along with 9 other modules and were able to identify 4 hub genes including TOP2A, TTK, CHEK1, and CENPA.

Muhan Zhang et al. (2018) [24] in their study, Graph Neural Network based Link Prediction, discuss a heuristic paradigm for link prediction. Previously, link prediction heuristics have been based on score functions such as the Katz index which are simple and interpretable but have a limitation when it comes to their effectiveness on networks where the hypothesis whereby two nodes on the same network are likely to link, fails. The theory proposed in this research brings together a couple of heuristics to a single framework as well as proving the assertion that these heuristics can be approximated well from local subgraphs.

Yue-Hua Feng, Shao-Wu Zhang and Jian-Yu Shi (2020) [34] in their research, introduce a Deep Predictor for Drug-Drug Interaction known as DPDDI. This used Graph Convolutional Networks [27] to learn the feature representation in low-dimension for every drug in the Drug-Drug Interaction network structures and adapt a Deep Neural Network [28] approach to train models. The results from experiments indicate that this method performs better than 4 other state-of-the-art methods. The small molecular drugs which are approved as well as the relationships between their interactions were extracted from the DrugBank 4.0 and used to construct the first dataset called DB1. The DB1 dataset contains 180,576 annotated drug-drug interactions [26] and 1562 drugs. A smaller dataset called DB2 was used for comparison with state-of-the-art methods. Further the DrugBank 5.0 was used to create another dataset DB3 to make more effective results. The drug-binding proteins (DBPs) and Anatomical Therapeutic Chemical classification (ATC) codes were downloaded from DrugBank to the developed network-based features to other properties derived from

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202201.0170.v1

diverse drug properties. The predicted codes in DB1 were adopted by SPACE because out of the 1562 drugs found in DB1, 138 of them had no ATC code. SPACE was able to use chemical structures to derive the ATC code.

Morteza Chalabi Hajkarim et al. (2019) [42] aim at identifying differentially mutated subnetworks of a large gene-gene interaction network by defining a computational problem then showing that this problem is a NP-Hard problem. The proposed method identifies subnetworks infrastructures that have a statistically significant difference in their frequency of mutation in cases where aa reasonable generative model is used as a data source. The model, called DAMOKLE, is implemented by first defining a computational problem for detecting all connected subgraphs that could be found in a network and proving, the computational problem as NP-hard. The statistical significance of the results gotten from the proposed method is measured by using the False Discovery Rate (FDR) or Family-Wise Error Rate (FWER) as well as through permutation testing. The permutation testing involves two steps: The first step evaluates if the differential coverage observed is acquired with the mutations in genes occurring independently and further by taking into consideration the null distribution whereby every gene is mutated in a randomly selected subset. For the second test, an evaluation for the observed differential coverage of a meaningful subnetwork structure is performed with the condition that under the observed marginal distribution and under independence between mutation and sample membership, the coverage of these modules can be obtained. Results of the DAMOKLE are tested with both real cancer genome data and simulated data. The cancer data include: Lung Cancer Data, Colorectal vs Ovarian Cancer Data, Esophagus-stomach Cancer data and the Diffuse Gliomas.

Zongliang Yue (2020) [45] in his thesis, presents a new framework in network-based gene module construction and visualization with three advanced aspects. (i) integration of experimental data (transcriptome) and knowledge-based data (interactome and gene ontologies) in network construction, (ii) layout optimization considering multiple biological factors, and (iii) Exploration of gene signals and insights in the constructed Terrain Knowledge Mappings (TKMs). The thesis is done in 3 parts, first is the Data Curation step where they focus on data cleaning and database integration to provide a pre-knowledge basis in TKM construction. The second step is the Network-based module construction where they applied two novel algorithms, the “Weighted in-Path Edge Ranking for Biomolecular Association Networks (WIPER)” enables Protein-Protein Interactions prioritization by evaluating initially edge weights and global topological structure. The other algorithm “distance-bounded energy-field minimization algorithm (DEMA)” is for quantitatively optimizing network layout with biological properties such as gene weights, gene-gene relationships and gene associations. The final step was the visual analytics where they used the GeneTerrain visualization technique for discovering disease markers.

An important study to identify meaningful infrastructures in biological networks is the study by Rasif Ajwad et al. (2021) [48], whereby they propose a method for discovering significantly mutated subnetwork structures in the breast cancer genome. They aim to evaluate whether the identified subnet or pathway can be implemented as a biomarker to predict survival of breast cancer patients, and also to provide improved mechanisms for cancer therapy. The model applied in this research uses two network analysis approaches, the HotNet2 and ClusterONE methods. These two algorithms are graph-based and are used to identify significantly mutated and still clinically and functionally relevant subnets from the Copy Number Exchange (CNA) database. In this study, patient-specific mutation data were obtained from the Copy Number Alteration (CNA) database of the METABRIC library. The proposed model defines mutated subnets as follows. Initial CNA-specific gene information is retrieved using the BiomaRt R package. Next, the gene mutation frequency for each gene is calculated and used as a "temperature score" to run the HotNet2 algorithm. Gene pair mutation similarity is also calculated using gene mutation frequencies. The calculated similarity scores of the gene pairs are taken as input data for the ClusterONE algorithm. Then, significantly mutated subnets are identified using HotNet2 and ClusterONE algorithms. The HotNet2 algorithm detects mutated subnets with a certain genomic scale, while the ClusterONE algorithm is used to identify meaningful clusters or interacting gene groups in the network.

Similarly, Le Yang et al. (2021) [47] introduce a method to identify meaningful infrastructures (pathways or subnetworks) in cancer called FDRnet. The research addresses the heterogeneity problem in mutation by solving a Linear programming problem through use of a False Discovery Rate (FDR) parameter for budget constraint. The experiments were carried out here on both real cancer data and simulated data. The proposed FDRnet algorithm has advantage over other existing approaches as they are able to perform homogeneous subnetwork identification in a scale-free biological network. It can also control the frequency discovery rate for genes in the detected subnetwork,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202201.0170.v1

integrate multi-omics data and improve computational efficiency. The real-world PPI network iRefIndex was used as well as protein complexes extracted from the CORUM database which have been found to take part in the promotion of breast cancer. FDRnet was applied to The Cancer Genome Atlas mutations as well as cancer data for detecting significantly mutated subnetworks in breast cancer. FDRnet could be also used for the evaluation of gene expression data. The authors demonstrated this by applying the algorithm to data gotten from the diffuse large-B-cell lymphoma (DLBCL) study. For each candidate, the expression data was identified by performing a t-test and calculating p-values for the statistical significance of individual genes.

When discussing identification of meaningful infrastructures in biological networks, it is important to mention the recent study by Di Zhang and Yannan Bin (2020) [48]. In this study, they propose a new approach for finding driver genes involved in cancer development, through a subnetwork enrichment analysis. The proposed method is known as DriverSubNet. This algorithm doesn't make use of any parameters and detects the driver genes through the effective mining of the mutation and gene expression information, with respect to subnetwork enrichment analysis. It ranks genes efficiently based on F1-score, precision and recall values. For testing this algorithm, 4 types of data were used. These include Somatic Copy Number Alterations (SCNA), Head and Neck Squamous Carcinoma (HNSC), Kidney renal clear cell Carcinoma (KIRC), Thyroid Carcinoma (THCA), Breast Cancer (BRCA) and RNA-seq expression data obtained from The Cancer Genome Atlas (TCGA). The information for undirected interaction network was gotten from the Human Protein Reference Database (HPRD). Genes obtained from the Cancer Gene Census were used for evaluation of the algorithm. As mentioned earlier, the evaluation criteria used here were the Recall, Precision and F1-score. The significance in mutated genes was evaluated using the mutation frequency and ranked according to this significance. Function Enrichment Analysis and Survival and Drug analysis was carried out on the results to understand the features discovered by the DriverSubNet algorithm and verify whether the identified genes are clinically relevant. After performing enrichment analysis on Gene Ontology (GO) and KEGG, it was discovered that the first 100 uncovered genes related to cancer had been significantly Enriched.

3. Algorithms for Meaningful Infrastructure Identification in Biological Networks

Over the years, there have been different methods proposed by researchers for detection communities or modules in biological networks. Most of these methods for Meaningful Infrastructure Identification are based on the concept of modularity. This chapter will review some of the algorithms which are applied for Meaningful Infrastructure Identification in biological networks. These algorithms include but are not limited to: The Genetic Algorithm, Differential Evolution Algorithm, Water Cycle Algorithm (WCA), Walktrap Algorithm and Firefly algorithms.

3.1. Genetic Algorithm (GA)

Out of all the meta-heuristic and evolutionary algorithms, the Genetic Algorithm is the most popularly used for module detection in networks. This is because it is easy to understand and it has a high accuracy. In [3] the steps in the GA were modified by Marwa Ben et al. whereby a new mutation operator was used and the population initialization step was also modified. The proposed Genetic Algorithm in this research was as follows:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202201.0170.v1

Table 1. GA by Marwa et al. [3].

- *Begin with a set of genes generated randomly.*
- *Parents from the current populattion are selected for mating.*
- *Crossover is applied to parents in order to develop new population.*
- *Implement the mutation on the new offspring generated.*
- *Evaluate and replace off-springs having the worse existing individuals in the population with these offsprings.*
- *If the stopping criteria are not satisfied, repeat process from the second step.*
  (Stopping Criteria: predefined number of generations is reached or maximum computational time is attained).

In the phase where the initial population is generated, the population is represented as a two-dimensional array of individuals. For population initialization, a gene group having the same size as those extracted from the KEGG database have to be recovered. Calculating the fitness function is an important part of implementing the Genetic Algorithm. This function tells how "fit" the obtained solution is in relation to the problem. Here the fitness function used was based on the computation of the average interaction score calculated for every pair of genes found in the community and an average similarity value. The mutation and crossover steps are implemented to obtain new solutions for the next generation.

Another Variant of the Genetic Algorithm used for Meaningful Infrastructure Identification is the Scatter Search-based Genetic Algorithm (SSGA) [14]. This method was proposed by Yilmaz Atay et al. in their research whereby transmissions to the next generation was based on a scatter search approach. The crossover and mutation were implemented for each reference set with respect to what was known as a Change Control Array (CCA).

In [16] MIGA (Modularity-based Improved Genetic Algorithm) was proposed to identify community structures in biological networks. This improved Genetic Algorithm made use of the parameter called the Modularity, Q. This parameter was analyzed and used to improve precision and complexity of the Meaningful Infrastructure Identification. The Crossover and mutation genetic operators were also used as well as a local search procedure. Due to its advantages over hill climb algorithm, the Simulated Annealing was used here as the local search procedure.

3.2. Differential Evolution (DE)

This is an evolutionary algorithm whereby the initial population contains randomly sampled NP individuals from the search space. Differential Evolution has been often used for solving hard mathematical or optimization problems and in 2012, it was used in [6] for meaningful infrastructure identification in complex biological networks.

The Differential Evolution algorithm is similar to the Genetic Algorithm except for the fact that in this algorithm the Mutation step comes before crossover and selection. Therefore, it starts with initialization, mutation, crossover, then applies selection. In the beginning of the population initialization step, DE puts every node into a random group or cluster or community by assigning IDs randomly. The mutation phase implements a mutation strategy known as the "rand/1" strategy. This strategy does not have any special search direction bias and uses random selection to choose a new search direction. Implementing this algorithm on real networks like the Protein-protein interaction network, and comparing with results from the Genetic algorithm has shown that Differential Evolution is an effective algorithm for meaningful infrastructure identification in biological networks.

3.3. Water Cycle Algorithm (WCA)

The Water Cycle Algorithm is an algorithm which was originally developed for handling continuous optimization problems. Like most other algorithms, the water cycle algorithm can be adapted to be used for meaningful infrastructure identification. The most important aspect to take into consideration in the design of the Water Cycle Algorithm is in the way rivers and streams flow into their corresponding leading sea or river. Eskandar H. et al [40] in their research propose a Water Cycle Algorithm, which just like many other meta-heuristic algorithms begins with

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202201.0170.v1

population initialization in which the original populations are initialized as raindrops. After this the cost of each raindrop is calculated and the flow intensity for seas and rivers is also calculated. Next, stream and river flow downhill are calculated and the positions of rivers are exchanged with the streams which give best solutions. Similarly, the position of the sea is replaced with that of the river which gives best results. Next evaporation conditions are checked and if satisfied, the raining process will occur. Finally, convergence criteria are also checked and if satisfied, the algorithm is going to stop. Osaba, E. et al [43] in their research used adapted the Water Cycle Algorithm to be used for meaningful infrastructure identification in networks using bio-inspired optimization.

3.4. Walktrap Algorithm (WA)

The Walktrap algorithm is a hierarchical clustering algorithm which uses the random walk method and is popularly used for meaningful infrastructure identification. Like other meaningful infrastructure algorithms, it can be adapted for identification of community structures in biological networks. Just like the random walk algorithm, the distance between vertices, are also calculated through random walks in the network. This algorithm works on the assertion that random walks through the given network are attracted towards a given area and create densely connected sections which can be referred to as communities. Furthermore, by use of random walks, it conglomerates different communities using a bottom-to-top mechanism.

3.5. FireFly Algorithm

The Firefly Algorithm is a very powerful approach which is used for solving constrained optimization problems which are NP-hard. Just like other algorithms, it is one of the algorithms which can be modified to suit the purpose of meaningful infrastructure identification, as demonstrated by Osaba E. et al [43] in their research. The firefly algorithm is very similar to the famous Bat Algorithm [18] where each firefly is a representation of a feasible solution for the given problem at hand. Here discrete adaptation is applied using the light absorption method, which is a very important and necessary method for adjusting the attractiveness of fireflies. Also, the NMI is used as the distance metrics to calculate the distance between two fireflies. This results in different flavors of the firefly algorithm used for meaningful infrastructure identification.

3.6. Overlapping Multiple Label Propagation Algorithm (OMLPA)

Label Propagation Algorithm (LPA) is an algorithm for meaningful infrastructure identification whereby every node found in the network is assigned a “Label” which indicates the network structure or community to which it belongs. It uses an inheritance concept whereby a node inherits the label given to most of the nodes in its neighborhood. It is necessary to remember that prior information about the given network is not necessary for LPAs. In 2019, Jyoti Shokeen et al. [49] in their study, proposed a novel approach for label propagation called the Overlapping Multiple Label Propagation Algorithm (OMLPA). This algorithm works by assigning to the nodes in a network structure, multiple labels which will permit the identification of overlapping community structures. In the algorithm, they use the Jaccard Index based Common Neighbor Similarity score (CNS) for the identification of prime community structures. The algorithm starts by randomly assigning to every node in the network a unique label and then placing these nodes in a decreasing order based on their respective degrees. Next the algorithm searches for nodes with common neighbors and then performs a calculation of closeness between neighbors that are common to both nodes. This degree of closeness is calculated in every iteration of the algorithm. Here when calculating the CNS score of two neighboring nodes, the node with a lower CNS score assumes the nodes label that has the higher CNS score and if both nodes have the same high CNS score, then they are placed in the same set. In this algorithm, there are some cases where the many small community structures are created. In this situation, the algorithm applies a method to the small communities whereby the small communities merge with larger communities until every node assumes the label of the of the its neighboring nodes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202201.0170.v1

3.7. Connect Intensity Iteration Algorithm (CIIA)

The Connect Intensity Iteration Algorithm is a modularity-based community detection approach that uses an indicator which depicts the difference in edge number count in the network when the edges are randomly placed and the actual edge number count between two vertices found in the original network. The proposed method here is a study by Zhang Renquan et al. [50] in 2021 for identifying community structures in complex networks. In the CIIA model, there exist an iterative process of self-selection and self-learning. In each iteration, the Connect Intensity, CI, is calculated and used as the weight in the next round so as to achieve continuous training. The algorithm is modelled as follows: first the CI of each edge is calculated and each edge taken as a separate network, next the weight of each point and the points adjacent to it is calculated. Now according to the degree of the weighted node, the connection strength of the connection strength of the network calculated again. Finally, the communities are merged on both ends with respect to the calculated CI from larger to smaller and if the calculated modularity gain is positive, the combination or merger is accepted and a new partition formed. If the calculated modularity gain is negative the subsequent partition is stopped. It is worth noting that the CIIA model's processing time is hugely affected by the average degree as well as the edge count found in the subnetwork structure and affected less by the node count.

3.8. Louvain Community Detection

The Louvain community detection algorithm was first proposed in 2008 as a fast community unlocking method for large networks. This approach is based on modularity, which tries to maximize the difference between the actual number of edges in an ensemble and the expected number of edges in the ensemble. However, optimizing modularity in a network is NP-hard, so it is necessary to use heuristics. The Louvain algorithm is divided into two iteratively iterative stages.

- Moving nodes locally
- Gathering the network

The Louvain community detection algorithm reveals community communities during the process. It is quite popular due to its ease of implementation and also because of the speed of the algorithm. However, an important limitation of the algorithm is the use of storage of the network in main memory.

3.9. Leiden Community Detection

In a recent study by (2019), by VA Traag et al. He showed that Louvain community detection tends to discover communities that are internally disconnected (poorly linked communities). In the Louvain algorithm, moving a node that acts as a bridge between two components in a community to a new community can disconnect the old community. This will not be a problem if the old community is further divided. But that won't be the case, according to Traag et al. Other nodes in the old community allow it to remain as a single community due to their strong connections. Also, according to them, Louvain has a tendency to discover interconnected communities every week. Therefore, they proposed the much faster Leiden algorithm, which guarantees good interconnection of communities. In addition to the stages used in the Louvain algorithm, Leiden uses one more stage that tries to improve the discovered sections. The three stages in the Leiden algorithm are:

- Moving nodes locally
- Improvement of partitions
- Aggregation of the network according to the developed sections

4. Performance Evaluation Parameters

4.1. Modularity

Modularity is one of the most popular fitness-function known in literature of Meaningful Infrastructure Identification research which has been around since 2004. According to [19], Modularity is a parameter that has the unique

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202201.0170.v1

privilege of being a general parameter for defining a community and at the same time the key function used for the popular graph clustering. In Meaningful Infrastructure Identification, the modularity function is a phenomenon based on the idea that, relationships between members belonging to the same community has to be completely maximized while minimizing the inter-member relationships of one community with members of a different community. This evaluation function is obtained non-randomly by a mathematical model. The most basic modularity is defined by the equation:

$$Q_{basic} = \sum_{k=1}(e_i^n - a_i^2) \quad (1)$$

Where, e_i signifies the edge probabilities of the binary vertices, a_i signifies the fraction of edges that have at least one endpoint within the group (percentage of edges with at least 1 end in module i).

It is well known that modularity works efficiently in many cases of module detection in biological networks. This notwithstanding, there have been some limitations encountered with it performance. First and foremost, it is noted that when it comes to vertex similarity, modularity tends to have a restricted interpretation. Another setback of modularity is the problem of resolution limit which is as a result of its null model. Modularity is a concept that compares the number of links that are found in a single subnetwork with the number of links found in the subnetwork if the network was randomly developed with similar number of vertices and also every node in the random network keeps its degree, but edges on the other hand are randomly attached. This results in the systematic combination of small community structures into larger ones, also in cases where the community structures are loosely connected to each other and well defined [20]. In attempt to find a solution to the problem of resolution limit, multiresolution versions of modularity were developed that permits end users to indicate the target scale of the community structures.

The maximization of modularity and other Meaningful Infrastructure Identification criteria in biological networks are often focused on unipartite networks which have denser edges inside communities.

On the other hand, in bipartite networks the links between different community structures are denser than the connections within these networks. Due to this, modularity maximization cannot find correct modules due to the fact that it aims to reduce the number of links found between the communities.

4.2. Normalized Mutual Information (NMI)

The NMI is and evaluation criterion that calculates the dissimilarity found within separate clusters or subnetworks and the total similarity within the clusters or subnetworks. It measures the similarity between two segments, ranging from 0 to 1 for community structures that are not similar, to similar community structures respectively. This NMI criterion is obtained from the information theory concept and also indicates the amount of information shared between two community structures. There exist situations where the community structures don't contain the same sets of nodes. Here nodes that are found to be common to both community structures are added to the calculation of mutual information. For Example: A graph G = (V, E) can be divided into two disjoint sets, (A, B), $A \cup B = V$, $A \cap B = \emptyset$, by removing the edges that connect the two parts. A measure of the dissimilarity degree between the two parts can be calculated as total amount of removed weight edges [40]. The Normalized Mutual Information is given by:

$$N(A,B) = \sum_{u\in A, v\in B} w(u,v) \quad (2)$$

4.3. Functional Enrichment

Gene Ontology (GO) in biological network analysis is used to show the functional content of a point found in the network [5]. Gene Ontology is a major initiative of bioinformatics in attempt to make the representation of genes as well as their product attributes unique across every species. Gene Ontology gives gene description a functional vocabulary in terms of Biological Process (BP), Cellular Component (CC) and Molecular Function (MF) [3]. With the aim of finding the enrichment annotations of a specific Gene Ontology term in a biological network, a probability is calculated which gives the percentage of the same or different higher number of nodes being annotated with the term in question.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202201.0170.v1

The application is done through the use of a statistical test which involves the upper tail of a hyper-geometric distribution. This test is called the one-tailed Fisher's exact test. It has one setback which is as a result of the inheritance problem [21] [22].

Gene Ontology analysis can be referred to as the need to compare raw count with expected values. There are various methods can be used to perform Gene Ontology Analysis including Fisher's Exact Text, Hypergeometric Test or the Bonferroni Correction. Some of the Tools used for this purpose include: PlantRegMap, PLAZA workbench, FunRich and Blast2GO. In performing functional enrichment analysis, a study set, a population set which must contain the study set, GO annotations which associate the genes in the population set to Gene Ontology terms and Gene Ontology with its terms and relationships are needed.

For example, functional enrichment analysis has been applied to detect cancer driver genes and pathways using their various proposed methods over the years. Some of these genes include CDK4, CDKN2A, CDKN2B, CYP27B1, and MTAP discovered by Junhua Zhang et al. [44] using the iMCMC method. In this work, they also identified CDKN2B and a metagene including CDK4 and TSPAN31. In the study by Fabio Vandin et al. [45] they analyzed the original HPRD network and discovered some gene pathways by using an FDR of less than 0.01. The discovered pathways included known high degree, highly mutated and important genes like the TP53. Some of these pathways are as follows: 10 gene pathway - WT1, CDKN2A, TP53, CCNG1, KLF6, ATR, CDKN2C, TP73L, TFDP1, CHEK1; 10 gene pathway - RAP2B, PIK3CA, HRAS, RASSF2, NRAS, MRAS, PIK3CG, BRAF, NF1, RHOB; 7 gene pathway - MAML2, MAML1, NOTCH4, NOTCH2, NOTCH3, JAG2, JAG1. Le Yang et al. during their research to detect significantly mutated subnetworks in breast cancer assessed genes detected by comparing them with the COSMIC cancer gene database. HotNet2 was able to identify up to 21 genes including MYB, TP53, CDH1, ERBB2, RB1 and PIK3CA. In the enrichment analysis performed by FDRnet, the first network consisted of a higher number of the TP53 gene, which is the most prevalently mutated gene in breast cancer. Other genes in breast cancer which are mutated frequently such as MAP3K1, PIK3CA, RB1 and PTEN were all detected within specific subnetwork structures. Evaluation also picked out substructures that had the estrogen receptor (ER) and ERBB2/HER2. FDRnet succeeded in identifying up to 59 subnetworks and effectively controlled the FDRs of these subnetworks. Next an enrichment analysis of gene-ontology terms of the identified subnetworks was applied and very importantly FDRnet succeeded in identifying the NF-kB pathway. This pathway maybe pivotal in the maintenance and initiation of lymphoma. The STAT3 was seen to be the most prevalent factor amongst the detected subnetworks, which is an important aspect in the JAK-STAT transcriptional regulation pathway.

4.4. Precision and Recall

These two parameters are very important parameters considered when solving optimization problems. The Recall can be defined as the ratio of total number of True Positives (TP) to the sum of True Positives (TP) and False Negatives (FN). This is given mathematically as:

$$\text{Recall (R)} = \text{TP} / (\text{TP} + \text{FN}) \quad (3)$$

Similarly, Precision is the ratio of the total number of True Positives (TP) to the sum of total number of True Positives (TP) and False Positives (FP). This is defined mathematically as:

$$\text{Precision (P)} = \text{TP} / (\text{TP} + \text{FP}) \quad (4)$$

With respect to biological networks, the True Positives can be considered as gene products which are annotated as belonging to a certain protein complex by Gene Ontology. In biological network clustering, Precision and Recall for every cluster can be calculated in relation to every annotated complex in the validation set. The validation set here could be Gene Ontology (GO). The True Positives (TP) for each annotated complex can be defined as those nodes belonging to a cluster that can be found in the annotated complex, the False Positives (FP) on the other hand are members of the cluster, not found in the annotated complex and finally, False Negatives (FN) are those which are not members of the cluster but are found in the annotated complex.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202201.0170.v1

4.5. Redundancy

The redundancy of a given biological network structure is a phenomenon whereby the nodes found in a given community structure or subnetwork tend to be found in another community structure or subnetwork. It is a phenomenon which features more often in multi-layered biological networks. The redundancy is directly proportional to the number of layers connecting every vertex pair found in the network structure. Meaning that, the greater the number of layers connecting each vertex pair in the network the greater the redundancy.

4.6. Purity and Surprise

In comparing and evaluating the performance of meaningful infrastructure identification algorithms, test for measuring clustering accuracy such as the purity are used. The purity of a multi-layered biological network gives the mathematical representation of the total number of proteins, genes or biological component which are placed in the correct clusters. This is based on the intersection of the network elements on the different layers of the multi-layered network. The Surprise parameter, S, is defined as the measure of the improbability of randomly discovering a subnetwork, community, or meaningful infrastructure with the same observed functional enrichment of links withing links belonging to these subnetworks or communities and links of a random graph. The Surprise, S, implicitly defines a community in a more complex way, and with improvement in studies over time, Surprise maximization techniques have been improved and developed to be used for identifying community structures in complex networks.

5. Conclusions and Future Research Directions

Meaningful Infrastructure Identification is an important computational technique that is used for the analysis of networks. This concept helps to derive very important insights that have to do with the structural organization of a network structure and serves as the start point for understanding the concept of correspondence between the network structure with its internal functions. This concept of Meaningful Infrastructure Identification is applied to every area of research including sociology, economics, biology and medicine, finance and science. In this article, we focus on Meaningful Infrastructure Identification in biological networks. Here communities are commonly referred to as modules, where the modules are made up of vertices (could be drugs, DNA, proteins) and the edges in the module represent the relationships between these components. Module detection in biological networks has evolved over time from being in single layered networks (unipartite networks) to detection in multilayered networks (bipartite networks) where it becomes more complicated. The datasets used for this Meaningful Infrastructure Identification are often defined with respect to the annotations of the Gene Ontology (GO). There have been various algorithms implemented by researchers to perform community/module detection in biological networks. Here 5 of these algorithms are discussed. These algorithms include: The Genetic Algorithm, the Differential Evolutionary Algorithm, Water Cycle Algorithm (WCA), Walktrap Algorithm (WA), Connect Intensity Iteration Algorithm (CIIA), FireFly Algorithm and Overlapping Multiple Label Propagation Algorithm. The performance measurement parameters discussed here are: Modularity, Mutual Information (NMI), Functional Enrichment, Precision and Recall, Redundancy, Purity and Surprise concepts are discussed. In future research, experimental analysis will be done and compared on the discussed algorithm for identifying meaningful infrastructures in biological networks. This experimental analysis is aimed to be carried out on both synthetic and real-life datasets such as the C. elegans metabolic reaction network, E. coli transcription Network, the Cattle protein–protein interaction network and the Helicobacter pylori protein–protein interaction network. A performance evaluation will be done on these different datasets for each of the algorithms to determine what algorithm produces best results for module detection. In future studies, the performance parameters including Modularity, Normalized Cut/Normalized Mutual Information, will be examined and algorithms will be graded based on the calculated fitness value and accuracy.

References

1. R. M. Ames, J. I. Macpherson, J. W. Pinney, Lovell and Robertson, Modular biological function is most effectively captured by combining molecular interaction data types, PloS One 8, e62670; doi: 10.1371/journal.pone.0062670, 2013.
2. Girvann and Newman, Community structure in Social and Biological Networks, Proc. Natl. Acad. Sci. U.S.A 99:7821, doi: 10.1073/pnas.122653799, 2002.
3. M. B. M'Barek, B. Amel, B. Walid and H. Sana Ben, Genetic Algorithm for Community Detection in Biological Networks, Procedia Computer Science 126, 195-204, 2018.
4. C. G and M. Kuijjer, Community Detection in Large-Scale Bipartite Biological Networks, Front. Genet. 12:649440. doi: 10.3389/fgene.2021.649440, 2021.
5. L. Bennett, Aristotelis Kittas, Gareth Muirhead, Lazaros G. Papageorgiou and Sophia Tsoka, Detection of Composite Communities in Multiplex BIological Networks, DOI: 10.1038/srep10345. (2015), 2015.
6. G. Jia, Zixing Cai, Mirco Musolesi, Yong Wang, Dan A. Tennant, Ralf. J.M. Weber, John K. Heath and Shan He, Community Detection in Social and Biological Network using Differential Evolution, DOI: 10.1007/9783-642-34413-8_6, 2012.
7. R. Henriques and Sara C. Madeira, BicNET: Flexible module discovery in large-scale biological networks using biclustering., DOI 10.1186/s13015-016-0074-8, 2016.
8. M. Serban, Exploring modularity in biological networks, Phil. Trans. R. Soc. B 375: 20190316. http://dx.doi.org/10.1098/rstb.2019.0316, 2020.
9. Rhrissorrakrai K. and Gunsalus K. C., MINE: Module Identification in Networks, BMC Bioinformatics, 12:192., 2011.
10. W. Chen, Zexuan Zhu and Shan Hec, MUMI: Multitask module identification for biological networks, IEEE Transactions on Evolutionary Computation. https://doi.org/10.1109/TEVC.2019.2952220, 2019.
11. Gilles Didier, Alberto Valdeolivas and Anais Baudot, Identifying communities from multiplex biological networks by randomized optimization of modularity., doi.org/10.12688/f1000research.15486.1, 2018.
12. E. Ravasz, Detecting Hierarchical Modularity in Biological Networks, Harvard Medical School, doi: 10.1007/978-1-59745-243-4_7, 2009.
13. K. Mitra, Anne-Ruxandra Carvunis, Sanath Kumar Ramesh and Trey Ideker, Integrative approaches for finding modular structure in biological networks., 14(10): 719–732. doi:10.1038/nrg3552, 2013.
14. Yilmaz A., Koc I., Babaoglu I., and Kodaz H., Community detection from biological and social networks: A comparative analysis of metaheuristic algorithms, Applied Soft Computing 50. 194–211., 2017.
15. R. Mall, Ehsan Ullah, Khalid Kunji, Michele Ceccarelli and Halima Bensmail, An unsupervised disease module identification technique in biological networks using novel quality metric based on connectivity, conductance and modularity [version 1; peer review: 2 approved with reservations], F1000 Research, 7:378 https://doi.org/10.12688/f1000research.14258.1, 2018.
16. R. Shang, Jing Bai, Licheng Jiao and Chao Jin, Community detection based on modularity and an improved genetic algorithm, Physica A 392, 1215–1231., 2013.
17. Sara Rahiminejad, Mano R. Maurya and Shankar Subramaniam, Topological and functional comparison of community detection algorithms in biological networks, BMC Bioinformatics 20(1).1-25, 2019.
18. X. Yang, A new metaheuristic bat-inspired algorithm, Berlin Heidelberg: Nature Inspired Cooperative Strategies for Optimization (NISCO), Springer. pp. 65–74., 2010.
19. S. Fortunato, Community detection in graphs., Physics Reports 486(3), 75–174, 2010.
20. Fortunato, S and Barthelemy, M, Resolution limit in community detection., Proceedings of the National Academy of Sciences 104(1), 36–41, 2007.
21. Bauer, S, Grossmann, S, Vingron, M and Robinson, P. N, Ontologizer 2.0 - a multifunctional tool for GO term enrichment analysis and data exploration., Bioinformatics 24, 1650–1651 (2008), 2008.
22. S. Grossmann, S. Bauer, Robinson, P. N and Vingron, M, Improved detection of overrepresentation of gene ontology annotations with parent child analysis, Bioinformatics 23, 3024–3031, 2007.
23. Matlock, M. K., Datta, A., Le Dang, N., Jiang, K., & Swamidass, S. J. (2019, July). Deep learning long-range information in undirected graphs with wave networks. In 2019 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.
24. Zhang, M., & Chen, Y. (2018). Link prediction based on graph neural networks. Advances in Neural Information Processing Systems, 31, 5165-5175.
25. Datta et al. "Machine Learning liver-injuring drug interactions with non-steroidal anti-inflammatory drugs (NSAIDs) from a retrospective electronic health record (EHR) cohort ", PLOS Computational Biology 2021.

26. Burkhardt HA, Subramanian D, Mower J, Cohen T. Predicting Adverse Drug-Drug Interactions with Neural Embedding of Semantic Predications. bioRxiv. 2019.
27. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics. 2018.
28. Ryu JY, Kim HU, Lee SY. Deep learning improves prediction of drug–drug and drug–food interactions. Proceedings of the National Academy of Sciences. 2018.
29. Grover et al, node2vec: Scalable Feature Learning for Networks, ACM SIGKDD, 2016.
30. Devikanniga, D., Vetrivel, K., & Badrinath, N. (2019, November). Review of meta-heuristic optimization based artificial neural networks and its applications. In Journal of Physics: Conference Series (Vol. 1362, No. 1, p. 012074). IOP Publishing.
31. Xia W, Yu Q, Li G, Liu Y, Xiao F, Yang L, Rahman ZU, Wang H, Kong Q. 2019. Identification of four hub genes associated with adrenocortical carcinoma progression by WGCNA. PeerJ 7: e6555, https://doi.org/10.7717/peerj.6555.
32. Yue-Hua Feng, Shao-Wu Zhang and Jian-Yu Shi. BMC Bioinformatics (2020). "DPDDI: a deep predictor for drug-drug interactions."
33. Li, T., Wernersson, R., Hansen, R. B., Horn, H., Mercer, J., Slodkowicz, G., et al. (2017). A scored human protein–protein interaction network to catalyze genomic interpretation. Nat. Methods 14:61. doi: 10.1038/nmeth.4083
34. Turei, D., Korcsmaros, T., and Saez-Rodriguez, J. (2016). Omni path: guidelines and gateway for literature-curated signaling pathway resources. Nat. Methods 13:966. doi: 10.1038/nmeth.4077.
35. Li, Y., Calvo, S. E., Gutman, R., Liu, J. S., and Mootha, V. K. (2014). Expansion of biological pathways based on evolutionary inference. Cell 158, 213–225. Doi: 10.1016/j.cell.2014.05.034.
36. Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics, 9(1), 1-13.
37. Leicht, E. A., & Newman, M. E. (2008). Community structure in directed networks. Physical review letters, 100(11), 118703.
38. Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence, 22(8), 888-905.
39. Sah, P., Singh, L. O., Clauset, A., & Bansal, S. (2014). Exploring community structure in biological networks with random graphs. BMC bioinformatics, 15(1), 1-14.
40. Eskandar, H., Sadollah, A., Bahreininejad, A., & Hamdi, M. (2012). Water cycle algorithm–A novel metaheuristic optimization method for solving constrained engineering optimization problems. Computers & Structures, 110, 151-166.
41. Osaba, E., Del Ser, J., Camacho, D., Bilbao, M. N., & Yang, X. S. (2020). Community detection in networks using bio-inspired optimization: Latest developments, new results and perspectives with a selection of recent meta-heuristics. Applied Soft Computing, 87, 106010.
42. Hajkarim, M. C., Upfal, E., & Vandin, F. (2019). Differentially mutated subnetworks discovery. Algorithms for Molecular Biology, 14(1), 1-11
43. Yue, Z. (2020). Network-Based Analytics for Discovering Gene Modules and Biomarkers in Complex Diseases (Doctoral dissertation, The University of Alabama at Birmingham).
44. Zhang, J., & Zhang, S. (2016). The discovery of mutated driver pathways in cancer: Models and algorithms. IEEE/ACM transactions on computational biology and bioinformatics, 15(3), 988-998.
45. Vandin, F., Upfal, E., & Raphael, B. J. (2011). Algorithms for detecting significantly mutated pathways in cancer. Journal of Computational Biology, 18(3), 507-522.
46. Ajwad, R., Domaratzki, M., Liu, Q., Feizi, N., & Hu, P. (2021). Identification of significantly mutated subnetworks in the breast cancer genome. Scientific reports, 11(1), 1-15.
47. Yang, L., Chen, R., Goodison, S., & Sun, Y. (2021). An efficient and effective method to identify significantly perturbed subnetworks in cancer. Nature Computational Science, 1(1), 79-88.
48. Bin, Y., & Zhang, D. (2020). DriverSubNet: A novel algorithm for identifying cancer driver genes by subnetwork enrichment analysis. Frontiers in genetics, 11, 1854.
49. Shokeen, J., Rana, C., & Sehrawat, H. (2019). A novel approach for community detection using the label propagation technique. In Integrated intelligent computing, Communication and Security (pp. 127-132). Springer, Singapore.
50. Renquan, Z., Yu, W., Xiaolin, W., Yuze, S., & Jilei, T. (2021), CIIA: A New Algorithm for Community Detection, arXiv:2110.15264.