Biological network analysis

Nikhitha Chavalam, Parvana J Kuruppal, Ankitha, Sudarshana , Hridhi Sethi
CB.EN.U4CSE18012,CB.EN.U4CSE18042,CB.EN.U4CSE18107,CB.EN.U4CSE18152,CB.EN.U4CSE18502
Dept. of Computer Science and Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham
Tamil Nadu, India

***Abstract*—Biological systems are often represented as networks which represent interactions or relations between different biological entities. In this document you will get a gist about network analysis, graph representation of biological networks, deep learning methods for analysis , visualization techniques, tools available like NetworkAnalyst, CytoScaPe, CentiScaPe and applications. It also discusses a few new inventions to better analyse biological network graphs according to topological as well as biological properties.**

***Index Terms*—Deep learning , Graph Theory,biological networks, protein interactions, drug development, drug-target prediction, NetworkAnalyst, Cytoscape, CentiScaPe, Network Centralities,**

I. Introduction

Biological network contains of biological entities like proteins, DNA, RNA as well as relation between them. The analysis of biological networks helps us in identifying genes or proteins, network construction network analysis and visualization.The amount of data obtained by new technologies, like micro or Chip-Chip arrays, and largescale "OMICS"-approaches are vast and biological data repositories are growing exponentially in size.The main challenge is analysing this large amount of data use them to answer some of the biological questions. An interactive visual representation [3] of information together with data analysis techniques is often the method for the better understanding of data.

II. Literature Survey

A. Using Graph Theory To Analyse Biological Networks

[1] This paper focuses on how graph theory concepts can be used to show various aspects of biology.

In Biology various components can be represented with the help of graphs such as genes or diseases and connected elements can be linked using edges, this paper dives deep into graph theory and shows how it can help us in gaining better understanding of the biological significance of a system under consideration.

Popular applications of biological networks looked into in the paper are protein-protein interaction, drug target identification, signal transduction etc. The insights gained on how the above applications can be represented using graph theory is as follows:

Protein-protein interaction network which shows physical link between proteins within a cell, the study shows that these PPI networks are scale- free that is follow power law distribution and these networks are highly dynamic making it difficult to gain general conclusion.

Regulatory networks which are linked to gene expression in cells and can be represented as sparsely connected network based on average number of upstream regulators linked to each gene, this is also largely scale -free

Signal transduction networks can be modelled using multi-edge directed graphs to show inter-protein or chemical interactions and have feedback loops, these are also sparsely connected as well as scale-free

Metabolic networks show correlation of pathways which hold information about biochemical events that are present in any metabolic process, by graph theory these can be analysed using the network diameter to gain insight and is known to form hierarchical structures

The drawback of using graph theory for this analysis is that these are highly static though the system as such is dynamic.

B. Biological network analysis using deep learning

[2] Biological network contains of biological entities like proteins, DNA, RNA as well as relation between them. Such kind can be represented in form of a graph in tackling and understanding bioinformatics. In this paper we will discuss about tackling these challenges with the use of deep learning aka multiple layer learning which is used to detect the complex patterns.

There are different kinds of networks which help in representing and understanding heterogeneous and complex data so different analysis can be jotted down and discover different disease pathways. Those network types are:

1. Protein protein interaction networks
2. Gene regulatory networks
3. Metabolic networks
4. Drug-drug interaction network

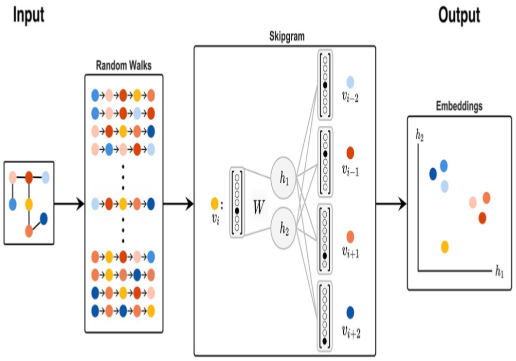
For analysing the information in the graph we need to learn tasks related to it and learning tasks could be categorised into different types like :

1. Node classification
2. Link prediction
3. Graph classification or regression
4. Graph embedding

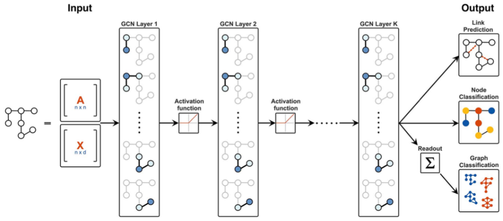
There are different graph neural network algorithms like graph embedding and graph convolutionary networks which could be useful for learning tasks. Graph embedding algorithm uses Deep Walk method via word2vector representation to learn the embeddings of each node in graph. These word embeddings can be useful to predict the surrounding context words. Formalised as :

$$\min_{\Phi} \; -\log P(v_{i-w},\dots,v_{i-1},v_{i+1},\dots,v_{i+w}|\Phi(v_i)), \qquad (1)$$

This shows the process of learning a simple graph embedding using DeepWalk



Apart from Deep Walk LINE is also used which takes first and second order proximity of nodes into consideration. Apart from this graph convolutional networks algorithm where training is done iteratively by loss function and then its propagated back via back propagation where W weight is calculated and adjusted accordingly.A visual depiction of a k-layer GCN is as follows.



There are different applications in biology that apply biological network analysis like proteomics where deep learning helps to understand which proteins interact i.e a link prediction problem , function of a given protein which could be done via node classification or graph classification and finally determine the structure of protein. Other application include drug development, discovery, polypharmacy and disease diagnosis.

C. Visualisation Tools for Biological Network Analysis

[3] The amount of data obtained by new technologies, like micro or Chip-Chip arrays, and largescale "OMICS"-approaches are vast and biological data repositories are growing exponentially in size. The main challenge is analysing this large amount of data use them to answer some of the biological questions. An interactive visual representation of information together with data analysis techniques is often the method for the better understanding of data.

Graphs represent biological interactions in entities as networks consisting of vertices(biological entities) and edges(Connection between entities). In most of the cases single edge is not sufficient to show the relations between two biological entities as two entities many be connected by different relations Data visualisation faces challenges like exponential increase of data,integration of heterogeneous data and the representation of multiple connections between nodes with heterogeneous biological meanings.This paper discussed the various visulaisation tools that can be used to extract patterns and structures in different situations. Ondex, Pivot or Medusa are used when there is heterogeneous data, Cytoscape or BioLayoutExpress3D are used when sheer mass of data but with less heterogenity and Medusa and other tools featuring multi-edged networks are used when there are multiple relations between entities

D. Protein Interaction Network Analysis and Visual Exploration using NetworkAnalyst

[4] This paper focuses on the developed tool 'NetworkAnalyst' which enables high performance network analysis. It is a combination of the three major steps in biological network analysis and exhibits it's results in an online network visualization tool. It mainly provides protein protein interaction. Among different molecular networks, protein–protein interaction (PPI) networks is one of the most important resource for understanding data as Protein interactions play fundamental roles in structuring and mediating essentially all biological processes. protein–protein interaction networks are usually undirected graphs with proteins nodes as and edges indicating interactions between two connecting proteins.

The 3 major steps in in network analysis are:

1) Identification of genes or proteins of interest
    a) A. Data processing step to identify significant genes The steps here include:
        i) Data formats and uploading INPUT: list of genes/proteins IDs uploaded as a .txt file.
        ii) Data processing and annotation: Includes conversion of feature ID to common ID [4]
        iii) Data normalization and analysis Data can be further normalized to log2 scale (microarray) or log2 counts per million (RNA-seq).
2) Network Construction step for mapping, building and refining networks
   The inputs(seed proteins) are used to search and retrieve binary interactions from a curated PPI database. For each protein, a search algorithm is performed to identify all the proteins that directly interact with seed proteins (first-order interactors). The output of this is used to build the default networks and will return one large subnetwork called continent with several smaller subnetworks called islands.
3) Network Analysis and Visualization step
    a) Network analysis: it has two approaches that are complementary
        i) The Topology Analysis
           This considers the whole network structure to search for important nodes which are useful as biomarkers.

ii) Module Analysis
This method breaks the complex network into small densely connected modules or units and aims to identify the ones showing more active hotspots. The results from network analysis will next be visually inspected and validated by processes like GO or pathway enrichment analysis.

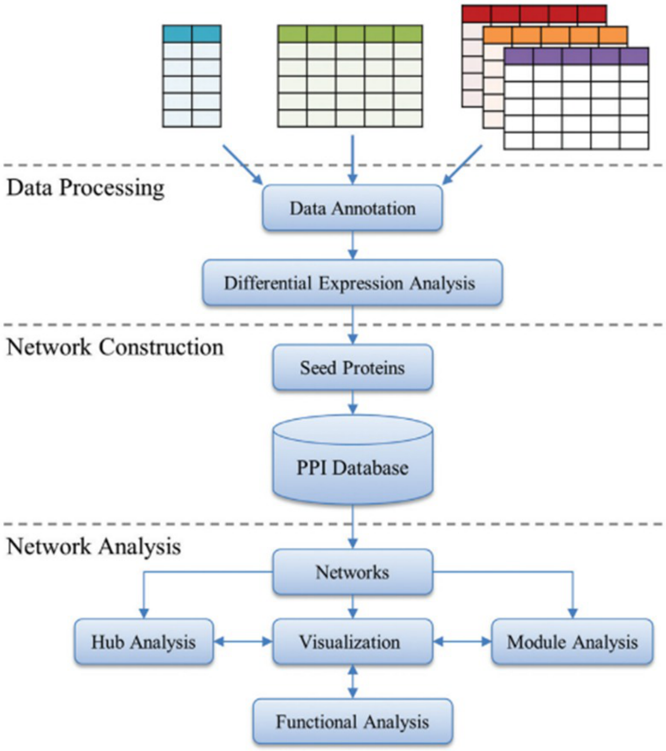
b) Visualization It has 4 different explorers

i) Network Explorer
This module shows all networks constructed from the seed proteins. The numbers of nodes, edges and seed proteins are summarized for each network.

ii) The Module Explorer This module is used to decompose the current network into smaller densely connected modules, which are considered as relatively independent components such as pathways or protein complexes using the WalkTrap algorithm and weighted network for module detection with edge weights

iii) Path Explorer
This module is used to detect the shortest paths between any two given nodes.

iv) Function Explorer
This module is used to perform functional enrichment analysis for currently highlighted nodes using different databases such as the GO, KEGG



This paper tells us that even though tools like Cytoscape is very powerful, it can work only for standalone systems or embedded applications due to it's inabilty to maintain compatabilty between different versions of the plugins. Hence. the requirement of tools like NetworkAnalyst.

E. Analyzing biological network parameters with CentiScaPe [5]

The large amount of available experimental data that produces annotated gene or protein complex networks has spurred the need for network analysis. Biological networks are usually represented in the form of graphs. The nodes of this graph are biological entities (such as cells, genes, proteins or metabolites) and the edges are functional and/or physical interactions between them. [5]

Now, there is a big need for Visualization and analysis tools to understand individual node functions which can often get masked by the overall network complexity of such graphs.

There are various methods suitable for network structure analysis, such as global network structure, network motif, network clustering, and network centrality analysis. Particularly, centralities are node parameters that can identify nodes having a relevant position in the overall network architecture. [5]

Cytoscape is an excellent visualization and analysis tool with the analysis features greatly enhanced by plug-ins. [5]. Other Plug-ins compute some node centralities but does not allow direct integration with experimental data. Even if they do compute these node centralities, they compute a very few of them.

CentiScaPe is a versatile and user-friendly bioinformatics tool for integrating centrality-based network analysis with experimental data. CentiScaPe is fully integrated with Cytoscape, and the ability to treat centrality as a regular attribute allows you to enhance your analysis with Cytoscape core features and other Cytoscape plugins.

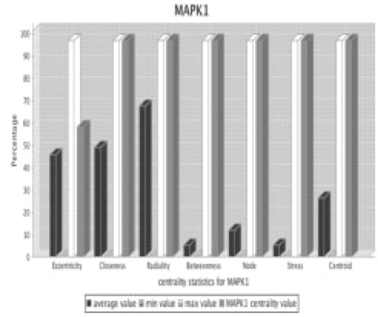
What makes CentiScaPe stand out is it's capability to compute several centralities at once. With CentiScaPe, you can easily correlate the calculated centrality with each other or with the biological parameters obtained from the experiment to identify the most important nodes according to both topological and biological characteristics. The output is a scatter plot with value options. This makes it easy to correlate node centrality values with user-defined experimental data.

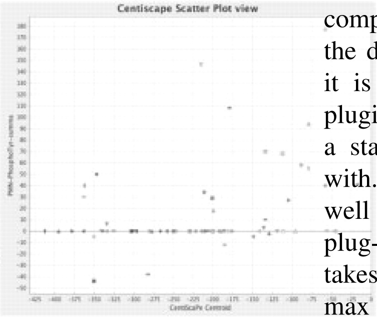
CentiScaPe can calculate multiple network centralities only for bidirectional networks. These parameters include: Stress, Closeness, Degree, Radiality, Betweenness, Centroid Value, Diameter, Average Distance and Eccentricity.

The approach is to compute Min, max and mean values for each computed centrality. The system also supports Multiple networks. The centrality values are displayed in the Cytoscape attribute browser and can be saved and loaded as a regular attribute for visualization using the Cytoscape mapping core function. After this, the actual analysis begins with the help of the graphical interface provided by the CentiScaPe.

There are two types of graphical outputs: plot by centrality and plot by node. Both enable analysis not possible with other central tools. Users can correlate centrality with each other or with experimental data to analyze all centrality values.

The plot by centrality visualization is a simple and convenient way to distinguish the most relevant nodes and node groups according to the combination of the two parameters you choose. It shows the correlation between centrality and / or other quantitative node attributes such as experimental data from genomic and / or proteome analysis. The result of the Plot by Centrality option is a diagram in which the individual nodes represented by the geometry are mapped to the Cartesian axis. The value of the selected attribute is displayed on the horizontal and vertical axes. Most of the related nodes can be easily identified in the upper right quadrant of the figure.





In plot by node option, the per-node plotting option, displays all centrality values calculated for each individual node in the form of a bar graph. The average, maximum, and minimum values are displayed in different colors. For ease of visualization, all values in the figure are normalized and the actual values are displayed when you hover your mouse over the bar.

Thus, by iterating in a similar manner for all centralities, CentiScaPe helps extract the most relevant nodes, selecting all nodes having all centrality values over the average.

F. Summary

Biological networks are of different types like protein protein interactions , gene regulatory networks etc and graph theory can be used to represent these networks,by using deep learning We have explored graph neural networks like graph embedding and graph convolutional networks [2] for analysing these complex biological networks and gain insights from it. Next we have looked into the the data visualisation which faces challenges of exponential increase in data,integration of heterogeneous data and the representation of multiple connections between nodes with heterogeneous biological meanings.We saw that different tools having specific functionalities can be used to address these challenges.The three steps required for biological network analysis to identify patterns of gene expressions are identification of genes or proteins which includes data formating, normalization, processing and annotations, network construction which includes mapping, building and refining networks network analysis and visualization which includes two approaches namely topological and module analysis. Different types of explores like Path, Network, Function are also used.Now coming to tools we explored CentiScaPe is a versatile and user-friendly bioinformatic tool to integrate centrality-based network analysis with experimental data. CentiScaPe is completely integrated into Cytoscape and the possibility of treating centralities as normal attributes permits to enrich the analysis with the Cytoscape core features and with other Cytoscape plug-ins, and finally To understand working of different biological networks We have seen different applications like Proteomics, Drug development, discovery , polypharmacy and Disease diagnosis.

III. Discussion and Analysis

A. Conclusion

Usage of graphs and deep learning networks [2] are advantageous because we are dealing with heterogeneous and complex data however representation via graph theory has the drawback that this assumes the system to be static while it is highly dynamic in nature [1]. Cytoscape [4] and its plugins are powerful but the visualisations are present on a standalone system and cannot be shared or collaborated with.For a deeper functional exploration of the sub-network as well as to support Multiple networks analysis , an additional plug-in named CentiScaPe [5] has been introduced which takes into consideration nodal centralities by computing min, max and mean values for each computed centrality. The plots obtained give meaningful results not accessible to other tools and allow easy categorization of nodes in large complex networks derived from experimental data.

References

[1] G. A. Pavlopoulos, M. Secrier, C. N. Moschopoulos, T. G. Soldatos, S. Kossida, J. Aerts, R. Schneider, and P. G. Bagos, “Using graph theory to analyze biological networks,” BioData Mining, vol. 4, no. 1, 2011.

[2] G. Muzio, L. O’Bray, and K. Borgwardt, “Biological network analysis with Deep Learning,” Briefings in Bioinformatics, vol. 22, no. 2, pp. 1515–1530, 2020.

[3] G. A. Pavlopoulos, A.-L. Wegener, and R. Schneider, “A survey of visualization tools for biological network analysis,” BioData Mining, vol. 1, no. 1, 2008.

[4] Jianguo Xia, Maia J. Benner, Robert E. W. Hancock, NetworkAnalyst - integrative approaches for protein–protein interaction network analysis and visual exploration, Nucleic Acids Research, Volume 42, Issue W1, 1 July 2014, Pages W167–W174, https://doi.org/10.1093/nar/gku443

[5] Giovanni Scardoni, Michele Petterlini, Carlo Laudanna, Analyzing biological network parameters with CentiScaPe, Bioinformatics, Volume 25, Issue 21, 1 November 2009, Pages 2857–2859, https://doi.org/10.1093/bioinformatics/btp517