Using biological networks to integrate, visualize and analyze gene-disease interactions

Hamza Hanafi[1,*], *Badr Dine Rossi Hassani*[2], *and M'hamed Aït Kbir* [1]

[1]LIST Laboratory, UAE University, Morocco
[2]LABIPHABE Laboratory, UAE University, Morocco

Abstract. Nowadays, data integration methods have been widely used to build models and to represent interactions between the data. They are showing high efficiency. Recent technologies permitted the research community to perform complex analysis on cell structures and it's functioning system. The tremendous amount of data collected from a biological system encouraged the exploration of new hypothesis. However, the manipulation of heterogenous data require additional efforts to find the model that handles perfectly data of different type. In this paper we present our method to create a unified model and to integrate gene-disease interactions. We will talk about stat of the art methods in data integration, and how we built our network based on omics layers. Moreover, we will present the overall framework we followed to extract important interactions by visually interpreting the generated graph, and the betweenness centrality of nodes. We compared our findings to the medical literature to explain the topology of our generated network. Some genes revealed as important nodes due to the fact holding many interactions and being connected to several syndromes.

1 Introduction

A biological system is a complex structure composed of distinctive molecules. Recent advances in biology stimulate the extraction of diverse functional aspects of cells which are represented in omics layers. This allowed scientists to better understand chemical reactions and the functional relations between the different components in a biological system. In addition, it supplies the analysis of the different interactions among Omics layers. Numerous technological tools; essentially microarrays, RNA sequencing tools [1–9], played a valuable role in the construction of various types of omics layers. For example, the genes in a genome layer show all the interactions between gene-gene entities it also describes outer interactions with the phenotype layer. A join analysis of separate Omics layers is essential to understand the functional interactions.

Networks allowed researchers to represent aspects of cells in a mathematical structure. It is a powerful tool that shows how different components interact with each other to accomplish their functional tasks. Nodes and edges are the essential units of a network. Nodes can be of diverse types, and edges can be directed or undirected, and used to link multiple nodes within a graph. Additionally, edges may be weighted to represent a specific aspect of the interaction. biological networks empowered studies of biological aspects from a system level perspective. Some networks handle information's about evolution and interactions of species; as example of these networks the protein-protein interactions network (PPIs) which convey data about how proteins collaborate together to perform their activity inside a cell. A commonly used repositories that stores PPI networks are BioGRID, MINT, BIND [10–13]. Sequence similarity networks (SSNs) are another type of biological networks which involve entities about proteins or genes, and the links representing the sequence similarity among nucleotide sequences. most powerful techniques used to obtain the sequence similarity are the LAST, BLAST and FASTA3 algorithms [14–16]. Metabolic interaction networks (MI) carry biochemical activities that contribute in the conversion of a metabolite to another [17]. Epidemiological networks are manipulated in public health to better study the spread of diseases (e.g., Covid-19) [18,19].

2 Data integration

Data integration methods handle diverse data coming from diverse sources to create a unified model. In the field of biology data integration can be used to generate a unified model for biological networks based on disparate Omics layers [27,28]. In network analysis, data integration can be divided into homogenous integration which deals with networks having same type of nodes but edges are of diverse types. Or heterogenous integration which concerns networks with diverse type of nodes and edges. Some examples of methods used in data integration are Network-based methods; which are very simple and used most of the time. A commons way they use to build an integrated graph is to merge edges of

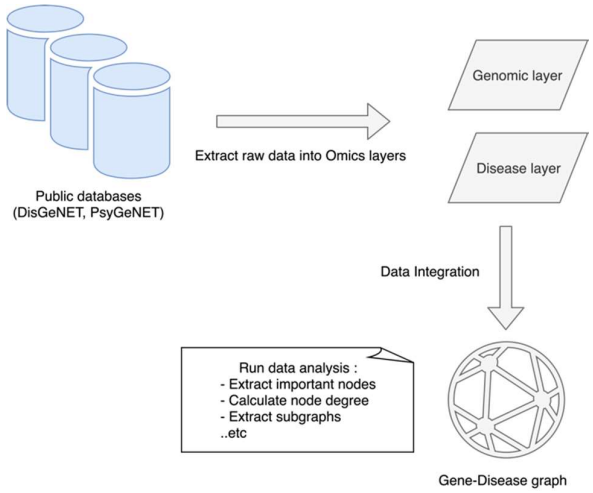
* Corresponding author: hamzahanafi1@gmail.com

all networks to the same set on nodes [20]. Other approaches rely on the adjacency matrix; by creating a weighted sum of the adjacency matrix of the wanted networks. Bayesian networks methods are founded on the concept of probabilities and graph theory; usually entities are random variables while the links are directed and show conditional probabilities among the variables [21]. Kernel based methods which are more sophisticated and complex. They go to the category of machine learning approaches and represent the data to a feature space to be then represented by a kernel matrix [22].

Our methodology relies on network-based methods to construct an integrated network. Figure 1 shows the overall framework we followed. We first inferred our data from different public data sources manly DisGeNET database, we oriented our data collection to sources containing curated human genes and variant disease interactions. In addition, the DisGeNET database offers some useful metrics about gene-disease interactions; such as scoring the interactions based on their evidence level, which enables ranking the interactions from strong to weak associations. The computed score is calculated for (GDA) gene-disease association as well as (VDA) variant disease association. We then represented the mined data into two Omics layers which are genomic layer and disease layer. The genomic layer shows all the characteristics of genes. It identifies the functional aspects of a gene alongside its relationships with other genes and diseases. The disease layer identifies phenotypes of a disease together with its linked diseases.

To build an integrated network that shows the interactions between genes and diseases we relied on the standard identifiers available in the public data sources to correctly map the data. Additionally the biodb.jp helped us to find the hyperlink among major biological databases when no identifier is found.

Fig. 1. Overall framework we followed to build Gene-Disease interactions graph.



3 Biological graph representation

There is several metrics of a graph that play an important role to better understand the data. Measures such as: the degree of a node, the distance between entities, and clustering coefficient can provide new insights and help researchers to make new hypothesis. For example, nodes with a higher degree connection should be given higher attention and be more studied in particular. Likewise, the diseases related to multiple genes may indicate a potential interaction between those genes. The topology of a network interests several measures; mainly node degree, centrality, betweenness, cluster coefficient, and shortest path. Additionally, the modularity which deals with the identification of clusters that may denotes molecules functionally linked to perform a certain task.

Graph algorithms are analytical tools used to analyze the interactions and establish the strength between entities that we manipulate. They also enable a pairwise relationship among the entities and the structural characteristics of a graph. In our work we have used the vis.js network javascript library (https://visjs.org/). It is a powerful tool to render a network for up to a few thousand nodes and edges. It handles a large amount of dynamic data, and facilitate the interaction and the manipulation of the data. The visualization supports custom shapes, styles, colors, sizes, and images when rendering a network. It also supports clustering to handle a large number of nodes, Moreover, it supplies APIs (Application Programming Interface) to handle the data with its hierarchical positioning. The network will use a force-directed drawing algorithm, named Kamada Kawai for initial layout. Its aim is to place the entities of a network in a dimensional space in a manner that all links are of more or less equal length, and there are as few crossing edges as possible. We interested to path analysis to establish the shortest distance between two nodes, and to connectivity analysis which helps to determine the strength of interconnected nodes; additionally, centrality analysis which helps to estimate the importance of a present node within the graph and its connectivity to others. This helps understand the most influential node and the connection it accesses.

4 Results and discussion

We've been able to collect a set of interactions between genes and diseases from public databases mainly DisGeNET. It stores curated human being gene-disease interactions. The process of collection stores the mined data into different Omics layers based on their type and structure. Then the interactions from each layer are mapped into a single set of nodes to create our gene-disease interactions; which we named GDI graph.

Fig. 2. Gene-Disease interactions (GDI).

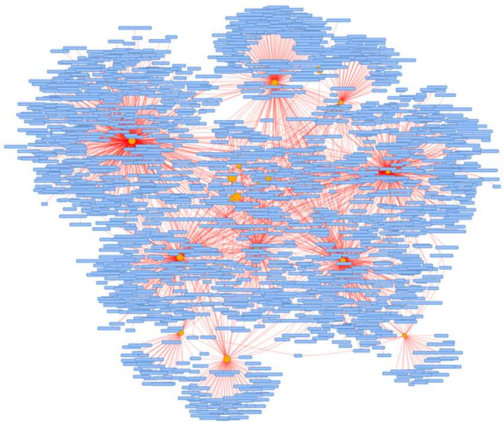
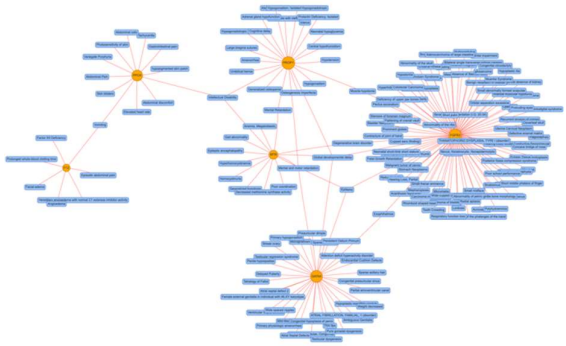


Figure 2 shows our GDI graph, genes are represented by orange circles while diseases are blue boxes. A link is drawn between a gene and a disease when an interaction is identified between gene-disease entities. Also, the size of a gene node is reflected by the in-degree node; so that big nodes have higher node degree. The visualization of the network is performed with the vis.js library; At first the visualized networks is very dense; however, we can easily determine some sub graphs inside it. Besides some nodes are being highly connected than others which may present some interconnected sub graphs within the GDI graph. We interested to the betweenness centrality to provide significant insights about the manipulated data, and helps extract the important nodes within our network. Graph analytics are hence easier to work with than the traditional techniques being used. Just by representing the data in a graph we could have insights about the interactions. Laying out the graph is a crucial stage in network analysis that makes the graph more intelligible and easier to understand, our main goal of these analysis is to show that the use of graphs can help to extract important interactions between genes and disease. Our network is being very large and complex, for that reason we provided Figure 3, it shows another version for our GDI network which is zooming into some important nodes that we are going to talk about and explain.

Fig. 3. A light version of our Gene-Disease interactions (GDI) network, zooming into some important nodes (e.g., Epilepsy, FGFR3 …).



Within our GDI graph that we represented in Figure 3; it is using the force-directed graph drawing algorithm. I is very important to better visualize the network to be able to understand the functional interactions between the entities, we have been able to visually determine some important nodes like the gene FGFR3 which have the higher connectivity degree, about 2130 edges connected to it. In addition, the diseases related to FGFR3 (fibroblast growth factor receptor 3) are mostly syndromes that involve thanatophoric dysplasia, achondroplasia, and hypochondroplasia [23]. Another important gene in our graph is the GATA4 gene which also have the second higher node degree with 487 links. It has been reported that a heterozygous G296S missense mutation of GATA4 is responsible of atrial and ventricular septal defects and also pulmonary valve stenosis in humans. GATA4 encodes a cardiac transcription factor, which may result in cardiac bifida and lethality by embryonic day (E)9.5 [24] when deleted in mice. We should also mention that the network holds some disease nodes that are linked to several genes; as an example Epilepsy, It is a neurological disorder that is related to abnormal activity in brain, causing seizures or periods of unusual behavior, sensations, and sometimes loss of awareness, additionally to other phenotypes. The causes are in some cases genetic, but in most cases they are not identified [25], within our GDI graph we found that the Epilepsy is linked to several genes such as FGFR3, GATA4 and MTR genes. Therefore these important nodes should be further investigated. Jie Wang. et al in their research article [25] about finding Epilepsy associated genes found 977 genes related to Epilepsy which they grouped into categories based on the manifestation of epilepsy' phenotype. They also discovered it is associated to the genes we found in our GDI graph. Therefor further attention should be provided to the associated genes due to the fact of their presence in multiple diseases. It has been known that activating mutations in the FGFR3 gene are responsible for several autosomal dominant syndromes; which have been found to be present in cancer. Moreover, CATSHL syndrome, characterized by pamprodactyly, tall stature and hearing loss, is caused by loss-of-function mutations of FGFR3 gene [26]. Based on the medical literature we could explain why the FGFR3 gene has several connections with different syndromes and diseases inside our GDI graph; which shows the importance of graphs for the analysis of such interactions.

5 Conclusion

In this article we explained how network analysis help the researcher community to extract hidden information's from gene-disease interactions. We first used data mining techniques to collect gene-disease interactions from multiple sources. We represented the different interactions into Omics layers. Using data integrations methods, we have been able to build a network of interactions between genes and diseases. by representing the network, we successfully extracted some important nodes, mainly genes with a high node degree like FGFR3 and GATA4 which should be further investigated.

References

1. R. Hawkins, G. Hon, B. Ren, *Next-generation genomics: an integrative approach*. Nat. Rev. Genet, **11**, 476–486 (2010)
2. R. Nielsen, J. Paul, A. Albrechtsen, Y. Song, Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. **12**, 443–451 (2011)
3. J. Hirschhorn, M. Daly, *Genome-wide association studies for common diseases and complex traits*. Nat. Rev. Genet. **6**, 95–108 (2005)
4. R. Duerr R, et al, *A genome-wide association study identifies IL23R as an inflammatory bowel disease gene*, Science **314**, 1461–1463 (2006)

5. J. Quackenbush, *Computational analysis of microarray data*. Nat. Rev. Genet. **2**, 418–427 (2001)
6. K. Dahlquist, et al, *GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways*. Nat. Genet. **31**, 19–20 (2002)
7. J. Marioni, et al, *RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays*. Genome Res. **18**, 1509–1517 (2008)
8. A. Mortazavi, B. Williams, K. McCue, L. Schaeffer, B. Wold, *Mapping and quantifying mammalian transcriptomes by RNA-seq* Nat Methods. **7**, 621–628 (2008)
9. Z. Wang, M. Gerstein, M. Snyder, *RNA-seq: arevolutionary tool for transcriptomics. Nat. Rev.Genet*. **10**, 57–63 (2009).
10. V. Rao, K. Srinivas, G. Sujini, and G. Kumar, *Protein-protein interaction detection: methods analysis*. Int. J. Proteomics 2014
11. C. Stark, et al, *BioGRID: a general repository for interaction datasets*. Nucleic Acids Res. **34**, D535–D539, (2006)
12. A. Chatraryamontri, et al, *MINT: the molecular INTerac-tion database*. Nucleic Acids Res. **35**, D572–D574, (2007)
13. G. Bader, D. Betel, C. Hogue, *BIND: the biomolecular interaction network database*. Nucleic Acids Res. **31**, 248–250, (2003)
14. S. Altschul, W. Gish, W. Miller, E. Myers, D. Lipman, *Basic local alignment search tool*. J. Mol. Biol. **215**, 403–410, (1990)
15. S. Kielbasa, R. Wan, K. Sato, P. Horton, M. Frith, *Adaptive seeds tame genomic sequence comparison*. Genome Res. **21**, 487–493, (2011)
16. W. Pearson, *Flexible sequence similarity searching with the FASTA3 program package*. Methods Mol. Biol. **132**, 185–219, (2000)
17. W. Reisig, Petri Nets: An Introduction. Berlin, NY: Springer-Verlag. 161 (EATCS monographs on theoretical computer science), (1985)
18. L. Danon, et al, *Networks and the epidemiology of infectious disease*. Interdis-cip. Perspect Infect. Dis. (2011), 1–28.
19. D. P. Croft, J. Krause, R. James, *Social networks in the guppy (poecilia reticulata)*. Proc. Biol. Sci. **271**, S516–S519, (2004)
20. J. Dutkowski, M. Kramer, Surma, R. Balakrishnan, J. Cherry, N. Krogan, T. Ideker, *A gene ontology inferred from molecular networks*. Nat. Biotechnol. **31**, 38–45 (2013).
21. I. BenGal, Bayesian networks, NewYork, NY: JohnWiley & Sons, Ltd (2008)
22. B. Scholkopf, K. Tsuda, J. Vert, *Kernel methods in computational biology*. Cambridge, MA: MIT Press (2004).
23. A. Joshua, E. Mary, H. Feltovich, E. Gratacós, D. Krakow, O. Anthony, D. Lawrence, B. Tutschek, FGFR3 Disorders: Thanatophoric Dysplasia, Achondroplasia, and Hypochondroplasia. Fetal Diagnosis and Care (Second Edition), Elsevier, **50**, 264-267, (2018)
24. C. Misra, N. Sachan, C. McNally, S. Koenig, H. Nichols, A. Guggilam, P. Lucchesi, W. Pu, D. Srivastava, V. Garg, *Congenital heart disease-causing Gata4 mutation displays functional deficits in vivo*. PLoS genetics, **8**, (2012),
25. J. Wang, et al. Epilepsy-associated genes. Seizure. Epub 2016
26. X. Sun, et al, Fgfr3 mutation disrupts chondrogenesis and bone ossification in zebrafish model mimicking CATSHL syndrome partially via enhanced Wnt/β-catenin signaling. Theranostics, **10**, 7111–7130, (2020)
27. I. Subramanian, et al, *Multi-omics Data Integration, Interpretation, and Its Application*. Bioinformatics and biology insights, **14,** (2020)
28. J. Yan, S. Risacher, L. Shen, A. Saykin, *Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data*, Briefings in bioinformatics, **19**, 1370–1381, (2018)