Integration of bio-medical information in a multimodal complex network for gene-disease prioritization.

Ingrid Heuer[1] and Ariel Chernomoretz[1,2]

[1]Physics Department, FCEN, University of Buenos Aires
[2]IFIBA (CONICET)

Keywords: gene-disease networks, data integration, disease-disease networks

One of the biggest challenges in current biomedical research is trying to bridge the gap between the different scales of organization that coexist in an organism, solving what is known as phenotype-genotype relationship. Ultimately, this translates into finding the molecular basis of different biological functions or pathologies and diseases. Identifying relationships between disease phenotypes and genetic alterations is essential to better understand disease etiology and to improve genome-based diagnostics. However, experimental methods designed to find these associations can be expensive and time consuming. To address this challenge, we took advantage of the vast amount of available biomedical data and integrated a multimodal complex network that could be used to prioritize novel gene-disease associations.

In this work, we construct and analyze a multimodal biomedical knowledge graph that contains data about gene-disease associations, complemented with protein-protein interactions, biological pathways , disease ontology relationships and natural language descriptions of the involved diseases. This multimodal network integrates quality resources such as DisGeNET [1], HIPPIE [2], PrimeKG [3], Reactome [4] and Signor [5] (table 1). The integrated network consists of 5 types of nodes organized in two layers: a disease layer and a gene/protein layer. The two layers are connected by gene-disease associations (fig. 1a). We found that 99% of nodes of our multimodal network were at less than three hops away from a node of the complementary layer.

We considered two disease ontologies in our network, MONDO [6] and UMLS[7], which we combined using vocabulary mapping. The integration of knowledge embedded in disease ontologies is a challenging task, since the definition of a unique disease is ambiguous and often inconsistent between databases. To address this issue, we incorporated BERT-group nodes in our network, which are disease concept groups obtained using the ClinicalBERT natural language processing model [3].

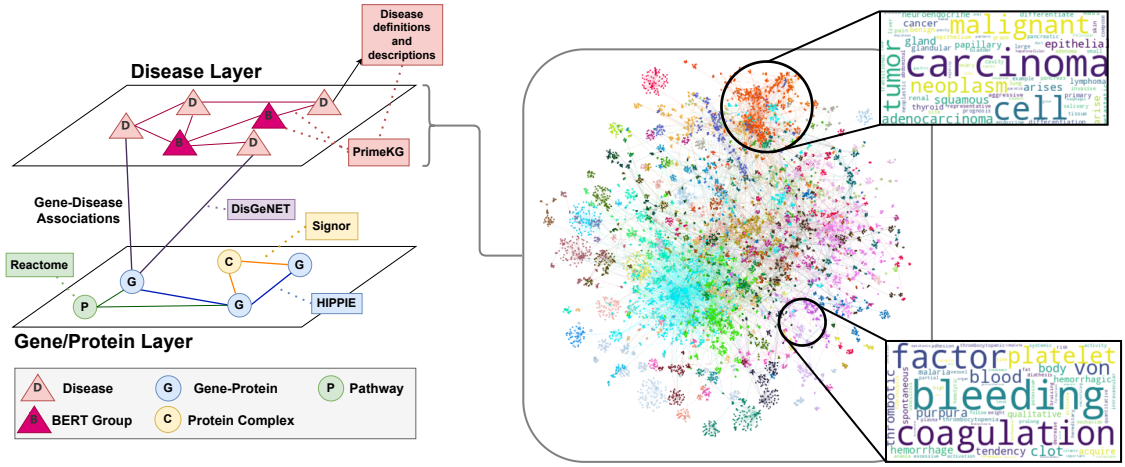
In order to probe the disease layer, we studied its mesoscale structure at two different resolutions considering two community detection algorithms: Infomap and Louvain (fig 1b). In each case we characterized the detected communities in terms of the homogeneity of their components considering two metrics. The first one was based on shared gene associations between diseases whereas the second one considered the semantic similarity of disease nodes inferred from a TF-IDF analysis of their natural language descriptions. To that end we used a measure of semantic specificity, $Spec$, based on the entropy of the TF-IDF distributions associated with descriptions of the components of a community:

$$Spec_j = 1 - H_j = 1 + \sum_{i=1}^{N} p_{i,j} \log p_{i,j} \tag{1}$$

where N is the number of terms in the network corpus and $p_{i,j}$ is the TF-IDF value of the term i in the community j. We compared the semantic specificity of these communities with a randomly generated control sample and found that the communities detected by both algorithms showed significant semantic specificity (fig 2).

We also studied the structural role of BERT-group nodes in the disease layer. Using set similarity metrics, we compared disease communities with groups of nodes that belong to BERT-groups, and found that communities tend to form around BERT nodes. To further understand the role of disease group nodes, we characterized them in terms of their participation coefficient and within-module degree [8]. We saw that BERT-groups tend to have a connective role within their communities and a non-zero participation coefficient, which indicates that they act as module connectors.

Overall, we were able to build a bio-medical network integrating more than 10000 diseases and 84000 high confidence gene-disease associations with protein-protein interaction and biological pathway information. In particular we analyzed different features of the disease layer and found that the observed connectivity patterns could provide a meaningful scaffold to implement message passing algorithms for link prediction and prioritization tasks.



(a) Multimodal network diagram

(b) Disease layer

Figure 1: (a) Diagram of the integrated multimodal network: The network is organized in two layers: a disease layer and a gene/protein layer. The disease layer contains disease nodes and BERT-group nodes. The gene/protein layer contains gene/protein, pathways and protein complex nodes.(b) Disease layer: Colors indicate communities detected with the Louvain algorithm. Wordcloud examples show the most relevant terms associated with a community, which were extracted from the descriptions of the diseases that belong to that community using a TF-IDF approach.

Node Type	Number	Source
Disease	15766	DisGeNET
BERT-Group	1067	PrimeKG
Gene/Protein	17363	DisGeNET-HIPPIE
Complex	422	Signor
Pathway	2020	Reactome
Total	36638	

Edge Type	Number	Source
Gene-Disease Association	84038	DisGeNET
Disease-Disease	17488	PrimeKG-MONDO
Protein-Protein Interaction	110062	HIPPIE
Pathway-Protein	42646	Reactome
Protein-forms-Complex	1888	Signor
Total	256122	

Table 1: Number of nodes and edges in the network, node/edge type and source database for each type.

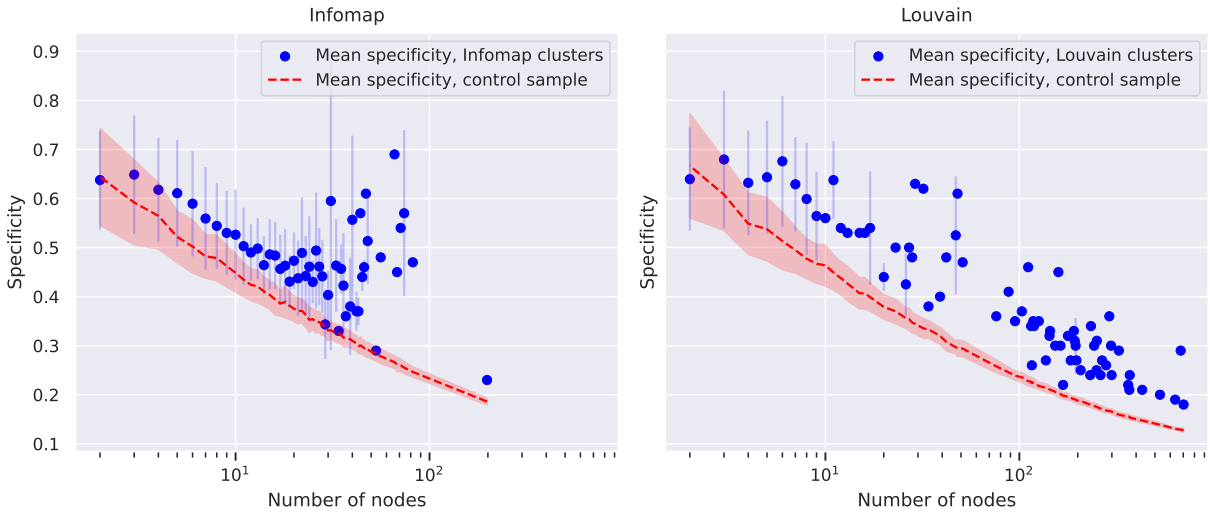


Figure 2: Semantic specificity of communities in the disease layer. Blue markers show the mean specificity of groups of communities of the same size. We compared this metric with a randomly generated sample of communities, shown in red. We found that the communities detected by both algorithms showed significant semantic specificity

References

[1] Janet Piñero et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1):D845–D855, 11 2019.

[2] Alanis-Lobato et al. HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Research*, 45(D1):D408–D414, 10 2016.

[3] Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *bioRxiv*, 2022.

[4] Marc Gillespie et al. The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50(D1):D687–D692, 11 2021.

[5] Luana Licata et al. SIGNOR 2.0, the SIGnaling Network Open Resource 2.0: 2019 update. *Nucleic Acids Research*, 48(D1):D504–D510, 10 2019.

[6] Christopher J. Mungall et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research*, 45(D1):D712–D722, 11 2016.

[7] O. Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(90001):267D–270, January 2004.

[8] Roger Guimerà and Luís A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, February 2005.