Contents

1 Introduction

1.1 Background

1.1.1 Heterogeneous Information Network

Heterogeneous information networks (HINs), such as bibliographical datasets, are widely used and discussed in the field of data mining [1, 2]. Nodes of HINs are labeled, providing more abundant semantic meanings than unlabeled graphs [3]. Compared to homogeneous information networks, HINs distinguish different types of nodes and edges in the networks, consisting of rich semantic meanings of structural types of nodes [4].

1.1.2 Motif

A motif is essentially a small subgraph pattern, which is a foundational building block of complex HINs [5, 6]. Also known as higher-order structure, motif provides a tool to discover higher-order semantics of HINs [7]. It is widely used in graph analysis problems, such as graph clustering [8, 9], social network analysis [10].

1.1.3 Clique

A clique is by definition a complete graph, i.e., every two nodes in the clique are adjacent. Thus, a clique represents a set of nodes that are closely relevant (e.g., a clique in a social network can represent a group of close friends). Cliques have been widely studied in both research and industry communities. Usages of cliques include social network detection [10], gene group detection [11], and transportation network analysis [12]. A maximal clique is a clique that is not a subgraph of any larger clique.

1.1.4 Motif-clique

Hu et al. [7] proposed a new concept, namely motif-clique or m-clique in short, which incorporates motifs to the clique definition. Recall a clique is a complete graph based purely on edges, i.e. it is complete since every two distinct vertices are connected by an edge. A motif-clique, as a generalization of a traditional clique, is a complete graph based on a user-defined pattern, i.e. motif, rather than edges. An m-clique is, therefore, a *higher-order* clique based on a user-given motif. Compared to traditional cliques, which treats nodes with different labels equally, an m-clique can capture the desired relationship among labeled nodes.

2 Methodology

We developed a fully functional web application for motif-clique discovery. Afterwards, we utilized the platform to conduct case studies using bioinformatic datasets. Bioinformatics, as an interdisciplinary field, focuses on understanding biological data, such as identification of genes [17]. Motif-clique could be used for discovering the relationship between genes and human diseases. Common bioinformatics datasets, including DisGeNET [13], Reactome

Pathway Database [14], NCBI E-utilities [15], DrugBank [16], and OPHID [18] would be utilized.

However, different datasets are in different formats. For example, as demonstrated in table 1 and 2, DisGeNET dataset and OPHID dataset have different schemas.

geneId	geneSymbol	diseaseId	diseaseName	score	source
10	NAT2	C0005695	Bladder Neoplasm	0.25	CTD_human

Table 1: Sample of DisGeNET

Dataset	SwissProt1	SwissProt2
SOURAV_MAPK_LOW	P63000	A0AUZ9

Table 2: Sample of OPHID

Consequently, those datasets cannot be combined unless they have compatible formats. Therefore, python would be used to clean and preprocess the datasets, merge them and import into the web platform. Python is a scripting language which is widely used as data processing and analysis tool due to its high readability. Useful data would be extracted from different datasets and merged into a single dataset so that researchers could see combined results.

3 Results

In the genetic world, genes are related to various kinds of diseases and drugs. These relationships can be represented by graphs, where an edge between two nodes represents some relationship. An edge between a gene and a drug usually means the drug has some effect on that particular gene. An edge between a gene and a disease usually means the gene is one of the causes of that specific disease.

3.1 Disease Subtyping

Disease subtyping is a new field which aims to discover hidden intrinsic characteristics of drugs and find subtypes of medicines. There is much medical research focusing on determining subtypes of certain diseases, but to our best knowledge, there is no generic framework that can be used to find subtypes for all illnesses. To resolve this open problem, we design suitable motifs and utilize the platform to study disease subtyping.

In a graph consisting of diseases, genes, and drugs, we would like to find subgraphs that conform to the following pattern:

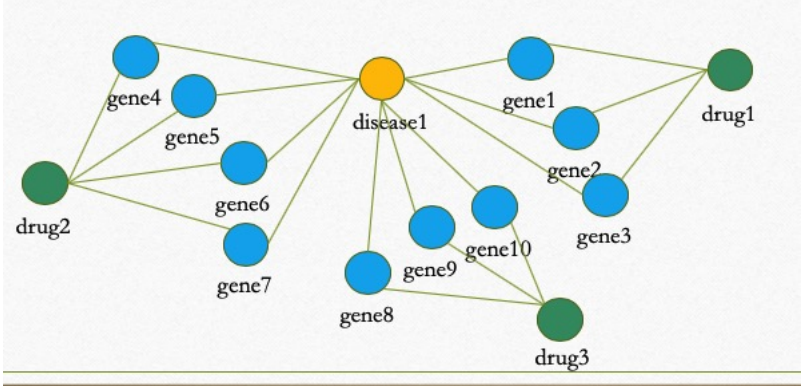

Figure 1: Target pattern

As Figure 1 shows, a disease is connected to a cluster of genes. We can classify these genes into three groups, i.e., gene 1-3, gene 4-7, and gene 8-10 if we consider genes connected to the same drug as members in the same group. Consequently, we can infer that the disease can be classified into three subtypes, and each subtype can be cured with a unique drug.

We designed a simple but useful motif: drug - gene - disease (Figure 2), and combine results with same disease node.

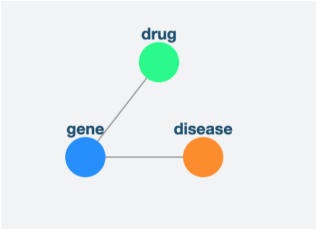

Figure 2: Drug - Gene - Disease motif

Cancer is an important research area in the bioinformatics field. It is believed that the complexity of cancer can be reduced to a small number of underlying principles [24]. Later, the classification develops and currently, hallmarks of cancer include ten types.

We try to use our platform to see if our findings are consistent with the subtypes concluded by the medical experts. We incorporate DisGeNet [13], DrugBank [16], and GeneOntology [26] [27] into a single dataset. Then we design a drug - gene - disease motif, and utilize the platform to search motif-cliques based on the disease breast carcinoma. Motif-cliques are combined based on the common disease, i.e. breast carcinoma. We find that 7 out of 10 hallmarks of breast cancer have corresponding motif-cliques. These 7 motif-cliques are extracted and included in Figure 4.

Moreover, we use disease - gene motif to investigate all genes related to breast carcinoma, and include genes related to the other three subtypes in figure 10 likewise. These genes include CTLA4 which is related to 'Avoiding immune destruction' hallmark, TERT which is related to 'Enabling replicative immortality' hallmark, PARP1 which is related to 'Genome instability and mutation' hallmark.

Out of ten hallmarks of cancer, our work shows that breast carcinoma has ten subtypes,

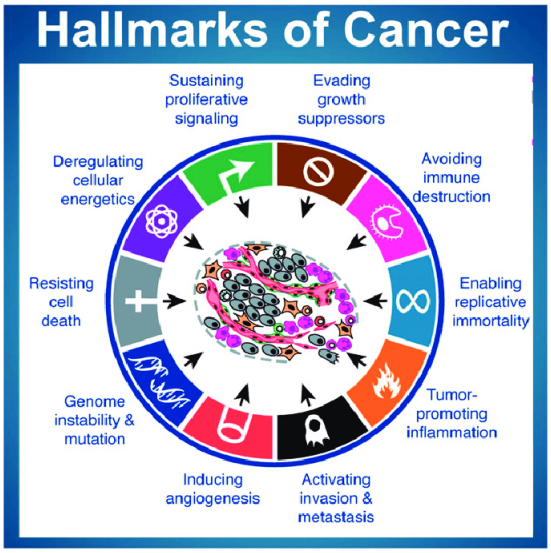

Figure 3: Hallmarks of cancer [25]

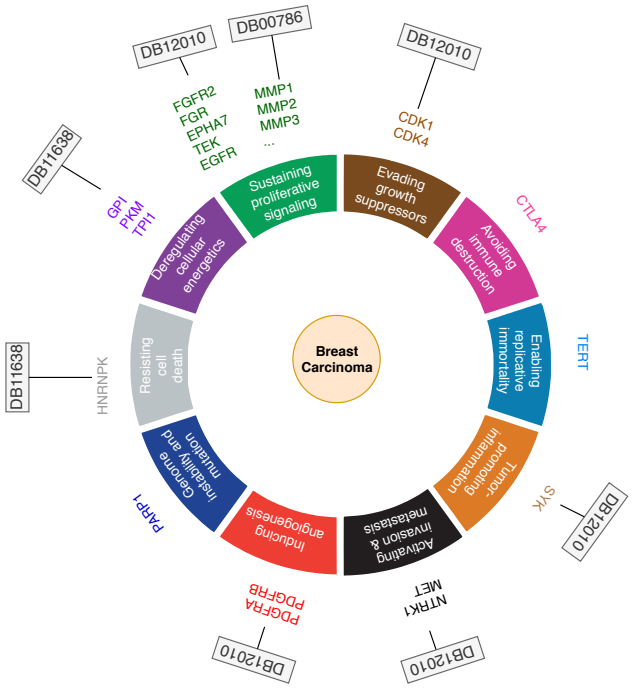

Figure 4: Breast carcinoma subtypes found

seven out of them have existing drugs. For example, one maximal motif-clique found is DB00786(Marimastat) - genes{MMP1, MMP2, MMP3, MMP7, MMP8, MMP9, MMP10, MMP11, MMP12, MMP13, MMP14, MMP17, MMP25, MMP36} - C0678222(Breast Carcinoma), as shown in Figure 5. These genes are metalloproteinases (MMP), and the inhibition of MMP leads to lowered TGFalpha and EGFR expression, i.e, the drug is targeting the sustaining proliferative signaling.

However, not every motif-clique can be classified into a subtype. For example, we have

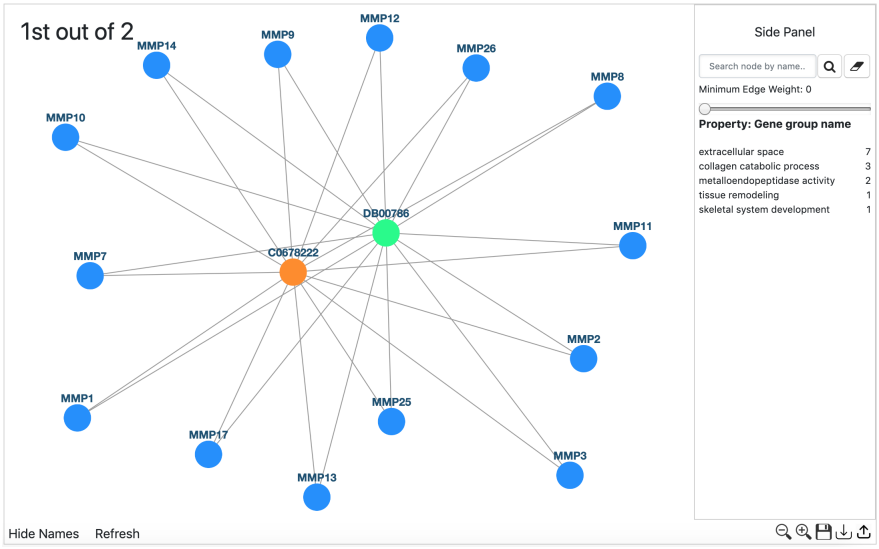

Figure 5: One motif-clique related to breast carcinoma

found a motif-clique DB11638 (Artenimol) - genes{HNRNPK, GPI, PKM, TPI1, etc} - C0678222(Breast Carcinoma), but these genes do not share similar characteristics. Based on the side panel which clearly shows properties for each gene, we classify them into 2 groups, i.e. DB11638 - gene{HNRNPK} - C0678222, which represents 'Resisting cell death' subtype, and DB11638 - genes{GPI, PKM, TPI1} - C0678222, which represents 'Deregulating cellular energetics' subtype, as shown in Figure 10. Recent research shows that Artenimol targets lysosome that cures 'Resisting cell death' subtype, and also targets mitochondria that cures 'Deregulating cellular energetics' subtype [30].

The breast carcinoma case verifies our system for knowledge discovery works because our findings are consistent with known knowledge. The discovery of new knowledge and making new hypothesis requires examining motif-cliques one by one by medical professionals, which is considered as a future step of our system and discoveries.

Besides breast carcinoma into which we did thorough investigation and cross-verification, we also found several results of other common cancers, including liver neoplasm subtypes (Figure 6), stomach neoplasm subtypes (Figure 8), lung cancer subtypes (Figure 7) and ovary neoplasm subtypes (Figure 9). Although we did not do thorough investigation on all diseases, we believe the samples we provide will be useful for future research.

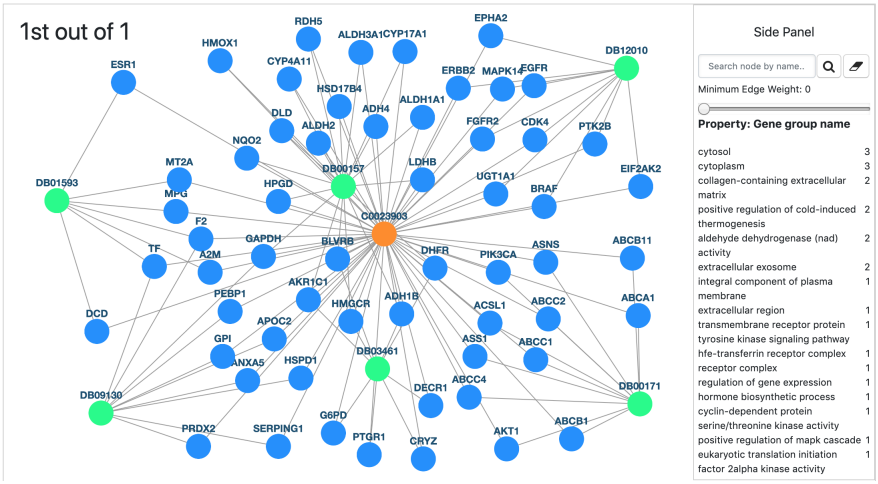

Figure 6: Liver Neoplasm Subtypes

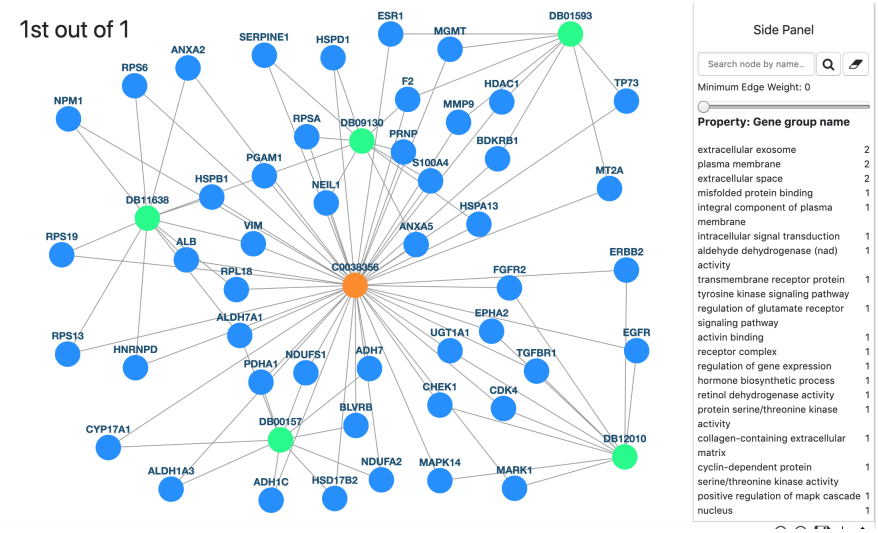

Figure 7: Stomach Neoplasm Subtypes

3.2 Drug Mechanism Investigation

By using the motif drug - gene - disease and restrict the number of drugs to be 1, we can do literature mining on characteristics of drugs. We make use of drug - gene - disease motif (Figure 2).

Some interesting results are found. For example, as shown in Figure 10, Loxapine (DB00408) is connected to a cluster of genes, which are connected to Schizophrenia (C0036341) and Obesity (C0028754). This shows that Loxapine might have a relationship between both Schizophrenia and Obesity. Loxapine is a common drug for Schizophrenia with obesity as a side effect. If we don't know obesity is a side effect, then this analysis on Loxapine will be useful.

Amoxapine (DB00543) is another example. In Figure 11, Amoxapine is connected to a cluster of genes, which are connected to Schizophrenia (C0036341) and Alcohol depen-

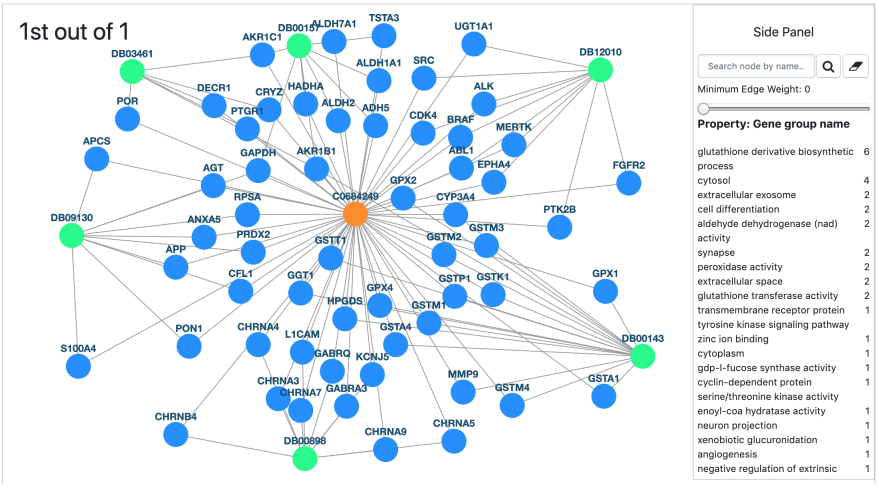

Figure 8: Lung Cancer Subtypes

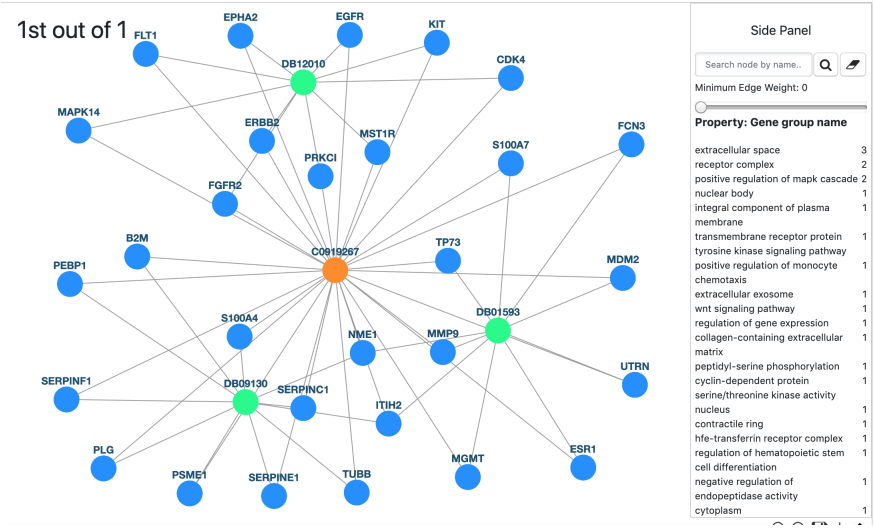

Figure 9: Ovary Neoplasm Subtypes

dence (C0001973). The ground truth is that Amoxapine is a common antidepressant for Schizophrenia. The result shows that there might be some relationship between Amoxapine and alcohol dependence as well. Recent search [28] shows that antidepressants might be useful for the treatment of depression and alcohol dependence, among patients with both diseases, which supports our assumption.

To conclude, we have shown our system can help researchers analyze drugs, including side effects and other potential usages in a novel approach. This method does not only apply in one or two drugs but every drug in general, as long as sufficient information exists in the dataset.

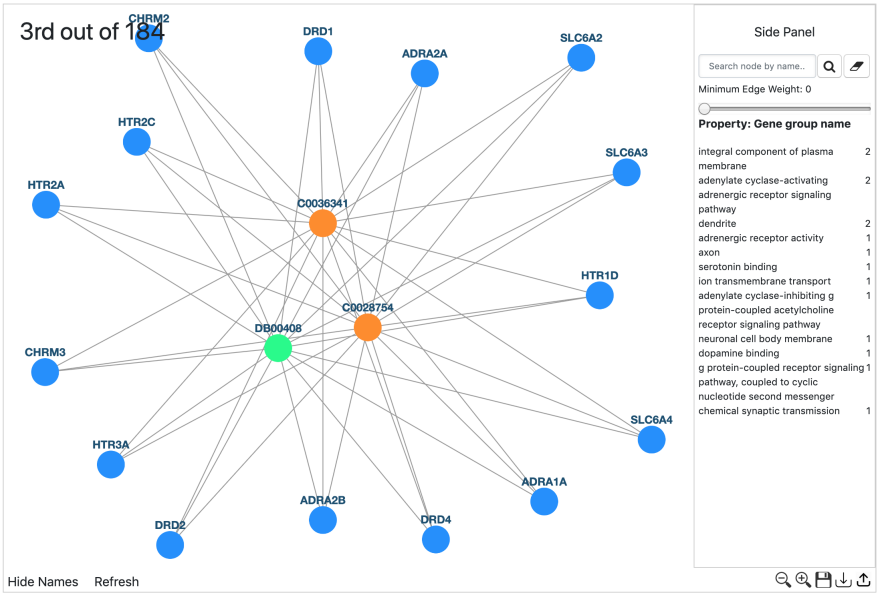

Figure 10: Mechanism Investigation of Loxapine

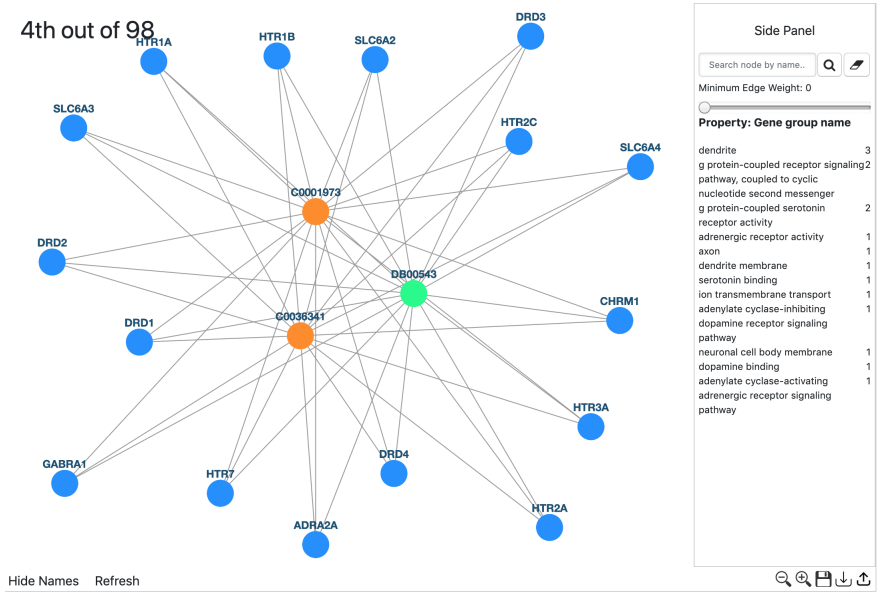

Figure 11: Mechanism Investigation of Amoxapine

3.3 Drug Repurposing

Drug repurposing is a huge industry and literature mining is one of the very first steps for establishing new ideas.

We propose an assumption that if a drug X that cures a disease A and if genes that associate A and X overlap with those associate A and another drug Y ($X \neq Y$), then Y might also be used to cure A (even if the original purpose of Y is not to cure disease A). This assumption makes sense because if we know X can cure A based on ground knowledge, and we find a motif-clique X - G (genes set) - A, then it is very likely that group of genes G

is the target of drug X, and the medical effect based on genes G is the underlying reason for the treatment of A. Thus, it is a reasonable guess that drug Y can also treat A due to the assumption that Y targets genes G and effects on genes G lead to cure of A.

It is easy to find such associations using our system. We still use the drug - gene - disease motif (Figure 2) and restrict the number of drugs to be 2, while the number of diseases to be 1.

As Figure 12 shows, drugs Loxapine (DB00408) and Doxepin (DB01142), and disease Schizophrenia (C0036341) are all connected to the same cluster of genes. Based on the ground truth that Loxapine is a common medicine for the treatment of Schizophrenia, we can guess that Doxepin may also have the potential to be used in the treatment of Schizophrenia. Although Doxepin is a common antidepressant used to treat major depressive disorders, there is also some research showing antidepressant drugs may be beneficial for people with depression and schizophrenia, though the treatment effect might be overestimated [29].

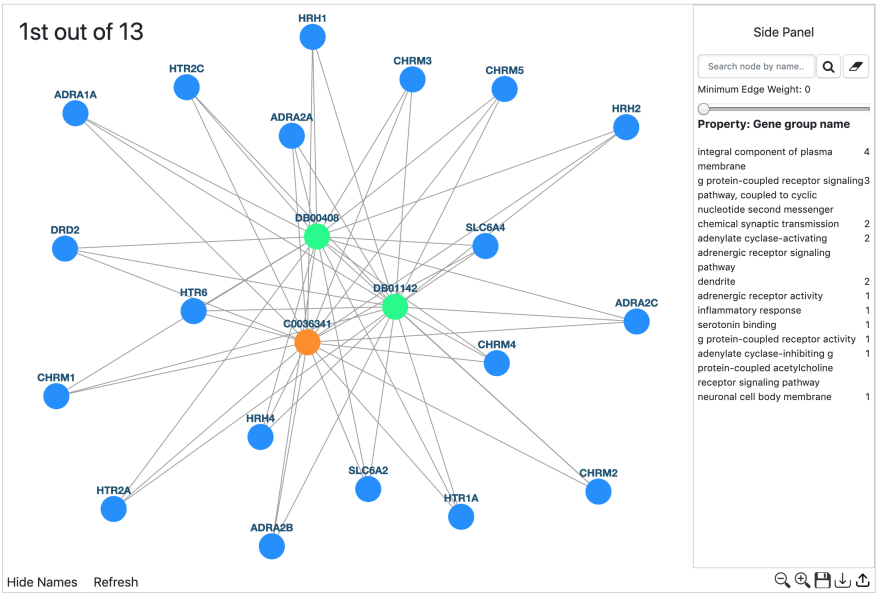

Figure 12: Doxepin Repurposing

We have shown that our system can be used to do literature mining for drug repurposing and may inspire medical researchers to find potential new use cases for drugs.

4 Conclusion

Motif-clique incorporates motif into clique definition, providing a new way to discover higher-order semantics of large heterogeneous information networks. We developed a fully functional web-based motif platform, with which it would be flexible and convenient for researchers to conduct data mining on heterogeneous information networks. The demo can be found on https://www.loom.com/share/8ffa0def75d14cf6a42ab0464561ecdb. We also collaborated with bioinformatics researchers and provided a generic web platform which can be utilized to detect and analyze abundant higher-order semantics in the bioinformatics field. We have

shown that the motif-clique platform is useful in Disease Subtyping, Drug Mechanism Investigation and Drug Repurposing areas. We plan to extend the project and make it easy for doctors to use as a future step.

References

[1] M. Ji et al. Graph regularized transductive classification on heterogeneous information networks. In ECML-PKDD, pages 570586, 2010.

[2] M. Ley. Dblp: some lessons learned. PVLDB, 2(2):14931500, 2009.

[3] Shi et al. A survey of heterogeneous information network analysis. IEEE Transactions on Knowledge and Data Engineering 29(1):17-37, 2017.

[4] Sun et al. Mining heterogeneous information networks: a structural analysis approach. Acm Sigkdd Explorations Newsletter 14(2):20-28, 2013.

[5] R. Milo et al. Network motifs: simple building blocks of complex networks. Science, 298(5594):824827, 2002.

[6] N. Przulj and N. Malod-Dognin. Network analytics in the age of big data. Science, 353(6295):123124, 2016.

[7] J. Hu et al. Discovering Maximal Motif Cliques in Large Heterogeneous Information Networks. In ICDE, 2019.

[8] H. Yin et al. Local higher-order graph clustering. In KDD, pages 555 564, 2017.

[9] A. R. Benson et al. Higher-order organization of complex networks. Science, 353(6295):163166, 2016.

[10] R. A. Hanneman and M. Riddle. Introduction to social network methods, chapter 11: cliques., 2005.

[11] G. A. Pavlopoulos et al. Using graph theory to analyze biological networks. BioData mining, 4(1):10, 2011.

[12] X. Yang et al. Bus transport network model with ideal n-depth clique network topology. Physica A: Statistical Mechanics and its Applications 390(23-24):4660-4672, 2011.

[13] J. Piero et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database 2015, 2015.

[14] Fabregat, Antonio, et al. The reactome pathway knowledgebase. Nucleic acids research 44.D1: D481-D487, 2015.

[15] Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK25501/

[16] V. Law et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. 42(1):D1091-7, 2014.

[17] Wikipedia contributors. Bioinformatics. In Wikipedia, The Free Encyclopedia. Available from: https://en.wikipedia.org/w/index.php?title=Bioinformatics&oldid=870192795

[18] K.R. Brown, et al. Online Predicted Human Interaction Database. Bioinformatics, 21(9):2076-82, 2005.

[19] Navlakha, Saket, and Carl Kingsford. The power of protein interaction networks for associating genes with diseases. Bioinformatics 26.8: 1057-1063, 2010.

[20] Li, Xia, et al. The implications of relationships between human diseases and metabolic subpathways. PloS one 6.6: e21131, 2011.

[21] Yang, Lei, et al. Predicting disease-related proteins based on clique backbone in Protein-Protein interaction network. International journal of biological sciences 10.7: 677, 2014.

[22] Yang, Lei, and Xianglong Tang. Protein-protein interactions prediction based on iterative clique extension with gene ontology filtering. The Scientific World Journal 2014, 2014.

[23] Matsunaga, Tsutomu, et al. "Clique-based data mining for related genes in a biomedical database." BMC bioinformatics 10.1: 205, 2009.

[24] Hanahan, Douglas, et al. "The hallmarks of cancer." cell 100, no. 1: 57-70, 2000.

[25] Hanahan, Douglas, et al. "Hallmarks of cancer: the next generation." cell 144, no. 5: 646-674, 2011.

[26] Ashburner et al. "Gene ontology: tool for the unification of biology". Nat Genet 25(1):25-9, 2000.

[27] The Gene Ontology Consortium. "The Gene Ontology Resource: 20 years and still GOing strong". Nucleic Acids Res 47(D1):D330-D338, 2019.

[28] Agabio et al. "Antidepressants for the treatment of people with cooccurring depression and alcohol dependence." Cochrane Database of Systematic Reviews 4, 2018.

[29] Whitehead et al. "Antidepressants for people with both schizophrenia and depression." Cochrane Database of Systematic Reviews 2, 2002.

[30] Konstat-Korzenny et al. "Artemisinin and its synthetic derivatives as a possible therapy for cancer." Medical Sciences 6(1): 19, 2018.