ICT Focus

https://www.ictfocus.org

JOURNAL ARTICLE OPEN ACCESS

A Heterogeneous Graph Model for Repurposing Drugs against Echinococcosis

Gantsooj Demberel[1], Temuulen Dorjsuren[2], Yansen Su[3], and Dolgorsuren Batjargal[4]

[1,4]Mongolian University of Science and Technology, Ulaanbaatar, Mongolia

[2]Mongolian National University of Medical Sciences, Ulaanbaatar, Mongolia

[3]School of Artificial Intelligence, Anhui University, Hefei, Anhui, China

e-mail: b.dolgorsuren@must.edu.mn

Abstract

Through this research work, we aimed to use a graph model to determine important drug substances that are proper for the *Echinococcosis* hydatid disease. Also, that model can be used in convergence with other similar diseases due to the genes of the *Echinococcus granulosus* that causes this parasitic disease, and to analyze it by applying graph algorithms to the established model. Furthermore, there is an urgent need to create a basic model for machine learning and artificial intelligence methods in health and pharmaceuticals. Also, the collection of genes and proteins registered in the official internationally recognized gene-disease databases, as well as the data of drugs and pharmaceutical products used for the specific disease, is the first step in the understanding and development of interdisciplinary science. In this paper, we define a heterogeneous graph with multiple types of nodes and edges as a basic model for future studies, as well as the methods and algorithms used in the study, considering the interrelationship between *Echinococcus granulosus* genes, proteins, diseases, and drugs.

Keywords: **echinococcus, drug interaction network, graph theory, data analysis**

Received: August 28, 2022.

Accepted: September 28, 2022.

Published: September 30, 2022.

Corresponding author: Dolgorsuren Batjargal

1. INTRODUCTION

Recently, zoonotic diseases have increased among Mongolians, and they are infectious diseases transmitted from animals to humans either directly or indirectly. Examples include foot and mouth disease, anthrax, bubonic plague, ringworm, and rabies. In modern veterinary science, there are more than 300 diseases that can be transmitted between animals according to the National Center for Zoonotic Disease Research [1]. It follows that about 160 of them are zoonotic diseases that can be transmitted from

animals to humans, which include rabies, anthrax, black death, FMS, and echinococcosis, which are caused by viruses, bacteria, parasites, fungi, and rickettsia.

In 2012, the Ministry of Health declared [25] echinococcus to be one of the ten leading zoonotic diseases in Mongolia. And the causative agent in infection of this disease is *Echinococcus granulosus*, a small tapeworm liver cyst parasite found in dogs, foxes, and wolves [2]. It has a 3-6 mm long body consisting of 2-3 generations, and lives by sucking nutrients from the intestine/gut of the animals. When the venous blood is filtered after nesting in the liver, the larvae remain filtered and begin to grow into liver flukes. Furthermore, it forms parasitic cysts mainly in the liver, lungs, and other organs.

The World Health Organization [26] reported that zoonoses result in 2.5 billion cases of human illness each year, and 2.7 million human deaths, but only a few drugs have been fully studied and used, therefore they are classified *Echinococcosis* as neglected diseases. There are 70 million farm animals in our country and about 40% of the working age population is involved with herds or animals. Therefore, the risk of occurrence and transmission of this types of disease is considered high. In Mongolia, a study conducted in slaughterhouses [3] showed that there is a high prevalence of echinococcosis among animals. It was reported that the prevalence of ringworm among slaughtered animals is 3.5 percent, and among goats, it is 9.2 percent.

Computer and medical scientists need to work together to develop and introduce new advanced methods and technologies in the research and analysis of new drugs for this disease. Therefore, time and resources to make progress and steps in new drug discovery and development are required too much, such as taking around 12 years and costing around a £1.15 billion in pharmaco-medical research. To save opportunities, we analyzed the data of *echinococcus granulosus* genes, proteins, and drugs using machine learning methods leveraging graph theory and methods. We implemented this research work in the following stages, and we will present the results in this article.

- collected necessary data and prepared for machine learning
- answered how to build a graphical model on the data to analyze the data
- studied and developed some useful graph metrics and algorithms for use in research
- tested and written to introduce and explain the test results.

The research work consists of four main parts. The first part is an introduction to hydatid disease and its causative echinococci. The structure and implementation methods of algorithms are included. The last or fourth chapter describes the experimental data and results.

2. RELATED WORKS

Animal husbandry is our main industry, but at the same time, it is the main factor in the transmission of many types of viral and bacterial infectious diseases that are harmful to human health and can lead to death. One of them is the echinococcosis, and there is no drug or preparation for the perfect treatment of the causative *Echinococcus* granulosus, and antiparasitic drugs are used to stop the growth of cysts, reduce their size, reduce recurrence, and prepare for surgery [4]. Also, Mebendazole and Albendazole 10mg/kg twice a day for three months or in case of rupture of the primary cyst, praziquantel and protosolicide group drugs are used to prevent secondary cyst formation.

Therefore, the main goal of this research is to extract new types of essential drugs based on the genetic information of *echinococcus granulosus*, or to confirm and prove whether existing drugs can be used together based on their similarity. Much research [14, 22-24] has

been done on medical and biological data, including biological networks, on processing and analysis using graph data structures. Recently, a significant amount of research has been conducted in this field along with the development of novel technologies such as machine learning and deep learning [10, 15-21].

For example, researchers Cheng et al. [10] researched the heterogeneous network to predict some interaction between one drug and an others based on drug composition, chemical elements, genomic characteristics, and therapeutic information. Also, network analysis works [15, 16] have applied deep learning methods to protein and protein interaction networks intending to accelerate new drug testing. Our work is unique in that it is done for a specific disease, and it is characterized by building networks, evaluating nodes, and predicting protein connections with the help of isolated seed genes.

Currently, there are 11,355 genes and 42,941 proteins related to *echinococcus granulosus* in the NCBI [5] gene database. There are only two drugs available in the DrugBank [6], Albendazole (Approved, Vet Approved), which is approved for human use, and Toluene (Vet Approved), which is approved for animal use.

3. THE PROPOSED METHODOLOGY

Based on the echinococcal gene data, we have identified protein-protein interaction (D), protein-drug interaction (PDI), drug-disease interaction (DDI), and drug-target interactions (DTI). An $G(V, E, \mathcal{A}, \mathcal{R})$ graph was constructed to show the relationship between genes, proteins, drugs, and diseases by combining the bipartite graphs of DTI.

Given graph $G(V, E, \mathcal{A}, \mathcal{R})$ is a heterogeneous graph because it contains different types of vertices and edges. Given that each node is $v \in V$ and each edge is $e \in E$ in the graph, the functions corresponding to their types are $\tau(v): V \to \mathcal{A}$ and $\phi(e): E \to \mathcal{R}$. The graph is undirected. The set of vertices of a heterogeneous graph $V = \{v_1, v_2, v_3, \ldots, v_n\}$, and E is the set of edges $E = \{e_{ij} = (v_i, v_j) \mid v_i \in V, v_j \in V\}$ were considered, respectively. In other words, depending on the connection of vertex v with other vertices or nodes and the topological structure of the generated network, it will be determined how important the gene, protein, drug, or disease is an important node in the graph $G(V, E, \mathcal{A}, \mathcal{R})$.

At the beginning of the research work, we need to collect the genes related to the selected disease, and these genes will be the seeds that can identify the *echinococcus* that causes the disease. So, *Echinococcus granulosus* is the causative agent that causes *echinococcosis*, is directly dependent on a certain number of genes, and it can be understood that humans and all other living organisms suffer from the disease due to the effect or mutation of genes. The 21 echinococcal genes we selected are all human/homo-sapiens genes. Since genes define functional chains of all proteins and RNAs, we further derive PPIs associated with the identified genes. Next, we can obtain DDI network's information from the pharmaceutical database, based on the fact that drugs are obtained by combining and synthesizing proteins from a medical point of view.

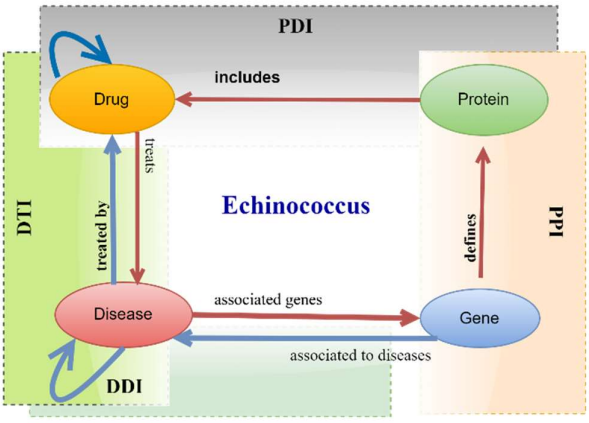


Figure 1. Overview of proposed model

In other words, we can also come up with drugs for other diseases that contain protein compounds in hydatid-related

drugs, and those drugs can be used alone or in combination for the treatment of that disease. Cheng et al [10] mentioned that DDIs are not identified during clinical trials in many cases, and are not reported until after the drug is approved. Therefore, we can identify drugs with similar ingredients available in advance from the constructed graph $G(V, E, A, R)$ with the help of an algorithm.

We analyzed the graph by considering several important node metrics while constructing the graph. The graph was analyzed by considering several important node metrics, and we built PPI, PDI, DDI, and DTI networks based on the main 21 genes of *Echinococcus granulosus* hydatid worm, and combined them into a graph with heterogeneous nodes. In the following subsections, we have introduced graph analysis methods and algorithms, including degree centrality, proximity centrality, path centrality, and spectral centrality measures, respectively.

Real networks generally consist of a large number of nodes and links. However, it does not mean every node is important to its network. Using the following measures, we have confidence in it is possible to calculate how important each node is to the network.

3.1 Degree centrality

The simplest, easy way is to measure the importance of a given vertex based on its degree. The number of connected nodes means that the node is important. Using Formula 1, we can directly calculate $G(V, E, A, R)$ based on the number of degrees of the vertex v.

$$C_D(v) = deg(v) \tag{1}$$

Degree centrality directly depends on how many nodes are connected to that vertex in the given graph. For example, if vertex v is the central node of a star-shaped network, the value of C_D is $C_D(v) = k - 1$. So, the value of degree centrality should be normalized, and depending on the size of the network, the value of centrality will be as follows.

$$\widetilde{C_D(v)} = \frac{C_D(v)}{|V| - 1} \tag{2}$$

Thus, the value of $\widetilde{C_D(v)}$ decreases to $0 \leq \widetilde{C_D(v)} \leq 1$.

3.2 Proximity centrality (Closeness centrality)

This type of centrality measure includes closeness, harmonic and eccentricity centrality measures, respectively. We only mention the closeness centrality measure here. A given node is connected to a small number of other vertices, but the value of closeness centrality may be relatively high. Because it depends on how many vertices can be connected by the fastest or shortest path. In other words, it depends on the length of the average shortest path. And on the contrary, it can be calculated depending on the size of the path. In other parts of the network or other vertices, depending on the average distance from v to the vertex, the centrality of the vertex is considered, and how important v is to the network is determined, and it is calculated by the following formula.

$$C_C(v) = \frac{1}{\sum_u d(v, u)} \tag{3}$$

In other words, if $\frac{1}{d(v,u)} = 0$, there is no path connecting the two vertices v and u, and the harmonic concentration is normalized by dividing by $|V| - 1$. Here, $|V|$ is the number of vertices in the graph. On larger networks, more paths are found, which reduces the numerical value of closeness centrality to a negligible number. Accordingly, we can also normalize as follows.

$$\widetilde{C_C(v)} = \frac{|V| - 1}{\sum_u d(v,u)} \tag{4}$$

3.3 Path centrality (Betweenness centrality)

The centrality measure is used to determine whether each vertex is important for the graph $G(V,E,A,R)$ we have constructed, since its value depends on how influential the participation of node v is in terms of the shortest paths connecting all other possible vertices connected to the given vertex. In other words, if there is a short path to all other vertices through that vertex, it means that the centrality value for the vertex v will be high. If $\sigma_{st}(v)$ is the shortest path from source s or starting vertex to target t or ending vertex, it means that it passes through the vertex v. Paths connecting vertices s and t without crossing this vertex are also found, and the probability that the shortest path passes through this vertex is $p_{st}(v)$, and for the graph $G(V,E,A,R)$. At least one short path passes through the vertex v, $\sigma_{st}(v) \neq 0$. The centrality value of vertex v is calculated by the following Formula 5 betweenness centrality measure.

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{5}$$

On the contrary, if there is no short path in the graph that crosses the vertex v, the centrality value of the vertex is $C_B(v) = 0$.

3.4 Spectral centrality (Eigenvector centrality)

For one node in the network, the node influence is calculated from how many other important nodes the node is connected to it. These types of measures include Eigenvector, PageRank, and Katz centrality measures. These spectral centrality measures are considered recursively, meaning that the centrality measure of each vertex affects the centrality measure of its connected neighbors. For a given graph $G(V,E,A,R)$, the number of vertices is $N = |V|$ and we create a neighborhood matrix $A = (a_{v,t})$. Here, $a_{v,t} = 1$ if vertex v is connected to vertex t and $a_{v,t} = 0$ if not. From this value, the centrality score x_v for the vertex v depends on the values of its neighboring vertices and is calculated by the following formula.

$$C_E(v) = x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t \tag{6}$$

Here, $M(v)$ is the set of neighboring vertices of v, and λ is a constant numerical measure. The eigen value λ can be found in the following equation, which is the formula of the eigen vector. where x is an eigenvector or vector quantity.

$$Ax = \lambda x \tag{7}$$

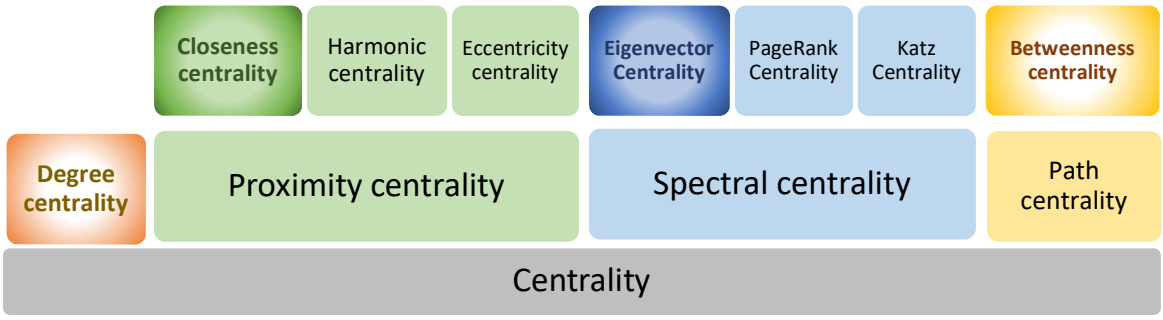


Figure 2. Centrality measure's Taxonomy

If the interrelationship network of drugs, proteins, and drugs is represented by a non-directional multivariate $G(V, E, A, R)$ graph, it is appropriate to choose this type of measure. Still, if it is expressed by a directed graph, other measures should be used.

From the four types of centrality measures shown in Figure 2, one of each type, namely degree, closeness, eigenvector, and betweenness centrality measures, was tested in the next group on the prepared data, and the test results were included.

4. EXPERIMENTAL RESULTS

4.1 Experimental setups

In our experiments, we constructed PPI, PDI, and PDI networks using the NetworkAnalyst [9] tool. Data from the open databases shown in Table I below were used to create these networks. NetworkAnalyst is a gene-centric online platform that can perform biological network analysis and visual analysis, and we chose it based on our belief that it would be suitable for further data extraction to work on specific *Echinococcus granulosus* genes. The tool is still under constant improvement and is a free online platform to use.

TABLE I

Source of experimental data

Bipartite Graph	Database	Description
PPI	InnateBD [7]	A repository containing information on the human, mouse, and bovine proteins, genes, and signaling pathways and a member organization of the International Molecular Exchange Consortium (IMEx).
PDI DTI	DrugBank [6]	Drugbank is an online open database containing detailed information about drugs and drug targets.
DDI	DisGenNET [9]	A large collection of human diseases and genes.

4.2 Experimental datasets

Data related to *echinococcosis* disease were obtained from the NCBI (National Center for Biotechnology Information) [5] gene pool. This library is a free downloadable database. The NCBI database contains a total of 11,355 genes, which belong to 11 species, including homo-sapiens(human), musculus(mouse), norvegicus(rat), elegans(roundworm), and melanogaster (fruit fly) so on. We chose the human genes because of intending to derive a drug for humans from it. Human genes associated with this disease have been less studied than other species, with only 21 genes found in the database and used in our experiment. See Table V in the Appendix for details on these genes.

4.3 Experimental results

Using 21 homo-sapiens genes of *Echinococcus granulosus*, the causative agent associated with hydatid disease, we constructed a PPI network (see Figure 3).

A total of 17 seeds or important genes, 899 nodes, and 1296 edges were formed in constructing the PPI network of *Echinococcosis*-related genes. Nodes represent proteins, and edges represent interactions between proteins. In a PPI network, proteins are directly or indirectly connected to each other. DTI subnetworks (Figures 5) were also constructed, showing which *Echinococcosis*

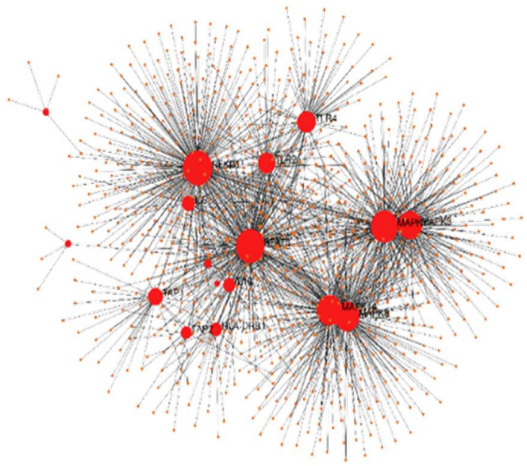
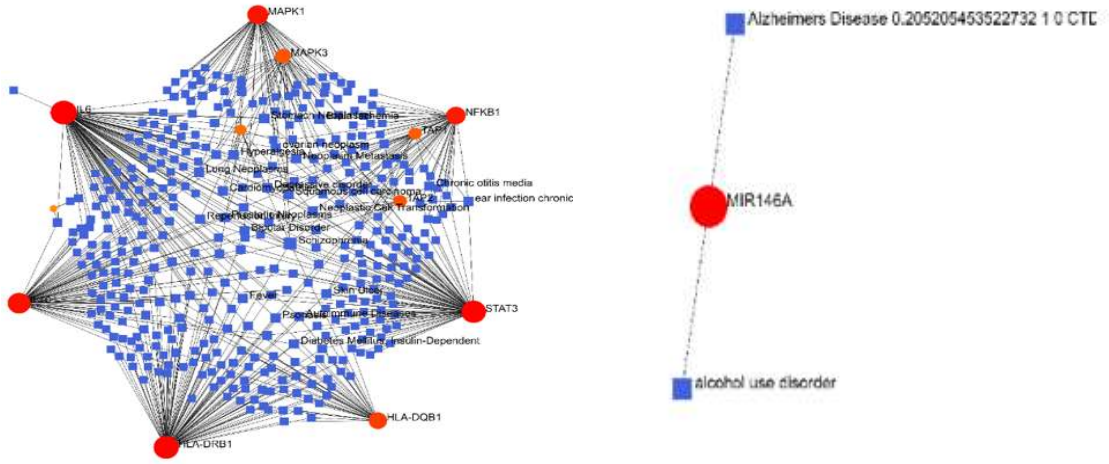


Figure 3. The PPI network is associated with Echinococcus

related 21 genes are associated with other diseases.

In the network shown in Figures 4a and 4b, square blue nodes represent diseases, and round red nodes represent proteins. For example, the protein coded MIR146A in Figure 4b is a protein associated with Alzheimer's disease and Alcohol use disorder. A total of 12 and 1 seeds, 414 and 3 nodes, and 621 and 2 edges were established in the DTI network. We have created PDI networks showing echinococcal disease proteins and which drugs they interact with, as illustrated in the Appendix section (see Figure 5 a-f). Table II also provides general information about the networks.



(a) (b)

Figure 4. DTI network of 21 genes associated with *Echinococcus granulosis*

For $G(V, E, A, R)$ spanning all subnets, the total number of nodes |V| = 2453; |E| = 5399, with five types of vertices (gene, seed, protein, drug, chemical) and four types of links (PPI, PDI, DDI, DTI), graph density is 0.002, diameter or longest path is 6, etc.

Algorithms were developed using the NetworkX[12] library in Python. The values of the centrality measures for the established network are listed for the first 5 and 5 most important nodes and are shown in Table III below. The most important genes are MARK1, as shown by the results of degree, closeness, and eigenvalue centrality (Table III a,b,d). However, considering betweenness centrality (Table III c), the CYP1A1 gene is a more important gene than the MARK gene.

TABLE III

Results of centrality measures

	v	$C_D(v)$
1	MARK1	970
2	MARK3	918
3	CYP1A1	693
4	IL6	662
5	MARK8	412

(a)

	v	$C_C(v)$
1	MARK1	0.516
2	MARK3	0.498
3	Arsenic	0.497
4	Lipopoly.	0.495
5	Estradiol	0.494

(b)

	v	$C_B(v)$
1	CYP1A1	906565.1
2	MARK1	778697.7
3	IL6	659035.1
4	MARK3	602836.8
5	NFKB1	381747.2

(c)

	v	$C_E(v)$
1	MARK1	1
2	MARK3	0.95
3	IL6	0.54
4	CYP1A1	0.48
5	MARK8	0.40

(d)

In the next section, we developed a graph-based protein interaction prediction model based on the previously established PPI network using machine learning methods and ran it on the data.

The PPI network we built before had a total of 802 edges or links, and randomly selected 80 percent (a total of 641) of links as training data and the remaining 161 links as testing data.

When choosing predictors, we chose CommonNeighbors predictors, Adamic Adar predictors based on neighboring vertices in unsupervised learning, and Jaccard coefficient predictors based on network topology, and Katz predictors based on pathway information between two proteins to predict protein interactions.

TABLE IV

Results of machine learning predictors

	s	*t*	score
1	5599 (MARK8)	5595 (MARK3)	32
2	6774 (STAT3)	1432 (MARK14)	24
3	5599 (MARK8)	4790 (NFKB1)	13
4	7099 (TLR4)	7097 (TLR2)	12
5	7099 (TLR4)	5594 (MARK1)	10

(a) Result using CommonNeighbor predictor

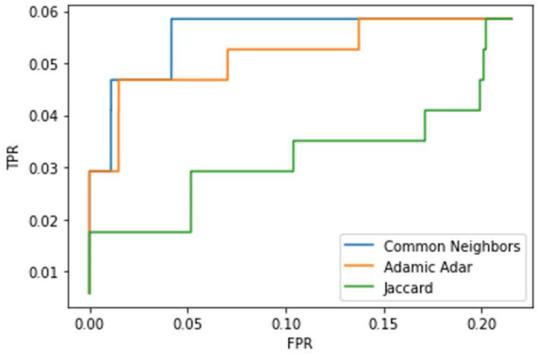
	s	*t*	score
1	998 (CDC42)	9043 (SPAG9)	1
2	998 (CDC42)	8061 (FOSL1)	1
3	998 (CDC42)	6416 (MAP2K4)	1
4	998 (CDC42)	5899 (RALB)	1
5	998 (CDC42)	5898 (RALA)	1

(b) Result using Jaccard predictor

	s	*t*	score
1	5599 (MARK8)	5595 (MARK3)	22.33
2	6774 (STAT3)	1432 (MARK14)	16.98
3	7099 (MARK8)	7097 (NFKB1)	10.56
4	5599 (TLR4)	4790 (TLR2)	9.22
5	7099 (TLR4)	5594 (MARK1)	5.69

(c)Result of Adamic Adam predictor

ROC curves are also used to evaluate how true or false a predictor is. We made our algorithms in the Python programming language and developed them using the LinkPred [13] library.




```
Area Under Roc Curve (AUROC)
Common Neigh.: 0.011905725890099469
Adamic Adar:    0.011110691472196408
Jaccard:        0.006849802269769901
```

Figure 6. ROC curve

As can be seen from the ROC curve shown in Figure 6, for the three selected predictors, the Common Neighbor predictor predicts slightly more efficiently by considering the common neighboring vertices. However, this type of data predicts a certain small percentage, so it is concluded that further improvement is necessary. In Figure 6, we have included the AUC curve, which is one of the main evaluation metrics to check the performance of our classification model.

5. CONCLUSION

In a particular disease, a certain number of genes are altered or mutated so that they are used for the development of therapeutic and pharmaceutical methods. *Echinococcus* hydatid disease, the subject of our research, has a high prevalence and risk in our country, so it is critical to develop appropriate new drugs quickly or to determine whether they can be used in combination with drugs used for similar diseases. We conducted this research with the aim of identifying important information by creating a graph model on the interaction data of genes, proteins, and drugs related to the disease. The experiment determined that the important genes of *echinococcus* are MARK1, MARK3, CYP1A1, and arsenic, lipopolysaccharide, and estradiol are among the important chemical ingredients that should be included in the drug's composition.

REFERENCES

[1] National Centre for Zoonotic Disease; Last Accessed: March. 29, 2022. [Online]. Available: https://nczd.gov.mn/?p=9882

[2] Yu, Y., Li, J., Wang, W., Wang, T., Qi, W., Zheng, X., ... & Duan, L. (2021). Transcriptome analysis uncovers the key pathways and candidate genes related to the treatment of Echinococcus granulosus protoscoleces with the repurposed drug pyronaridine. BMC genomics, 22(1), 1-9.

[3] A study of Echinococcosis in animal and livestock; Last Accessed: March. 29, 2022. [Online]. Available: https://old.legalinfo.mn/annex/details/9124?lawid=13970

[4] P. Nyamdavaa., (2010) Directory of Infectious Disease Control, American Public Health Association, World Health Organization, (United Book Press Inc.)

[5] National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2022 Sep 28]. Available from: https://www.ncbi.nlm.nih.gov/

[6] Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. Drugbank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006 Jan 1;34 (Database issue): D668-72. 16381955. Last Accessed: March. 29, 2022. [Online]. Available: https://go.drugbank.com/

[7] Breuer et al., InnateDB: systems biology of innate immunity and beyond - recent updates and continuing curation. Nucl. Acids Res. (2013) 41 (D1); Last Accessed: March. 29, 2022. [Online]. Available: https://www.innatedb.com/index.jsp

[8] Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, Laura I Furlong. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucl. Acids Res. (2019) doi:10.1093/nar/gkz1021; Last Accessed: March. 29, 2022. [Online]. Available: https://www.disgenet.org/

[9] Zhou, G., Soufan, O., Ewald, J., Hancock, REW, Basu, N. and Xia, J. (2019) "NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis" Nucleic Acids Research 47 (W1): W234-W241.

[10] Cheng, F., & Zhao, Z. (2014). Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. Journal of the American Medical Informatics Association, 21(e2), e278-e286.

[11] Zhou, F., Malher, S., & Toivonen, H. (2010, December). Network simplification with minimal loss of connectivity. In 2010 IEEE international conference on data mining (pp. 659-668). IEEE.

[12] Hagberg, A., Swart, P., & S Chult, D. (2008). Exploring network structure, dynamics, and function using NetworkX (No. LA-UR-08-05495; LA-UR-08-5495). Los Alamos National Lab.(LANL), Los Alamos, NM (United States)

[13] Kerrache, S. (2021). LinkPred: a high performance library for link prediction in complex networks. PeerJ Computer Science, 7, e521.

[14] Nushi, E., Popescu, V.-B., Angel Sanchez Martin, J., Ivanov, S., & Czeizler, E. (2021). Network modeling methods for precision medicine.

[15] Muzio, G., O'Bray, L., & Borgwardt, K. (2021). Biological network analysis with deep learning. Briefings in bioinformatics, 22(2), 1515-1530.

[16] Jamasb, A. R., Day, B., Cangea, C., Liò, P., & Blundell, T. L. (2021). Deep learning for protein–protein interaction site prediction. In Proteomics Data Analysis (pp. 263-288). Humana, New York, NY.

[17] Yuan, Q., Chen, J., Zhao, H., Zhou, Y., & Yang, Y. (2022). Structure-aware protein–protein interaction site prediction using deep graph convolutional network. Bioinformatics, 38(1), 125-132.

[18] Gaudelet, T., Day, B., Jamasb, A. R., Soman, J., Regep, C., Liu, G., ... & Taylor-King, J. P. (2021). Utilizing graph machine learning within drug discovery and development. Briefings in bioinformatics, 22(6), bbab159.

[19] Raimondi, D., Simm, J., Arany, A., & Moreau, Y. (2021). A novel method for data fusion over entity-relation graphs and its application to protein–protein interaction prediction. Bioinformatics, 37(16), 2275-2281.

[20] Coşkun, M., & Koyutürk, M. (2021). Node similarity-based graph convolution for link prediction in biological networks. Bioinformatics, 37(23), 4501-4508.

[21] Yang, J., Xu, Z., Wu, W. K. K., Chu, Q., & Zhang, Q. (2021). GraphSynergy: a network-inspired deep learning model for anticancer drug combination prediction. Journal of the American Medical Informatics Association, 28(11), 2336-2345.

[22] Hu, L., Zhang, J., Pan, X., Yan, H., & You, Z. H. (2021). HiSCF: leveraging higher-order structures for clustering analysis in biological networks. Bioinformatics, 37(4), 542-550.

[23] Rekik, I., & Gurbuz, M. B. (2021). MGN-Net: a multi-view graph normalizer for integrating heterogeneous biological network populations. arXiv preprint arXiv:2104.03895.

[24] Awan, Z., Alrayes, N., Khan, Z., Almansouri, M., Bima, A. I. H., Almukadi, H., ... & Banaganapalli, B. (2022). Identifying significant genes and functionally enriched pathways in familial hypercholesterolemia using integrated gene co-expression network analysis. Saudi Journal of Biological Sciences, 29(5), 3287-3299.

[25] List of priority diseases for zoonotic disease surveillance and research, (2012) Last Accessed: September. 26, 2022. https://moh.gov.mn/uploads/files/7e67621d23cdc96a989ec8f3400e24bc.pdf

[26] Gebreyes WA, Dupouy-Camet J, Newport MJ, Oliveira CJ, Schlesinger LS, Saif YM, et al. The global One Health paradigm: challenges and opportunities for tackling infectious diseases at the human, animal, and environment interface in low-resource settings. PLoS Negl Trop Dis. 2014;8:e3257. 10.1371/journal.pntd.0003257

ACKNOWLEDGEMENTS

This work was supported by the Mongolian Foundation of Science and Technology, Research project SHUT-2021/350. We thanks to Prof. Tseren-Onolt Ishdorj whose valuable guiding and support of this research work.

BIOGRAPHIES

Dolgorsuren Batjargal received B.S and M.S degrees in Computer Science from the Mongolian University of Science and Technology in 2010 and 2012, respectively. Currently, she is working as an associate professor in the Department of Computer Science at the Mongolian University of Science and Technology since 2010 up to now. She did her Ph.D. in the Dept. of Computer Science and Engineering at Kyung Hee University, South Korea in 2019. Her research interests are in Streaming Graph Mining and Distributed Computing. She has received Best Paper Awards at BigDAS 2016 and Best Presentation Award at BigComp 2017 International Conference for papers related to streaming graph mining and graph algorithms, respectively.

Gantsooj Demberel is currently a Ph.D. candidate in the Dept. of Computer Science at the Mongolian University of Science and Technology, Ulaanbaatar, Mongolia. Her research interests are in network data structures and bioinformatics.

Temuulen Dorjsuren is a baccalaureate (1994) from Mongolian State University of Education, a master's (1996) with Mongolian University of Life Sciences, and a doctor (2003) with Mongolian National University of Medical Sciences, respectively. She was a lecturer between 1994 and 1998 in the department of biology at the Mongolian University of Life Sciences. She is currently a full professor at the Mongolian National University of Medical Sciences. Her research interest is in Molecular diagnostic and epidemiology of human parasitology.

Yansen Su received her Ph.D. degree in computer science from the Huazhong University of Science and Technology, Wuhan, China, in 2014. She is currently an associate professor in the School of Artificial Intelligence, Anhui University, Hefei, China. Her main research interests include bioinformatics, systems biology, complex networks, and multi-objective optimization.

APPENDIX

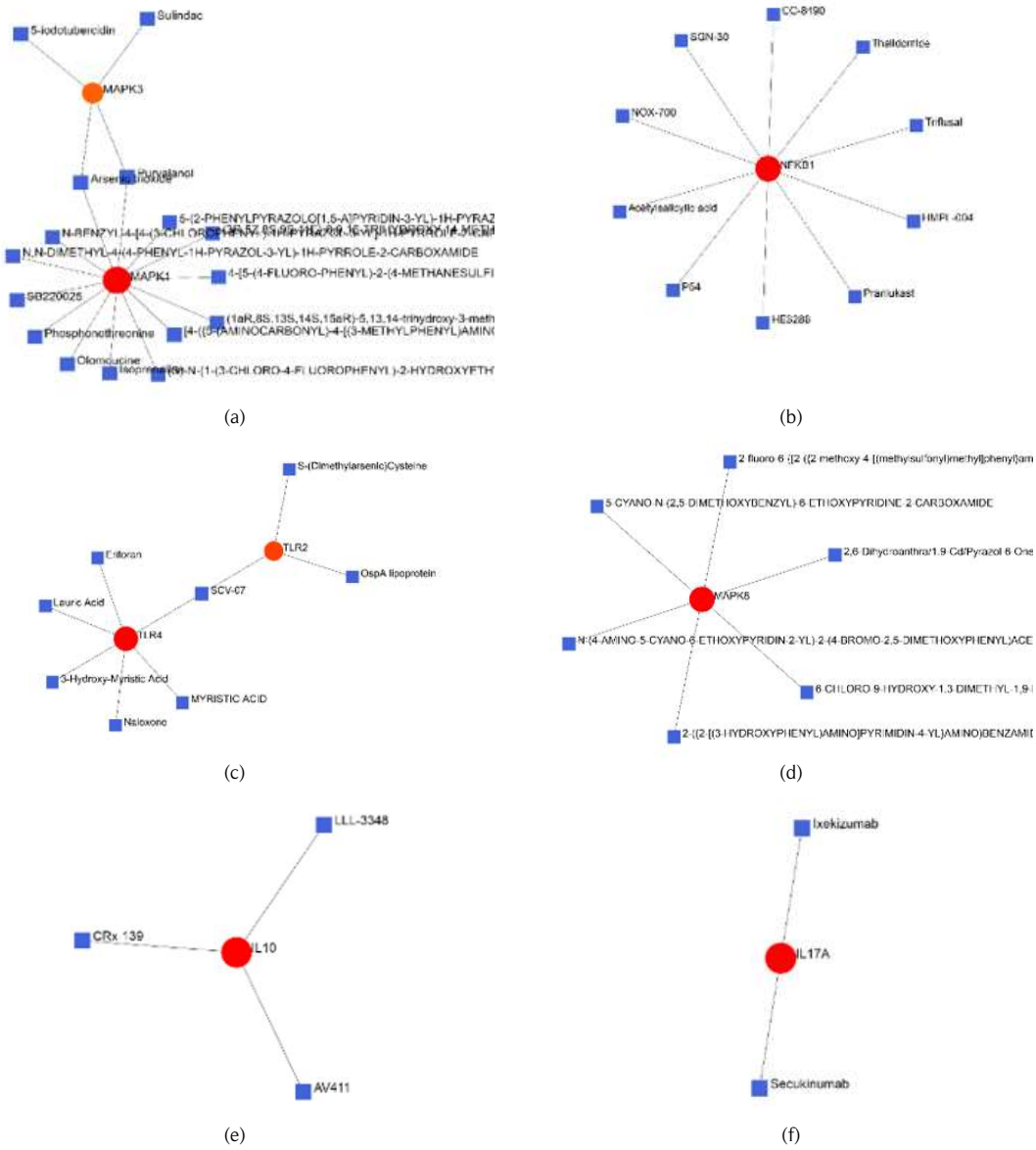


Figure 5. PDI network of 21 Echinococcus-related genes

TABLE II

Details of subgraphs

Subgraphs	Nodes	Edges	Seeds
Figure 5a	18	18	2
Figure 5b	11	10	1
Figure 5c	10	9	2
Figure 5d	7	6	1
Figure 5e	4	3	1
Figure 5f	3	2	1

TABLE V

Details of selected 21 homo-sapiens genes from NCBI

Gene ID	Symbol	Aliases	Description	Map location	Ehromosome	start_position_on_the_genomic_accession	end_position_on_the_genomic_accession	Exon count
1543	CYP1A1	AHH, AHRR, CP11, CYP1, CYPIA1, P1-450, P450-C, P450DX	cytochrome P450 family 1 subfamily A member 1	15q24.1	15	74719542	74725528	7
3119	HLA-DQB1	CELIAC1, HLA-DQB, IDDM1	major histocompatibility complex, class II, DQ beta 1	6p21.32	6	32659467	32666657	6
3123	HLA-DRB1	DRB1, HLA-DR1B, HLA-DRB, SS1	major histocompatibility complex, class II, DR beta 1	6p21.32	6	32578775	32589848	6
3586	IL10	CSIF, GVHDS, IL-10A, TGIF, IL10	interleukin 10	1q32.1	1	206767602	206772494	7
3605	IL17A	CTLA-8, CTLA8, IL-17, IL-17A, IL17, ILA17	interleukin 17A	6p12.2	6	52186375	52190638	3
246778	IL27	IL-27, IL-27AA, IL27p28, IL30, p28, IL27	interleukin 27	16p12.1-p11.2	16	28499362	28526730	6
3569	IL6	BSF-2, BSF2, CDF, HGF, HSF, IFN-beta-2, IFNB2, IL-6	interleukin 6	7p15.3	7	22725889	22732002	6
3578	IL9	HP40, IL-9, P40	interleukin 9	5q31.1	5	135892246	135895841	5
5594	MAPK1	ERK, ERK-2, ERK2, ERT1, MAPK2, NS13, P42MAPK, PRKM1, PRKM2, p38, p40, p41, p41mapk, p42-MAPK	mitogen-activated protein kinase 1	22q11.22	22	21759657	21867680	9
1432	MAPK14	CSBP, CSBP1, CSBP2, CSPB1, EXIP, Mxi2, PRKM14, PRKM15, RK, SAPK2A, p38, p38ALPHA	mitogen-activated protein kinase 14	6p21.31	6	36027711	36122964	22
5595	MAPK3	ERK-1, ERK1, ERT2, HS44KDAP, HUMKER1A, P44ERK1, P44MAPK, PRKM3, p44-ERK1, p44-MAPK	mitogen-activated protein kinase 3	16p11.2	16	30114105	30123309	10
5599	MAPK8	JNK, JNK-46, JNK1, JNK1A2, JNK21B1/2, PRKM8, SAPK1, SAPK1c	mitogen-activated protein kinase 8	10q11.22	10	48306673	48439360	16
406938	MIR146A	MIRN146, MIRN146A, miR-146a, miRNA146A	microRNA 146a	5q33.3	5	160485352	160485450	1
406947	MIR155	MIRN155, miRNA155, mir-155	microRNA 155	21q21.3	21	25573980	25574044	1
192343	NEWENTRY	Record to support submission of GeneRIFs for a gene not in Gene (human).						
4790	NFKB1	CVID12, EBP-1, KBF1, NF-kB, NF-kB1, NF-kappa-B1, NF-kappaB, NF-kappabeta, NFKB-p105, NFKB-p50, NFkappaB	nuclear factor kappa B subunit 1	4q24	4	102501266	102617302	27
6774	STAT3	ADMIO, ADMIO1, APRF, HIES	signal transducer and activator of transcription 3	17q21.2	17	42313324	42388502	24
6890	TAP1	ABC17, ABCB2, APT1, D6S114E, PSF-1, PSF1, RING4*0102N, TAP1N, TAP1	transporter 1, ATP binding cassette subfamily B member	6p21.32	6	32845209	32853704	12
6891	TAP2	ABC18, ABCB3, APT2, D6S217E, PSF-2, PSF2, RING11	transporter 2, ATP binding cassette subfamily B member	6p21.32	6	32821831	32838739	13
7097	TLR2	CD282, TIL4	toll like receptor 2	4q31.3	4	153684080	153710643	5
7099	TLR4	ARMD10, CD284, TLR-4, TOLL	toll like receptor 4	9q33.1	9	117704403	117724735	4