Studying Efficacy of Drug Repurposing using Biological Networks

Archa Jain
Department of Computer Science
Stanford University
archa@stanford.edu

1 Introduction

1.1 Problem Statement

The process for drug development is extremely limiting right now, being both prohibitively expensive, and very slow to take from start to finish. Promising compounds that might be effective for a certain disease are identified in the laboratory through various different techniques – either identified through the specific shape of the molecule required by the biology, or testing many compounds from a library of compounds in the lab to screen for desirable chemical properties. All these methods tend to be time consuming, with a fairly small hit rate. Drugs then undergo laboratory and animal testing to answer basic questions about safety. If this is promising, the drugs are then tested on people to make sure they are safe and effective. At this point, companies apply for approval for the drug, which in itself is usual a very long and expensive process.

Even through this whole process, the likelihood of developing a successful drug is still very low, making the industry very high barrier to entry, and only viable for large pharmaceutical companies. This make drug repurposing a very promising field. Drug repurposing involves identifying new uses for existing compounds, that are known to be chemically stable, and not harmful in humans, which significantly reduces the cost of overall drug development. It eliminates the process of developing a stable compound, doing toxicity and metabolic screens (determine whether they are toxic, and how long they take to be digested etc.), and setting up early clinical trials and post consumption statistics.

Recently, many new techniques for drug repurposing have been developed that could be very promising to apply to existing data sets. In this project, I will explore one such approach – Hetionets [1] with an existing open dataset, LINCS 1000 [2], and gauge the viability of using this approach for real world drug repurposing.

1.2 Background

Hetnets were introduced by Himmelstein et al in "Rephetio: Repurposing drugs on a hetnet". They are networks with multiple types of nodes and relationships, developed by integrating knowledge and experimental findings from existing biomedical research. Figure 1 shows the schema of Helionet, with the various node types represented in different colors, and the edges representing relationships between them. We adapted social network analysis algorithms and applied it to Hetionet to identify patterns of efficacy and predict new uses for drugs. Hetionet also introduced Rephetio, an algorithm that performs edge prediction through a machine learning framework that accommodates the breadth and depth of information contained in Hetionet. This represents an *in silico* implementation of network pharmacology that natively incorporates polypharmacology (use of pharmaceutical agents that act on multiple biological and chemical targets or disease pathways) and high-throughput phenotypic screening (screening of drugs that show the desired effect in animals).

They incorporate the effects of any biological relationship into the prediction of whether a drug treats a disease. By doing this, we were able to capture a multitude of effects that have been suggested as

influential for drug repurposing including drug-drug similarity, disease-disease similarity, genetic association, drug side effects etc (see figure 2). Rephetio learns which types of compound–disease paths discriminate treatments from non-treatments in order to predict the probability that a compound treats a disease.

To build these prediction, the Rephetio project first built 'metapaths' from drugs to disease, for isolating potential mechanism of action of the drugs. They evaluated all 1206 metapaths that traverse from compound to disease and have length of 2–4. To control for the different degrees of nodes, they used the degree-weighted path count which downweights paths going through highly connected nodes. Predicting the probability of treatment for a drug and a disease relies on the 755 known treatments as positives and 29,044 non-treatments as negatives to train a logistic regression model. The features consisted of a prior probability of treatment, node degrees for 14 metaedges, and 123 metapaths that were well suited for modeling.

Figure 1: Hetionet.

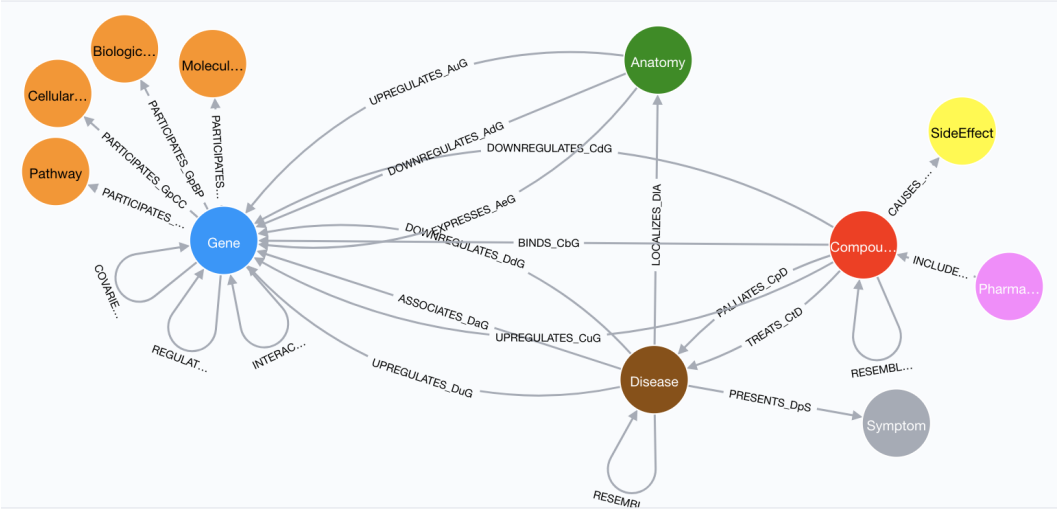


Figure 2: Hetionet Relationships

Metaedge	Abbr	Edges	Sources	Targets
Anatomy-downregulates-Gene	AdG	102,240	36	15,097
Anatomy-expresses-Gene	AeG	526,407	241	18,094
Anatomy-upregulates-Gene	AuG	97,848	36	15,929
Compound-binds-Gene	CbG	11,571	1389	1689
Compound-causes-Side Effect	CcSE	138,944	1071	5701
Compound-downregulates-Gene	CdG	21,102	734	2880
Compound-palliates-Disease	CpD	390	221	50
Compound-resembles-Compound	CrC	6486	1042	1054
Compound-treats-Disease	CtD	755	387	77
Compound-upregulates-Gene	CuG	18,756	703	3247
Disease-associates-Gene	DaG	12,623	134	5392
Disease-downregulates-Gene	DdG	7623	44	5745
Disease-localizes-Anatomy	DlA	3602	133	398
Disease-presents-Symptom	DpS	3357	133	415
Disease-resembles-Disease	DrD	543	112	106
Disease-upregulates-Gene	DuG	7731	44	5630
Gene-covaries-Gene	GcG	61,690	9043	9532
Gene-interacts-Gene	GiG	147,164	9526	14,084
Gene-participates-Biological Process	GpBP	559,504	14,772	11,381
Gene-participates-Cellular Component	GpCC	73,566	10,580	1391
Gene-participates-Molecular Function	GpMF	97,222	13,063	2884
Gene-participates-Pathway	GpPW	84,372	8979	1822
Gene→regulates→Gene	Gr > G	265,672	4634	7048
Pharmacologic Class-includes-Compound	PCiC	1029	345	724

There have been many similar works to predict drug efficacy in previously untested drug-disease pairs. Starting from the premise that similar drugs treat similar diseases, PREDICT trained a classifier that incorporates 5 types of drug-drug and 2 types of disease-disease similarity. A 2014 study compiled 890 treatments between 152 drugs and 145 diseases with transcriptional signatures [3]. The authors found that compounds triggering an opposing transcriptional response to the disease were more likely to be treatments, although this effect was weak and limited to cancers. WHAT IS THIS

2 Datasets and Discussion

In this study, I will work with the publicly available Hetionet dataset, available through both neo4j `https://neo4j.het.io/browser/`, and language specific APIs. The hetnet contains 47,031 nodes of 11 types and 2,250,197 relationships of 24 types, including 1,552 small molecule compounds and 137 complex diseases, as well as genes, anatomies, pathways, biological processes, molecular functions, cellular components, perturbations, pharmacologic classes, drug side effects, and disease symptoms.

To validate the techniques presented with Hetionet, I propose using the LINCS L1000 dataset, which contains collected gene-expression profiles for thousands of perturbagens at a variety of time points, doses, and cell lines. Each small molecule in the dataset is associated with a mechanism of action, and a curated list of diseases it is effective on, which makes it a great, curated, test set to use to assess Rephetio on Hetionet, and understand the characteristics of the prediction algorithm.

2.1 Exploration

Figure 3: Hetionet.

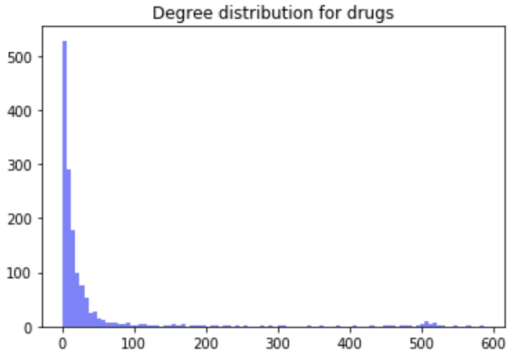
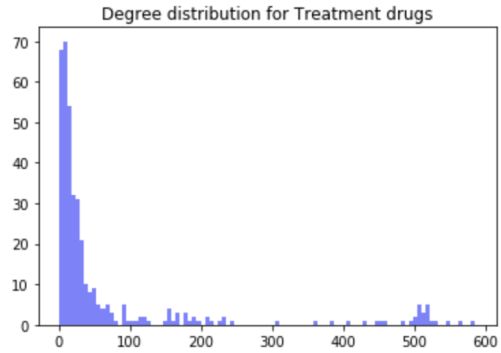


Figure 4: Hetionet Relationships



Figures 3 - 6 show my exploration of the degree distribution of the drug compounds in Hetionet. As the figures show, while the degree distribution of the overall set of drugs suggests that there are some highly connected nodes – which might indicate more known relationships between drugs, and so

Figure 5: Hetionet Relationships

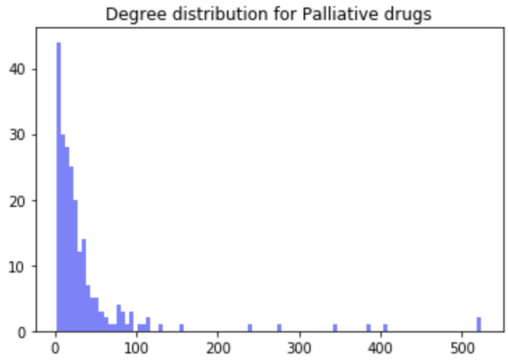
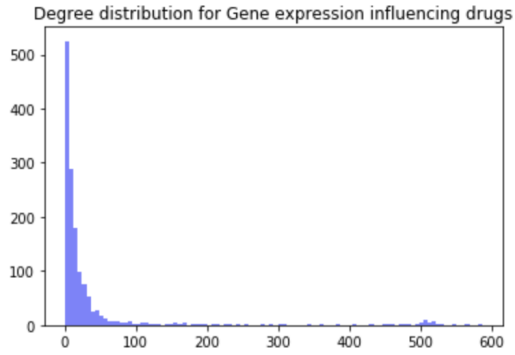


Figure 6: Hetionet Relationships



more information about these compounds. However, the degree distributions of just disease specific relationships (figures 4,5) show the maximum out degree nodes have significantly fewer connections, with the large overall out degree being accounted for by the drug-gene relationships (figure 6). This is promising for using Rephelio for drug repurposing, since it suggests that while there is not a lot of information known for drug-disease relationships, the relationship between drugs and gene expression has a lot more data, and could be used to generalize our knowledge of drug-disease relationships.

3 Method and Results

3.1 Prediction

To validate the use of Rephetio for the LINCS L1000 dataset, I ran the Rephetio prediction algorithm (using the browser at `http://het.io/repurpose/`) on both the maximally connected (figure 7) and minimally connected (figure 8) set of small molecules from LINCS L1000.

For a larger dataset, I tested Rephetio on the top 50 approved drugs that have annotations in LINCS, and have high confidence predictions in Rephetio. I limited the search to approved drugs because they high quality annotation in LINCS, and well researched mechanisms of action, I expected to lead to high quality predictions.

3.1.1 Prediction Results

For each drug, I ran the compound through Rephetio, and measured the results in terms of Recall@5, which measures whether the correct disease was in the top 5 predictions.

The maximally connected set had a 100 percent recall@5 for the highest effected disease in the curated LINCS L1000 list. However, the minimally connected set performed much worse, with about 54 percent recall@5.

Figure 7: Hetionet Relationships

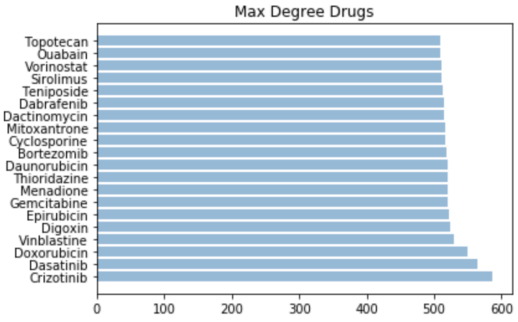
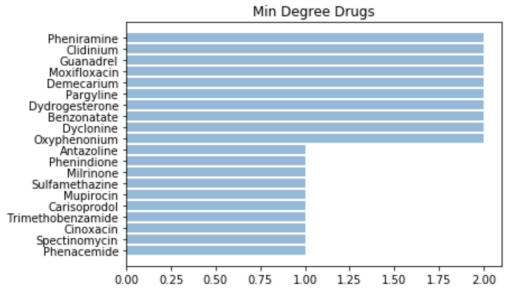


Figure 8: Hetionet Relationships



For the larger dataset, there was 78% recall. For the cases where Rephetio predicted the wrong disease for a drug, 60% were incorrectly classified as having an effect on cancer, while no such annotation existed in the LINCS database. The other mispredictions were random, with no repeated patterns.

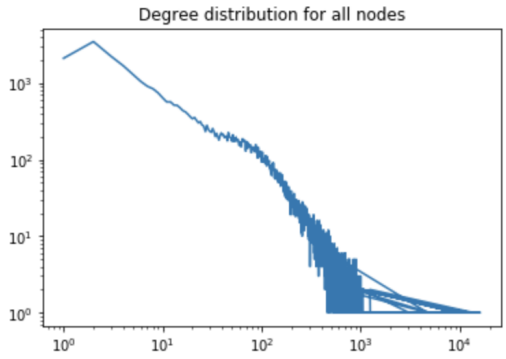
This suggests that the prediction algorithm has a systemic bias towards predicting certain diseases over other, with cancer being a potential outlier.

3.2 Graph Analysis

Cancer is highly connected in Hetionet, and so potentially the degree and connectedness of diseases and related biological mechanisms could cause this bias. To try to understand what might cause this, I conducted the following analyses.

3.2.1 Power Law Distribution

Figure 9: Distribution of Degrees



A power law is a functional relationship between two quantities, where one quantity varies as a power of another, independent of the initial size of those quantities. A power law network is one whose degree distribution follows a power law, at least asymptotically. Many real world networks have been reported to be power law networks.

To tell if a network is a power law network, we can plot the degree distribution histogram on a log-log scale, with the node degrees on the x-axis, and their respective counts on the y-axis. Such a "log-log" plot thus provides a quick way to see if one's data exhibits an approximate power-law: it is easy to see if one has an approximately straight line, which suggests that the network is a power law network.

I created such a plot for Hetionet, shown in Figure 9. The straight line in the plot suggests that there is a power law relationship for this network. Considering this under the rich-get-richer phenomena, this could be caused by biases towards certain diseases in literature and research, causing more information to exist about certain biological processes, and certain popular compounds over others. In this study conducted by Journal of the American Medical Association (JAMA), analyzing medical research funding worldwide between 1994 and 2012, the results suggest that cancer and HIV/AIDS receive disproportionate support of research funding. This is consistent with the above hypothesis.

The following two analyses were conducted to see if the network topography suggests a bias towards cancer, that could cause the results I saw in section 3.1.1.

3.2.2 Hubs

A hub is a component of a network with a high-degree node. Hubs have a significantly larger number of links in comparison with other nodes in the network. Hubs tend to exist more often in Power Law networks, than in random networks. Random networks have nodes all with the a comparable degree k, while in power-law networks, some nodes (hubs) have a much higher degree than average, and most nodes have a low degree.

The hubs in a network are also responsible for effective spreading of material on network. This could apply for both information flow in human networks, or for studying disease spreading modeling using hubs.

I calculated the top ten hubs for the Helionet network, and the results are summarized below. Note that this was run on the whole network, and not just for Anatomy nodes. Given this, its quite remarkable that all the hubs in the network are anatomical components or organs of the body. While this is inconclusive towards pointing towards disease biases, it quite informative of the meta pathways in the network.

1	endocrine gland
2	midbrain
3	brain
4	liver
5	prostate gland
6	urethra
7	adipose tissue
8	vagina
9	adrenal gland
10	nervous system

3.2.3 Page Rank

PageRank is a way of measuring the importance of website pages. PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.

In this case, Page Rank can serve as a measure of the importance of a node in the graph, in a more holistic manner that just the immediate degree of the node, since the influence of the whole network is factored into Page Rank.

For the Hetionet, I calculated the top 10 ranking nodes, which are listed in the table below. Note that all of these are high level biological functions, that are central to a lot of basic function. For example,

Gene expression is the process by which DNA dictates which proteins are synthesis in a cell, and G-protein coupled receptor activity" is a mechanism through which a large percentage of major drugs attach to receptors in the body.

However, both "G-protein coupled receptor activity"[5] and "nucleolus"[6] have recently been shown to be a very promising target for cancer drugs, and implicated in a the disease's mechanism of action. Since they are part of the disease pathway, this could potentially explain the overall bias of the network towards cancer.

1	G-protein coupled receptor activity
2	nucleolus
3	multi-organism reproductive process
4	Generic Transcription Pathway
5	RNA processing
6	Gene Expression
7	nuclear outer membrane-endoplasmic reticulum membrane network
8	hydrolase activity, acting on ester bonds
9	polymeric cytoskeletal fiber
10	RPS4Y1

4 Discussion

Drug repurposing using Rephetio is very promising, and could be used to identify new compounds for disease models. Overall Rephetio performed very well for the test set I derived from LINCS1000 with approved drugs, but it might be subject to some systematic biases towards some diseases. Recall that Rephetio uses prior probability of treatment, node degrees for metaedges, and metapaths, where metapaths are normalized by the degree of each node. However, the above analysis might suggest that just normalizing by the degree of each node is not enough to remove the desired effects, and it might be necessary to use a more complete metric, like Page Rank.

5 Future Work

The test data I presented in this work only use approved drugs, which could potentially be less useful for repurposing that less known compounds. It would be interesting to do a deeper dive on the performance of the network on relatively unstudied compounds, and evaluate using the compounds method of action.

References

[1] Daniel Himmelstein, Antoine Lizee, Chrissy Hessler, Leo Brueggeman, Sabrina Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, Sergio Baranzini (2016) Rephetio: Repurposing drugs on a hetnet [report]. Thinklab. doi:10.15363/thinklab.a7

[2] Koleti A, Terryn R, Stathias V, Chung C, Cooper DJ, Turner JP, Vidovic D, Forlin M, Kelley TT, D'Urso A, Allen BK, Torre D, Jagodnik KM, Wang L, Jenkins SL, Mader C, Niu W, Fazel M, Mahi M, Pilarczyk M, Clark N, Shamsaei B, Meller J, Vasiliauskas J, et al. 2017. Data Portal for the Library of Integrated Network-based Cellular Signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. Nucleic Acids Research. gkx1063.

[3] Systematic evaluation of connectivity map for disease indications Jie Cheng, Lun Yang, Vinod Kumar, Pankaj Agarwal (2014) Genome Medicine. doi:10.1186/s13073-014-0095-1

[4] Hamilton Moses, David H. M. Matheson, Sarah Cairns-Smith, Benjamin P. George, Chase Palisch, E. Ray Dorsey. The Anatomy of Medical ResearchUS and International Comparisons. JAMA. 2015;313(2):174–189. doi:10.1001/jama.2014.15939

[5] Dorsam, R. T., Gutkind, S. (n.d.). G-protein-coupled receptors and cancer. Nature. Retrieved from https://www.nature.com/articles/nrc2069

[6] Montanaro L, Treré D, Derenzini M. Nucleolus, Ribosomes, and Cancer. The American Journal of Pathology. 2008;173(2):301-310. doi:10.2353/ajpath.2008.070752.