Abstract

Computational drug repurposing aims at finding new medical uses for existing drugs. The identification of novel drug target interactions (DTIs) can be a useful part of such a task. Computational determination of DTIs is a convenient strategy for systematic screening of a large number of drugs in the attempt to identify new DTIs at low cost and with reasonable accuracy. This necessitates development of accurate computational methods that can help focus on the follow-up experimental validation on a smaller number of highly likely targets for a drug. Although many methods have been proposed for computational DTI prediction, they suffer the high false positive prediction rate or they do not predict the effect that drugs exert on targets in DTIs. In this report, first, we present a comprehensive review of the recent progress in the field of DTI prediction from data-centric and algorithm-centric perspectives. The aim is to provide a comprehensive review of computational methods for identifying DTIs, which could help in constructing more

reliable methods. Then, we present DDR, an efficient method to predict the existence of DTIs.

1.Introduction

1.1 Motivation

In the past most drugs have been discovered either by identifying the active ingredient from traditional remedies or by serendipitous discovery. A new approach of Drug discovery has been to understand how disease and infection are controlled at the molecular and physiological level and to target specific entities based on this knowledge.

Although some adverse drug reactions (ADR) are not very serious, others cause the death of more than 2 million people in the United States each year, including more than 100,000 fatalities. Drug development time frame can range from 3–20 years and costs can range between several billion to tens of billions of dollars

1.2 Problem Definition

- The number of newly approved drugs by the FDA is decreasing due to in the unacceptable toxicity and adverse side-effects for those drug candidates.
- Recent research definitely indicates that harmful side effects is due to namely off-target effects in addition to the primary therapeutic targets.
- Studies also showed that most of the FDA-approved drugs can have interaction with multiple targets (proteins). Aiming to reduce the spent cost and time of bringing a new drug to the market. The purpose of drug repositioning is the detection for new clinical uses for existing drugs that have already been strictly verified for their safety and bioavailability and narrow down the scope of search of candidate medications .so drug repositioning, is another important part in drug discovery. As The —multi-target, multi-drug‖ in place of —one target, one drug‖ model has been widely accepted in order to speed up the drug development process.
- Serendipity is one of the many factors that may contribute to drug discovery, it means finding of one thing while looking for something else.
- The known drug-target interactions based on wet-lab experiments are limited to a very small number due to cost and time.

1.3 Objectives

- Motivated by the limitations of the exiting methods for the prediction of the potential DTIs, the types of the effects a drug exert on the target, and aiming to further improve their prediction accuracy, we present a method that utilize a heterogeneous drug-target graph that contains information about DTIs, effect types of DTIs, as well as multiple similarities between drugs and multiple similarities between target proteins.
- The main objective of our idea is to find computational strategy to identify drug–target interactions (DTIs) types at low cost with reasonable accuracy in order to speed the drug development process.
- Avoid the side effects of a drug and Link the newly identified DataTarget interactions of a known drug to the treatment of diseases that are different from diseases for which the drug has been originally developed.

Project development methodology

Iterative Planning is the process to adapt as the project unfolds by changing the plans. Plans are changed based on feedback from the monitoring process or changes in the project assumptions.

It's a Team Effort - It is important to involve the team in the planning process. The people doing the work should be actively involved in planning the project. When they get involved in the decision, they become motivated to get it right. After all, they were hired and they have the skills to understand the dependencies. Once they complete the plans, they will own it and will accept the schedule. That's why everyone in the team worked on a certain thing during the whole project period and exchanged many tasks.

We used this developmental approach as we monitored the performance and changed some algorithms and methods for delivering some valuable improvements or additional features in each increment.

First of all we went through the process of breaking down big problems into smaller tasks, where we divided the project into 3 ways which are: - features, similarities and interactions. At the first iteration, we collected dataset from DrugBank which was CSV file containing 13,581 drugs with their features which were: type, state, groups, kingdom, superclass, subclass, direct parent and class. At the second iteration we started calculating Euclidean distance and Manhattan distance between the 13,581 drugs. Unfortunately, we found that at this iteration that the features we extracted didn't show true similarity for the collected drugs. At the third iteration, we downloaded another whole new dataset from DrugBank which was SDF file containing 11,160 drugs and we extracted some features which were: ID and chemical structure (for the fingerprint to be calculated) and started calculating the fingerprint for each molecule by using Tanimoto similarity which is the most popular similarity measure for comparing chemical structures represented by means of fingerprints, two structures are usually considered similar if $T > 0.85$, but some pairs of compounds didn't get a high score, even they look very similar with each other. Below is one example: 6037 (Pubchem CID) ------ DB00988 (the similarity is only 0.61). At the fourth iteration we tried another fingerprint method which was the Morgan fingerprint; Morgan fingerprints are circular fingerprints, which are also topological fingerprints. They are obtained by modifying the standard Morgan algorithm. It can be roughly equivalent to Extended-Connectivity Fingerprints (ECFPs). This type of fingerprint has many advantages, such as fast calculation speed, no predefined (which can represent an infinite number of different molecular characteristics), can contain chiral information, each element in the fingerprint represents a specific substructure, and can be easily Carry out analysis and interpretation, and make corresponding modifications according to different needs. The initial purpose of this type of fingerprint design is to search for molecular features related to activity, not substructure search. In addition, it can also be used in similarity search, clustering, virtual screening and other directions.

So we found at this iteration better results for similarity than the past one as the Morgan fingerprint is well suited to capture structural differences (atom type, bond type, connectivity etc).

So we used the Dice coefficient is the number of features in common to both molecules relative to the average size of the total number of features present, The equation for this concept is: $2 * |X \cap Y| / (|X| + |Y|)$.

After finishing the drug part we went through collecting dataset for protein from Uniprot as a FASTA file which had 563975 records and extracted these records by using BioPython to deal only with protein id and sequence, then we tried to get the similarity between these proteins by using Pairwise Sequence Alignment which is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid).

After finishing the similarities part, we collected a new dataset containing interactions between drugs and proteins which held 11,060 record in a CSV file and we used only the drug id and the protein id in the data.

At the end of the fourth iteration we managed to get the best score for the drug similarity using the Dice similarity and the best score for protein similarity using Blast similarity.

At the fifth iteration we went through a new process which is the graph node embedding. Node embedding algorithms compute low-dimensional vector representations of nodes in a graph. These vectors, also called embeddings, can be used for machine learning. In our project, we created a graph using NetworkX library in python and used DeepWalk algorithm on the graph we created.

DeepWalk is an algorithm that is used to create embeddings of the nodes in a graph. The embeddings are meant to encode the community structure of the graph. To obtain node embeddings, we first need to arrange for sequences of nodes from the graph. How do we get these sequences from a graph? Well, there is a technique for this task called Random Walk.

Random Walk is a technique to extract sequences from a graph. We can use these sequences to train a skip-gram model to learn node embeddings. So we extracted sequences from the nodes in this graph.

After generating node-sequences, we have to feed them to a skip-gram model to get node embeddings. That entire process is known as DeepWalk.

This leads us to another new concept which is Word2vec, Word2vec is a two-layer neural net that processes text by “vectorizing” words. Its input is a text corpus and its output is a set of vectors: feature vectors that represent words in that corpus. While Word2vec is not a deep neural network, it turns text into a numerical form that deep neural networks can understand. The output of the Word2vec neural net is a vocabulary in which each item has a vector attached to it, which can be fed into a deep-learning net or simply queried to detect relationships between words. The Word2vec model is dependent on many parameters which are: skip gram, word-size and hierarchical softmax.

Skip gram; the aim of skip-gram is to predict the context given a word. In our project we used skip gram in order to predict the next or previous similar protein or drug. Two separate errors are calculated with respect to the two target variables and the two error vectors obtained are added element-wise to obtain a final error vector which is propagated back to update the weights. The weights between the input and the hidden layer are taken as the word vector representation after training.

The idea for using skip gram model:

-Skip-gram model can capture two semantics for a single word

-Skip-gram with negative sub-sampling outperforms every other method generally.

The softmax layer is a core part of many current neural network architectures. When the number of output classes is very large, such as in the case of language modeling, computing the softmax becomes very expensive. This post explores approximations to make the computation more efficient. H-Softmax essentially replaces the flat softmax layer with a hierarchical layer that has the words as leaves. This allows us to decompose calculating the probability of one word into a sequence of probability calculations, which saves us from having to calculate the expensive normalization over all words. Replacing a softmax layer with H-Softmax can yield speedups for word prediction tasks of at least 50× and is thus critical for low-latency tasks.

After finishing the whole Word2vec process, the result is a new dataset containing all the node embeddings with their features and holding a record of 4864 rows x 64 columns holding features.
Using our drug-target interactions dataset, we extract the vectors representing each interaction between a drug and protein whether it’s expressing a positive or negative interaction. After this process, we get a new dataset which is interaction embedding dataset.

At the sixth iteration we'll go through our machine learning phase, in this phase we tried 29 different models and we'll refer to these models in the testing chapter. The best model we figured was the SVM model.

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text. Compared to newer algorithms like neural networks, they have two main advantages: higher speed and better performance with a limited number of samples (in the thousands). This makes the algorithm very suitable for text classification problems, where it's common to have access to a dataset of at most a couple of thousands of tagged samples.

A support vector machine takes the data points and outputs the hyperplane (which in two dimensions it's simply a line) that best separates the tags. This line is the decision boundary: anything that falls to one side of it we will classify as blue, and anything that falls to the other as red.

A kernel is a method of placing a two dimensional plane into a higher dimensional space, so that it is curved in the higher dimensional space. (In simple terms, a kernel is a function from the low dimensional space into a higher dimensional space.)

One of the powerful kernel tricks is the RBF kernel, RBF kernel is a function whose value depends on the distance from the origin or from some point.

The seventh and the last iteration is the GUI part, in this part we used streamlit framework, Streamlit is an open-source app framework for Machine Learning and Data Science teams.

We created a front-end demo using this framework to show the whole functionalities for our project like drug-drug similarity, protein-protein similarity and prediction of new drug or new protein using our built SVM model.

And we'll refer to this at Chapter 5.

Tools

- Streamlit
- Lucidchart
- Creately
- Microsoft (gantt chart template)
- Git and Github

- Google collab
- Anaconda

2.Background

2.1 Biological Background

2.1.1 Protein Structure and Function

The building blocks of proteins are amino acids; the defining feature of an amino acid is its side chain. When connected together by a series of peptide bonds, amino acids form a polypeptide, another word for protein. The polypeptide will then fold into a specific conformation depending on the interactions forming ionic bonds, hydrogen bonds, van der Waals interactions, covalent bonds (By Cysteines) between its amino acid side chains.

Assigning the same function to things that look similar is innate to human nature. The same process can be used for protein sequences:

This is called **homology annotation** and the principle that enables scientists to use similarity to infer function is based on the conservation of a given sequence or slight variations of it throughout evolution. In general terms, the more similar two sequences are, the more likely they are to be related. Consequently, homology annotation is based on the comparison of DNA or proteins at the sequence level - that is, by comparing the similarity of nucleotides or amino acids sequences between related sequences. Protein sequences that confer function are often found in blocks of conservation called protein domains. These regions have a defined three-dimensional structure or motif (shape) that can function and evolve independently from the rest of the protein sequence. These blocks of conservation are found in proteins throughout nature, and any given protein sequence can have more than one protein domain. The key to using motif similarity to infer function relies on the principle that when two proteins have a conserved function, although their sequence similarity at the amino acid level can be lost, their protein domain conservation must remain.

2.1.2 Drug Structure and Function

All drugs are chemical compounds, prepared from chemical reactions like condensation reaction, cyclization reaction, elimination reaction, substitution reaction. The chemical structure for drugs determines the shape of the drug and how to interact with the binding sites of the protein by chemical bonds like ionic, hydrogen, and covalent bonds, and van der Waals forces. Hydrogen and ionic are the most common types of drugs and proteins

bonds require little energy and are made and broken easily the chemical structure of the drug determines these bonds

Pharmaceutical drugs are often classified into drug classes—groups of related drugs that have similar chemical structures, the same mechanism of action (binding to the same biological target), a related mode of action, and that are used to treat the same disease. The Anatomical Therapeutic Chemical Classification System (ATC), the most widely used drug classification system, assigns drugs a unique ATC code, which is an alphanumeric code that assigns it to specific drug classes within the ATC system. Another major classification system is the Biopharmaceutics Classification System. This classifies drugs according to their solubility and permeability or absorption properties.

Figure-2 Smile format

2.1.3 Drug design

Drug Design:

In the most basic sense, drug design involves the design of small molecules that are complementary in shape and charge to the bio molecular target with which they interact and therefore will bind to it. Drug design frequently but not necessarily relies on computer modeling techniques. This type of modeling is often referred to as **computer-aided drug design (LB-CADD)**.

Finally, drug design that relies on the knowledge of the three-dimensional structure of the bio molecular target is known as **structure-based drug design**. The phrase —drug designǁ is to some extent a misnomer. What is really meant by drug design **is ligand design** (i.e., design of a small molecule that will bind tightly to its target). Drugs may be designed that bind to the active region and inhibit this key molecule. Another approach may be to enhance the normal pathway by promoting specific molecules in the normal pathways that may have been affected in the diseased state.

2.1.4 Drug Development Challenges

- Drug development is a lengthy, complex, and costly process, entrenched with a high degree of uncertainty that a drug will actually succeed.
- The unknown pathophysiology for many nervous system disorders makes target identification challenging. • Animal models often cannot recapitulate an entire disorder or disease.
- Challenges related to heterogeneity of the patient population might be alleviated with increased clinical phenotyping and endotyping.
- Greater emphasis on human data might lead to improved target identification and validation.
- FDA organization doesn’t approve on new drugs as it need long time till it makes sure that these drugs have no dangerous effects on people As a result, to these challenges we need fast and safe way to discover if there are drugs already made and can target more than protein and this can be made by DDR method.

2.1.5 Common Approaches of Targeted Drug Delivery (smart drug delivery)

1) Controlling the distribution of drug by incorporating it in a carrier system.
2) Altering the structure of the drug at molecular level.
3) Controlling the input of the drug into biological environment to ensure a programmed and desirable bio distribution.

2.1.6 Properties of Ideal Targeted Drug Delivery

1) Nontoxic, biocompatible and physicochemical stable in vivo and in vitro. 2) Restrict drug distribution to target cells or tissue or organ or should have uniform capillary distribution.
3) Controllable and predictable rate of drug release.
4) Minimal drug leakage during transit.
5) Carrier used must be biodegradable or readily eliminated from the body without any problem. (decomposed)
6) Its preparation should be easy or reasonably simple, reproductive and cost effective.

2.2 Computational Background

2.2.1 Computational Approaches for DTI prediction

There are different approaches that have proposed for addressing the problem of predicting new DTIs; some major prediction approaches are docking simulation, ligand-based approaches, machine learning, network inference and text mining approaches. Although many methods have been proposed for computational DTI prediction, they suffer the high false positive prediction rate. In the following sections we characterize the commonly used computational techniques and how they tackled the problem of predicting new DTIs.

Figure-3 Summary of the computational approaches for the DTI prediction

2.2.2 Early attempts in the DTI prediction

Early attempts in computational prediction of DTIs can be categorized into two main groups and include docking simulations and ligand-based approaches. These approaches typically focus on one particular target of interest (i.e., the 3D structure of the target (or the compound). Docking methods consider the three-dimensional structure of target proteins. However, this approach is extensively time-consuming and the structural information of targets is not available for all target proteins. Ligand based methods compare a query ligand with a set of known ligands with target proteins. However, it may not perform well in cases the number of known ligands with target proteins is small.

2.2.3 The use of network topology in the DTI prediction

The uses of the topology of DTI network, as the only source of information for the prediction of new DTIs links is capable of predicting true DTIs with reasonable accuracy. Some of the DTI prediction methods are based on

only considering known DTIs using techniques based on graph theory and network analysis. For instance, the prediction model built using only interaction profiles from known DTIs, compared to the model that uses additional information about the chemical structure and genomic sequence similarities, can be used as an accurate tool for prediction of DTIs. The relevance of using DTI topology network as a source of information for predicting new DTI is based on the assumption that drugs exhibiting a similar pattern of interaction and non-interaction with the targets of a DTI network are likely to show similar interaction behavior with respect to new targets. (homology annotation)

2.2.4 DDR Method

The heterogeneous DTI graph is a weighted graph that is constructed with m nodes from the drug set and n nodes from the set of target proteins. The edge between two drug nodes or two target protein nodes represents the similarity between them and is weighted by the similarity value obtained

from the similarity calculation. The edge between a drug and a target protein represents a known DTI and is weighted by 1.

Figure-4 DTI Graph

2.2.5 Similarity Measures

-Similarity Matrix between drugs and each other, its size is m*m, where m is number of drugs. ranges between [0,1].

-Similarity matrix between proteins and each other, its size n*n where n is the number of targets, ranges between [0,1]. the closer to zero the less similar.

2.2.6 Inferring interaction profile

Drugs (or target proteins) with high similarities to a new drug (or a new target protein) are said to be the neighbors of the drug (the target protein), the inferred value of interaction for a new drug with a specific target protein is represented as the ratio of the sum of similarity values for drug. neighbors interacting with this target protein relative to the total sum of all neighbors‟ similarity values.

2.2.7 Graph Mining

Is the set of tools and techniques used to (a) analyze the properties of real-world graphs, (b) predict how the structure and properties of a given graph might affect some application, and (c) develop models that can generate realistic graphs that match the patterns found in real-world graphs of? interest.

Graph Mining has played an important role in a range or real world applications:

- Medicines: structure of molecules.
- Bioinformatics: biological networks. Social Science: social networks.

We formalize data mining and machine learning challenges as graph **problems.**

Graph mining, which has gained much attention in the last few decades, is one of the novel approaches for mining the dataset represented by graph structure.

Graph mining finds its applications in various problem domains, including: bioinformatics, chemical reactions, Program flow structures, computer networks, social networks.

Different data mining approaches are used for mining the graph-based data and performing useful analysis on these mined data.

Various graph mining approaches have been proposed. Each of these approaches is based on either classification; clustering or decision trees data mining techniques.

2.2.8 Path-category-based features

A path structure of a path that starts at a D node and ends up at a T node describes a subgraph that sequentially links drug and target protein nodes. For example, a path Drug1-Drug2-Target1 connects the Drug1 node with the Target1 node through the similarity edge between Drug1 and Drug2 and via the interaction edge between Drug2 and Target1. The path structure of this path is D-D-T. All paths with more than one edge and without loops, starting at a D node and ending at a T node, and having the same path structure define a path-category on the heterogeneous DTI graph.

2.2.9 Predicting drug–target interactions

We will consider these path-categories through which drug nodes could connect to target protein nodes. To do this we start with a given drug di to reach a given target protein tj through a specific path-category. We restrict traversing the graph to retrieve all paths passing only through the K-nearest neighbors of drugs to di and only through the K-nearest neighbors of target proteins to tj. Next, for each path we calculate an edge-weight product value obtained by multiplying all weights of edges of these paths.

2.2.10 Machine Learning

- **KNN:**

  The k-nearest neighbors (KNN) algorithm is a simple, easy-to implement **supervised machine learning algorithm** that can be used to solve both classification and regression problems.

  The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

- **Decision Tree**

  Covering both **classification and regression as supervised machine learning algorithm**. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Though a commonly used tool in **data mining** for deriving a strategy to reach a particular goal.

- **SVM (Support-vector machine)**

  More formally, a support-vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin, the lower the generalization error of the classifier.

- **Random Forest**

  Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. **"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."** Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

1-method for the large-scale prediction of drug indications (PREDICT) that can handle both approved drugs and novel molecules

This method is based on the observation that similar drugs are indicated for similar diseases, and utilizes multiple drug–drug and disease–disease similarity measures for the prediction task on cross-validation, We validate our predictions by their overlap with drug indications that are currently under clinical trials, and by their agreement with tissue-specific expression information on the drug targets we further show that disease-specific genetic signatures can be used to accurately predict drug indications for new diseases.

2-Network-based Random Walk with Restart on the Heterogeneous network (NRWRH) is developed to predict potential drug–target interactions on a

large scale under the hypothesis that similar drugs often target similar target proteins and the framework of Random Walk. Compared with traditional supervised or semi-supervised methods, NRWRH makes full use of the tool of the network for data integration to predict drug–target associations.

It integrates three different networks (protein–protein similarity network, drug–drug similarity network, and known drug–target interaction networks) into a heterogeneous network by known drug–target interactions and implements the random walk on this heterogeneous network.

When applied to four classes of important drug–target interactions including enzymes, ion channels, GPCRs and nuclear receptors, NRWRH significantly improves previous methods in terms of cross-validation and potential drug–target interaction prediction.

3-A simple machine learning method that uses the drug-target network as the only source of information is capable of predicting true interaction pairs with high accuracy we introduce interaction profiles of drugs (and of targets) in a network, which are binary vectors specifying the presence or absence of interaction with every target (drug) in that network. We define a kernel on these profiles, called the Gaussian Interaction Profile (GIP) kernel, and use a simple classifier, (kernel) Regularized Least Squares (RLS), for prediction drug-target interactions.

1-Proteins Database: (UniProt)

Which is online database containing proteins features such as Protein’s id, name, sequence which will be used to find similarity between proteins

2-Drugs Database: (DrugBank)

Which is online database containing drugs features that we will use to find the similarity between them such as group, status, targeted proteins, class, sub-class.

3-Drugs and Targets positive (known) interactions Database: (DrugBank)

Which is online database containing the already known interactions between the drug database and the proteins database.

3.2 Stakeholders

- Pharmacists
- Drug companies
- Researchers
- Bioinformaticians

Reference:

1- DRUG TARGET PREDICTIONS BASED ON HETEROGENEOUS GRAPH INFERENCE:

(https:/www.ncbi.nlm.nih.gov/pmc/articles/PMC3605000/#S2title)

2- Prediction of Drug–Target Interactions from Multi-Molecular Network Based on Deep Walk Embedding Model:

(https://www.frontiersin.org/articles/10.3389/fbioe.2020.00338/full)

3- A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information:

(https://www.nature.com/articles/s41467-017-00680-8#Sec24)

4- Drug–target interaction prediction by learning from local information and neighbors:

(https://academic.oup.com/bioinformatics/article/29/2/238/203064)

5- Machine learning basics with the k-nearest neighbors algorithm

(https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest neighbors-algorithm-6a6e71d01761) 6- Graph mining:

(https://www.researchgate.net/publication/228387780_Graph_mining_An_overview)

7- Random forest algorithm:

)https://www.javatpoint.com/machine-learning-random-forest-algorithm)

8-Pre-study in time plan

(https://www.veronicastenberg.com/how-to-make-a-clear-and-visual-time-plan/)

9- Utility class in class diagram

(https://www.uml-diagrams.org/classreference.html#:~:text=Utility%20is%20class%20that%20has%20only%20class%20scoped%20static%20attributes%20and%20operations.&text=Abstract%20class%20was%20defined%20in,instance%20of%20an%20abstract%20class.)