Computational Approaches Towards Drug Repositioning by Combining Heterogeneous Network Model and Random Walk Algorithm for Medications in Neurological Disorders

Bipin Nair B J  
Department of computer science  
Amrita School of Arts and Science, Mysuru Campus  
Amrita VishwaVidyapeetham, India  
bipin.bj.nair@gmail.com

Abstract: Available drug information including the drug disease relationships are collected from data sources like Drug Bank, ChEMBL etc. A heterogeneous network is formed using collected data set of drugs and diseases from various data sources. Relationships among the elements in the network are identified. Random walk algorithm is applied in the two layer drug-disease network to identify new potential targets for existing drugs. A combined approach of Heterogeneous Network model and Random walk algorithm on this network was able to identify new potential targets for existing drugs.

Key words: SMILES, Drug Repositioning, Random Walk Algorithm

I. INTRODUCTION

Cheminformatics is an area which integrates Computer science, Machine learning, and Chemistry. It helps in storage, retrieval, and manipulation of chemical data such as drugs 2D chemical structures. It provides tools and techniques for drug discovery and Drug Repositioning. Cheminformatics tools help in plotting Drugs 2D, 3D structures, comparing drugs structures, manipulating structures etc. Cheminformatics helps in fastening the drug discovery process by providing computational methods for handling and manipulating chemical data.

Bioinformatics is an area which helps in managing biological data. It provides tools and techniques for manipulating protein structures, visualization of protein-ligand binding etc. Bioinformatics helps in storage, retrieval, and manipulation of large amount of biological data.

Drug Repositioning or Drug Repurposing is the area in which new applicable targets for existing drugs are found. It uses the tools and techniques from both Bioinformatics and Cheminformatics. Drug related information such as chemical structure are manged using Cheminformatics tools and Bioinformatics tools are used to handle biological data such as protein structures. Computational Drug Repositioning helps in reducing the time and cost for the drug discovery process.

II. LITERATURE SURVAY

Luo, H. et al. [1] The drug relocation based on the comprehensive similarity measure and the double random walk algorithm shows that compared with other algorithms, the double random walk algorithm is the best in the effective prediction of drug-disease relationship. It uses a multi-layered network of relationships between drugs and disease. Wang W, Li et al. [2] Drug Relocation by Integrating Target Information Through Heterogeneous Network Models: This article presents a two-layer network model and a three-layer network model. In the two-layer network model there will be a network of drug similarity and disease similarity networks. There is an additional target network in the three-tier network model. Cheng & Tang, Y. et al. [3] Prediction of Drug Target Interactions and Drug Repositioning Via Web-Based Inference Description - Finding DTI (Drug Targeting Interaction) NBI (web-based inference) or web-based drug discovery is the most effective of other approaches such as DBSI. Deshpande T & Butte et al [4] Abusing drug-disease associations for computational Drug Transposition explains about the two types of Drug Repositioning approaches one is the drug based Repositioning and other is the disease based Repositioning. Li, Y. & Jones S. J. et al. [5] Drug Transposition for modified medicine says that - The two approaches Personalized medicine and Drug Repositioning both aim to increase the efficiency of current drug discovery methods. This paper explains about the current research developments going on in these fields. Jin, G & et al. [6] Achieve better drug repositioning: Prioritize and integrate existing approaches into effective pipelines: This article describes a variety of available drug repositioning methods such as blinded search or screening methods, goal-based approaches, pathways, or based on Network method. The authors explain how to improve the methods available for effective drug relocation. Sirotà & Butte, A. J. et al. [7] Discovery and preclinical validation of drug indications using a public gene expression database: The authors sought to determine a new relationship between drug elements and disease through the integration of available gene expression data on available drugs and diseases. Public gene expression data is obtained from a data source such as NCBI. GEO. Xu, H., Tao Y & Huang, L. et al. [8] Yuanhu analgesic side of the calculation of drug target network: Yuanhu analgesic side (YZP) is a traditional Chinese medicine methodology. In this article, they are now calculating the new target for the use of YZP drugs with the help of the drug repositioning method. Nair, B. B., Bhaskaran, V., & Arumjigit, K. (2017). [9] Structure Design of Autism Spectrum Disease Suppressor Using Molecular
Dynamics Sketches. International Journal of Drug Delivery, 8 (4), 142-146: In this paper, by means of computational methods, the authors are mapping the chemical design of drugs for autism spectrum disorders. Compounds such as fluoxetine, risperidone, melatonin structure successfully produced. Bipin Nair B.J., Arunjit K., ViheshBhaskaran et al.[10]Melatonin and fluoxetine interaction with shank3 protein gene for autism spectrum disorder: In this paper authors are visualizing the interaction of SHANK3 gene (which is related to Autism spectrum disorder) with the drugs Fluoxetine, Melatonin. This method can be used to check the efficiency of binding among drugs and diseases. Namboori, P.K.(2017)[11].Evaluation of Colorectal Cancer (CRC) Epidemiology A Pharmacogenomic Approach. Journal of Young Pharmacists, 9(1), 36: In this paper an analysis of various ACP gene (Adenomatous polyposis coli) mutations which are responsible for CRC (colorectal cancer) is carried out. A population wise analysis of ACP gene mutations is done. A research on factors which cause colorectal cancer is carried out with genetic, epigenetic, metagenomic and environmental factors. The results of this work shows the impact of these factors in ACP gene mutations.

III. PROBLEM FORMULATION

Existing drug development methods or traditional drug discovery methods are mainly based on cell-based or target-based analysis of elements to find out a small subset of ‘hits’, result of the process are then analyzed to improve their affinity, efficiency and selectivity.

Traditional drug development methods which are currently following is both costly and time consuming. Traditional drug development methods can cost more than 1 Billion dollars and will take around 12 years to bring a new drug to market. On the other hand Computational Drug Repositioning will take only about 3 years of time and about 250K dollars as an expense to bring new drugs to the market. When we compare the success rates traditional drug development methods shows a very low success rate, where 1 in 10000 drugs released is successful. Drug Repositioning approach have a very high success rate, where 3 in 10 drugs released are successful.

IV. PROBLEM DEFINITION

Autism spectrum, Depression, Alzheimer’s disease, Parkinson’s disease are all neurological disorders which mainly affects the nervous system of the body. The number of drugs available for these diseases are very less. These available drugs lack accuracy and efficiency. The traditional drug development methods which are currently being followed is both costly and time consuming. Computational drug development methods are both cost efficient and time efficient. It have a very high success rate. Our work is based on a Computational approach towards Drug Repositioning. The overall work can be divided to the following stages: Database collection, Clustering, Classification, Feature selection, Drug Repositioning.

Database is collected from sources like Drug Bank, Protein Data Bank, ChEMBL, NCBI etc. We collected information about 60 drugs which are used to treat neurological diseases such as Autism spectrum, Depression, Parkinson’s disease, and Alzheimer’s disease. Collected drug data set contains about 30 different properties for each drugs. It include properties like Drug SMILES, InChI key, molecular weight, water solubility, biological half life, protein binding etc. The data base is stored using SQL data base server.

The data set is obtained from sources like drug bank, Protein Data Bank, ChEMBL etc. We collected information about 60 drugs which are used to treat neurological diseases such as Autism spectrum, Depression, Parkinson’s disease, and Alzheimer’s disease. Collected drug data set contains about 30 different properties for each drugs. It include properties like Drug SMILES, InChI key, molecular weight, water solubility, biological half life, protein binding etc.

V. METHODOLOGY

A. Clustering

Clustering is an unsupervised machine learning method, where a set of data is given as an input. In unsupervised machine learning the input data is grouped into different clusters. But
the criteria for clustering is not fixed. Clusters are made using analyzing the input set of data. Here we are using a modified version of K-Means clustering algorithm.

**Data input**: Molecular Weight, Water solubility.

```python
def cal_mean() :
    for i in range(p) :
        m[i] = 0
    for i in range(p) :
        cnt = 0
        for j in range(n-1) :
            if k[i][j] != -1 :
                m[i] += k[i][j][1]
                cnt += 1
        m[i] /= cnt
```

The algorithm clusters given data into various groups based on some common criteria. We considered two chemical properties of the drugs for the clustering purpose: Molecular Weight, and Water Solubility. In which the Molecular Weight based clustering was able to equally cluster the input drugs into different groups.

![Fig.3 K-Means clustering output](image)

The algorithm assigns mean values of the clusters. Then for each input data it measures the distance from the mean value, the input will be assigned to a cluster in which the measured value is minimum. Algorithm will recompute the mean value by totaling all the values in the cluster and dividing by the number of elements in the cluster.

Re-computation of the mean values will be continued until the values become constant.

**B. Classification**

Classification is a supervised machine learning. In which the data is grouped into various classes. In our work we are using a classification approach similar to Scaffold Tree classification. In this approach data is grouped based on common substructures.

**Data input**: SMILES (Simplified molecular input level Entries).

```python
for i in enumerate(suppl2) :
    if i.HasSubstructMatch(Chem.MolFromSmiles('CC(N)CC1=C
C=CC=C1')) :
        baseclass1 += [i]
        baseclass1Name += db_names[j]
```

The SMILES (Simplified molecular input level Entries) for the drugs which are obtained from data sources mentioned above. These values are used to determine the chemical structure of the drug. Three common Base Structures are fixed and accordingly drugs are classified into groups based on the common base structure they reflect.

![Fig4: Sampleclass1](image) ![Fig 5: Sample class2](image) ![Fig 6: Sample class 3](image)

**C. Feature selection**

In the feature selection phase a two layer Heterogeneous network is constructed, which contains DSN and a disease network. DSN contains the relationship among drugs. These relationships are made by comparing the chemical structure of the drugs. SMILES (Simplified molecular input level Entries) are used to generate the chemical structure of the drugs. An algorithm is developed to compare the structures. Drug similarity network and disease network are connected using existing drug disease relationships.

**Data input**: SMILES (Simplified molecular input level Entries).

### a) Constructing Drug similarity network

The SMILES are first converted to mol files in order to find the sub structure match. The algorithm iterate through the mol files for each drug in the data set and identifies the sub structure match. If there is a match then a relationship is drawn between the two drugs. Drug similarity network is constructed.

```python
for i in enumerate(suppl1) :
```
for j, mol2 in enumerate(suppl2):
    if mol1.HasSubstructMatch(mol2):
        i = j :
        if array[j][i] != 1 :
            array[i][j] = 1
suppl2.reset()

Algorithm was able to identify 26 drug similarities in the given set. The list is filtered by removing the drugs used for same disease. These filtered relationships are stored in drug_sim[] array. The filtered list contains 12 drug-drug relationships.

<table>
<thead>
<tr>
<th>Drug 1</th>
<th>Related Drug 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dexmethylphenidate</td>
<td>Amoxapine</td>
</tr>
<tr>
<td>Buspirone</td>
<td>Escitalopram</td>
</tr>
<tr>
<td>Clozapine</td>
<td>Carbipoda</td>
</tr>
</tbody>
</table>

Table 1 Sample List of Drug relations

b) Constructing Two Layer DDN

To the constructed drug similarity network a disease network layer is added. The Disease layer contains 4 neurological disorders Autism spectrum, Depression, Parkinson’s disease, Alzheimer’s disease. These two graphs are connected by the Drug-Disease relationships obtained from data set

Plotting the graphs

Networkx is a graph plotting library in python. All the graphs used in this paper are plotted using Networkx.

a) Plotting Drug similarity network

Links between drugs are identified and stored in drug_sim[] array at the feature selection phase. This array is given as an input to the graph plotting function.

```
lklist = []
for i in range(58):
    for j in range(58):
        if dbsim[i][j] == 1 :
            lklist += [(i, j)]
        graph = [(dbnamearray[i][0], dbnamearray[j][1]) for i in lklist]
        draw_graph(graph)
```

Output:

Fig. 7 Drug similarity network

b) Plotting Two Layer Drug-Disease network

Known Drug-Disease relationships are given as input to the graph plotting function. There are 58 known drug disease relationships in the data set.

```
lklist = []
for i in range(58):
    lklist += [(dbnamearray[i][0], dsnamearray[j][0]) for i in lklist]
    print(graph1)
    # you may name your edge labels
    labels = map(chr, range(65, 65+len(graph1)))
    draw_graph(graph, labels)
    # if edge labels is not specified, numeric labels (0, 1, 2...) will be used
    draw_graph(graph1)
```

Output:

Fig. 8 Drug similarity network

Drug Repositioning

A arbitrary stride on both the drug similarity network and Disease network is carried out. New potential targets for the existing drugs are found in this phase. The base concept of the method is that similar drugs can act on similar targets and similar targets can be bonded with similar drugs.

Data input: Drug similarity array, Disease array

```
newlklist = []
for i in range(58):
    for j in range(58):
        if dbsim[i][j] == 1 :
            if dsnamearray[i] != dsnamearray[j] :
                newlklist += [(dbnamearray[i][0], dsnamearray[j][1])]
        # you may name your edge labels
        labels = map(chr, range(65, 65+len(graph1)))
        draw_graph(graph, labels)
        # if edge labels is not specified, numeric labels (0, 1, 2...) will be used
        draw_graph(graph2)
```

The algorithm walks on the drug similarity network and disease network simultaneously. If there is a structural
matching between two drugs that are used for two different diseases, then the algorithm draws a new relationship between the two different drugs and the diseases. In our work 14 new potential targets for existing drugs are found.

<table>
<thead>
<tr>
<th>Drug</th>
<th>Identified Potential Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amoxapine</td>
<td>Sodium-dependent dopamine transporter</td>
</tr>
<tr>
<td>Escitalopram</td>
<td>Hydroxytryptamine receptor 1A</td>
</tr>
<tr>
<td>Carbidopa</td>
<td>Multidrug resistance protein 1</td>
</tr>
</tbody>
</table>

Table 2 Sample Set of newly identified potential drug targets.

The graph connecting the potential drug targets are drawn using Networkx python library.

Fig.9 Graph for newly identified potential drug targets

VI. COMPUTING EFFICIENCY

The identified 14 new potential targets for existing drug are tested for efficiency. iGEMDOCK is a protein-ligand docking software. All the identified 14 drugs are docked with potential targets are tested for efficiency. The results shows that identified drugs can be used for the potential targets.

VII. RESULT AND CONCLUSION

The structure based classification carried out in this work was able to identify 26 drug-drug structural similarities. A combined method of Network based clustering and Random walk algorithm was able to identify 14 new potential targets for existing drugs. A combined approach of Network Based Inference and Random walk algorithm can be used for an efficient Drug Repositioning purpose. Using the two layer heterogeneous network model a drug-disease network can be formed with drug-drug relationships. A Random walk algorithm can be used to iterate through the drug similarity network and drug-disease relationship to identify new potential targets for existing drugs.

VII. FUTURE ENHANCEMENT

The chemical structure of identified drugs can be modified to increase the efficiency of Docking with the target protein.

REFERENCES