Classification and Predicting Drug based on Chemical Dataset

Pradip Sarkate[1], Prof. A. V. Deorankar[2]

P.G. Scholar, Department of Computer Sci. & Engg., Govt. College of Engineering, Amravati, India[1]
Associate Professor, Department of Computer Sci. & Engg.,
Govt. College of Engineering,Amravati, India[2]

ABSTRACT:

Research and developing new drugs using computational approach is an important to predict drug and drug target interaction. It will identify drug similarity and drug target in the database or data set. The drug or chemical medicine will approve by State Food and Drugs Administration (FDA) research and developing anew drugs. identify drug and drug target interaction and similar search drug of chemical data can advanced search on system level drug molecules and its structure properties that efficiently integrates molecules, proteins and its contents information used to drug discovered. Developing new drug of a large scale based on two approaches are Random Forest algorithms and Support Vector Machine. These methods concatenate the chemical structural and properties will express the drug target in training data sets.

Keywords: *Chemical datasets, SVM, Random forest algorithm, KNN algorithms.*

I. INTRODUCTION:

The growing productivity of a drug need for innovative approach of drug target prediction and similarity of a medicine drugs. Predicting drug target interaction between drug and target method identify similar drug and target in the database. The drug target used a classification features that searching similarity between medicines.

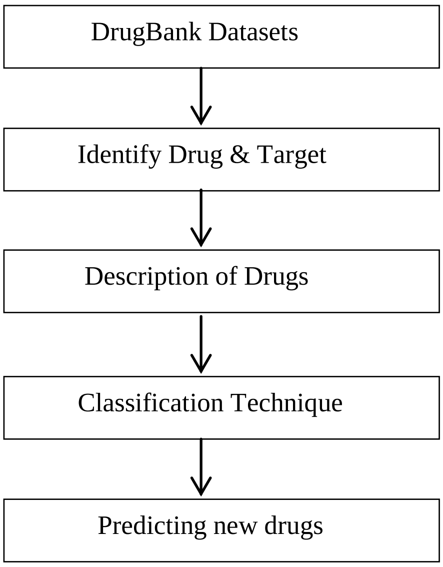

Fig. predicting new drug

The DrugBank database is unique drugs information resources that details bioinformatics and chemical informatics drugs data. Classified drug data and target data of training data sample datasets. It will identify relationship between drug description and target description, based on these descriptions which give the positive or negative samples data. Using sample data point it test data which entirely available on new data sample datasets. Classification of chemical drugs which used randomized algorithms are classified data in training samples data, support vector machine algorithm based on cosine formula to calculating similarity between medicines.

II. RELATED WORK:

In a related work drug target interaction challenging and expensive it is difficulties understanding between chemical space and biological system [1]. Drug target interaction understanding as structurally different drug to express an similar activities and bind to the same proteins [2,3]. The chemical genomics research an chemical space with the genomic space to identify target interaction [4,5,6]. Predicting drug target interaction various computational approach to developed to analyses and predict drug target such as docking simulation and literature text mining[7,8,9].text mining usually used based on keyword searching , the docking does not applied to protein 3D structures are unknown[9] .

The chemical genomics approaches includes ligand target based method used predict interaction between the chemical compound and protein target [10-16], which aims exploiting the whole chemical space interaction between small molecules. The ligand based approach combines chemical space, target space and drug target information whose purposed the predict the drug target [11]. Prediction which assumes chemically similar compound should share common target and target share similar ligands in binding sites [12].

III.PROPOSED WORK:

Drug dataset and target interaction extracted from DrugBank database which included drug FDA approved drugs or medicines. Predicting drug target we concatenate the chemical structural and physicochemical properties of a drugs or medicines. In classification, prediction of drug target interaction typically performed on basic several parameters such as chemical structure, compound protein interaction, molecules structure, function and its properties. Providing drugBank database which includes drugs and target, which includes information of a drug descriptors and target descriptors, it will evaluate the drug descriptors and drug target relationship.

A. Random forest algorithm:

Classification of drug target we used a random forest algorithms and support vector machine to classify drug target or medicines. Random forest generally proposed by Leo Breiman which is ensemble learning method to generate the classifiers technique main features of random forest is does not overfit problem to number of sample in datasets. It robust against the boosting method and it will high efficiently on high dimensional multiclass datasets. Random forest algorithm have many decision trees, in each trees choose the randomly N samples dataset points of a classifier as bootstrap sample datasets. Number of dataset decision tree split the decision tree attribute feature selected and best split attributes used to split the node. Random forest algorithm is used to predicting test data, it has *k* decision tree formed by repeating *k* times.

B. Support Vector Machine:

Support vector machine proposed by Vapnik, as an supervised machine learning algorithms. The SVM mode widely used in bioinformatics and chemical informatics its remarkable generalization performance in non-separable problem. It will implement a cosine formula based on input vector consisting of various features. The SVM model classifiers are generated sample data vector space and then it find hyperplane in a space. The major advantage SVM is a training data is redundant and minimization the structure.

C. KNN Classification:

K Nearest Neighbors is most popular classification algorithm in data mining. *k* -NN is supervised classification algorithm to similarity based on distance measurement in databases. *k* -NN used instance based training sample data set points which define sample closest data to new data that predict the label from these instances. It will predict new sample data calculated upon k value. *k* -NN classification algorithm test as sample data class to nearest neighbor, it used Euclidean distance of nearest neighbor. In *k* -NN algorithm k means clustering technique which forms cluster or group of object which define the number of iterations.

IV.CONCLUSION:

In drug discovery predicting drug and drug target interaction is an important of developing a new drug. Developing new drug target interaction prediction used to investigate drug diseases and integrating the information chemical structure and drug target. We used classification algorithm random forest model and *k* -NN algorithm to predicting drug and drug target. The drug and drug target are binding pattern can be extracted datasets and identify structure chemical similarity of drug. We proposed drug method and target method interaction including nearest neighbor and random forest algorithms. It will identify similarity calculation between drug and target chemical datasets.

REFERENCES:

[1]. Lipinski C, Hopkins A (2004) Navigating chemical space for biology and medicine. Nature 432: 855–861.

[2]. MacDonald ML, Lamerdin J, Owens S, Keon BH, Bilter GK, et al. (2006) Identifying off-target effects and hidden phenotypes of drugs in human cells. Nature Chemical Biology 2: 329–337.

[3]. Paolini GV, Shapland RHB, van Hoorn WP, Mason JS, Hopkins AL (2006) Global mapping of pharmacological space. Nature Biotechnology 24: 805–815.

[4]. Dobson, C.M. (2004) Chemical space and biology. *Nature*, **432**, 824–828.

[5]. Kanehisa,M. *et al.* (2006) From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res.*, **34**(Database issue), D354–D357.

[6]. Stockwell,B.R. (2000) Chemical genetics: ligand-based discovery of gene function. *Nat. Rev. Genet.*, **1**, 116–125.

[7]. Cheng,A.C. *et al.* (2007) Structure-based maximal affinity model predicts small molecule drug ability. *Nat. Biotechnol.*, **25**, 71–75.

[8]. Zhu,S. *et al.* (2005) A probabilistic model for mining implicit 'chemical compound gene' relations from literature. *Bioinformatics*, **21** (Suppl 2), ii245–ii251.

[9]. Rarey,M. *et al.* (1996)Afast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, **261**, 470–489.

[10]. Balakin KV, Tkachenko SE, Lang SA, Okun I, Ivashchenko AA, et al. (2002) Property-based design of GPCR-targeted library. Journal of Chemical Information and Computer Sciences 42: 1332–1342.

[11]. Nagamine N, Sakakibara Y (2007) Statistical prediction of protein chemical interactions based on chemical structure and mass spectrometry data. Bioinformatics 23: 2004–2012.

[12]. Frimurer TM, Ulven T, Elling CE, Gerlach LO, Kostenis E, et al. (2005) A physicogenetic method to assign ligand-binding relationships between 7TM receptors. Bioorganic & Medicinal Chemistry Letters 15: 3707–3712.

[13]. Klabunde T (2006) Chemogenomics approaches to ligand design. Ligand Design for G Protein-coupled Receptors ch7: 115–135.

[14]. He Z, Zhang J, Shi XH, Hu LL, Kong X, et al. (2010) Predicting drug-target interaction networks based on functional groups and biological features. PloS one 5: e9603.

[15]. Xia Z, Zhou X, Sun Y, Wu L (2009) Semi-supervised Drug-Protein Interaction Prediction from Heterogeneous Spaces. The Third International Symposium on Optimization and Systems Biology 11: 123–131.

[16]. Yamanishi Y, Pauwels E, Saigo H, Stoven V (2011) Extracting sets of chemical substructures and protein domains governing drug-target interactions. Journal of Chemical Information and Modeling 51: 1183–1194.