Drug Target protein interaction prediction by PSSM and LOOP method

**Team Details:**

C. Vijayabaskar 2017103048

G. Velmurugan 2017103047

K. Sathish Kumar 2017103588

**Guided by**

G. Sudhakaran

Teaching Fellow

Department of Computer Science and Engineering

College Of Engineering Guindy

Anna University

**Introduction:**

Drug-Target Interaction Prediction (DTI) is an important application of machine learning in medicine industry, the importance is coming from the fact that we need to save the time and cost of the drugs development. The prediction of interactions between drugs and target proteins is a critical part of drug discovery pipeline as it can help and a novel drug candidate and understand side effects. Although modern medicine is aligned with antibiotic treatment, the discovery of new and potential drugs is declining, as there is an increase of the misuse of the current available medicine, causing a resistance effect to these kinds of agents.

Computational methods for DTI prediction are divided into 3 main approaches: ligand based, docking simulation and chemogenomic. Ligand based approaches are built upon the concept that similar molecules have similar properties and therefore should bind to the same group of proteins. Docking Simulation approaches are used for structure based drug design, where the interaction between a protein and a drug is simulated and scored, according to the intermolecular interaction energy, using 3D structures. Chemogenomic approaches are based on the chemical space of compounds, genomic space of target proteins and/or the pharmacological space (interactions between proteins and drugs) to predict new potential interactions.

**Overall Objective:**

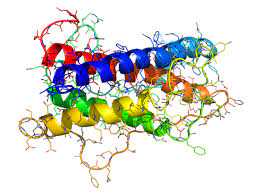
The main objective of the project is to predict the interaction of drug and target. So, we propose a novel computational method based on target protein sequence and drug substructure fingerprints. It uses the Chemogenomic approach of predicting the interaction. The method combines **local optimal oriented pattern** (LOOP), **position specific scoring matrix** (PSSM) and **rotation forest** (RF) for predicting DTIs. Specially, we first transform the target protein sequence into PSSM in order to retain biological evolutionary information, and consider molecular substructure fingerprints are considered as the feature of drugs. We then applied local optimal oriented pattern (LOOP) to extract the 256 dimension feature vectors from PSSM. Finally, we utilize rotation forest to predict the DTIs.

**Literature survey:**

With the great explosion number of publicly-available data in biology and chemistry, several different types of related databases, such as Therapeutic Target Database (TTD) Super Target and Matador Kyoto Encyclopedia of Genes and Genomes (KEGG) , and Drug Bank have been established. Until now, there have been proposed many computational methods based on machine learning in order to solve the limitations of traditional computational method. For example, Nidhi proposed a multiple-category Laplacian modified naïve Bayesian model to train 964 target categories in the (World of Molecular Bio Activity) WOMBAT database and predicted the top three most potential compound targets in the MDDR database. Liu proposed a novel prediction algorithm, namely neighbourhood regularized logistic matrix factorization (NRLMF), which focus on predicting the probability whether a drug would interact with a target and also study local structure of drug-target pairs for further improving the accuracy of DTIs. Wang developed a computational method which combines auto covariance (AC) and rotation forest for predicting potential drug-target interactions. Mei developed a novel approach namely BLM-NII, which integrated neighbour-based interaction profile inferring (NII) into bipartite local model (BLM), the method achieved excellent improvement in inferring unknown drug-target interaction. Huang proposed a computational model using extremely randomized trees for predicting drug-target interactions, the improvement of this work mainly come from the protein sequence is converted into pseudo substitution matrix representation (Pseudo-SMR) descriptor that can retain evolutionary information. Chen and Zhang proposed Net CBP, a semi-supervised learning based model for identifying DTIs by using labelled and unlabelled interaction information. Sridhar report a probabilistic approach which fully utilizes multiple drug-based similarities and known interactions, this method obtained excellent performance and find five novel interactions validated by external sources.

**Block diagram:**

**Protein Sequence**



**Gapped BLAST search**

**Conversion to low-complexity sequence**

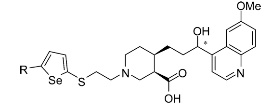
r

**2D Matrix (PSSM)**

**Alignment**

PSI BLAST

PSSM Matrix conversion

****

**Non-Linear amalgamation of LBP & LDP**

**Encoding**

**Drug Molecular structure**

**Fingerprint vector Conversion**

**Feature description**

Drug Target Vectorization

LOOP representation

**Fingerprint vector Conversionn**

**Rotation Forest Classification**

**Details of modules:**

1. **PSSM Formation**

Position-specific scoring matrix (PSSM) is one of the descriptor which carries the evolutionary information of sequence and gives probability scores of any given amino acid for a specific position. In order to convert the target protein sequence, the Position-Specific Iterated Basic Local Alignment Search Tool (PSI -BLAST) which can search and compare the homologous sequence of each target protein sequence is adopted to create PSSM of each target protein sequence. PSI blast first scan for the presence of low complexity regions. The program then initially operates on a single query sequence by performing a gapped BLAST search. It then takes local alignments found, constructs a multiple alignment (master-slave alignment) and creates PSSM.

1. **Fingerprint Vectorization**

There have been proposed different kinds of drug compounds descriptors, such as constitutional, quantum chemical properties, topological and geometrical. Moreover, the structure properties of the drug molecules were encoded in binary bits, which can directly know whether specific substructure fragments in the drug molecules exist or not. It reduce the workload of molecular descriptor calculation and screening.

1. **LOOP formation:**

Local optimal oriented pattern s texture descriptors which encode repeated local patterns in images as binary codes, and it is a popular type of feature used for classification in computer vision. Because of the disadvantage of local binary pattern (LBP) and local derivative pattern (LDP) is the arbitrary sequence of binarization weights that adds dependency to orientation. Thus, LOOP presents a nonlinear amalgamation of LBP and LDP that overcomes these drawbacks while preserving these strengths. Here, the input signal PSSM is a N x 20. Each target protein sequence would be represented by 256 feature vectors after using LOOP feature descriptor.

1. **Classification**

The Rotation Forest algorithm focuses on improving the difference and accuracy of the base classifier. In this work, we adopt RF as a classification model for predicting DTIs. Specifically, the RF randomly divides entire sample set into K subsets, and principal component analysis (PCA) method is adopted to transform the subsets which make the difference between each subset.

**Performance Measures:**

In order to evaluate the performance of the propose method, we use the evaluation measures such as the overall prediction accuracy (Accu.), sensitivity (Sens.), precision (Prec.), and Matthews correlation coefficient (MCC). When predicting the DTIs datasets of enzyme, ion channels, GPCRs and nuclear receptor, five-fold cross-validation would be adopted in this work in order to avoid the over-fitting of the prediction model.

**Datasets:**

In this study, we execute the experiment for predicting DTIs on four golden standard datasets, namely enzyme, ion channels, GPCRs, and nuclear receptor, respectively. These datasets are collected from DrugBank, KEGG BRITE, SuperTarget & Matador, and BRENDA which were considered as high-reliability databases.

Drugbank- <https://go.drugbank.com/releases/latest>

**References:**

1. Xinkie Zhan , Zhu-Hong You (Member, IEEE), Jinfan Cai, Liping Li, Chanqing Yu, Jie Pan, and Jiangkun Kong “Prediction of Drug-Target Interactions by Ensemble Learning Method from Protein Sequence and Drug Fingerprint”.

2. Nelson R. C. Monteiro, Bernardete Ribeiro, Joel P. Arrais “Drug-Target Interaction Prediction: End-to-End Deep Learning Approach”.

3. Y. Yamanishi, E. Pauwels, and M. Kotera, ``Drug side-effect prediction based on the integration of chemical and biological spaces, Nov. 2012.

4. M. Kuhn, M. Al Banchaabouchi, M. Campillos, L. J. Jensen, C. Gross, A. Gavin, and P. Bork, ``Systematic identification of proteins that elicit drug side effects,'' Jan. 2013.

5. A. Ezzat, M. Wu, X.-L. Li, and C.-K. Kwoh, ``Drug-target interaction prediction via class imbalance-aware ensemble learning,''Dec. 2016.

6. S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht, ``How to improve R&D productivity: The pharmaceutical industry's grand challenge,''Feb. 2010.

7. F. Zhu, B. Han, P.Kumar, X. Liu, X. Ma, X.Wei, L.Huang,Y. Guo, L. Han,C. Zheng, and Y. Chen, ``Update of TTD: Therapeutic target database,''Jan. 2010.

8. Y. Wang, S. Zhang, F. Li, Y. Zhou, Y. Zhang, Z.Wang, R. Zhang, J. Zhu,Y. Ren, Y. Tan, C. Qin, Y. Li, X. Li, Y. Chen, and F. Zhu, ``Therapeutic target database 2020: Enriched resource for facilitating research and early development of targeted therapeutics,'' Jan. 2020.

9. S. Gunther, M. Kuhn, M. Dunkel, M. Campillos, C.Senger, E. Petsalaki, J. Ahmed, E. G. Urdiales, A. Gewiess, L. J. Jensen, R. Schneider, R. Skoblo, R. B. Russell, P. E. Bourne, P. Bork, and R. Preissner, ``Super-Target and matador: Resources for exploring drug-target relationships,''Jan. 2007.

10. M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki Kinoshita, M. Itoh,S. Kawashima, T. Katayama, and M. Hirakawa, ``From genomics to chemical genomics: New developments in KEGG,''Jan. 2006.

11. D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, ``DrugBank: A knowledgebase for drugs, drug actions and drug targets,''Jan. 2008.

12. D. S.Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J.Woolsey, ``DrugBank: A comprehensive resource for in silico drug discovery and exploration,''Jan. 2006.

13. D.-L. Ma, D. S.-H. Chan, and C.-H. Leung, ``Drug repositioning by structure-based virtual screening,''2013.

14. B. Waszkowycz, D. E. Clark, and E. Gancia, ``Outstanding challenges in proteinfiligand docking and structure-based virtual screening,''Mar. 2011.

15. Y.-F. Dai and X.-M. Zhao, ``A survey on the computational approaches to identify drug targets in the postgenomic era,''Jan. 2015.

16. P. R. Caron, M. D. Mullican, R. D. Mashal, K. P. Wilson, M. S. Su, and M. A. Murcko, ``Chemogenomic approaches to drug discovery,''Aug. 2001.