**Computational drug repositioning for ischemic stroke: Neuroprotective drug discovery**

Yunong Li1#, Jingbo Yang1#, Yan Zhang1, Qingkang Meng1, Andreas Bender2\*, Xiujie Chen1\*

1. Harbin Medical University, Harbin, Heilongjiang, China.

2. Department of Chemistry, University of Cambridge, Cambridge, United Kingdom.

#These authors contribute equally to this work.

**\*Corresponding authors:**

1. **Xiujie Chen, Harbin Medical University, Harbin 150001, Heilongjiang, China. Telephone: +86-451-86352422, Email: chenxiujie@ems.hrbmu.edu.cn.**
2. **Andreas Bender, Cambridge university, Cambridge CB2 1 EW, United Kingdom. Mobile Phone: +44(79) 06158533, Email: ab454@cam.ac.uk.**

（the submission contains 5 Figures, 5 Supplementary Tables and 2 Supplementary Figures）

**Materials and Methods**

**Identification and quantification of gene-domain-substructure-drug relationships**

Gene-domain, substructure-drug and domain-substructure relationships were extracted by the following steps, respectively. Step 1: we identified the proteins corresponding to the prioritized candidate genes in DAVID database and further utilized the method of web crawler to extract domain data related to proteins in UniProt database. Thus, we obtained the gene-domain relationships. Step 2: we searched DrugBank database and PubChem database for the relationships between the approved small molecule drugs and their substructures. Step 3: drug-target interaction pairs from DrugBank database and “L1LOG, L1SVM” two methods provided in the literatures were used to predict the relationships between drug substructures and protein domains 1. Step 4: we integrated gene-domain pairs (Step 1), domain-substructure pairs (Step 2) and substructure-drug pairs (Step 3) to obtain gene-domain-substructure-drugs relationships.

**Obtaining drugs with strong similarity** In Mcode algorithm, the default parameters were set: Find Cluster: in whole network, Degree Cutoff: 2, Node Score Cutoff: 0.2, K-Core: 2, and Max. Depth: 100. In K-means clustering, we treated drugs as a 526-dimensional vector, which was defined as 1 if there was a direct association with gene, domain or substructure and was defined as 0 if there was no association. We replaced Euclidean distance method for calculating distance by default with the binary distance algorithm that is more suitable for 0-1 eigenvectors. By Elbow method, we determined the number of clusters. If two or more drugs are clustered into one class using Mcode algorithm and K-means clustering, these drugs are considered to be functionally similar. Finally, we obtained drugs that were clustered into the same class by two methods and established the sub-networks.

**Identification and quantification of the correlation between substructures and drug function**

Score3(S*i*) represents the proportion of drugs containing the *i*-th substructure in all approved drugs. The higher the score is, the stronger the specificity of the substructure is. Score4(S*i*) represents the proportion of drugs containing the *i*-th substructure in a drug group with a specific function. The higher the score is, the stronger the commonality of the substructure is. In order to obtain the substructures whose score3(S*i*) and score4(S*i*) values are both high, we screened score3(S*i*) and score4(S*i*), respectively taking the top 60% of the substructures and taking the intersection. The screened substructure was further evaluated by score5(S*i*). Score5(S*i*) is used to define the correlation intensity between the *i*-th substructure and drug function. The higher the score is, the stronger the correlation intensity of the substructure is.

**Results**

**Identification of substructure-domain relationships**

Like our previous study, we used “L1LOG and L1SVM” two methods and drug-target interaction pairs from DrugBank database (3310 drug-target pairs) to extract the substructure-domain relationships. Using the 10-fold cross-validation, the corresponding best C values were obtained as the negative/positive set ratio changed (from 1:1 to 8:1). C values were screened from 1 to 1000 at 10-fold intervals. The AUC values were got by 10-fold cross-validation of each negative/positive set ratio and best C value combination. The AUC maximums of both methods are all greater than 0.86. We identified 62400 substructure-domain features shared in two methods for the higher accuracy.

In order to test the predictive accuracy of the shared substructure-domain features by two methods, we reconstructed drug-target relationships as validation process. Based on 912 target proteins derived from the initial 3310 drug-target pairs and the shared domain-substructure features, we finally predicted target-domain-substructure-drug relationships. The prediction efficiency was 100% because the predicted drugs completely covered 943 drugs derived from the initial 3310 drug-target pairs, indicating that the domain-substructure features shared in two methods can be used as a feature set for subsequent predictions.

**Reference**

1. Tabei, Y.; Pauwels, E.; Stoven, V.; Takemoto, K.; Yamanishi, Y., Identification of chemogenomic features from drug–target interaction networks using interpretable classifiers. *Bioinformatics* **2012,** *28* (18), i487-i494.

**Supplementary Figure**

****

# Supplementary Figure S1. Brain-protective gene-domain-substructure-drug network. It contains 22 genes, 24 domains, 480 substructures, and 2297 drugs. The red node represents the gene, the blue node represents the drug, and the green node represents the domain-substructure.

****

# Supplementary Figure S2. Elbow method (clustering). This method is used to find the appropriate number of clusters in the brain-protective gene-domain-substructure-drug dataset of K-means clustering. According the “elbow criterion”, the number of clusters chosen should be 8. The "elbow" is indicated by the red dot.