DrugE-Rank: predicting drug-target interactions by Learning to Rank

Jieyao Deng[1,2], Qingjun Yuan[1,2], Hiroshi Mamitsuka[3,4], Shanfeng Zhu[1,2,5,*]

[1]School of Computer Science and [2]Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai, China,
[3]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji 611-0011, Japan, [4] Department of Computer Science, Aalto University, 02150, Espoo, Finland and [5]Center for Computational System Biology, Fudan University, Shanghai 200433, China

Email: zhusf@fudan.edu.cn;

*Corresponding author

DrugE-Rank: Drug-target Interaction Prediction by Learning to Rank

Jieyao Deng[1,2], Qingjun Yuan[1,2], Hiroshi Mamitsuka[3,4], Shanfeng Zhu[1,2,5,*]

[1]School of Computer Science and [2]Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai, China,
[3]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji 611-0011, Japan, [4]Department of Computer Science, Aalto University, Espoo 02150 Finland, and [5]Center for Computational System Biology, Fudan University, Shanghai 200433, China

Abstract

Identifying drug-target interactions is crucial for the success of drug discovery. Approaches based on machine learning for this problem can be divided into two types: feature-based and similarity-based methods. By utilizing the 'Learning to rank' framework, we propose a new method, DrugE-Rank, to combine these two different types of methods for improving the prediction performance of new candidate drugs and targets. DrugE-Rank is available at http://datamining-iip.fudan.edu.cn/service/DrugE-Rank/.

Keywords: DrugE-Rank, learning to rank, drug discovery,

1. Introduction

Identifying drug-target interactions is crucial for the success of drug discovery. It can facilitate the understanding of drug side effect [1-3], disease pathology, as well as the drug action mechanism. Compared with using biochemical experiments to identify drug target interaction, computational approaches are more efficient and economical. Approaches based on machine learning for this problem can be divided into two types: feature-based and similarity-based methods [4-6]. By utilizing the 'Learning to rank' (LTR) [7-8] framework, we propose a new method, DrugE-Rank [9], to combine these two different types of methods for improving the prediction performance of new candidate drugs and targets.

We are interested in the problem of predicting drug-target interactions for new drugs or new targets. This problem is especially challenging, due to three main reasons. Firstly, since there are no known interactions for new drug or target, the training of prediction models is difficult. Secondly, existing computational methods based on LTR do not consider the connections among different drugs or targets very well. Thirdly, the prediction of drug target interaction is a challenging multi-label learning problem, where a new target (or drug) has multiple interacting drugs (or targets).

Compared with previous computational approaches, DrugE-Rank has multiple advantages. Firstly, by utilizing the LTR paradigm, DrugE-Rank can solve this multi-label learning problem naturally and provide the most powerful performance. Secondly, DrugE-Rank integrates diverse cutting-edge techniques in the framework of LTR, which include both similarity-based and feature-based methods. Thirdly, DrugE-Rank only considers the top drug (or target) candidates recommended by each component method, which can greatly reduce the computational burden.

2. Materials

The performance of DrugE-Rank was examined by using DrugBank [10], a manually annotated drug target interactions database. We carried out three rounds of experiments. I) cross validation over DrugBank data with FDA approved drugs before March 2014; 2) Independent test over DrugBank data with new targets and FDA approved drugs after March 2014; 3) Independent test over FDA experimental drugs. The experimental results demonstrate that DrugE-Rank outperformed all competing methods, being statistically significant. The improvement is especially promising for new drugs. Finally, we train DrugE-Rank with DrugBank data by the end of 2015. It consists of 1324 human protein targets, 1242 FDA approved drugs, and altogether 5484 known interactions.

3. Methods

Six cutting-edge similarity-based methods are used in DrugE-Rank as component methods in the LTR framework: Bipartite Local Model with support vector classification (BLM-svc) [11], Bipartite Local Model with support vector regression (BLMsvr), k-Nearest Neighbor (k-NN) [12], Weighted Nearest Neighbor-based Gaussian Interaction Profile classifier (WNN-GIP) [13], Laplacian regularized least squares (LapRLS) [14], Network-based Laplacian regularized least squares(NetLapRLS). In addition, we extract drug features using RDKit (see Note 1), and target features from PROFEAT [15].

4. Usage

4.1 New Drug

Given a new drug, DrugE-Rank returns the top 20 targets as the predicted result. The input interface is shown in Figure 1.

1. Choose input format. You can input the drug profile by DrugBank ID, SMILES or MOL Format Text. An example of input is shown in Figure 2.

[1] http://www.rdkit.org/

2. Click the "Send" button. Click the button at the bottom of the page and your task will be in processing. The process takes about 10 minutes, and the server will return the top 20 predictions for each method (DrugE-Rank and six similarity-based methods). The result can help you to prioritize the most promising targets (Figure 3).

4.2 New target

Given a new target, DrugE-Rank returns the top 20 drugs as the predicted result. The input interface is shown in Figure 4.

1. Choose input format. You can input the target profile by UniProt ID or Amino Acid Sequence (Fasta format). An example of input is shown in Figure 5.
2. Click the "Send" button. Click the button at the bottom of the page and your task will be in processing. The process takes around 10 minutes, and the server will return the top 20 predictions for each method (DrugE-Rank and six similarity-based methods). The result may help you to prioritize the most promising drugs (Figure 6).

5. Notes

1. http://www.rdkit.org/

Acknowledgements

This work has been partially supported by National Natural Science Foundation of China (Grant Nos: 61572139), MEXT KAKENHI #16H02868 and FiDiPro by Tekes.

References

[1] Keiser MJ, Setola V, Irwin JJ et al. (2009). Predicting new molecular targets for known drugs. Nature, 462(7270), 175–181.

[2] Lounkine E, Keiser MJ, Whitebread S, et al. (2012). Large-scale prediction and testing of drug activity on side-effect targets. Nature, 486(7403), 361–367

[3] Nunez S, Venhorst J, Kruse CG. (2012). Target-drug interactions: first principles and their application to drug discovery. Drug Discov Today, 17, 10–22

[4] Ding H, Takigawa I, Mamitsuka H, Zhu S. (2014). Similarity-based machine learning methods for predicting drug–target interactions: a brief review. Briefings in Bioinformatics, 15(5), 734–747.

[5] Zheng X, Ding H, Mamitsuka H, Zhu S. (2013). Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1025-1033. ACM.

[6] Takigawa, I and Mamitsuka, H. (2013). Graph Mining: Procedure, Application to Drug Discovery and Recent Advance. Drug Discov Today, 18 (1-2), 50-57

[7] Liu T. (2009). Learning to rank for information retrieval. Foundations and Trends in Information Retrieval, 3(3), 225-331.

[8] Li H. (2011). A short introduction to learning to rank. IEICE Transactions, 94-D(10), 1854–1862.

[9] Yuan Q, Gao J, Wu D, et al. DrugE-Rank: improving drug-target interaction prediction of new candidate drugs or targets by ensemble learning to rank. Bioinformatics, 2016, 32(12):i18-i27.

[10] Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski, A., Arndt D,Wilson M, Neveu V, et al. (2014). Drugbank 4.0: shedding new light on drug metabolism. Nucleic acids research, 42(D1), D1091-D1097.

[11] Bleakley K, Yamanishi Y. (2009). Supervised prediction of drug-target interactions using bipartite local models. Bioinformatics, 25(18), 2397-2403.

[12] Van LT, Marchiori E. (2013). Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. PloS one, 8(6), e66952.

[13] Van LT, Nabuurs SB, Marchiori E. (2011). Gaussian interaction profile kernels for predicting drug-target interaction. Bioinformatics, 27(21), 3036-3043.

[14] Xia Z, Zhou X, Sun Y, Wu L. (2009). Semi-supervised drug-protein interaction prediction from heterogeneous spaces. In The Third International Symposium on Optimization and Systems Biology, volume 11, pages 123-131.

[15] Rao H, Zhu F, Yang G, Li Z, Chen Y.(2011). Update of profeat: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic acids research, 39(suppl 2), W385–W390.

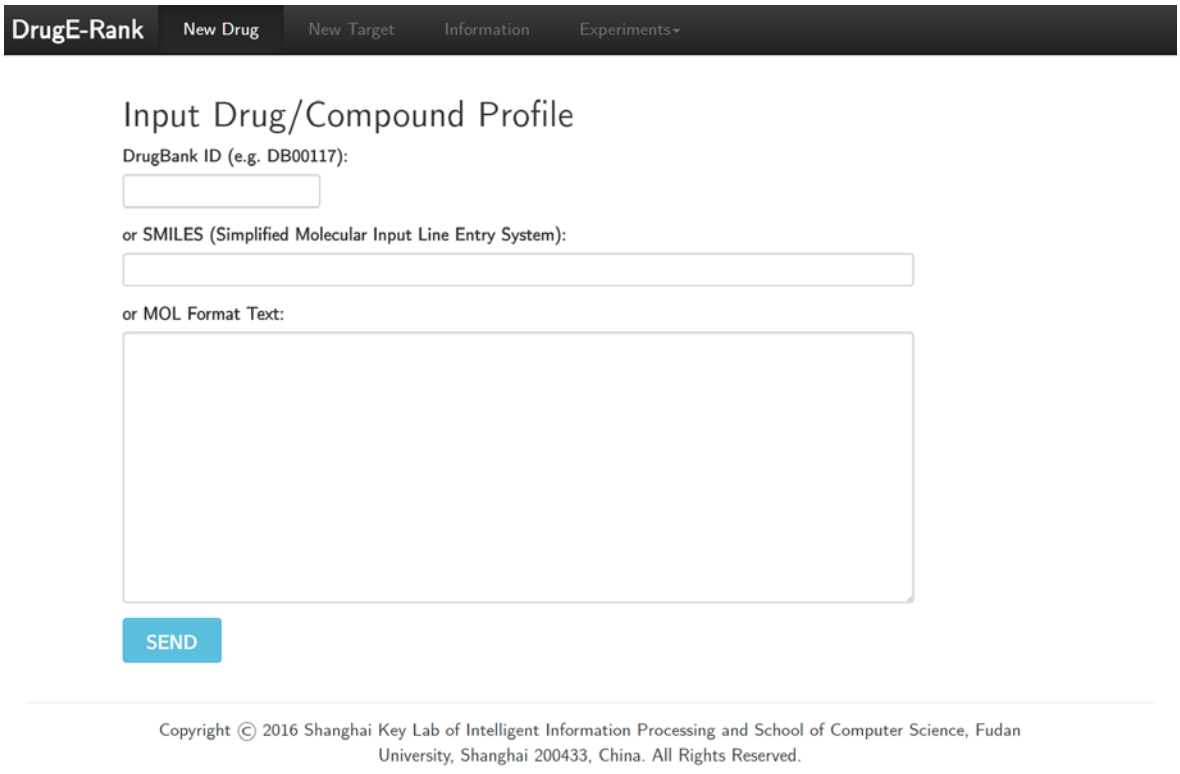

Figure 1: Input interface for New Drug

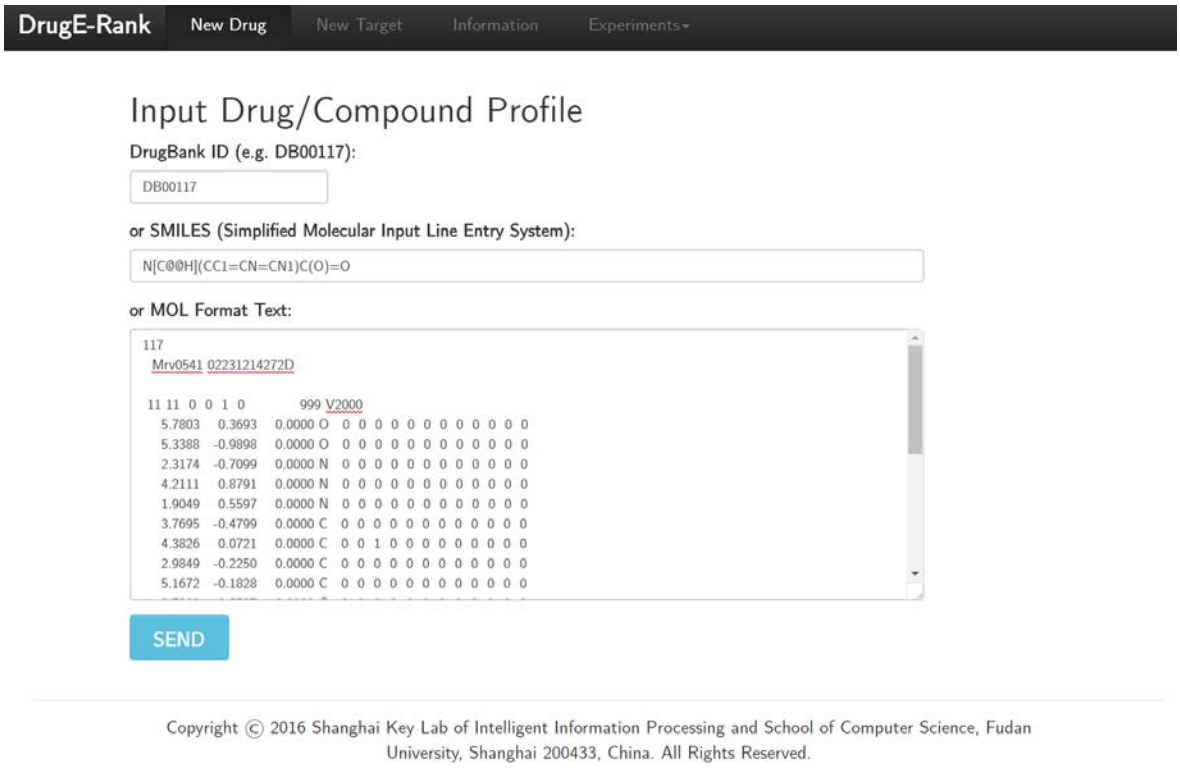

Figure 2: Input example for New Drug

Input Drug/Compound Profile

DrugBank ID (e.g. DB00117):

DB00117

or SMILES (Simplified Molecular Input Line Entry System):

or MOL Format Text:

SEND

Predicting Result (Top 20 predictions for each method)

#	DrugE-Rank	kNN	BLM-svc	BLM-svr	LapRLS	NetLapRLS	WNN-GIP
1	P19113	Q99624	Q99624	Q99624	Q99624	Q99624	Q99624
2	P42357	P19113	P12081	P19113	P19113	P19113	P12081
3	P12081	P42357	P19113	P42357	P42357	P42357	P42357
4	Q99624	P12081	P42357	P12081	P12081	P12081	P19113
5	P20309	O43246	P11229	P05981	P20309	P20309	P20711
6	P08172	P30825	P08172	P17812	P08172	P08172	P29474
7	P11229	Q8WY07	P35348	P35218	P11229	P11229	P20309
8	O43246	P78540	P20309	P00734	P18089	P29474	P35228
9	P29474	P08243	P08913	P37288	P08913	P35228	O43246
10	P30825	P15104	P18089	P23297	P18825	P78540	P30825
11	P78540	P52569	P14416	P04271	P23219	O43246	Q8WY07
12	Q8WY07	P21918	P18825	P29034	P28222	P30825	P08172
13	P35228	P21917	P08908	O00329	P29474	Q8WY07	P78540
14	P08913	P21728	P21728	Q13370	P08908	P00966	P11229
15	P18089	P35462	P08173	Q9HCR9	P35462	Q96A70	P04424
16	P23219	P14416	P08912	P42338	P21728	P04424	Q16850
17	P18825	P23219	P35368	P42336	P41595	P28222	P23219
18	P21728	P08913	P28335	Q14643	P21917	P23219	P08243
19	P20711	P18089	P28223	Q13315	P35228	P32297	Q96A70
20	P05981	P18825	P21917	P78527	P28221	P41595	P00374

Figure 3: Output example for New Drug

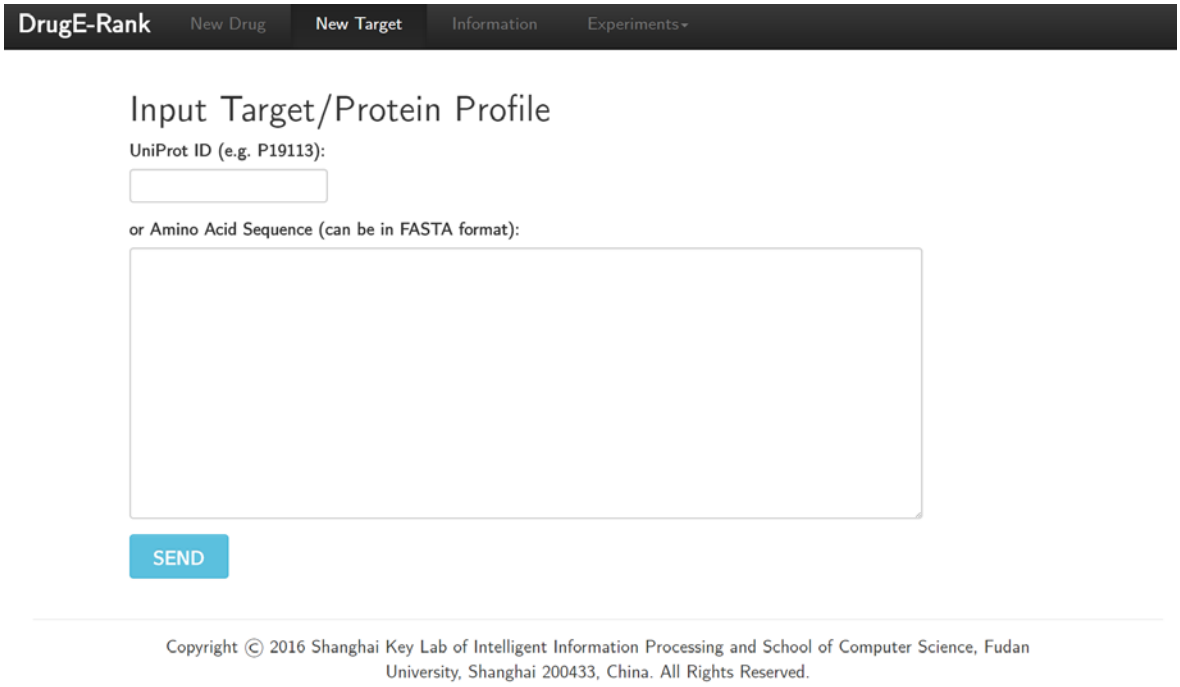

Figure 4: Input interface for New Target

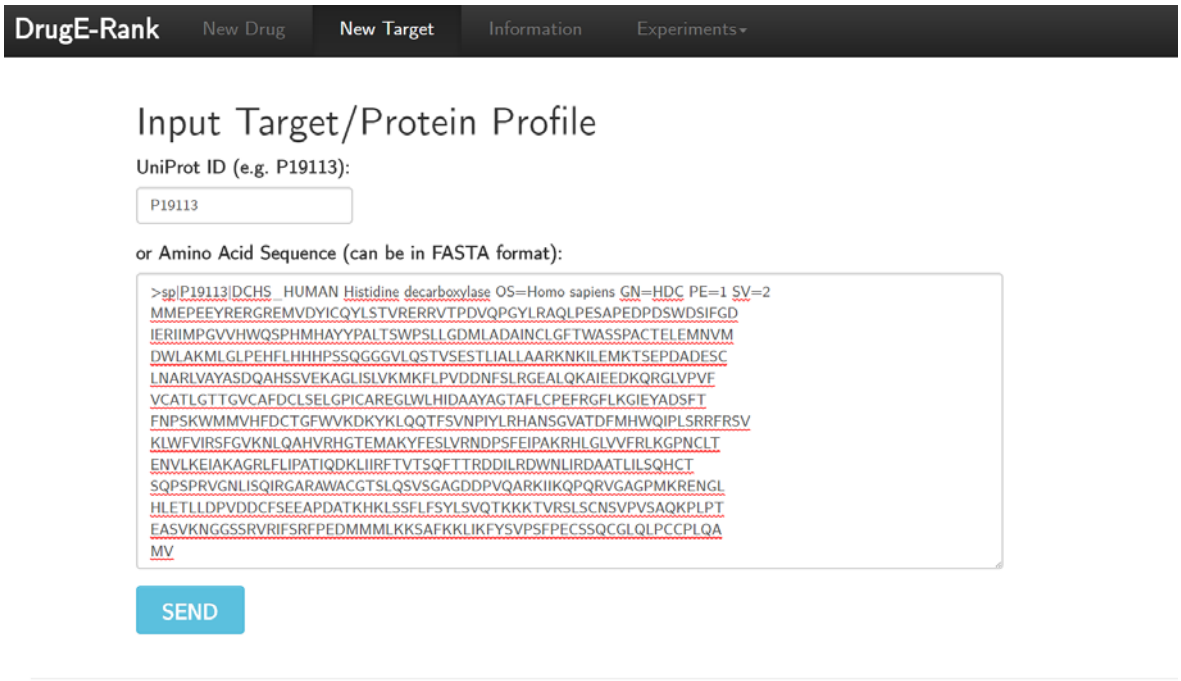

Figure 5: Input example for New Target

Input Target/Protein Profile

UniProt ID (e.g. P19113):

P19113

or Amino Acid Sequence (can be in FASTA format):

SEND

Predicting Result (Top 20 predictions for each method)

#	DrugE-Rank	kNN	BLM-svc	BLM-svr	LapRLS	NetLapRLS	WNN-GIP
1	DB00142	DB00117	DB05266	DB00117	DB00157	DB00157	DB01235
2	DB00157	DB00968	DB05260	DB00619	DB00142	DB00142	DB00765
3	DB00898	DB00190	DB00830	DB00193	DB00898	DB00898	DB00667
4	DB00201	DB00142	DB00843	DB01034	DB00334	DB00334	DB01085
5	DB00117	DB00151	DB00849	DB00201	DB00201	DB00201	DB00782
6	DB00334	DB00780	DB00847	DB01219	DB01049	DB01049	DB00997
7	DB01049	DB00149	DB06795	DB01043	DB00171	DB00171	DB00376
8	DB00171	DB00160	DB06335	DB00202	DB00143	DB00143	DB00416
9	DB00143	DB00242	DB06589	DB00606	DB00543	DB00543	DB06262
10	DB00543	DB00157	DB00411	DB00869	DB00909	DB00909	DB00221
11	DB00421	DB00116	DB00653	DB00880	DB00321	DB00321	DB00174
12	DB00909	DB00653	DB00412	DB01194	DB00139	DB00139	DB00181
13	DB05266	DB00661	DB00651	DB00626	DB00421	DB00421	DB01253
14	DB00321	DB00421	DB00894	DB00875	DB00786	DB00786	DB00723
15	DB01235	DB00622	DB00656	DB05246	DB00408	DB00408	DB00988
16	DB05260	DB00401	DB00419	DB00347	DB01159	DB01159	DB00191
17	DB00653	DB04855	DB00418	DB00593	DB00145	DB00145	DB00286
18	DB00619	DB01115	DB00661	DB01196	DB00514	DB00514	DB00157
19	DB00968	DB00270	DB00422	DB00562	DB00312	DB00312	DB01337
20	DB00765	DB00381	DB00668	DB01624	DB00128	DB00128	DB01043

Figure 6: Output example for New Target