Drug-Target Interaction Prediction by Integrating Chemical, Genomic, Functional and Pharmacological Data[†] (Supplementary Material)

Fan Yang[1]
Jinbo Xu[2]
Jianyang Zeng[3,*]

August 1, 2013

Section S1 shows descriptive statistics of two datasets that were tested in the paper. Section S2 presents sensitivity analysis of parameter K for constructing the underlying graph of our CRF model. Section S3 describes the comparison results on constructing the underlying graph of our CRF model using different criteria. Section S4 shows the results of the 5-fold cross-validation test that was performed in the paper. Section S5 provides the descriptions of different approaches tested in Section 3.3 of the paper.

S1 Descriptive Statistics of Two Datasets

Table S1 shows descriptive statistics of two datasets that were tested in the paper.

Statistics	**First Dataset**	**Second Dataset**
Number of drugs	875	357
Number of proteins	249	220
Number of interactions	2596	1174
Average degree for a drug	3.0	3.3
Average degree for a protein	10.4	5.3

Table S1: Descriptive statistics of two datasets tested in the paper.

[1] Department of Mathematical Sciences, Tsinghua University, Beijing, 100084, P. R. China

[2] Toyota Technological Institute at Chicago, 6045 S. Kenwood Ave. Chicago, IL 60637, USA

[3] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, 100084, P. R. China

[†] This work was supported in part by the National Basic Research Program of China Grant 2011CBA00300, 2011CBA00301, the National Natural Science Foundation of China Grant 61033001, 61061130540.

[*] Corresponding authors: Jianyang Zeng, zengjy321@tsinghua.edu.cn.

S2 Sensitivity Analysis of Parameter K for Constructing the Underlying Graph

To choose a proper value of parameter K for constructing the underlying graph, we tested our algorithm on the first dataset using different values of K. Our test was based on target-based CRFs with sequence similarity using 10-fold cross-validation. As shown in table S2, the results did not vary much for different K values.

K	**Total Number of Edges**	**AUC**	**AUPR**
1	177	96.9	79.1
2	355	97.3	80.7
3	540	97.3	80.7
4	731	97.3	80.7
5	925	97.3	80.8
6	1130	97.3	81.0

Table S2: The 10-fold cross-validation results on the first dataset with different choices of parameter K. The test was based on target-based CRFs with sequence similarity.

S3 Comparison Results on Constructing the Underlying Graph Using Different Criteria

We compared the performance of our algorithm using different approaches for constructing the underlying graph of our CRF model. We first connected two nodes if their similarity score was larger than a chosen threshold. Then we checked if degree of each node was at least K. For any node whose degree was less than K, we added more edges according to the similarity score until its degree was up to K. The case when $K = 0$ in fact corresponded to the threshold-based approach, as described in the paper (Sec 3.1). Our test was performed on target-based CRFs with sequence similarity using 10-fold cross-validation. Table S3 shows the AUPR results in this comparison test.

K	**Threshold**		
	0.2	0.4	0.6
0	81.0	74.9	45.8
2	81.0	80.6	80.0
4	80.8	80.8	80.3
6	81.1	81.0	80.3

Table S3: The AUPR results on constructing the underlying graph of our CRF model using different criteria. The test was performed on target-based CRFs with sequence similarity using 10-fold cross-validation.

S4 5-fold Cross-validation on the First Dataset

Table S4 shows the results of the 5-fold cross-validation test on the first dataset using different approaches. Compared to the results of the 10-fold cross-validation test described in the paper, only a slight decrease in

AUC and AUPR was observed.

Approach		Evaluation Criterion	
		AUC	AUPR
Target-based CRF	GEN	97.1	80.6
	FUN	97.7	80.8
	IGF	97.9	83.4
Drug-based CRF	CHEM	96.8	80.0
	PHAR	96.1	76.5
	ICP	97.9	84.7
Full Integration Approach (FI)		**99.2**	**94.6**

Table S4: The 5-fold cross-validation results on the first dataset using different approaches. The best result is shown in bold.

S5 Descriptions of Different Approaches Tested in Section 3.3 in the Paper

- AERS-freq-based pharmacogenomic approach (AERS-freq): The test was performed on genomic and AERS-freq-based pharmacological data.
- AERS-bit-based pharmacogenomic approach (AERS-bit): The test was performed on genomic and AERS-bit-based pharmacological data.
- SIDER-based pharmacogenomic approach (SIDER): The test was performed on genomic and SIDER-based pharmacological data.
- JAPIC-based pharmacogenomic approach (JAPIC): The test was performed on genomic and JAPIC-based pharmacological data.
- Chemogenomic approach (CHEM): The test was performed on genomic and chemical data.
- Integrated pharmacogenomic approach (INTEG-P): The test was performed on genomic, AERS-freq-based, SIDER-based and JAPIC-based pharmacological data.
- Integrated pharmaco-chemogenomic approach (INTEG-PC): The test was performed on genomic, chemical, AERS-freq-based, SIDER-based and JAPIC-based pharmacological data.