An accurate and confident prediction of drug-target binding affinity

Tong He,[†] Marten Heidemeyer,[†] Fuqiang Ban,[‡] Artem Cherkasov,[‡] and Martin Ester[†]

Simon Fraser University, Canada, and Vancouver Prostate Center

E-mail:

Abstract

Computational prediction of the interaction between drugs and targets is a standing challenge in drug discovery. High performance on binary drug-target benchmark datasets was reported for a number of methods. As previously reported, a possible drawback of binary datasets is that missing values and non interacting drug-target pairs are not differentiated. In this paper we present a method that utilizes Continuous Conditional Random Fields (CCRF) to predict the continuous binding affinities of compounds and proteins and thus incorporates the whole interaction spectrum from true negative to true positive interactions in the learning phase. Additionally, our method computes a confidence score for each compound-protein pair in order to assess the confidence of the predicted affinity. We evaluate our model on three continuous drug-target dataset and we compare the performance of our model to a recently published similarity-based method for continuous binding affinity prediction.

*To whom correspondence should be addressed
[†]Simon Fraser University, Canada
[‡]Vancouver Prostate Center

1. Introduction

Drug effects are caused by the interaction of drug compounds with proteins. Drug side effects occur when a drug compound is interacting with proteins other than its primary target. Knowledge about the interaction of drugs and targets (or more general drug candidate compounds and candidate target proteins) is therefore necessary in the process of drug development. Predicting the interaction between compounds and targets with computational methods is a hot topic in drug development because the in vitro validation of compound-protein interaction strengths is extremely costly and time consuming. A number of machine learning methods for large scale drug-target interaction prediction have been published so far and high accuracies in terms of AUC and AUPR were reported on binary drug target benchmark datasets. Ding et al. reviews 6 methods, all of which were evaluated on binary datasets compiled from KEGG BRITE, BRENDA, SuperTarget, DrugBank, DCDB and ChEBI. These datasets record only validated interacting drug-target pairs and do not differentiate between missing measurements and non interacting drug target pairs. Ding et al. and Pahikkala et al. suggest that incorporating true negative interactions (observed non-interacting drug target pairs) could improve the performance of prediction models. Additionally, Pahikkala et al. suggests to build models that predict the continuous binding affinities of drugs and targets and present a kernel method for this task. In this study we present a new method for the task of predicting the continuous binding affinities between drugs and targets which is based on Continuous Conditional Random Fields. Binary Conditional Random Fields were previously applied by Yang et al. for the classification task of predicting if a drug target pair is interacting or not. We evaluate our model on four datasets of different sparsity and compare our model to the kernel method presented in Pahikkala et al.

Models and Method

Predictive Model

Conditional Random Fields are a type of Markov Network that encode a conditional distribution $P(Y|X)$, where Y is a set of target variables and X is a set of observed variables. For the task of drug target interaction prediction, the vector Y stands for the drug-target binding affinities that we wish to predict and the vector X stands for a first prediction of the binding affinity. We utilized the CRF notation to integrate drug and target similarity matrices to improve the initial prediction X. In the used CRF model, the probability of a vector Y given the initial prediction X and a similarity score for all drug-target pairs is defined as follows:

$$P(Y|X) = \frac{1}{Z} exp(\alpha \sum_{i} f(y_i, X_i) + \beta \sum_{i,j} g(y_i, y_j))$$

Here, the term $\alpha \sum_{i} f(y_i, X_i)$ penalizes the difference of the prediction from the initial prediction and the term $\beta \sum_{i,j} g(y_i, y_j)$ penalizes the difference of two similar drug-target pairs. Formally, f is defined as $f(y_i, X_i) = -(y_i - X_i)^2$ and g is defined as $g(y_i, y_j, X_i) = -\frac{1}{2} S_{i,j}(y_i - y_j)^2$. Figure 1 illustrates the graphical structure of the CRF.

Parameter Learning

Inference

Experimental Settings

The performance of the model was evaluated by 5 fold Cross Validation on the drug-target pairs. Therefore the drug-target matrix was randomly partitioned into 5 parts, of which each was removed in turn from the training set and used as test data. The approach corresponds to the use case where it is the aim to predict the interaction affinities for the missing drug-target pairs where both, the drug and the target, have been encountered in the training set.

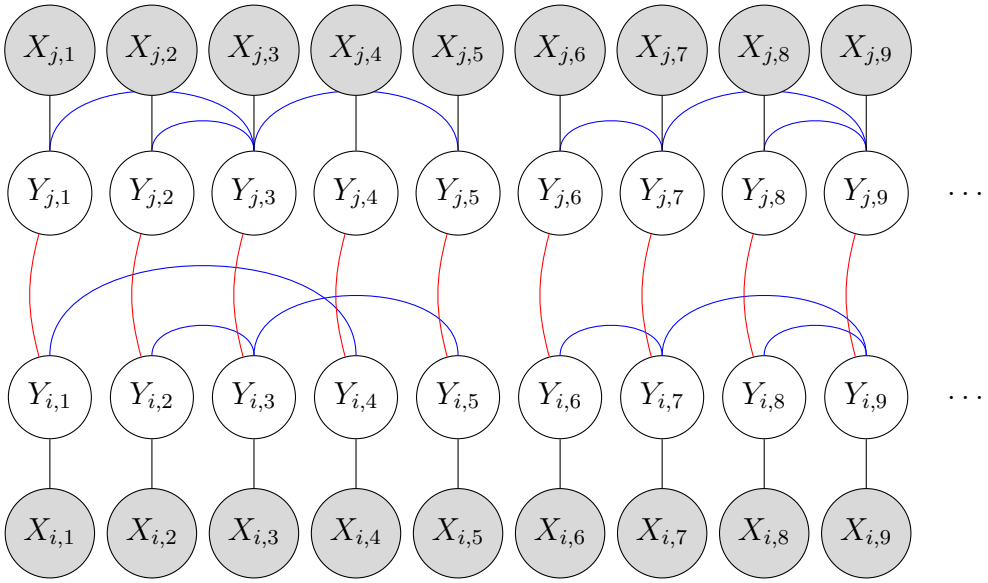

Figure 1: CRF over drugs d_i and d_j and targets $t_i, \ldots, t_9$: The intuition behind the CRF is to predict binding affinities, s.th. (a) the predictions are not too far from the predicted affinity of Matrix-Factorization (X), (b) the predicted binding strengths of similar targets (blue edges) are similar and (c) the predicted binding strengths of similar drugs (red edges) are similar. The importance of the first prediction X, and the drug- and target similarities is weighted with parameters α, β_1 and β_2 which are learned in the training step. In order to keep the inference step tractable, the number of nodes in the CRF is kept low by clustering the drugs (targets) and learn separate parameters for each drug cluster (target cluster). The above example illustrates a CRF over the two similar drugs d_i and d_j. In the experiments on the datasets that are reported in this paper, CRFs were build for single drugs, single targets, drug clusters and target clusters. When a CRF was build for a drug-cluster (target-cluster), it consisted of nodes for the respective drugs (targets) and all targets (drugs). Suppose a dataset with 3000 drugs and 400 targets: A CRF over a drug cluster with 10 drugs, consists of $10 \times 400 = 4000$ nodes and the matrix that needs to be inverted in the inference step is of size $16.000.000 \times 16.000.000$. Thus the number of drugs (targets) in the clusters need to be kept rather small.

Model Evaluation

Evaluation Datasets

In addition to the *Davis* and *Metz* (Davis et al. and Metz et al.) that were already used in the study by Pahikkala et al. two sparser datasets (Kiba 1 and Kiba 2) were compiled from the kinase dataset that was originally compiled from a number of sources by Tang et al. This dataset contains originally 52.498 compounds and 467 targets. The evaluation dataset Kiba 1 was obtained by removing all drugs and targets with less than 10 observations from this dataset. A fourth evaluation dataset (Kiba 2) with the highest sparsity was created by sampling half of the observations from dataset Kiba 1.

Table 1: statistics of the used evaluation datasets.

Dataset	Drugs	Targets	Density
Davis	68	442	100%
Metz	1421	156	42.1%
Kiba 1	2116	229	24.4%
Kiba 2	2116	229	12.21%

Evaluation Metrics

To evaluate the continuous predictions, we used the concordance index (CI) as an evaluation metric for the prediction accuray as suggested by Pahikkala et al.. As described by Pahikkala et al., the CI over a set of paired data is the probability that the predictions for two randomly drawn drug-target pairs with different label values are in the correct order and is defined as:

$$CI = \frac{1}{Z} \sum_{y_i > y_j} h(f_i - f_j)$$

where $h(u)$ is the step function, returning 1.0, 0.5 and 0.0 for $u > 0$, $u = 0$ and $u < 0$ respectively and f_i is the prediction for the larger binding affinity y_i and f_j is the prediction for the smaller binding affinity y_j.

Binary vs. Continuous Prediction

Results

Dataset	CI					
	KronRLS			CCRF		
Davis	2D	**88.3**		2D	87.1	
	δ			δ		
		SW	δ		SW	δ
	KronRLS			CCRF		
Metz	2D	79.3		2D	**81.8**	
	δ			δ		
		SW	δ		SW	δ
	KronRLS			CCRF		
Kiba 1	2D	78.3		2D	**81.9**	
	δ			δ		
		SW	δ		SW	δ
	KronRLS			CCRF		
Kiba 2	2D	76.1		2D	**78.7**	
	δ			δ		
		SW	δ		SW	δ

Dataset	AUC					
	KronRLS			CCRF		
Davis	2D	**0.952**	0.729	2D	0.937	0.873
	δ	0.927		δ	0.933	0.869
		SW	δ		SW	δ
	KronRLS			CCRF		
Metz	2D	0.934	0.868	2D	**0.947**	
	δ	0.846		δ		0.885
		SW	δ		SW	δ
	KronRLS			CCRF		
Kiba 1	2D	**0.908**		2D	0.87	
	δ			δ		0.833
		SW	δ		SW	δ
	KronRLS			CCRF		
Kiba 2	2D	0.908		2D	0.87	
	δ			δ		0.833
		SW	δ		SW	δ

Dataset	AUPR								
Davis	KronRLS			CCRF					
	2D	**0.67**	0. 279		2D	0.63			
	δ	0.649			δ				
		SW	δ			SW	δ		
Metz	KronRLS			CCRF					
	2D	0.572	0.44		2D	**0.577**	0.515		
	δ	0.28			δ	0.431	0.331		
		SW	δ			SW	δ		
Kiba 1	KronRLS			CCRF					
	2D	**0.754**			2D	0.736			
	δ				δ				
		SW	δ			SW	δ		
Kiba 2	KronRLS			CCRF					
	2D				2D				
	δ				δ				
		SW	δ			SW	δ		

References

(1) Ding, H.; Takigawa, I.; Mamitsuka, H.; Zhu, S. *Briefings in Bioinformatics* **2013**, bbt056.

(2) Pahikkala, T.; Airola, A.; Pietilä, S.; Shakyawar, S.; Szwajda, A.; Tang, J.; Aittokallio, T. *Briefings in bioinformatics* **2014**, bbu010.

(3) Davis, M. I.; Hunt, J. P.; Herrgard, S.; Ciceri, P.; Wodicka, L. M.; Pallares, G.; Hocker, M.; Treiber, D. K.; Zarrinkar, P. P. *Nature biotechnology* **2011**, *29*, 1046–1051.

(4) Metz, J. T.; Johnson, E. F.; Soni, N. B.; Merta, P. J.; Kifle, L.; Hajduk, P. J. *Nature chemical biology* **2011**, *7*, 200–202.

(5) Yang, F.; Xu, J.; Zeng, J. Drug-target interaction prediction by integrating chemical, genomic, functional and pharmacological data. Pacific Symposium on Biocomputing. 2013; pp 2304–2310.

(6) Tang, J.; Szwajda, A.; Shakyawar, S.; Xu, T.; Hintsanen, P.; Wennerberg, K.; Aittokallio, T. *Journal of Chemical Information and Modeling* **2014**, *54*, 735–743, PMID: 24521231.