BIOINFORMATICS

Vol. 00 no. 00 2011
Pages 1–8

Gaussian interaction profile kernels for predicting drug–target interaction

Twan van Laarhoven [1,*], Sander B. Nabuurs [2] and Elena Marchiori [1]

[1]Department of Computer Science, Radboud University Nijmegen, The Netherlands,
[2]Computational Drug Discovery, Center for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Center, The Netherlands

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: The *in silico* prediction of potential interactions between drugs and target proteins is of core importance for the identification of new drugs or novel targets for existing drugs. However, only a tiny portion of all drug–target pairs in current datasets are experimentally validated interactions. This motivates the need for developing computational methods that predict true interaction pairs with high accuracy.

Results: We show that a simple machine learning method that uses the drug–target network as the only source of information is capable of predicting true interaction pairs with high accuracy. Specifically, we introduce interaction profiles of drugs (and of targets) in a network, which are binary vectors specifying the presence or absence of interaction with every target (drug) in that network. We define a kernel on these profiles, called the Gaussian Interaction Profile (GIP) kernel, and use a simple classifier, (kernel) Regularized Least Squares (RLS), for prediction drug–target interactions. We test comparatively the effectiveness of RLS with the GIP kernel on four drug–target interaction networks used in previous studies. The proposed algorithm achieves area under the precision-recall curve (AUPR) up to 92.7, significantly improving over results of state-of-the-art methods. Moreover, we show that using also kernels based on chemical and genomic information further increases accuracy, with a neat improvement on small datasets. These results substantiate the relevance of the network topology (in the form of interaction profiles) as source of information for predicting drug–target interactions.

Availability: Software and supplementary material are available at http://cs.ru.nl/~tvanlaarhoven/drugtarget2011/.

Contact: tvanlaarhoven@cs.ru.nl, elenam@cs.ru.nl

1 INTRODUCTION

The *in silico* prediction of interaction between drugs and target proteins is a core step in the drug discovery process for identifying new drugs or novel targets for existing drugs, in order to guide and speed up the laborious and costly experimental determination of drug–target interaction (Haggarty *et al.*, 2003).

Drug–target interaction data are available for many classes of pharmaceutically useful target proteins including enzymes, ion channels, GPCRs and nuclear receptors (Hopkins and Groom, 2002). Several publicly available databases have been built and maintained, such as KEGG BRITE (Kanehisa *et al.*, 2006), DrugBank (Wishart *et al.*, 2008), GLIDA (Okuno *et al.*, 2007), SuperTarget and Matador (Günther *et al.*, 2008), and BRENDA (Schomburg *et al.*, 2004), and ChEMBL (Overington, 2009), containing drug–target interaction and other related sources of information, like chemical and genomic data.

A property of the current drug–target interaction databases is that they contain a rather small number of drug–target pairs which are experimentally validated interactions. This motivates the need for developing methods that predict true interacting pairs with high accuracy.

Recently, machine learning methods have been introduced to tackle this problem. They can be viewed as instances of the more general link prediction problem, see Lü and Zhou (2011) for a recent survey of this topic. These methods are motivated by the observation that similar drugs tend to target similar proteins (Schuffenhauer *et al.*, 2003; Klabunde, 2007). This property was shown for instance for chemical (Martin *et al.*, 2002) and side effect similarity (Campillos *et al.*, 2008), and motivated the development of an integrated approach for drug–target interaction prediction (Jaroch and Weinmann, 2006). A desirable property of this approach is that it does not require the 3D structure information of the target proteins, which is needed in traditional methods based on docking simulations (Cheng *et al.*, 2007).

The current state-of-the-art for the *in silico* prediction of drug–target interaction is formed by methods that employ similarity measures for drugs and for targets in the form by kernel functions, like Bleakley and Yamanishi (2009); Jacob and Vert (2008); Wassermann *et al.* (2009); Yamanishi *et al.* (2008, 2010). By using kernels, multiple sources of information can be easily incorporated for performing prediction (Schölkopf *et al.*, 2004).

In Yamanishi *et al.* (2008) different settings of the interaction prediction problem are explored.

The authors make the distinction between 'known' drugs or targets, for which at least one interaction is in the training set; and 'new' drugs or targets, for which there is not. There are then four possible settings, depending on whether the drugs and/or targets are known or new. In this paper we focus on the setting where both the drugs and targets are known. That is, we use known interactions for predicting novel ones.

*to whom correspondence should be addressed

Table 1. The number of drugs and target proteins, their ratio, and the number of interactions in the drug–target datasets from Yamanishi *et al.* (2008).

Dataset	Drugs	Targets	$n_\mathrm{d}/n_\mathrm{t}$	Interactions
Enzyme	445	664	0.67	2926
Ion Channel	210	204	1.03	1476
GPCR	223	95	2.35	635
Nuclear Receptor	54	26	2.08	90

We want to analyze the relevance of the topology of drug–target interaction networks as source of information for predicting interactions. We do this by introducing a kernel that captures the topological information. Using a simple machine learning method we then compare this kernel to kernels based on other sources of information.

Specifically, we start from the assumption that two drugs that interact in a similar way with the targets in a known drug–target interaction network, will also interact in a similar way with new targets. We formalize this property by describing each drug with an interaction profile, a binary vector describing the presence or absence of interaction with every target in that network. The interaction profile of a target is defined in a similar way. From these profiles we construct the Gaussian Interaction Profile kernel.

We show that interaction profiling can be effectively used for accurate prediction of drug–target interaction. Specifically, we propose a simple regularized least square algorithm incorporating a product of kernels constructed from drug and target interaction profiles. We test the predictive performance of this method on four drug–target interaction networks in humans involving enzymes, ion channels, GPCRs and nuclear receptors. These experiments show that using *only* information on the topology of the drug–target interaction, in the form of interaction profiles, excellent results are achieved as measured by the area under the precision-recall curve (AUPR) (Davis and Goadrich, 2006). In particular, on three of the four considered datasets the performance is superior to the best results of current state-of-the-art methods which use multiple sources of information.

We further show that the proposed method can be easily extended to also use other sources of information in the form of suitable kernels. Results of experiments where also chemical and genomic information on drugs and targets is included show excellent performance, with AUPR score of 91.5, 94.3, 79.0 and 68.4 on the four datasets, achieving an improvement of 7.4, 13.0, 12.3 and 7.2 over the best results reported in Bleakley and Yamanishi (2009). A thorough analysis of the results enable us to detect several new putative drug–target interactions, see http://cs.ru.nl/~tvanlaarhoven/drugtarget2011/new-interactions/.

2 MATERIALS

We used four drug–target interaction networks in humans involving enzymes, ion channels, G-protein-coupled receptors (GPCRs) and nuclear receptors; first analyzed by Yamanishi *et al.* (2008). We worked with the datasets provided by these authors, in order to facilitate benchmark comparisons with the current state-of-the-art algorithms that do the same. These datasets are publicly available at http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/. Table 1 lists some properties of the datasets.

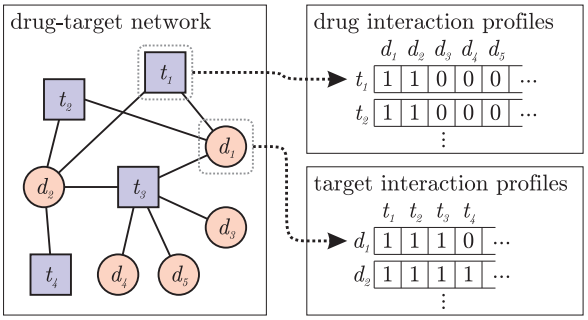


Figure 1. An illustration of the construction of interaction profiles from a drug–target interaction network. Circles are drugs, squares are targets. In this example the interaction profile of target t_1 indicates that it interacts with drugs d_1 and d_2, but not with d_3, d_4 or d_5.

Drug–target interaction information was retrieved from the KEGG BRITE (Kanehisa *et al.*, 2006), BRENDA (Schomburg *et al.*, 2004), SuperTarget (Günther *et al.*, 2008) and DrugBank (Wishart *et al.*, 2008) databases. Chemical structures of the compounds was derived from the DRUG and COMPOUND sections in the KEGG LIGAND database (Kanehisa *et al.*, 2006). The chemical structure similarity between compounds was computed using SIMCOMP (Hattori *et al.*, 2003). This resulted in a similarity matrix for the denoted by S_c, which represents the chemical space. Amino acid sequences of the target (human) proteins were obtained from the KEGG GENES database (Kanehisa *et al.*, 2006). Sequence similarity between proteins was computed using a normalized version of Smith–Waterman score (Smith and Waterman, 1981), resulting in a similarity matrix denoted S_g, which represents the genomic space.

3 METHODS

3.1 Problem formalization

We consider the problem of predicting new interactions in a drug–target interaction network. Formally we are given a set $X_\mathrm{d} = \{d_1, d_2, \ldots, d_{n_\mathrm{d}}\}$ of drugs and a set $X_\mathrm{t} = \{t_1, t_2, \ldots, t_{n_\mathrm{t}}\}$ of target proteins. There is also a set of known interactions between drugs and targets. If we consider these interactions as edges, then they form a bipartite network. We can characterize this network by the $n_\mathrm{d} \times n_\mathrm{t}$ adjacency matrix Y. That is, $y_{ij} = 1$ if drug d_i interacts with target t_j, and $y_{ij} = 0$ otherwise. Our task is now to rank all drug–target pairs (d_i, t_j) such that highest ranked pairs are the most likely to interact.

3.2 Gaussian Interaction Profile Kernel

Our method is based on the assumption that drugs exhibiting a similar pattern of interaction and non-interaction with the targets of a drug–target interaction network are likely to show similar interaction behavior with respect to new targets. We use a similar assumption on targets. We therefore introduce the (target) *interaction profile* $y_{\mathrm{d}i}$ of a drug d_i to be the binary vector encoding the presence or absence of interaction with every target in the considered drug–target network. This is nothing more than row i of the

adjacency matrix Y. Similarly, the (drug) interaction profile $y_{\mathrm{t}j}^T$ of a target protein t_j is a vector specifying the presence or absence of interaction with every drug in the considered drug–target network. The interaction profiles generated from a drug–target interaction network can be used as feature vectors for a classifier. Figure 1 illustrates the construction of interaction profiles.

Following the current state-of-the-art for the drug–target interaction prediction problem, we will use kernel methods, and hence construct a kernel from the interaction profiles. This kernel does not include any information beyond the topology of the drug–target network.

One of the most popular choices for constructing a kernel from a feature vector is the Gaussian kernel, also known as the Radial Basis Function (RBF) kernel. This kernel is, for drugs d_i and d_j,

$$K_{\mathrm{GIP,d}}(d_i, d_j) = \exp(-\gamma_{\mathrm{d}}\|y_{\mathrm{d}i} - y_{\mathrm{d}j}\|^2).$$

A kernel for the similarities between target proteins, $K_{\mathrm{GIP,t}}$, can be defined analogously. We call these kernels Gaussian Interaction Profile (GIP) kernels.

The parameter γ_{d} controls the kernel bandwidth. We set

$$\gamma_{\mathrm{d}} = \tilde{\gamma}_{\mathrm{d}} \Big/ \Big(\frac{1}{n_{\mathrm{d}}}\sum_{i=1}^{n_{\mathrm{d}}} |y_{\mathrm{d}i}|^2\Big).$$

That is, we normalize the parameter by dividing it by the average number of interactions per drug. With this choice the kernel values become independent of the size of the dataset. In principle the new bandwidth parameter $\tilde{\gamma}_{\mathrm{d}}$ could be set with cross-validation, but in this paper we simply use $\tilde{\gamma}_{\mathrm{d}} = 1$.

There are other ways to construct a kernel from interaction profiles. For example, Basilico and Hofmann (2004) propose using the correlation of interaction profiles. We have performed brief experiments with these other kernels, which show that Gaussian Interaction Profile kernels consistently outperform kernels based on correlation or inner products. The detailed results of these experiments are included in Supplementary Table S1.

3.3 Integrating Chemical and Genomic Information

We construct kernels containing information about the chemical and genomic space from the similarity matrices S_d and S_g. Because these similarity matrices are neither symmetric nor positive definite we apply a simple transformation to make them symmetric with $S_{sym} = (S + S^T)/2$ and add a small multiple of the identity matrix to enforce the positive definite property. We denote the resulting kernels for drugs and targets by $K_{\mathrm{chemical,d}}$ and $K_{\mathrm{genomic,t}}$ respectively.

To combine the interaction profile kernel with these chemical and genomic kernels, we use a simple weighted average,

$$K_{\mathrm{d}} = \alpha_{\mathrm{d}} K_{\mathrm{chemical,d}} + (1 - \alpha_{\mathrm{d}}) K_{\mathrm{GIP,d}}$$
$$K_{\mathrm{t}} = \alpha_{\mathrm{t}} K_{\mathrm{genomic,t}} + (1 - \alpha_{\mathrm{t}}) K_{\mathrm{GIP,t}}.$$

For the reported results of our evaluation we use simply the unweighted average, for both drugs and targets, i.e. $\alpha_{\mathrm{d}} = \alpha_{\mathrm{t}} = 0.5$. In section 4.2 we further analyze the effect of these parameters on the predictive performance of the method.

3.4 RLS-avg classifier

In principle we could use the Gaussian Interaction Profile kernels with any kernel based classification or ranking algorithm. We choose to use a very basic classifier, the (kernel) Regularized Least Squares (RLS) classifier. While Least Squares is primarily used for regression, when a good kernel is used it has classification accuracy similar to that of Support Vector Machines (Rifkin and Klautau, 2004). Our own experiments confirm this finding. In the RLS classifier, the predicted values $\hat{y}$ with a given kernel K have a simple closed form solution,

$$\hat{y} = K(K + \sigma I)^{-1} y,$$

where σ is a regularization parameter. Higher values of σ give a smoother result, while for $\sigma = 0$ we get $\hat{y} = y$, and hence no generalization at all. The value $\hat{y}$ is a real valued score, which we can interpret as a confidence.

The RLS classifier is sensitive to the encoding used for y. Here we use 1 for encoding interacting pairs and 0 for non-interacting ones. Brief experiments have shown that the classifier is not sensitive to this choice, as long as the value used for non-interactions is close to 0. Using a value very different from 0, like -1, would place too much weight on non-interactions. The classifier would then try to avoid predicting pairs that look like non-interactions, rather than predicting pairs that look like interactions.

In the previous sections we defined kernels on drugs and kernels on target proteins. There are several ways in which we can use kernels in both of these dimensions. Following other works, like Bleakley and Yamanishi (2009); Zheng Xia and Wong (2010), a simple and effective approach is to apply the classifier for each drug independently using, only the target kernel; and also for each target independently using only the drug kernel. Then the final score for a drug–target pair is a combination of the two outputs.

Here we use the average of the output values, and denote the resulting method by RLS-avg. Observe that in the formulation of the RLS classifier we use, performing independent prediction amounts to replacing the vector y with the matrix Y, hence the prediction of RLS-avg is

$$\hat{Y} = \frac{1}{2}\big(K_{\mathrm{d}}(K_{\mathrm{d}} + \sigma I)^{-1} Y\big) + \frac{1}{2}\big(K_{\mathrm{t}}(K_{\mathrm{t}} + \sigma I)^{-1} Y^T\big)^T.$$

Note this model is slightly different from using the Kronecker sum kernel (Kashima *et al.*, 2009a). Because regularization is performed for drugs and targets separately in the RLS-avg method, rather than jointly.

3.5 RLS-Kron classifier

A better alternative is to combine the kernels into a larger kernel that directly relates drug–target pairs. This is done with the Kronecker product kernel (Basilico and Hofmann, 2004; Ben-Hur and Noble, 2005; Oyama and Manning, 2004; Hue and Vert, 2010). The Kronecker product $K_{\mathrm{d}} \otimes K_{\mathrm{t}}$ of the drug and target kernels is

$$K((d_i, t_j), (d_k, t_l)) = K_{\mathrm{d}}(d_i, d_k) K_{\mathrm{t}}(t_j, t_l).$$

With this kernel we can make predictions for all pairs at once,

$$\mathrm{vec}(\hat{Y}^T) = K(K + \sigma I)^{-1} \mathrm{vec}(Y^T),$$

where $\mathrm{vec}(Y^T)$ is the a vector of all interaction pairs, created by stacking the columns of Y^T. We call this method RLS-Kron.

Using the Kronecker product kernel directly would involve calculating the inverse of an $n_d n_t \times n_d n_t$ matrix, which would take $\mathcal{O}((n_d n_t)^3)$ operations, and would also require too much memory. We use a more efficient implementation based on eigendecompositions, previously presented in Raymond and Kashima (2010).

Let $K_d = V_d \Lambda_d V_d^T$ and $K_t = V_t \Lambda_t V_t^T$ be the eigendecompositions of the two kernel matrices. Since the eigenvalues(vectors) of a Kronecker product are the Kronecker product of eigenvalues(vectors), for our Kronecker product kernel we have simply $K = K_d \otimes K_t = V \Lambda V^T$, where $\Lambda = \Lambda_d \otimes \Lambda_t$ and $V = V_d \otimes V_t$. The matrix that we want to invert, $K + \sigma I$ has these same eigenvectors V, and eigenvalues $\Lambda + \sigma I$. Hence

$$K(K + \sigma I)^{-1} = V \Lambda (\Lambda + \sigma I)^{-1} V^T.$$

To efficiently multiply this matrix with $\mathrm{vec}(Y^T)$ we can use a further property of the Kronecker product, namely that $(A \otimes B)\mathrm{vec}(X) = \mathrm{vec}(BXA^T)$. Combining these facts we get that the RLS prediction is

$$\hat{Y} = V_d Z^T V_t^T,$$

where

$$\mathrm{vec}(Z) = (\Lambda_d \otimes \Lambda_t)(\Lambda_d \otimes \Lambda_t + \sigma I)^{-1} \mathrm{vec}(V_t^T Y^T V_d).$$

So, to make a RLS prediction using the Kronecker product kernel we only need to perform the two eigendecompositions and some matrix multiplications, bringing the runtime down to $\mathcal{O}({n_d}^3 + {n_t}^3)$. The efficiency of this computation could be further improved yielding a quadratic computational complexity by applying recent techniques for large scale kernel methods for computing the two kernel decompositions (Kashima *et al.*, 2009b; Wu *et al.*, 2006).

3.6 Comparison methods

In order to assess globally the performance of our method, we compare it against current state-of-the-art algorithms. To the best of our knowledge, the best results on these datasets obtained so far are those reported by Bleakley and Yamanishi (2009), where the Bipartite Local Models (BLM) approach was introduced. These results were achieved by combining the output scores of the Kernel Regression Method (KRM) (Yamanishi *et al.*, 2008) and BLM by taking their maximum value. We briefly recall these methods here.

In the KRM method, drugs and targets are embedded into a unified space called the 'pharmacological space'. A regression model is learned between the chemical structure (respectively, genomic sequence) similarity space and this pharmacological space. Then new potential drugs and targets are mapped into the pharmacological space using this regression model. Finally, new drug–target interactions are predicted by connecting drugs and target proteins that are closer than a threshold in the pharmacological space.

The BLM method is similar to our RLS-avg method. In the BLM method, the presence or absence of a drug–target interaction is predicted as follows. First, the target is excluded, and a training set is constructed consisting of two classes: all other known targets of the drug in question, and the targets not known to interact with that drug. Second, a Support Vector Machine that discriminates between the two classes is constructed, using the available genomic kernel for the targets. This model is then used to predict the label of the target, and hence the interaction or non-interaction of the considered drug–target pair. A similar procedure is applied with the roles of drugs and targets reversed, using the chemical structure kernel instead. These two results are combined by taking the maximum value.

Table 2. Results on the drug target interaction datasets. The AUC and AUPR scores are normalized to 100. For each dataset, * indicates the the highest AUC/AUPR score.

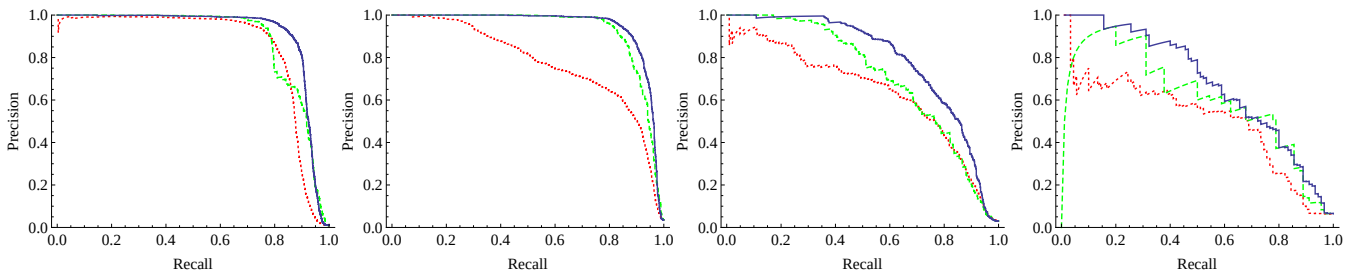
Dataset	Method	Kernel	AUC	AUPR
Enzyme	BY09 (AUC)	chem/gen	97.6	83.3
	BY09 (AUPR)	chem/gen	97.3	84.1
	RLS-avg	GIP	98.2	88.1
	RLS-avg	chem/gen	96.6	84.5
	RLS-avg	avg.	97.9	90.5
	RLS-Kron	GIP	98.3*	88.5
	RLS-Kron	chem/gen	96.6	85.6
	RLS-Kron	avg.	97.8	91.5*
Ion Channel	BY09 (AUC)	chem/gen	97.3	78.1
	BY09 (AUPR)	chem/gen	93.5	81.3
	RLS-avg	GIP	98.5	91.8
	RLS-avg	chem/gen	97.1	80.7
	RLS-avg	avg.	98.1	93.2
	RLS-Kron	GIP	98.6*	92.7
	RLS-Kron	chem/gen	97.1	77.5
	RLS-Kron	avg.	98.4	94.3*
GPCR	BY09	chem/gen	95.5*	66.7
	RLS-avg	GIP	94.5	70.0
	RLS-avg	chem/gen	94.7	66.0
	RLS-avg	avg.	95.0	77.1
	RLS-Kron	GIP	94.7	71.3
	RLS-Kron	chem/gen	94.8	63.8
	RLS-Kron	avg.	95.4	79.0*
Nuclear Receptor	BY09	chem/gen	88.1	61.2
	RLS-avg	GIP	88.7	60.4
	RLS-avg	chem/gen	86.4	54.7
	RLS-avg	avg.	92.5*	67.0
	RLS-Kron	GIP	90.6	61.0
	RLS-Kron	chem/gen	85.9	51.1
	RLS-Kron	avg.	92.2	68.4*

4 EVALUATION

In order to compare the performance of the methods, we performed systematic experiments simulating the process of bipartite network inference from biological data on four drug–target interaction networks. These experiments are done by full leave-one-out cross-validation (LOOCV) as follows. In each run of the method, one drug–target pair (interacting or non-interacting) is left out by setting its entry in the Y matrix to 0. Then we try to recover its true label using the remaining data.

Note that when leaving out a drug–target pair the Y matrix changes, and therefore the GIP kernel has to be recomputed.

We also performed a variation of these experiments using 5 trials of 10-fold cross-validation. We recomputed the GIP kernels for each



(a) Enzyme (b) Ion Channel (c) GPCR (d) Nuclear Receptor

Figure 2. Precision–recall curves for the RLS-Kron method. The red dotted line corresponds to using only the chemical and genomic kernels. The green dashed line corresponds to using only the GIP kernels. The blue solid line corresponds to the average of the two types of kernels. On all datasets the average kernel shows a small improvement over either other kernel type alone.

fold, also for 10-fold cross-validation. So no information about the removed interactions was leaked in this way.

The results can be found in Supplementary Table S2; we observed no large differences compared to the results obtained using LOOCV.

In all experiments we have chosen the values for the parameters in an uninformative way. In particular, we set the regularization parameter $\sigma = 1$ for both RLS methods; and as stated before, we set the kernel bandwidths $\tilde{\gamma}_d = \tilde{\gamma}_t = 1$ for both the drug and target interaction profile kernels.

We assessed the performance of the methods with the following two quality measures generally used in this type of studies: AUC and AUPR. Specifically, we computed the ROC curve of true positives as a function of false positives, and considered the area under the ROC curve (AUC) as quality measure (see for instance Fawcett, 2006). Furthermore, we considered the the precision–recall curve (Raghavan *et al.*, 1989), that is, the plot of the ratio of true positives among all positive predictions for each given recall rate. The area under this curve (AUPR) provides a quantitative assessment of how well, on average, predicted scores of true interactions are separated from predicted scores of true non-interactions. For this task, because there are few true drug–target interactions, the AUPR is a more significant quality measure than the AUC, as it punishes much more the existence of false positive examples found among the best ranked prediction scores (Davis and Goadrich, 2006).

Table 2 contains the results for the two RLS-based classifiers, RLS-avg and RLS-Kron, each with three different kernel combinations:

- GIP: Using only the Gaussian Interaction Profile kernels, i.e. $K_d = K_{GIP,d}$ and $K_t = K_{GIP,t}$, corresponding to $\alpha_d = \alpha_t = 1$.
- chem/gen: Using only the chemical structure and genomic sequence similarity, so $K_d = K_{chemical,d}$ and $K_t = K_{genomic,t}$, corresponding to $\alpha_d = \alpha_t = 0$.
- avg: Using the average of the two types of kernels, corresponding to $\alpha_d = \alpha_t = 0.5$.

For comparison, we have also included in the table as BY09 (AUC) and BY09 (AUPR) the best results from the combined BML and KRM methods from Bleakley and Yamanishi (2009). For the GPCR and Nuclear Receptor datasets, the method with the highest AUC is the same as the one with the highest AUPR, therefore it is included only once, as BY09.

4.1 Analysis

Using only the GIP kernel, our Kronecker product RLS method has AUPR scores of 88.5, 92.7, 71.3 and 61.0 on the Enzyme, Ion Channel, GPCR and Nuclear Receptor datasets respectively. These results are superior to the results from using only the chemical and genomic kernels.

Overall the RLS-Kron and RLS-avg methods have comparable AUC scores. However, the RLS-Kron has a better AUPR when using the GIP kernel, and a worse AUPR when using the chemical and genomic kernels. We believe that this problem is due to the poor quality of the chemical similarity kernel, to which the RLS-Kron method is more sensitive.

Note also that the RLS-avg method is comparable to Bleakley and Yamanishi's bipartite local model (BLM) approach. The differences are that whereas we use a RLS classifier, they use Support Vector Machines; and whereas we use the average to combine results, they use the maximum value. It is therefore not surprising that when using the chemical and genomic kernels the results of the RLS-avg method are very similar to their results.

In all cases the best results are obtained when the GIP kernels are combined with the chemical and genomic kernels. With the RLS-Kron method we then obtain AUPR scores of 91.5, 94.3, 79.0 and 68.4 on the four datasets, which is an improvement of 7.4, 13.0, 12.3 and 7.2 over the best results reported by Bleakley and Yamanishi (2009). Figure 2 shows the precision–recall curves for the RLS-Kron method. Compared to other methods, the RLS-Kron method with the average kernels achieves a good precision also at higher recall values, especially on the larger datasets (Enzyme and Ion Channel).

4.2 Kernels' relevance

In the previous section we have shown that using a mix of the GIP kernels and the chemical and genomic kernels gives results superior to either type of kernel alone. In order to determine the relative importance of the network topology compared to chemical and sequence similarity, we have investigated the change in prediction performance when varying the parameters α_d and α_t between 0 (chemical/genomic kernels only) and 1 (interaction profiles kernels only). For computational reasons we have used 10-fold cross-validation instead of leave-one-out.

In figure 3 we have plotted the AUPR and AUC scores on the GPCR dataset for the different parameter values. Lighter colors correspond to higher values. Because of space limitations, plots for the other datasets are included in Supplementary Figures S1 and S2. For all datasets the optimal AUPR is obtained using a mix of the drug and target kernels. Using the parameters $\alpha_d = \alpha_t = 0.5$, as we did in the previous section, seems to be a good choice across the datasets. Also note that the choice of α_d is more important than the choice of α_t. This seems to indicate that the sequence similarity for targets is more informative than the chemical similarity for drugs. A similar observation was also made in Bleakley and Yamanishi (2009). The poor performance of the RLS-Kron method when using only chemical and genomic kernels that we observed in the previous section appears to be due entirely to this uninformative chemical similarity.

On the larger datasets (Enzyme and Ion Channel) the optimal AUC is obtained with $\alpha_d = 1$, while that choice gives the worst results on the smaller datasets. This can be explained by noting that when there are few drugs, there is less information available for each entry of GIP target kernel, and hence this kernel will be of a lower quality. We have confirmed this hypothesis by testing different sized subsets of the Ion Channel dataset, where we observe the same effect on small subsets. The full results of that experiment are available in Supplementary Figure S3.

4.3 New predicted interactions

In order to analyze the practical relevance of the method for predicting novel drug–target interactions, we conducted an experiment similar to that described by Bleakley and Yamanishi (2009). We ranked the non-interacting pairs according to the scores computed for leave-one-out cross-validation experiments. We estimate the most highly ranked drug–target pairs as most likely to be putative interactions. A list of the top 20 new interactions predicted for each of the four data sets can be found in Supplementary Tables S3–S6.

Table 3 lists the top 10 new interactions predicted for the GPCR dataset. We have looked up these predicted interactions in ChEMBL (Overington, 2009) (version 9), DrugBank (Wishart *et al.*, 2008) and the latest online version of KEGG DRUG (Kanehisa *et al.*, 2006). A significant fraction of the predictions (4 out of 10) is found in one or more of these databases. One should bear in mind that a large fraction of the interactions in these databases are already included in the training data, and hence are not counted as new interactions. Moreover these databases are incomplete, so if a predicted interaction is not present in one of the used databases, this does not necessarily mean it does not exist. For this dataset, we started with only 635 known drug–target interactions and 20550 drug–target pairs not known to interact. Of these 20550 we selected 10 as putative drug–target interaction, and found that at least 4 of them are experimentally verified. These findings support the practical relevance of the proposed method.

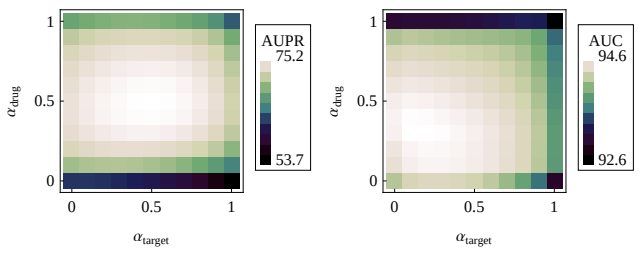


Figure 3. AUPR and AUC scores for the GPCR dataset with different weightings of the kernels. Lighter colors are better. For all datasets $\alpha_d = \alpha_t = 0.5$ gives near optimal results.

Table 3. The top 10 new interactions predicted in the GPCR dataset, 4 have been confirmed. Interactions that appear in the ChEMBL database are marked with "[C]", interactions in Drugbank are marked with "[D]", and interactions in Kegg are marked with "[K]". The NN column gives the similarity to the nearest drug interacting with the same target, and to the nearest target interacting with the same drug.

Rank	Pair	Description	NN
1	**D00283**	**Clozapine**	0.769
[C,D]	**hsa1814**	**DRD3: dopamine receptor D3**	0.455
2	**D02358**	**Metoprolol**	0.750
[C,D]	**hsa154**	**ADRB2: beta-2 adrenergic receptor**	0.434
3	D00604	Clonidine hydrochloride	0.933
	hsa147	ADRA1B: alpha-1B adrenergic receptor	0.435
4	D03966	Eglumegad	0.036
	hsa2914	GRM4: glutamate receptor, metabotropic 4	0.768
5	D00255	Carvedilol	0.380
	hsa152	ADRA2C: alpha-2C adrenergic receptor	0.489
6	**D04625**	**Isoetharine**	0.737
[K]	**hsa154**	**ADRB2: beta-2 adrenergic receptor**	0.434
7	D03966	Eglumegad	0.036
	hsa2917	GRM7: glutamate receptor, metabotropic 7	0.758
8	**D02340**	**Loxapine**	0.769
[D]	**hsa1812**	**DRD1: dopamine receptor D1**	0.205
9	D00503	Perphenazine	0.857
	hsa1816	DRD5: dopamine receptor D5	0.529
10	D00682	Carboprost tromethamine	0.914
	hsa5739	PTGIR: prostaglandin I2 receptor (IP)	0.150

Table 4. The number of highly ranked new interactions that are found in at least one of the three considered databases (ChEMBL, DrugBank or KEGG DRUG).

Dataset	Method	Top 20	Top 50	Top 80
Enzyme	BY09	6 (30%)	15 (30%)	17 (21%)
	RLS-Kron-avg	11 (55%)	15 (30%)	22 (28%)
Ion Channel	BY09	11 (55%)	14 (28%)	18 (22%)
	RLS-Kron-avg	8 (40%)	12 (24%)	22 (28%)
GPCR	BY09	13 (65%)	22 (44%)	30 (38%)
	RLS-Kron-avg	9 (45%)	28 (56%)	40 (50%)
Nuclear Receptor	BY09	5 (25%)	15 (30%)	22 (28%)
	RLS-Kron-avg	9 (45%)	20 (40%)	22 (28%)

We compared the newly predicted interactions generated by RLS-Kron-avg and those generated by Bleakley and Yamanishi (2009), here referred to as BY09. Specifically, given a dataset, for each method we extracted from its top x new predictions those that have been experimentally validated (that is, that could be found in ChEMBL, DrugBank or KEGG DRUG). Table 4 contains a summary of the results for $x = 20, 50, 80$. Looking at the top 20 predictions it seems that the two methods perform best on different datasets. For the top 50 and top 80 predictions, the results indicate

the capability of RLS-Kron-avg to predict successfully more new interactions than BY09.

We then compared the resulting two sets of confirmed new predictions among the top 50, by looking at common predictions and at interactions uniquely predicted by only one of the two methods. The results for the four datasets can be found in Supplementary Tables S7–S10.

On the Enzyme dataset BY09 and RLS-Kron-avg both successfully predicted 15 new interactions, with 10 common predictions. On the Ion Channel dataset, BY09 and RLS-Kron-avg successfully predicted 14 and 12 new interactions, respectively, of which only 1 interaction was predicted by both methods. Although BY09 found slightly more confirmed interactions they were less diverse, since 11 of them involve interactions between (different types of) the voltage-gated sodium channel alpha subunit target and only 2 drugs: Prilocaine and Tocainide. On the other hand, RLS-Kron-avg found interactions 4 different classes of targets and 10 different drugs. On the GPCR dataset, BY09 and RLS-Kron-avg successfully predicted 22 and 28 new interactions, respectively, with 14 common predictions. Finally, on the Nuclear Receptor dataset, BY09 and RLS-Kron-avg successfully predicted 15 and 20 new interactions, respectively. Among them, 13 were in common.

In general, the two methods seem to differ in the type of new predictions made. While there is always an overlap of new interactions between the two methods, there is also always a subset of new interactions which RLS-Kron-avg can successfully predict but BY09 fails to predict and vice-versa. Moreover, there seems to be a slight tendency of BY09 to generate new successful predictions that are less diverse than those generated by RLS-Kron-avg. However, we were not able to identify any differential biological bias of the methods towards the detection of specific types of interactions.

4.4 Surprising interactions

A closer inspection shows that many of the predicted interactions are not very surprising. For example, the GPCR dataset contains the interaction between Clozapine and Dopamine receptor D1. The drug Loxapine is very similar to Clozapine, and it is therefore to be expected that our method also predicts Loxapine to interact with Dopamine receptor D1. An analogous thing happens with very similar target proteins. In order to provide a quantitative measure of how surprising these predictions are, we computed the similarity of a the drug and target in an interaction pair to their Nearest Neighbor (NN), that is, the most similar drug (with respect to chemical structure similarity) and target (with respect to sequence similarity) in the training set, respectively. These similarities, which we call surprise scores, are listed in the NN column of table 3. An inspection of the surprise scores shows that the majority of the drug–target pairs predicted by our method consist of a drug and a target very similar to a drug and a target already known to interact, and therefore they are not very surprising. This phenomenon is common to any computational approach that uses similarity between objects for inferring interaction.

To assess the ability of our method to also predict more surprising interactions, we have looked specifically at the predicted interactions where there is no similar drug interacting with the same target or similar target interacting with the same drug in the dataset. We pick a threshold value and consider drugs (targets) to be dissimilar if their chemical (genomic) similarity is less than this threshold. We have used the threshold 0.5 for the chemical similarity and 0.25 for the genomic similarity.

When only these 'surprising' pairs are considered, we find, as expected, that fewer of them are present in the ChEMBL, DrugBank and KEGG databases. But we still find more interactions among the highly ranked 'surprising' pairs compared to to those that are ranked lower. For example, on the GPCR dataset, 89 of the 500 highest ranked pairs were surprising, and 10 of them (11%) were found in one of the databases. See the online Suplementary Material for details.

5 DISCUSSION

We have presented a new kernel that leads to good predictive performance as measured by AUPR on the task of predicting interactions between drugs and target proteins. An interesting aspect of our Gaussian Interaction Profile kernel is that it uses no properties beyond the interactions themselves. This means that knowing the sequence of proteins and chemical structure of drugs is perhaps not as important for this task as previously thought. For example, on the Ion Channel dataset our method with only the GIP kernel has an AUC score of 98.6 and an AUPR score of 92.7, which improves upon the state-of-the-art, while using less prior information.

Besides the GIP kernel we have also introduced the RLS-Kron algorithm that combines a kernel on drugs and a kernel on targets using the Kronecker product. Compared to previous methods that do prediction with the two kernels independently and then combine the results, this new method represents a small but consistent improvement.

By combining the GIP kernel with chemical and genomic information we get a method with excellent performance. This method has AUPR scores of 91.5, 94.3, 79.0 and 68.4 on four datasets of drug–target interaction networks in humans, representing an average improvement of 10 points over previous results. The AUPR is a particularly relevant metric for this problem, because it is very sensitive to the correctness of the highest ranked predictions. The large improvement in AUPR suggests that the top ranked putative drug–target interactions found by our method are more likely to be correct than those found with previous methods.

A limitation of all machine learning methods for finding new drug–target interactions is that they are sensitive to inherent biases contained in the training data. It would be interesting to try and analyze the bias of existing datasets of drug–target interaction, but this is out of the scope of the present paper. Note also that the datasets by Yamanishi *et al.* (2008) used in this paper do not include any singletons: each drug interacts with at least one target, and each target interacts with at least one drug. This property could affect the cross-validation results, by allowing a limited form of cheating. However, the experiments in section 4.3 show that our method also works when tested in other ways.

A further limitation of the approach used in this paper is that it can only be applied to detect new interactions for a target or a drug for which at least one interaction has already been established. Therefore, biologists can use the method as guidance for extending their knowledge about the interaction of a drug or of a target, not for discovering interactions of a new drug or target (that is, one for which no interaction is known). In particular, our method is useful for experimentalist to aid in experimental design and interpretation,

especially in solving problems related to drug-target selectivity and polypharmacology (Metz and Hajduk, 2010; Merino *et al.*, 2010).

There are several ways in which the result might further be improved. So far we have used uninformative choices of the parameters: $\tilde{\gamma} = 1$, $\sigma = 1$ and $\alpha = 0.5$. Of these choices we have only investigated the last one. Perhaps with tuning of the other parameters better predictions are possible, although one has to be careful not to over-fit them to the data.

Another avenue for improvement is in using more information about drugs and targets. Since combining the GIP kernel with chemical and genomic kernels leads to a better predictive performance, perhaps adding different information in the form of additional kernels would yield further improvements. These kernels could be interaction profile kernels based on other types data, such as protein–protein interaction networks. Similarly, for each pair of interacting drug and target more information is known beyond the fact they interact. For example, the type of interaction, the binding strength, the mechanism of discovery and its uncertainty might all be known. In this paper we have made no use of this additional information, nor did we attempt to predict the type or strength of interactions.

REFERENCES

Basilico, J. and Hofmann, T. (2004). Unifying collaborative and content-based filtering. In *ICML '04: Proceedings of the 21st International Conference on Machine learning*, pages 65–72. ACM.

Ben-Hur, A. and Noble, W. S. (2005). Kernel methods for predicting protein–protein interactions. *Bioinformatics*, **21**(suppl 1), i38–i46.

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

Campillos, M. *et al.* (2008). Drug target identification using side-effect similarity. *Science*, **321**(5886), 263–266.

Cheng, A. C. *et al.* (2007). Structure-based maximal affinity model predicts small-molecule druggability. *Nature Biotechnology*, **25**(1), 71–5+.

Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *ICML '06: Proceedings of the 23rd International Conference on Machine learning*, pages 233–240. ACM.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, **27**(8), 861–874.

Günther, S. *et al.* (2008). SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic acids res.*, **36**(Database issue), D919–D922.

Haggarty, S. J. *et al.* (2003). Multidimensional chemical genetic analysis of diversity-oriented synthesis-derived deacetylase inhibitors using cell-based assays. *Chemistry & Biology*, **10**(5), 383–396.

Hattori, M. *et al.* (2003). Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem Soc.*, **125**(39), 11853–65.

Hopkins, A. L. and Groom, C. R. (2002). The druggable genome. *Nature reviews. Drug discovery*, **1**(9), 727–730.

Hue, M. and Vert, J.-P. (2010). On learning with kernels for unordered pairs. In J. Fürnkranz and T. Joachims, editors, *ICML '10: Proceedings of the 27th International Conference on Machine Learning*, pages 463–470, Haifa, Israel. Omnipress.

Jacob, L. and Vert, J.-P. (2008). Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics*, **24**(19), 2149–2156.

Jaroch, S. E. and Weinmann, H., editors (2006). *Chemical Genomics: Small Molecule Probes to Study Cellular Function*. Ernst Schering Research Foundation Workshop. Springer, Berlin.

Kanehisa, M. *et al.* (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic acids res.*, **34**(Database issue), D354–357.

Kashima, H. *et al.* (2009a). On pairwise kernels: An efficient alternative and generalization analysis. In *PAKDD '09: Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 1030–1037. Springer.

Kashima, H. *et al.* (2009b). Recent advances and trends in large-scale kernel methods. *IEICE Transactions*, **92-D**(7), 1338–1353.

Klabunde, T. (2007). Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br. J. Pharmacol.*, **152**, 5–7.

Lü, L. and Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, **390**(6), 1150 – 1170.

Martin, Y. C. *et al.* (2002). Do structurally similar molecules have similar biological activity? *J. Med. Chem.*, **45**(19), 4350–4358.

Merino, A. *et al.* (2010). Drug profiling: knowing where it hits. *Drug Discovery Today*, **15**(17-18), 749–756.

Metz, J. T. and Hajduk, P. J. (2010). Rational approaches to targeted polypharmacology: creating and navigating protein-ligand interaction networks. *Current Opinion in Chemical Biology*, **14**(4), 498–504. Next Generation Therapeutics.

Okuno, Y. *et al.* (2007). GLIDA: GPCR ligand database for chemical genomics drug discovery database and tools update. *Nucleic Acids Res.*, **36**(Database issue), D907–D912.

Overington, J. (2009). ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI).

Oyama, S. and Manning, C. D. (2004). Using feature conjunctions across examples for learning pairwise classifiers. In *ECML '04: Proceedings of the 15th European Conference on Machine Learning*, volume 3201, pages 322–333. Springer.

Raghavan, V. V. *et al.* (1989). A Critical Investigation of Recall and Precision as Measures of Retrieval System Performance. *ACM Transactions on Information Systems*, **7**(3), 205–229.

Raymond, R. and Kashima, H. (2010). Fast and scalable algorithms for semi-supervised link prediction on static and dynamic graphs. In *Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part III*, ECML PKDD'10, pages 131–147, Berlin, Heidelberg. Springer-Verlag.

Rifkin, R. and Klautau, A. (2004). In defense of one-vs-all classification. *Journal of Machine Learning Research*, **5**, 101–141.

Schölkopf, B. *et al.*, editors (2004). *Kernel Methods in Computational Biology*. MIT Press.

Schomburg, I. *et al.* (2004). BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.*, **32**(suppl_1), D431–433.

Schuffenhauer, A. *et al.* (2003). Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.*, **43**(2), 391–405.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.*, **147**(1), 195–197.

Wassermann, A. M. *et al.* (2009). Ligand prediction for orphan targets using support vector machines and various target-ligand kernels is dominated by nearest neighbor effects. *J. Chem. Inf. Model*, **49**(10), 2155–2167.

Wishart, D. S. *et al.* (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids res.*, **36**(Database issue), D901–906.

Wu, G. *et al.* (2006). Incremental approximate matrix factorization for speeding up support vector machines. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 760–766. ACM.

Yamanishi, Y. *et al.* (2008). Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, **24**, i232–i240.

Yamanishi, Y. *et al.* (2010). Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*, **26**(12), i246–i254.

Zheng Xia, Ling-Yun Wu, X. Z. and Wong, S. T. (2010). Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Systems Biology*, **4**(Suppl 2).