BIOINFORMATICS

Vol. 26 ISMB 2010, pages i246–i254
doi:10.1093/bioinformatics/btq176

Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework

Yoshihiro Yamanishi[1,2,3]*, Masaaki Kotera[4], Minoru Kanehisa[4,5] and Susumu Goto[4]

[1]Mines ParisTech, Centre for Computational Biology, 35 rue Saint-Honore, F-77305 Fontainebleau Cedex, [2]Institut Curie, F-75248, [3]INSERM U900, F-75248, Paris, France, [4]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011 and [5]Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai Minato-ku, Tokyo 108-8639, Japan

ABSTRACT

Motivation: *In silico* prediction of drug–target interactions from heterogeneous biological data is critical in the search for drugs and therapeutic targets for known diseases such as cancers. There is therefore a strong incentive to develop new methods capable of detecting these potential drug–target interactions efficiently.

Results: In this article, we investigate the relationship between the chemical space, the pharmacological space and the topology of drug–target interaction networks, and show that drug–target interactions are more correlated with pharmacological effect similarity than with chemical structure similarity. We then develop a new method to predict unknown drug–target interactions from chemical, genomic and pharmacological data on a large scale. The proposed method consists of two steps: (i) prediction of pharmacological effects from chemical structures of given compounds and (ii) inference of unknown drug–target interactions based on the pharmacological effect similarity in the framework of supervised bipartite graph inference. The originality of the proposed method lies in the prediction of potential pharmacological similarity for any drug candidate compounds and in the integration of chemical, genomic and pharmacological data in a unified framework. In the results, we make predictions for four classes of important drug–target interactions involving enzymes, ion channels, GPCRs and nuclear receptors. Our comprehensively predicted drug–target interaction networks enable us to suggest many potential drug–target interactions and to increase research productivity toward genomic drug discovery.

Supplementary information: Datasets and all prediction results are available at http://cbio.ensmp.fr/~yyamanishi/pharmaco/.

Availability: Softwares are available upon request.

Contact: yoshihiro.yamanishi@ensmp.fr

1 INTRODUCTION

The identification of drug–target interactions (interactions between drugs and target proteins) is a key area in genomic drug discovery. Interactions with ligands can modulate the function of many classes of pharmaceutically useful protein targets including enzymes, ion channels, G protein-coupled receptors (GPCRs) and nuclear receptors. Owing to the completion of the human genome sequencing and the development of various biotechnologies, we are beginning to analyze the 'genomic space' populated by these protein classes. At the same time, the high-throughput screening (HTS) of large-scale chemical libraries is enabling us to explore the entire 'chemical space' of possible compounds. However, our knowledge about the relationship between the chemical space and the genomic space is very limited.

In recent years, the importance of chemical genomics is growing fast to relate the chemical space with the genomic space (Dobson *et al*., 2004; Kanehisa *et al*., 2006; Stockwell, 2000). The genome-wide detection of compound–protein interactions is a key issue in chemical genomics research, which can lead to identification of new drug leads and therapeutic targets for known diseases such as cancers. Although various biological assays are becoming available, experimental determination of compound–protein interactions remains challenging and very expensive even nowadays. There is therefore a strong incentive to develop new *in silico* methods capable of detecting these potential compound–protein interactions efficiently.

Traditional computational approaches are categorized into ligand-based approach and docking approach. Ligand-based approach like QSAR (Quantitative Structure Activity Relationship) compares a candidate ligand with the known ligands of a target protein to predict its binding using machine learning methods (Butina *et al*., 2002; Byvatov *et al*., 2003). However, the performance of the ligand-based approach is poor when the number of known ligands for a target protein of interest decreases. The docking is a powerful approach, but the docking cannot be applied to proteins whose 3D structures are unknown (Rarey *et al*., 1996). This limitation is serious for membrane proteins. For example, there are only two GPCRs with 3D structure information as of writing. Therefore it is difficult to use the docking on a genome-wide scale.

Recently, a variety of statistical methods have been developed to predict compound–protein interactions on a genome-wide scale, following the spirit of chemical genomics. The underlying idea is that similar ligands are likely to interact with similar proteins, and the prediction is performed based on compound chemical structures, protein sequences and the currently known compound–protein interactions. A straightforward statistical approach is to use binary classification methods where they take compound–protein pairs as an input for machine learning classifiers such as neural network and support vector machine (SVM) (Bleakley and Yamanishi, 2009; Bock and Gough, 2005; Erhan *et al*., 2006; Faulon *et al*., 2008; Jacob and Vert, 2008; Nagamine and Sakakibara, 2007). The other statistical approach is the distance learning in the framework of supervised bipartite graph inference (Yamanishi, 2009; Yamanishi *et al*., 2008).

Another promising approach is to use pharmacological information. The use of side-effect similarity has been recently proposed, which is based on the assumption that drugs with

*To whom correspondence should be addressed.

similar side-effects are likely to interact with similar target proteins (Campillos *et al*., 2008). However, the method requires drug package inserts that describe the detailed side-effect information, so it is applicable only to marketed drugs for which side-effect information is available. Therefore, it is not possible to infer potential interactions between new drug candidate compounds and target proteins.

In this article, we investigate the relationship between the chemical space, the pharmacological space and the topology of drug–target interactions networks. We then develop a new method to predict unknown drug–target interactions from chemical structure information, genomic sequence information and pharmacological effect information on a large scale. The proposed method consists of two steps: (i) prediction of pharmacological effects from chemical structures of given compounds and (ii) inference of unknown drug–target interactions based on the pharmacological effect similarity in the framework of supervised bipartite graph inference. The algorithm proposed in the first step enables us to obtain pharmacological information about not only marketed drugs but also any compounds, based on the correlation between chemical structures and pharmacological/adverse effects (Scheiber *et al*., 2009), which makes it possible to perform screening of any drug candidate compounds against many target candidate proteins. To our knowledge, there are no methods which predict drug–target interactions based on chemical, genomic and pharmacological data simultaneously. In the results, we make predictions for four classes of important drug–target interactions involving enzymes, ion channels, GPCRs and nuclear receptors. A comprehensive prediction of drug–target interaction networks enables us to suggest new potential drug–target interactions.

2 MATERIALS

In this study, we focus on drugs targeting four pharmaceutically useful target classes: enzymes, ion channels, GPCRs and nuclear receptors.

2.1 Chemical data

Chemical structures of drugs and other compounds were obtained from the KEGG DRUG and KEGG LIGAND databases (Kanehisa *et al*., 2008). We computed the chemical structure similarities between compounds using SIMCOMP (Hattori *et al*., 2003), a program that finds the common substructures between two compounds and outputs the global similarity score based on a graph alignment algorithm. The similarity between two compound structures $\mathbf{x}$ and $\mathbf{x}'$ is evaluated by Tanimoto coefficient defined as $s_{\text{chem}}(\mathbf{x},\mathbf{x}')=|\mathbf{x}\cap\mathbf{x}'|/|\mathbf{x}\cup\mathbf{x}'|$. The similarity score is referred to as 'chemical structure similarity' in this study. Applying this operation to all compound pairs, we construct a similarity matrix denoted as $\mathbf{C}$. The similarity matrix $\mathbf{C}$ is considered to represent 'chemical space'.

2.2 Pharmacological data

Pharmacological effect keywords for drugs (pharmaceutical molecules) were obtained from the JAPIC (Japan Pharmaceutical Information Center) database (http://www.japic.or.jp/). JAPIC manages all package insert information of pharmaceutical products in Japan, under the approval of Health and Welfare Minister of Japan. We used the JAPIC entries (package inserts) of ethical drugs described in natural Japanese language, which were morphologically analyzed to obtain the nouns or phrases using the MeCab program (http://mecab.sourceforge.net/). The resulted set of keywords were translated into English followed by the unification of synonymous words, using life science dictionary (http://lsd.pharm.kyoto-u.ac.jp/en/index.html). Since a pharmaceutical molecule is usually involved in various commercial products, each KEGG DRUG entry of a drug molecule is represented as a logical sum of the presence/absence (1 or 0, respectively) of the unified keywords found in the corresponding JAPIC entries. We obtained 18 653 keywords in total, representing the pharmaceutical effects, adverse effects, caution, usage, properties, etc.

We also performed a simple investigation of the context of the keywords. JAPIC entries are described in an XML format, where the sentences are tagged by the category words such as 'effect', 'side-effect', 'caution' and 'warning'. Unnecessary information in terms of analyzing pharmacological effects, such as manufacturers, are removed using the corresponding XML tag. Various types of profiles can be generated for a drug using different set of the XML tags. We tested using every tag independently to generate a profile, although we found it ineffective. Then we tested grouping the similar XML tags to form five tag groups: 'caution' (unwanted characteristics of the drug, such as adverse event, caution for application or handling, overdose and warning), 'interaction' (the combined use of drugs), 'patient' (the types of patients, such as children, pregnant, elder people, or the people having chronic diseases), 'pharmaceutical effect' (efficacy, usage and pharmacology) and 'property' (such as partition coefficient, pharmacokinetics, melting point and solubility). The number of keywords with the 'caution', 'interaction', 'patient', 'pharmaceutical effect' and 'property' tags are 16 849, 14 223, 16 362, 17 109, and 17 142, respectively. All the keywords for each tag are put on the supplementary website. In this study, we used keywords with the 'pharmaceutical effect' tag as pharmacological keywords.

Each drug is represented by a profile (binary vector) $\mathbf{y}=(y_1,y_2,\ldots,y_K)^\top$ in which a pharmacological keyword is coded 1 or 0, respectively, across the 17 109 keywords. The similarity between two drugs $\mathbf{y}$ and $\mathbf{y}'$ is evaluated by the weighted cosine correlation coefficient between the above profiles as follows:

$$s_{\text{phar}}(\mathbf{y},\mathbf{y}')=\frac{\sum_{k=1}^{K} w_k y_k y'_k}{\sqrt{\sum_{k=1}^{K} w_k y_k^2}\sqrt{\sum_{k=1}^{K} w_k y_k'^2}} \quad (1)$$

where w_k is the weight function for the k-th keyword defined as

$$w_k=\exp(-d_k^2/\sigma^2 h^2), \quad k=1,2,\ldots,K,$$

where d_k is the frequency of the k-th keyword in the data, and K is the total number of keywords in the data, σ is the SD of $\{d_k\}_{k=1}^K$, and h is a parameter (set to 0.1 in this study). The weight function is introduced to put more emphasis on infrequent keywords rather than frequent keywords across different drug package inserts, because rare keywords (e.g. 'cytopenia', 'pancytopenia', 'photosensitivity', 'teratogenic') are more informative than common keywords (e.g. 'disease', 'receptor', 'stability', 'biological') in terms of characteristics of drugs.

The similarity score is referred to as 'pharmacological effect similarity' or 'pharmacological similarity' in this study. Applying this operation to all drug pairs, we construct a similarity matrix denoted as $\mathbf{P}$. The similarity matrix $\mathbf{P}$ is considered to represent 'pharmacological space'.

2.3 Genomic data

Amino acid sequences of proteins coded in the human genome were obtained from the KEGG GENES database (Kanehisa *et al*., 2008). We computed the sequence similarities between two proteins $\mathbf{z}$ and $\mathbf{z}'$ using a normalized version of Smith–Waterman scores (Smith and Waterman, 1981). The similarity score is denoted as $s_{\text{geno}}(\mathbf{z},\mathbf{z}')$ and referred to as 'genomic sequence similarity' in this study. Applying this operation to all protein pairs, we construct a similarity matrix denoted as $\mathbf{G}$. In this study the similarity matrix $\mathbf{G}$ is considered to represent 'genomic space'.

2.4 Drug–target interaction data

The information about the interactions between drugs and target proteins was obtained from the KEGG BRITE (Kanehisa *et al*., 2008), BRENDA

(Schomburg *et al*., 2004), SuperTarget (Gunther *et al*., 2008) and DrugBank (Wishart *et al*., 2008) databases. According to our survey, the numbers of known drugs with pharmacological information in JAPIC are 212, 99, 105 and 27, for their targets enzymes, ion channels, GPCRs and nuclear receptors, respectively. The numbers of the corresponding target proteins in these classes are 664, 204, 95 and 26, respectively. The numbers of the corresponding interactions are 1515, 776, 314 and 44, respectively.

The set of known drug–target interactions is regarded as the 'gold standard' data in this study, and is used for evaluating the performance of the proposed method in the cross-validation experiments as well as training data in the comprehensive prediction.

3 METHODS

Suppose that we are given drug candidate compounds and we want to predict unknown interactions between the compounds and target proteins on a genome-wide scale. The proposed method consists of two steps: (i) prediction of potential pharmacological effects from chemical structures of given compounds and (ii) inference of unknown drug–target interactions based on the pharmacological effect similarity in the framework of supervised bipartite graph inference. The details of each step of the proposed method are described below.

3.1 Prediction of pharmacological effects from compound chemical structures

If pharmacological information is available for given compounds, this process can be skipped. In this subsection, we assume that given compounds do not have any pharmacological information.

3.1.1 Formulation of the problem Let us now consider the situation where chemical structure data is available for all the N compounds $\{\mathbf{x}_i\}_{i=1}^N$, while the pharmacological data is available for the first n compounds $\{\mathbf{y}_i\}_{i=1}^n$ and unavailable for the remaining $(N-n)$ compounds $\{\mathbf{y}_i\}_{i=n+1}^N$. We refer to the first n compounds as *training set*, and we refer to the remaining $N-n$ compounds as *prediction set* below.

For the prediction set, we want to predict a pharmacological profile $\mathbf{y}$ (K-dimensional binary vector) from a chemical structure $\mathbf{x}$ (chemical graph). A straightforward approach would be to apply a binary classification method such as SVM in order to individually predict whether each element y_k in $\mathbf{y}$ is 1 or 0. However, this strategy needs to construct K individual classifiers for K pharmacological keywords, which will require prohibitive computational burden, because K is quite huge in practical applications (K is 17 109 in this study).

Note that the inputs of the supervised bipartite graph inference method in the next step are similarity scores for compounds and proteins. Therefore, we propose to consider predicting the pharmacological similarity scores involving compounds rather than predicting the pharmacological profile itself directly. The key idea here is to reformulate the problem of predicting unknown high-dimensional binary vectors for the prediction set by the problem of predicting unknown similarity scores $s_{\mathrm{phar}}(\mathbf{y}_i,\mathbf{y}_j)$ involving the prediction set.

Let $s_{\mathrm{chem}}(\cdot,\cdot)$ and $s_{\mathrm{phar}}(\cdot,\cdot)$ be chemical structure and pharmacological effect similarity functions, respectively. When we compute the chemical structure similarity scores for $\{\mathbf{x}_i\}_{i=1}^N$, we obtain an $N\times N$ similarity matrix $\mathbf{C}$, where $(\mathbf{C})_{ij}=s_{\mathrm{chem}}(\mathbf{x}_i,\mathbf{x}_j)\,(1\le i,j\le N)$. On the other hand, when we compute the pharmacological similarity scores for $\{\mathbf{y}_i\}_{i=1}^n$, we obtain an $N\times N$ similarity matrix $\mathbf{P}$, where $(\mathbf{P})_{ij}=s_{\mathrm{phar}}(\mathbf{y}_i,\mathbf{y}_j)\,(1\le i,j\le n)\ (n<N)$. Note that $\mathbf{P}$ contains in fact missing values for all entries $(\mathbf{P})_{ij}$ with $\max(i,j)>n$. We want to estimate the missing part of $\mathbf{P}$ using full similarity matrix $\mathbf{C}$, taking into account a form of correlation between the two similarity functions.

In this study, we express each similarity matrix by splitting the matrix into four parts. We denote by $\mathbf{C}_{tt}$ (resp. $\mathbf{P}_{tt}$) the $n\times n$ similarity matrix for the *training set* versus itself, $\mathbf{C}_{pt}$ (resp. $\mathbf{P}_{pt}$) the $(N-n)\times n$ similarity matrix for the *prediction set* versus the *training set* and $\mathbf{C}_{pp}$ (resp. $\mathbf{P}_{pp}$) the $(N-n)\times(N-n)$ similarity matrix for the *prediction set* versus itself:

$$\mathbf{C}=\begin{pmatrix}\mathbf{C}_{tt} & \mathbf{C}_{pt}^\top\\ \mathbf{C}_{pt} & \mathbf{C}_{pp}\end{pmatrix},\quad \mathbf{P}=\begin{pmatrix}\mathbf{P}_{tt} & \mathbf{P}_{pt}^\top\\ \mathbf{P}_{pt} & \mathbf{P}_{pp}\end{pmatrix} \tag{2}$$

Note that $\mathbf{C}_{pt}$ and $\mathbf{C}_{pp}$ are known, while $\mathbf{P}_{pt}$ and $\mathbf{P}_{pp}$ are unknown. The goal is to predict $\mathbf{P}_{pt}$ and $\mathbf{P}_{pp}$ from $\mathbf{C}$ and $\mathbf{P}_{tt}$.

3.1.2 Algorithm The ordinary regression model between an explanatory variable $\mathbf{x}$ and a response variable y can be formulated as $y=f(\mathbf{x})+\epsilon$, where f is a regression function and ϵ is a noise term. By analogy we propose to regard $(\mathbf{x},\mathbf{x}')$ as an explanatory variable and $s_{\mathrm{phar}}(\mathbf{y},\mathbf{y}')$ as a response variable in our context.

Assuming the underlying feature space in which each $\mathbf{x}$ can be represented by m features $u^{(1)}(\mathbf{x}),u^{(2)}(\mathbf{x}),\ldots,u^{(m)}(\mathbf{x})$, we formulate a variant of the regression model as follows:

$$s_{\mathrm{phar}}(\mathbf{y},\mathbf{y}')=f(\mathbf{x},\mathbf{x}')+\epsilon=\mathbf{u}(\mathbf{x})^\top\mathbf{u}(\mathbf{x}')+\epsilon, \tag{3}$$

where $\mathbf{u}(\mathbf{x})=(u^{(1)}(\mathbf{x}),u^{(2)}(\mathbf{x}),\ldots,u^{(m)}(\mathbf{x}))^\top$. We refer to this model as similarity matrix regression model.

We consider features that possess an expansion of the form

$$u(\mathbf{x})=\sum_{j=1}^n s_{\mathrm{chem}}(\mathbf{x},\mathbf{x}_j)\beta_j, \tag{4}$$

where $\boldsymbol{\beta}=(\beta_1,\beta_2,\ldots,\beta_n)^\top$ is a weight vector and n is the number of compounds in the training set.

In order to represent the set of features for all the compounds, we define feature score matrices $\mathbf{U}_t(\mathbf{x})=[\mathbf{u}(\mathbf{x}_1),\mathbf{u}(\mathbf{x}_2),\ldots,\mathbf{u}(\mathbf{x}_n)]^\top$ for the training set and $\mathbf{U}_p(\mathbf{x})=[\mathbf{u}(\mathbf{x}_{n+1}),\mathbf{u}(\mathbf{x}_{n+2}),\ldots,\mathbf{u}(\mathbf{x}_N)]^\top$ for the prediction set. In the matrix form, we can actually compute the feature score matrices as $\mathbf{U}_t=\mathbf{C}_{tt}\mathbf{B}$ for the training set and $\mathbf{U}_p=\mathbf{C}_{pt}\mathbf{B}$ for the prediction set, where $\mathbf{B}=[\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)},\ldots,\boldsymbol{\beta}^{(m)}]$.

We consider predicting $s_{\mathrm{phar}}(\mathbf{y},\mathbf{y}')$ by the inner products of the feature vectors of $\mathbf{x}$ and $\mathbf{x}'$ based on the regression model (3). Since all the compound–compound similarities in the feature space can be represented as $\hat{s}_{\mathrm{phar}}(\mathbf{y}_i,\mathbf{y}_j)=\mathbf{u}(\mathbf{x}_i)^\top\mathbf{u}(\mathbf{x}_j)$ for $1\le i,j\le N$, the missing entries in $\mathbf{P}$ are to be estimated as

$$\textit{Training set}\ \text{versus}\ \textit{Training set}:\hat{\mathbf{P}}_{tt}=\mathbf{U}_t\mathbf{U}_t^\top=\mathbf{C}_{tt}\mathbf{B}\mathbf{B}^\top\mathbf{C}_{tt}^\top,$$

$$\textit{Prediction set}\ \text{versus}\ \textit{Training set}:\hat{\mathbf{P}}_{pt}=\mathbf{U}_p\mathbf{U}_t^\top=\mathbf{C}_{pt}\mathbf{B}\mathbf{B}^\top\mathbf{C}_{tt}^\top,$$

$$\textit{Prediction set}\ \text{versus}\ \textit{Prediction set}:\hat{\mathbf{P}}_{pp}=\mathbf{U}_p\mathbf{U}_p^\top=\mathbf{C}_{pt}\mathbf{B}\mathbf{B}^\top\mathbf{C}_{pt}^\top.$$

Here, we want to find the $n\times m$ weight matrix $\mathbf{B}$ such that $\hat{\mathbf{P}}_{tt}$ fits $\mathbf{P}_{tt}$ as much as possible. If we set $\mathbf{A}=\mathbf{B}\mathbf{B}^\top$, this problem can be replaced by finding $\mathbf{A}$ which minimizes the difference between $\mathbf{P}_{tt}$ and $\hat{\mathbf{P}}_{tt}$. It means that, this enables us to avoid considerable computational burden for computing $\mathbf{B}$ itself, even if m is infinite. Therefore, we attempt to find $\mathbf{A}(=\mathbf{B}\mathbf{B}^\top)$ which minimizes

$$L=\|\mathbf{P}_{tt}-\mathbf{C}_{tt}\mathbf{A}\mathbf{C}_{tt}^\top\|_F^2, \tag{5}$$

where $\|\cdot\|_F$ indicates the Frobenius norm. We can rewrite the above equation in the trace form as

$$L=\mathrm{tr}\left\{(\mathbf{P}_{tt}-\mathbf{C}_{tt}\mathbf{A}\mathbf{C}_{tt}^\top)(\mathbf{P}_{tt}-\mathbf{C}_{tt}\mathbf{A}\mathbf{C}_{tt}^\top)^\top\right\}. \tag{6}$$

From setting $\frac{\partial L}{\partial\mathbf{A}}=0$, the solution is analytically obtained by

$$\mathbf{A}=\mathbf{B}\mathbf{B}^\top=\mathbf{C}_{tt}^{-1}\mathbf{P}_{tt}\mathbf{C}_{tt}^{-1}. \tag{7}$$

Therefore, we can compute the feature-based similarity matrix $\hat{\mathbf{P}}$ involving the prediction set as follows:

$$\hat{\mathbf{P}}_{pt}=\mathbf{U}_p\mathbf{U}_t^\top=\mathbf{C}_{pt}\mathbf{C}_{tt}^{-1}\mathbf{P}_{tt},\ \hat{\mathbf{P}}_{pp}=\mathbf{U}_p\mathbf{U}_p^\top=\mathbf{C}_{pt}\mathbf{C}_{tt}^{-1}\mathbf{P}_{tt}\mathbf{C}_{tt}^{-1}\mathbf{C}_{pt}^\top. \tag{8}$$

3.2 Inference of drug–target interactions

We perform the inference of potential drug–target interactions based on pharmacological information about compounds and genomic information about proteins in the framework of supervised bipartite graph inference. Among several algorithms for the supervised bipartite graph inference mentioned in the introduction section, we use an algorithm based on distance learning (Yamanishi *et al*., 2008), because this method is known to work the best in terms of prediction accuracy and computational efficiency (Lodhi and Yamanishi, 2010).

The procedure of the method for drug–target interaction prediction in this context is briefly explained as follows:

(1) Embed compounds and proteins on the known interaction network into a unified feature space, where interacting compounds and proteins are close to each other.

(2) Learn a correlation model between the pharmacological space and the unified feature space with respect to compounds, and learn a correlation model between the genomic space and the unified feature space with respect to proteins.

(3) Map any compounds onto the unified feature space based on the pharmacological similarities, and map any proteins onto the unified feature space based on the genomic sequence similarities.

(4) Predict potential compound–protein interactions by connecting compounds and proteins which are closer than a threshold in the unified feature space, following the spirit of the nearest neighbor.

The details of each step can be found in the original article.

The resulting prediction score for any new compound $\mathbf{y}$ and protein $\mathbf{z}$ in the fourth process is formulated as

$$g(\mathbf{y},\mathbf{z}) = \sum_{i=1}^{n_y}\sum_{j=1}^{n_z} \alpha_{ij} s_{\text{phar}}(\mathbf{y}_i,\mathbf{y}) s_{\text{geno}}(\mathbf{z}_j,\mathbf{z}) \quad (9)$$

where n_y (resp. n_z) is the number of compounds (resp. proteins) in the training set, $s_{\text{phar}}(\cdot,\cdot)$ is pharmacological similarity function for compounds, $s_{\text{geno}}(\cdot,\cdot)$ is genomic sequence similarity function for proteins and $\alpha_{ij}(i=1,\ldots,n_y, j=1,\ldots,n_z)$ are the parameters learned. If $g(\mathbf{y},\mathbf{z})$ is higher than a threshold, compound $\mathbf{y}$ and protein $\mathbf{z}$ are predicted to interact to each other.

The pharmacological similarity for compounds and the genomic sequence similarity for proteins are used as inputs of the bipartite graph inference method. The use of $s_{\text{phar}}(\cdot,\cdot)$ is a unique feature of this study, while the use of $s_{\text{chem}}(\cdot,\cdot)$ corresponds to the previous study (Yamanishi *et al*., 2008). Note that the method is also able to infer potential interactions involving new target candidate proteins as well as new drug candidate compounds, but we focus on predicting potential interactions involving new drug candidate compounds, because the objective of this article is to investigate the effect of introducing pharmacological information about new drug candidate compounds.

4 RESULTS

4.1 Relationship between chemical and pharmacological spaces with respect to drug targets

We investigated the relationship between the chemical space and the pharmacological space about the same drugs. Each panel in Figure 1 shows the scatter-plot of pharmacological effect similarity scores against chemical structure similarity scores for drugs targeting enzymes, ion channels, GPCRs and nuclear receptors, respectively. The Pearson's correlation coefficients are 0. 321, 0.420, 0.344 and 0.391, respectively (the corresponding *P*-value is almost zero in each case).

It seems that chemical structure similarities are correlated with pharmacological effect similarities to some extent. However,

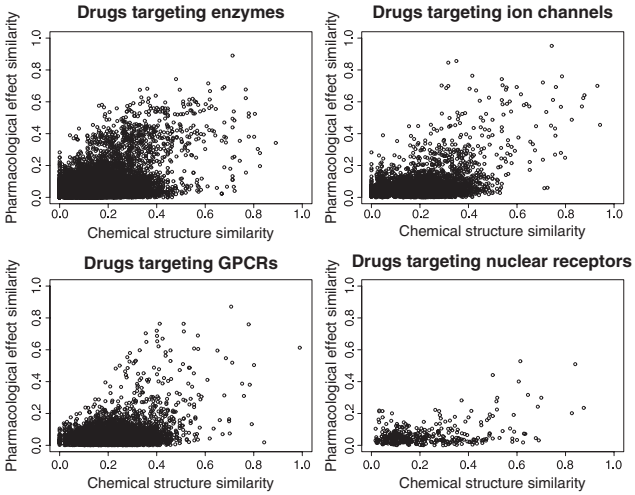

Fig. 1. Scatter-plots of pharmacological effect similarity scores and chemical structure similarity scores for drugs targeting enzyme, ion channel, GPCR and nuclear receptor, respectively.

there are many exceptions. For example, there exist many drug pairs which share high structure similarity but do not have similar pharmacological effects. These results suggest that chemical structures similarity does not always correspond to pharmacological effect similarity.

We investigated the relationship between the chemical space, the pharmacological space and the topology of drug–target interactions networks. We constructed the drug–target interaction network for each protein class using a bipartite graph representation (Yildirim *et al*., 2007). In the bipartite graph, the heterogeneous nodes correspond to either drugs or target proteins, and edges correspond to interactions between them. The edge is placed between a drug node and a target node if the protein is a known target of the drug.

Figure 2 shows the distributions of chemical structure similarity scores and pharmacological effect similarity scores against the network distance for drugs targeting enzymes, ion channels, GPCRs and nuclear receptors. The top four panels in Figure 2 show the box-plots of drug–drug chemical structure similarities, and the bottom four panels in Figure 2 show the box-plots of drug–drug pharmacological similarities. The network distance means the shortest path between drugs on the bipartite graph representation of each drug–target interaction network. From the figure, we observe several tendencies.

Firstly, the larger the network distance between drugs, the smaller the variability of chemical structure similarities and pharmacological similarities, respectively. Also, the larger the network distance between drugs, the lower the scores of the chemical structure similarities and the drug pharmacological similarities, respectively. These observations suggest that two drugs sharing high chemical structure similarity or high pharmacological similarity tend to interact with similar target proteins.

Secondly, the above tendency is much clearer in the pharmacological similarity than in the chemical structural similarity. It seems that most pharmacological similarity scores are almost zero at larger distances, while many chemical similarity scores are relatively high even at larger distances. The difference of the distributions between 'distance 2' and 'distance 4' is important,

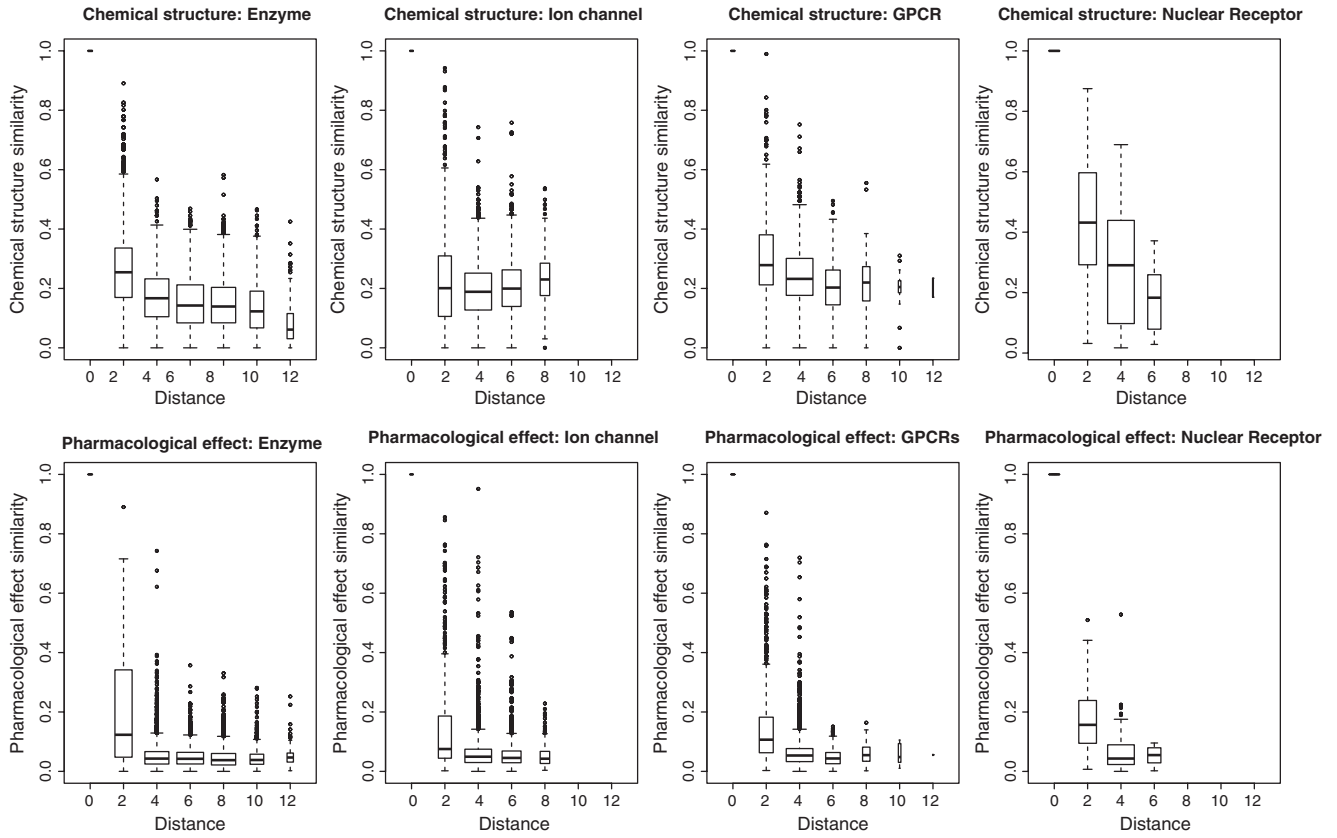

Fig. 2. Distributions of chemical structure similarity scores (top four panels) and pharmacological effect similarity scores (bottom four panels) against the network distance of drugs targeting enzymes, ion channels, GPCRs and nuclear receptors.

because 'distance 2' corresponds to drug-drug pairs which share the same target proteins, while 'distance 4' corresponds to drug–drug pairs which do not share the same target proteins. These observations suggest that pharmacological similarity is more correlated with drug targets than with chemical structure similarity, and the pharmacological similarity information is a more useful source for drug–target identification.

4.2 Performance evaluation of the proposed method

We tested the three different inputs: (i) chemical structure similarity, (ii) true pharmacological similarity, and (iii) predicted pharmacological similarity on their abilities to reconstruct four classes of drug–target interactions involving enzymes, ion channels, GPCRs and nuclear receptors. Note that input (i) corresponds to the previous method (Yamanishi *et al*., 2008), and input (ii) and input (iii) correspond to the proposed method in this study. Input (ii) reflects the situation where all compounds in the prediction set have pharmacological information, so we can skip the process of pharmacological effect prediction. Input (iii) reflects the situation where all compounds in the prediction set do not have any pharmacological information.

We performed the following 5-fold cross-validation procedure: drugs in the gold standard set were split into five subsets of roughly equal size, each subset was then taken in turn as a test set, and we performed the training on the remaining four sets. To obtain robust results and accurate comparison, we kept the same experimental conditions, where the same training drugs and test drugs are used across the three different inputs in each cross-validation. We repeated the above cross-validation experiment five times.

Table 1 shows the averages of the AUC [area under the receiver operating curve (ROC)], sensitivity, specificity and PPV (positive predictive value). The ROC (Gribskov and Robinson, 1996) is the plot of true positives as a function of false positives based on various thresholds, where true positives are correctly predicted interactions and false positives are predicted interactions that are not present in the gold standard interactions. The upper one percentile in the prediction score is chosen as a threshold for computing sensitivity, specificity and PPV, because high-confidence prediction results are interesting in practical applications.

It seems that the true pharmacological similarity-based method outperforms the chemical structure similarity-based method in all the four protein classes. Especially, the use of pharmacological information is effective in the case of enzyme and ion channel data. It seems that the predicted pharmacological similarity-based method also outperforms the chemical similarity-based method, but the performance is a little worse than that of the true pharmacological similarity-based method. In practical applications, it is rare to obtain the detailed pharmacological information about all compounds to be tested, so the result suggests that the predicted pharmacological information is useful for identification of unknown drug–target

Table 1. Statistics of the prediction performance

Class	Statistics	Input		
		Chemical structure similarity	True pharmacological similarity	Predicted pharmacological similarity
Enzyme	AUC	0.821	0.892	0.845
	Sensitivity	0.239	0.356	0.245
	Specificity	0.993	0.995	0.993
	PPV	0.358	0.527	0.369
Ion channel	AUC	0.692	0.812	0.731
	Sensitivity	0.134	0.137	0.142
	Specificity	0.996	0.996	0.997
	PPV	0.704	0.714	0.742
GPCR	AUC	0.811	0.827	0.812
	Sensitivity	0.147	0.172	0.164
	Specificity	0.994	0.996	0.995
	PPV	0.519	0.614	0.581
Nuclear receptor	AUC	0.814	0.835	0.830
	Sensitivity	0.067	0.057	0.077
	Specificity	0.995	0.994	0.996
	PPV	0.560	0.480	0.640

The AUC (ROC score) is the area under the ROC, normalized to 1 for a perfect inference and 0.5 for a random inference. The sensitivity is defined as TP/(TP+FN), the specificity is defined as TN/(TN+FP) and the PPV (positive predictive value) is defined as TP/(TP+FP), where TP, FP, TN, FN are the number of true positives, false positives, true negatives and false negatives, respectively.

interactions even when pharmacological information is not available for compounds of interest. These results serve to highlight the significant performance of the proposed method.

We also made a simple check of the effectiveness of grouping the keywords into the five tag groups. Figure 3 shows the AUC scores of the predicted pharmacological similarity-based method for the five tag groups (caution, interaction, patient, pharmaceutical effect and property) and the combination of the five groups, respectively, where 'caut', 'inte', 'pati', 'phar', 'prop' and 'comb' indicate the five tag groups and the combination, respectively. The low predictive performances of the inte profile is that the number of drugs having the inte keywords is much fewer than those of other types of keywords. It is notable that the remaining four types of keywords (caut, pati, phar and prop) outperformed the comb profiles, indicating the usefulness of discriminating the context of the keywords. It is natural, for example, that the drugs for high blood pressure and the drugs that cause high blood pressure have to be distinguished. These results suggest that appropriate selection of informative keywords and discriminating context will improve the predictive performance.

4.3 Comprehensive prediction for unknown drug–target interactions

After confirming the usefulness of our method, we conducted a comprehensive prediction of interactions between all possible compounds and proteins for the four classes of target proteins studied: enzymes, ion channels, GPCRs and nuclear receptors.

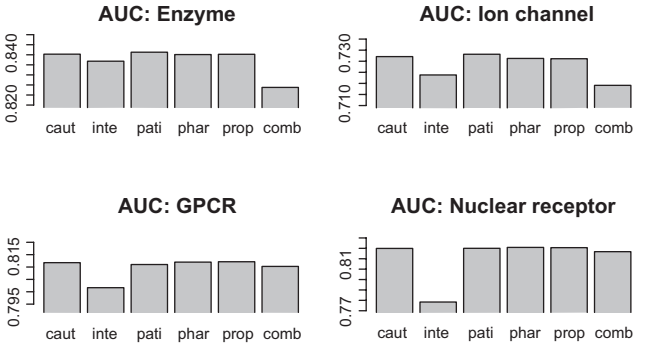

Fig. 3. Barplot of AUC score for the five tag groups (caution, interaction, patient, pharmaceutical effect and property) and their combination.

In the inference process for these predictions, we used all the known drugs and target proteins in the gold standard data as training data, and predicted potential interactions for all compounds in KEGG LIGAND and all the other drugs in KEGG DRUG (the drugs are absent from the gold standard data). Note that there remain many marketed drugs whose target proteins have not been identified yet. The total number of compounds including drugs in the prediction set is 15 383 in each case. Note that most of the compounds and drugs in the prediction set are not assigned any pharmacological information, so the pharmacological effect prediction is required. All the prediction results for each target protein class can be obtained from the web-supplement. Because of space limitations, we focused on the results for enzyme data below.

We focused on the top 1000 scoring predictions for the enzyme data. We investigated the validity of the predicted pairs based on the databases (e.g. KEGG BRITE, SuperTarget, DrugBank), because they contain information about interactions involving compounds which do not have any pharmacological information. Recall that in the Section 2 we constructed the gold standard set for drug–target interactions involving drugs for which the pharmacological information (by JAPIC package inserts) is available. As a result, we confirmed that 223 out of the top 1000 predictions are now annotated in at least one database. On the other hand, in the case of comprehensive prediction based on chemical structure information only, we confirmed that 140 out of the top 1000 predictions are now annotated in at least one database. We take this result as strong evidence supporting the practical relevance of our approach. Table 2 shows 10 examples of high scoring compound–protein pairs which were not predicted by chemical structure similarity but predicted by pharmacological similarity.

Next, we manually investigated the validity of the predicted pairs which were not confirmed in the databases, based on the literatures. We take some analgesic and antipyretic agents as examples, as shown in Figure 4. Salicylamide (D01811) and acetaminophen (D00217) are both known to act on prostaglandin-endoperoxide synthase 1/2 (PTGS1/2) (Aronoff *et al*., 2003). Based on these known interactions, some compounds are suggested to interact with PTGS1/2: ethenzamide (D01466), actarit (D01395), *N*-acetylphenylethylamine (C06746) and *N*-ethylphenylacetamide (C11487). Among these, D01466 is also an analgesic and antipyretic agent (Darias *et al*., 2006), although we could not find the target in the databases we used. On the other hand, D01395 is an anti-rheumatic agent (Ye *et al*., 2008). The JAPIC entry including

Table 2. Examples of compound–protein pairs predicted by the proposed method for enzyme data

	Pair	Annotation
1	C04000 5743	Benzyl 2-methyl-3-oxobutanoate prostaglandin-endoperoxide synthase 2
2	C04000 5742	Benzyl 2-methyl-3-oxobutanoate prostaglandin-endoperoxide synthase 1
3	D05868 5742	Sodium phenylbutyrate (USAN) prostaglandin-endoperoxide synthase 1
4	C07773 43	Ambenonium acetylcholinesterase (Yt blood group)
5	D05619 5742	Prodolic acid (USAN) prostaglandin-endoperoxide synthase 1
6	D05868 5743	Sodium phenylbutyrate (USAN) prostaglandin-endoperoxide synthase 2
7	D02587 476	Metildigoxin (JP15) ATPase, Na+/K+ transporting, alpha 1 polypeptide
8	C02505 5743	2-Phenylacetamide prostaglandin-endoperoxide synthase 2
9	C15513 5743	Benzyl acetate prostaglandin-endoperoxide synthase 2
10	C02505 5742	2-Phenylacetamide prostaglandin-endoperoxide synthase 1

Because of space limitation, all the prediction pairs are put on the supplemental website.

D01811 describes that this drug also has an effect on rheumatism (Frankl , 1953). We could not find any information about the pharmaceutical effects for other two compounds (C06746 and C11487), but they are structurally similar with the other drugs (D01811, D00217, D01466 and D01395). Therefore, it seems possible that these compounds act on PTGS1/2.

On the other hand, PTGS1 has some other interacting analgesic and antipyretic drugs, such as mofezolac (D01718) (Goto *et al*., 1998), from which tangeretin (C10190) (Hirano *et al*., 1995) is suggested as another potential drug. The structural commonality between these two compounds seems only that they both contain some *O*-methyl groups on aromatic rings, therefore this result might not be convincing. As the other questionable example, sodium lactate (D02183) is suggested to act on PTGS2 based on the known interacting drug sodium salicylate (D00566), an analgesic agent (Preston *et al*., 1989). However, this result seems not convincing at all, because their common substructures are only sodium ion and carboxylate group, and D02183 is an electrolyte replenisher.

The other group of analgesic and antipyretic drugs may possibly share a different target protein. Fluocinolone acetonide (D01825) (Emerit *et al*., 1983) and fluocinonide (D00325) (Schlessinger *et al*., 2006) are known to act on human cytosolic calcium-dependent phospholipase A2 (PLA2G4A), which is involved in lipid metabolism and related to various signal transductions (Balsinde *et al*., 1999). From resemblance to these two drugs, triamcinolone acetonide (D00983) (Keele, 1969) and diflorasone diacetate (D01327) (Bluefarb *et al*., 1976) are suggested to act on PLA2G4A. These four drugs are all corticosteroids, and are all known to act as analgesic and antipyretic drugs. Therefore we assume these results are convincing.

There are other possible drug–target interactions that belong to different therapeutic categories. For example, Metildigoxin (D02587) is predicted to have an interaction with a human Na^+/K^+ transporting ATPase (ATP1A1), based on the reported interaction of digoxin (D00298). D00298 is a digitalis-like cardiotonic substance that acts directly on heart muscle (Cumberbatch *et al*., 1981). D02587 is the methylated derivative of D00298, and many reports suggest that D02587 has no significant difference from D00298 in terms of their effects on heart functions (Kaufmann *et al*., 1981). Therefore, there is no wonder the two compounds share the same target protein.

5 DISCUSSION AND CONCLUSION

In this article, we investigated the relationship between the chemical space, the pharmacological space and the topology of drug–target interaction networks, and showed that drug–target interactions are more correlated with pharmacological effect similarity than with chemical structure similarity. We then developed a new statistical method to predict unknown drug–target interactions from chemical structure information, genomic sequence information and pharmacological effect information simultaneously on a large scale. The originality of the proposed method lies in prediction of pharmacological effects from chemical structures of given compounds, and its use for identification of unknown drug–target interactions in the framework of supervised bipartite graph inference. To our knowledge, this is the first report to predict drug–target interactions from the integration of chemical, genomic and pharmacological spaces in a unified framework.

One previous research related with this study is the use of side-effect similarity for drug–target identification (Campillos *et al*., 2008). However, the method is applicable only to marketed drugs for which detailed side-effect information is available. Therefore, newly detectable interactions were limited to the linkage between known marketed drugs assigned with side-effect information and known target proteins. To overcome these problems, we established a procedure to obtain pharmacological information about not only marketed drugs but also any drugs or any drug candidate compounds based on their chemical structures. The proposed procedure makes it possible to perform screening of any chemical compounds against many target candidate proteins.

In practice, there are four possible classes for predictable compound–protein pairs: (i) new drug candidate compounds versus known target proteins, (ii) known drugs versus new target candidate proteins, (iii) new drug candidate compounds versus new target candidate proteins, and (iv) known drugs versus known target proteins, where compounds and proteins with interaction partner information are called ‘known’, otherwise called ‘new’. Note that in this study we focus on class (i), because the objective of this article is to investigate the effect of introducing pharmacological information about new drug candidate compounds. Recently, the bipartite local model approach has been proposed to detect missing interactions between known drugs and known target proteins based

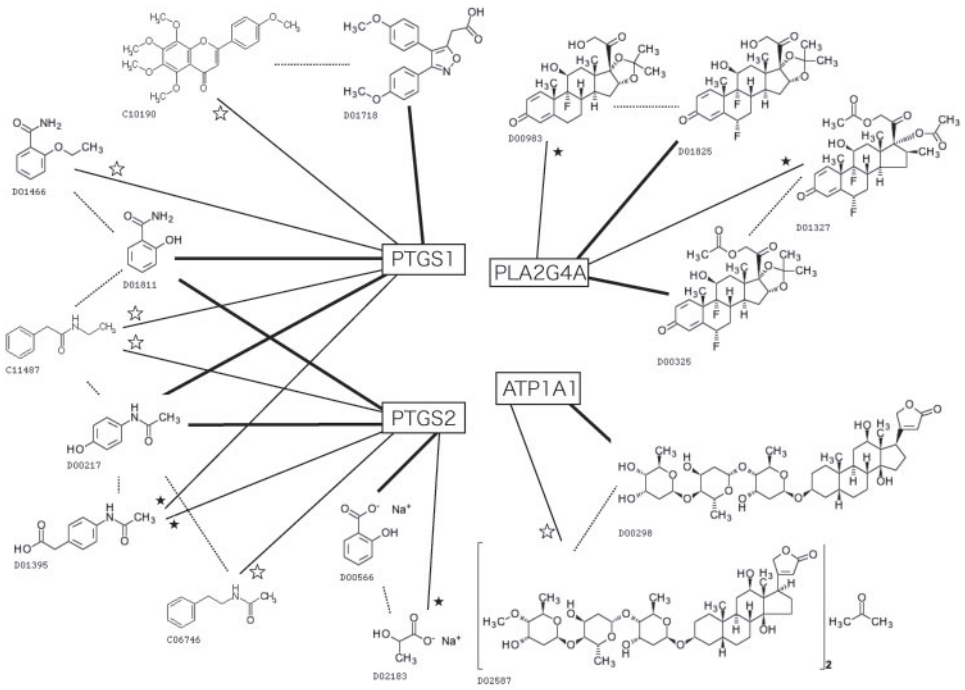

Fig. 4. Examples of the proposed drug–target interactions. Four boxes in the center of the figure are the target proteins, and bold lines indicate the known drug–target interactions. Solid lines represent the proposed interactions based on the resemblance to the known interacting drugs indicated by the dashed lines. Black stars indicate the interactions predicted by the previous method. White stars indicate the interactions additionally predicted by the proposed method.

on chemical and genomic data (Bleakley and Yamanishi, 2009). The approach works similarly with other bipartite graph inference methods for classes (i) and (ii) in terms of accuracy, but the approach with an aggregation scheme is quite powerful for class (iv) (Bleakley and Yamanishi, 2009), so the approach with pharmacological information could detect missing interactions in class (iv) with high accuracy.

From a technical viewpoint, the performance of our method could be improved by using more sophisticated similarity functions for compounds and proteins, such as kernel functions designed for genomic sequences and chemical structures (Schölkopf *et al*., 2004). In this study, we evaluated the drug pharmacological similarity based on all available pharmacological keywords categorized into each tag in the package insert of each drug. There remain many unimportant keywords to be filtered and there might exist some correlation between related keywords or hierarchy among medical vocabulary. To deal with these problems, the use of sophisticated text mining approaches is an important research direction.

The proposed method is expected to be useful for virtual screening of chemical libraries. To detect new biological findings and find potentially useful drug leads, we are currently working with collaborators on biological assays. We believe that our method is able to increase research productivity toward genomic drug discovery.

ACKNOWLEDGEMENTS

Computational resources were provided by the Bioinformatics Center, Institute for Chemical Research and the Super Computer Laboratory, Kyoto University.

Funding: Ministry of Education, Culture, Sports, Science and Technology of Japan, the Japan Science and Technology Agency; Japan Society for the Promotion of Science; the bi-national JSPS/INSERM grant Japan-France Research Cooperative Program.

Conflict of Interest: none declared.

REFERENCES

Aronoff,D.M. *et al*. (2003) Inhibition of prostaglandin h2 synthases by salicylate is dependent on the oxidative state of the enzymes. *J. Pharamacol. Exp. Thera.*, **304**, 589–595.

Balsinde,J. *et al*. (1999) Regulation and inhibition of phospholipase. *Annu. Rev. Pharmacol. Toxicol.*, **39**, 175–189.

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

Bluefarb,S.M. *et al*. (1976) Diflorasone diacetate: vasoconstrictor activity and clinical efficacy of a new topical corticosteroid. *J. Int. Med. Res.*, **4**, 454–461.

Bock,J.R. and Gough,D.A. (2005) Virtual screen for ligands of orphan g protein-coupled receptors. *J. Chem. Inf. Model*, **45**, 1402–1414.

Butina,D. *et al*. (2002) Predicting ADME properties in silico: methods and models. *Drug Discov. Today*, **7**, S83–S88.

Byvatov,E. *et al*. (2003) Comparison of support vector machine and articial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci.*, **43**, 1882–1889.

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

Cumberbatch,M. *et al*. (1981) The early and late effects of digoxin treatment on the sodium transport, sodium content and Na^+K^+-ATPase or erythrocytes. *Br. J. Clin. Pharmacol.*, **11**, 565–570.

Darias,V. *et al*. (2006) Synthesis and preliminary pharmacological study of thiophene analogues of the antipyretic and analgesic agent ethenzamide. *Arch. Pharm.*, **325**, 83–87.

Dobson,C.M. (2004) Chemical space and biology. *Nature*, **432**, 824–828.

Emerit,I. *et al*. (1983) Suppression of tumor promoter phorbolmyristate acetate-induced chromosome breakage by antioxidants and inhibitors of arachidonic acid metabolism. *Mutat. Res.*, **110**, 327–335.

Erhan,D. *et al*. (2006) Collaborative ltering on a family of biological targets. *J. Chem. Inf. Model*, **46**, 626–635.

Faulon,J.L *et al*. (2008) Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor. *Bioinformatics*, **24**, 225–233.

Frankl,R. (1953) Intravenous salicylamide therapy in rheumatic diseases. *Munch. Med. Wochensch.*, **95**, 512–513.

Goto,K. *et al*. (1998) Analgesic effect of mofezolac, a non-steroidal anti-inflammatory drug, against phenylquinone-induced acute pain in mice. *Prostaglandins Other Lipid Mediat.*, **56**, 245–254.

Gribskov,M. and Robinson,N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.*, **20**, 25–33.

Gunther,S. *et al*. (2008) SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res.*, **36**, D919–D922.

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**, 11853–11865.

Hirano,T. *et al*. (1995) Citrus flavone tangeretin inhibits leukaemic hl-60 cell growth partially through induction of apoptosis with less cytotoxicity on normal lymphocytes. *Br. J. Cancer*, **72**, 1380–1388.

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

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

Kanehisa,M. *et al*. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D485.

Kaufmann,B. *et al*. (1981) Pharmacokinetics of metildigoxin and digoxin in geriatric patients with normal and elevated serum creatinine levels. *Clin. Pharmacokinet.*, **6**, 463–468.

Keele.C.A. (1969) Sites and modes of action of antipyretic-analgesic drug. *Proc. R. Soc. Med.*, **62**, 535–539.

Lodhi,H. and Yamanishi,Y. (2010) *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques*. IGI Global, Hershey, PA.

Nagamine,N. and Sakakibara,Y. (2007) Statistical prediction of protein-chemical interactions based on chemical structure and mass spectrometry data. *Bioinformatics*, **23**, 2004–2012.

Preston,S.J. *et al*. (1989) Comparative analgesic and anti-inflammatory properties of sodium salicylate and acetylsalicylic acid (aspirin) in rheumatoid arthritis. *Br. J. Clin. Pharmacol.*, **27**, 607–611.

Rarey,M. *et al*. (1996) A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, **261**, 470–489.

Scheiber,J. *et al*. (2009) Mapping adverse drug reactions in chemical space. *J. Med. Chem.*, **52**, 3103–3107.

Schlessinger,J. *et al*. (2006) An open-label adrenal suppression study of 0.1% fluocinonide cream in pediatric patients with atopic dermatitis. *Arch. Dermatol.*, **142**, 1568–1572.

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

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

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

Stockwell,B.R. (2000) Chemical genetics: ligand-based discovery of gene function. *Nat. Rev. Genet.*, **1**, 116–125.

Wishart,D.S. *et al*. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, **36**, D901–D906.

Yamanishi,Y. (2009) Supervised bipartite graph inference. In Koller,D. *et al.* (eds), *Advances in Neural Information Processing Systems 21*, MIT Press, Cambridge, MA, pp. 1841–1848.

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

Ye,J. *et al*. (2008) Injectable actarit-loaded solid lipid nanoparticles as passive targeting therapeutic agents for rheumatoid arthritis. *Int. J. Pharmaceutics*, **352**, 273–179.

Yildirim,M.A. *et al*. (2007) Drug-target network. *Nat. Biotechnol.*, **25**, 1119–1126.